NISTIR 4941

Selected Publications for the Advanced Mass Measurements Workshop

Georgia L. Harris
Editor

U.S. DEPARTMENT OF COMMERCE
Technology Administration
National Institute of Standards
and Technology
Office of Weights and Measures
Gaithersburg, MD 20899

NIST

NISTIR 4941

Selected Publications for the Advanced Mass Measurements Workshop

Georgia L. Harris
Editor

U.S. DEPARTMENT OF COMMERCE
Technology Administration
National Institute of Standards
and Technology
Office of Weights and Measures
Gaithersburg, MD 20899

November 1992

U.S. DEPARTMENT OF COMMERCE
Barbara Hackman Franklin, Secretary

TECHNOLOGY ADMINISTRATION
Robert M. White, Under Secretary for Technology

NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY
John W. Lyons, Director

Editor's Note

The Office of Weights and Measures of the National Institute of Standards and Technology (NIST) provides training in the form of Basic, Intermediate, and Advanced seminars in mass, length, and volume to State and industry metrologists in support of the New State Standards Program which was funded by Congress in 1965 and provided standards and equipment to State weights and measures laboratories.

To meet the advancing technological needs of the State laboratories and private industry calibration laboratories, an Advanced Mass Measurements Workshop has been developed. This 5-day workshop aims to disseminate NIST expertise in mass measurements and calibration to professionals and senior technical personnel. Calibration designs and quality control techniques are emphasized for use in their own laboratories.

This volume is a collection of publications that has been found to be essential and useful in the conduct of an Advanced Mass Measurements Workshop under the sponsorship of the Office of Weights and Measures.

Publications included in this volume appeared originally as articles scattered in journals, NIST reports, technical notes, monographs, and special publications. Some of these papers are now out of print. It is anticipated that this publication will also serve as a useful reference on mass calibrations in addition to its use during the workshop. The pagination for this publication follows the same format as Selected Publications For the EMAP Workshop, namely: a unique alphabetic designation for each reprint is followed by the page number of the original article, e.g. A-14.

The author wishes to thank M. Carroll Croarkin for bringing the EMAP publication to her attention and for her assistance in the development of subject material for the workshop.

Georgia L. Harris

Contents

Measurement Assurance

Mass Metrology

Calibration Design and Statistical Analysis

Practical Representation of the Mass Unit, Effective January 1990

Bibiliography

Precision Measurement and Calibration: Statistical Concepts and Procedures
H.H. Ku - Editor
NBS Special Publication 300, Vol. I, 1969, 436 pages

Experimental Statistics
M.G. Natrella
NBS Handbook 91, 1966

ASTM Manual on Presentation of Data and Control Chart Analysis, MNL 7, American Society for Testing and Materials, 1990, 106 pages

Experimentation and Measurements
W.J. Youden
NBS Special Publication 672, 1984, 127 pages

MEASUREMENT ASSURANCE

NBSIR 77-1240

MEASUREMENT ASSURANCE

J. M. Cameron

Institute for Basic Standards
National Bureau of Standards
Washington, D.C. 20234

April 1977

Final

Prepared for
Office of Measurement Services
Institute for Basic Standards
National Bureau of Standards
Washington, D.C. 20234

U.S. DEPARTMENT OF COMMERCE, Juanita M. Kreps, *Secretary*
NATIONAL BUREAU OF STANDARDS, Ernest Ambler, *Acting Director*

Measurement Assurance

Introduction

A single measurement can be the basis for actions taken to maintain our health, safety or the quality of our environment. It is important therefore that the errors of measurement be small enough so that the actions taken are only negligibly affected by these errors. We realize this necessity on a personal basis when we consider medical measurements, or our exposure to radioactivity. In any government regulatory action or measurement involved in legal actions it is also obvious that the shadow of doubt surrounding the measurements should be suitably small. But this is no less true for all other measurements in science and industry and even though legal action may not be involved, the validity of scientific inference, the effectiveness of process control, or the quality of production may depend on adequate measurements [2].

Allowable Limits of Measurement Error

How does one achieve this condition--that the measurements are "good enough" for their intended use? It would seem obvious that one has to start with the need--i.e., deciding upon what is "good enough". There are a number of cases where physiological restraints provide the definition such as in the allowable error in exposure to cobalt radiation in cancer treatment or in the amount of pollutant entering a lake. In nuclear materials control the allowable error is a function of the amount of material which would pose a hazard if diverted. In industrial production or commercial transactions, the error limit is determined by a balance between the cost of better measurement and the possible economic loss from poorer measurement.

By whatever path such requirements are arrived at, let us begin with the assumption that the allowable error should not be outside the interval (-a, +b) relative to the quantity being measured. Our problem is one of deciding whether the uncertainty of a single measurement is wholly contained in an interval of that size. We therefore need a means of assigning an uncertainty to a single isolated measurement and, in fact, we need a perspective (i.e., physical and mathematical model) in which to view measurement so as to give operational meaning to the term "uncertainty."

Reference Base to Which Measurements Must Be Related

It is instructive to contemplate the possible "cross-examination" of a measurement if it were to become an important element in a legal controversy. Two essential features emerge. First, that the contending parties would have to agree on what (actually realizable) measurement would be mutually acceptable. The logic of this seems unassailable--if one cannot state what measurement system would be

accepted as "correct," then one would have no defensible way of developing specifications or regulations involving such measurements. Second, the scientific cross-examination by which one establishes the "shadow of doubt" relative to this acceptable value gives one the uncertainty to be attached to the measurement.

The consensus or generally accepted value can be given a particularly simple meaning in dealing with measurements of such quantities as mass, volt, resistance, temperature, etc. One may require that uncertainties be expressed relative to the standards as maintained by a local laboratory or, when appropriate, to the national standards as maintained by NBS. In other cases, nationally accepted artifacts, standard reference materials or in some cases a particular measurement process may constitute a reference base. One basic quality should not be overlooked--all are operationally realizable. The confusion engendered by introducing the term "true value" as the correct but unknowable value is thus avoided.

Properties of Measurement Processes

In discussing uncertainty, we must account for two characteristics of measurement processes. First, repeated measurements of the same quantity by the same measurement process will disagree and, second, the limiting means of measurements by two different processes will disagree. These observations lead to a perspective from which to view measurement namely that the measurement be regarded as the "output" of a process analogous to an industrial production process. In defining the process, one must state the conditions under which a "repetition" of the measurement would be made, analogous to defining the conditions of manufacture in an industrial process.

The need for this specification of the process becomes clear if one envisions the "cross-examination" process. One would begin with such questions as

> Within what limits would an additional measurement by the same instrument agree when measuring some stable quantity?
>
> Would the agreement be poorer if the time interval between repetitions were increased?
>
> What if different instruments from the same manufacturer were used?
>
> If two or more types (or manufacturers) were used, how much disagreement would be expected?

To these can be added questions related to the conduct of the measurement.

What effect does geometry (orientation, etc.) have on the measurement?

What about environmental conditions--temperature, moisture, etc.?

Is the result dependent on the procedure used?

Do different operators show persistent differences in values?

Are there instrumental biases or differences due to reference standards or calibrations?

The questions serve to define the measurement process--the process whose "output" we seek to characterize.

The current understanding of a scientific or industrial process or of a measurement process is embodied in a physical model which explains the interactions of various factors, corrections for environmental or other effects, and the probability models necessary to account for the fact that repetitions of the same event give rise to nonidentical answers. For example, in noise level measurement one is involved with assumptions regarding frequency response, weighing networks, influence of procedures and geometry, and an accepted theory for making corrections for temperature and other environmental factors. In mass the properties of the comparator (balance) the environmental effects, and the procedure used all enter into the description of the method.

One thus begins with the specification of a measurement method--the detailed description of apparatus, procedures and conditions by which one will measure some quantity. Once the apparatus is assembled and checked out, one has a measurement process whose output can be studied to see if it conforms to the requirement for which it was created.

In industrial production one tries to produce identical items but usually a measurement process is set up to measure a variety of quantities and ordinarily one does not measure the same quantity over and over. One thus has the problem of sampling the output of the measuring process so as to be able to make statements about the health of the process relative to the needs. The needed redundancy can sometimes be achieved by remeasuring some of the items, or by measuring a reference artifact periodically. It is essential that the repetitions be done under the same diversity of conditions as the regular measurements, and that the items being measured be typical of the regular workload.

As an example, a sequence of measurements was made using two sound level meters to measure a sound of nominally 90 dB re 20 μPa. The sound was generated by a loudspeaker fed broadband noise. On 16

different days measurements were made outdoors and over grass with the loudspeaker in the same orientation and location relative to a building 2 m behind the loudspeaker. The sound level meter was always the same distance (10 m) from the loudspeaker and on a line perpendicular to the face of the loudspeaker. Other than the grass, the person holding the sound level meter, and the building to the rear of the loudspeaker, there were no other reflecting surfaces or obstacles within 50 m. No measurements were made in the rain or in winds exceeding a few km/hr. The results from these 16 repetitions are shown in Figure 1. Typically, had duplicate measurements been made on the same day they would have given results as shown in Figure 2.

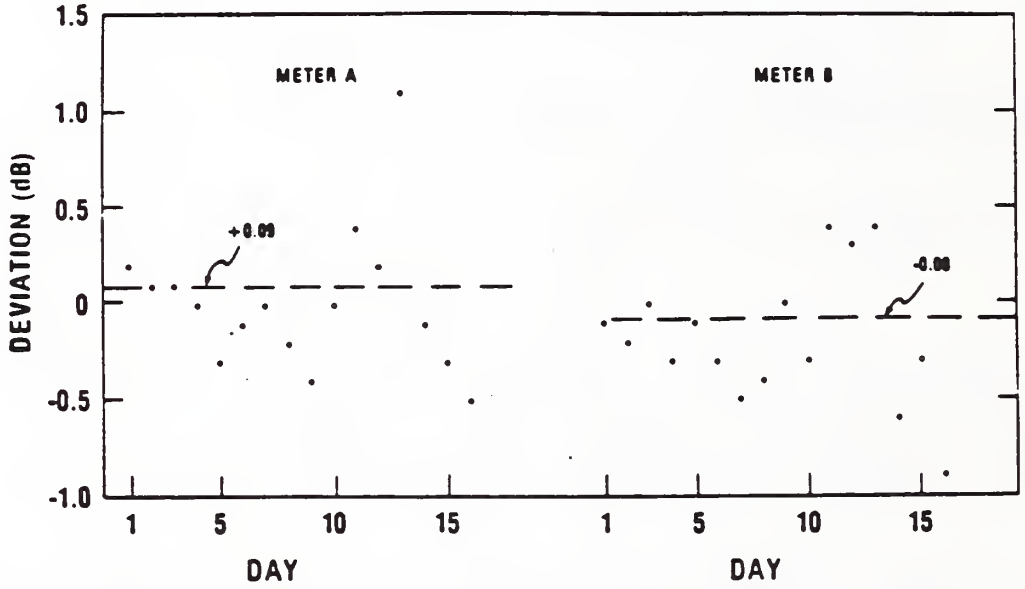


FIGURE 1: DAY-TO-DAY VARIATION IN METER READINGS.

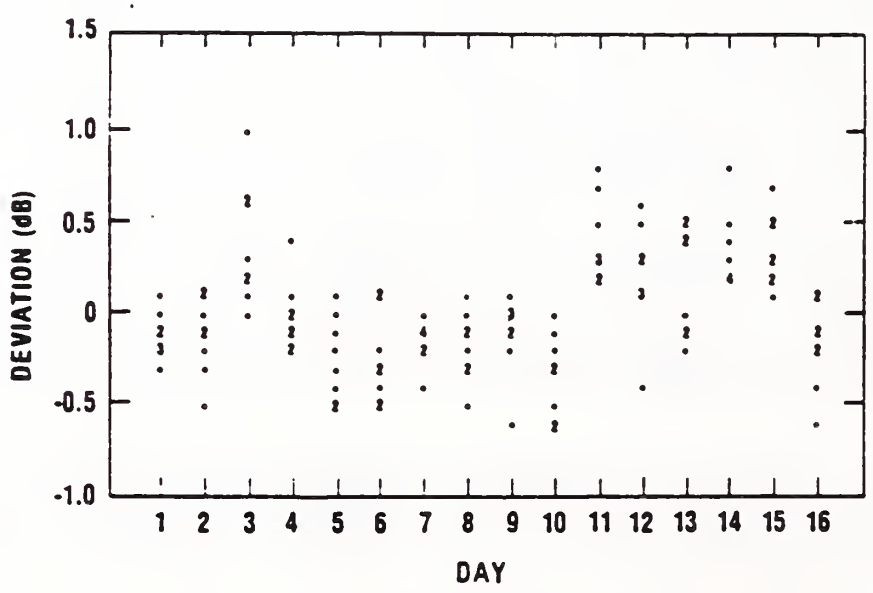


FIGURE 2: DAY-TO-DAY VARIATION IN METER READINGS WITH MULTIPLE VALUES PER DAY. (COINCIDENT POINTS INDICATED BY NUMBERS.)

One now faces the question of how to describe the variation that exists. Obviously there will be a different level of agreement expected between pairs on the same day, but this variation in no way predicts that encountered from day-to-day. The issue is not so much the statistical procedures to be used--these will follow after one defines the set of repetitions over which his conclusions must apply. For measuring the short term change in noise level, the difference between duplicates would apply; for any regulatory action, the day-to-day variation would have to be considered.

The crucial step in assessing the effects of random error is that of defining the set of repetitions over which the measurement is to apply. In the context of legal proceedings, one arrives at the degree of credibility of evidence by questions designed to find out how far the statement could be in error. In measurement, the uncertainty is arrived at by determining the amount of disagreement expected in the set of repetitions that would be appropriate in the context of the intended use of the measurement.

The Concept of a Repetition of a Measurement

Every measurement has a set of conditions in which it is presumed to be valid. At a very minimum, it is the set of repeated measurements with the same instrument-operator procedure-configuration. (This is the type of repetition one would envision in some process control oerations.) If the measurement is to be interchangeable with one made at another location, the repetition would involve different instrument-operator-procedure-environment configurations. (This type of repetition is involved in producing items to satisfy a specification and of manufacturing generally.) When the measurement is to be used for conformance to a health, safety, or environmental regulation even different methods may be involved in a "repetition."

To evaluate a measurement process some redundancy needs to be built into the system to determine the process parameters. This redundancy should be representative of the set of repetitions with which the uncertainty statement is to apply. In NBS' measurements of mass, a check standard is measured in parallel with the unknowns submitted for calibration. One thus generates a sequence of measurements of the same object covering an extended time period. From these results one can answer questions relating to the agreement expected in a recalibration and the operating characteristics of the measurement process. In this simple case the check standard is treated exactly the same way as the unknowns so that the properties of the process related to it are transferrable to the unknown.

The essential characteristic in establishing the validity of measurement is predictability that the variability remains at the same level and that the process has not drifted of shifted abruptly from its established values. One must build in redundancy in the form of a

control--the measurement of a reference quantity of known value--or by remeasuring some values by a reference method (or by an instrument with considerably smaller uncertainty). In cases where the phenomenon can be repeated, one can learn about random errors by remeasuring at a later time sufficiently far removed to guarantee independence.

In measuring an "unknown" one gets a single value, but one still is faced with the need to make a statement that allows for the scatter of the results. If we had a sufficiently long record of measurements, we could set limits within which we were fairly certain that the next measurement would lie. Such a statement should be based on a collection of independent determinations, each one similar in character to the new observation, that is to say, so that each observation of the collection and also the new observation can be considered as random drawings from the same probability distribution. These conditions will be satisfied if the collection of points is from a sufficiently broad set of environmental and operating conditions to allow all the random effects to which the process is subject to have a chance to exert their influence on the variability. Suitable collections of data can be obtained by incorporating an appropriate reference measurement into routine measurement procedures, provided they are representative of the same variability to which the "unknown" is subject. The statistical procedures for expressing the results will depend on the structure of the data but they cannot overcome deficiencies in the representativeness of the values being used.

The results from the reference item provide the basis for determining the parameters of the measurement process and the properties are transferable. One is saying, in effect, if we could have measured the "unknown" again and again, a sequence of values such as those for the reference item would have been obtained. Whether our single value is above or below the mean we cannot say, but we are fairly certain it would not differ by more than the bounds to the scatter of the values on the reference item.

The bound $\pm R$, to be used for the possible effect of random errors may be as simple as ± 3 (standard deviation) or may involve the combination of many components of variance. Once the set of repetitions over which one's conclusions must apply is defined, the structure of the random error bound can be determined.

Possible Offset of the Process

Once one has established that his measurement process is "in control" from the point of view of random variation, there remains the question of the possible offset of the process relative to other processes. It is not helpful to speak of the offset from a "true value" which exists only in the mathematical or physical model of the process. The usefulness of considering measurement in the context of legal proceedings helps clear away some of the classical confusion about

errors of measurement. In a legal or regulatory setting, one is forced to state what would be accepted as correct such as comparison (by a prescribed process) with national standards or with the results from a designated laboratory or consensus of many laboratories.

The idea of defining uncertainty as the extent to which a measurement is in doubt relative to a standard or process defined as correct finds expression in the recent Nuclear Regulatory Commission statement [12]:

> 70.57(a) "Traceability" means the ability to relate *individual measurement results* to national standards or nationally accepted measurement systems ... (italics added)

One could measure the offset of his process relative to the accepted process, and make suitable corrections to eliminate the offset. However, for most processes, one is content with setting bounds to the possible offset due to factors such as:

> Errors in the starting standards
>
> Departures from sought-after instrumentation (e.g., geometrical discrepancies)
>
> Errors in procedures, environment, etc.

and other effects which are persistent. From properly designed experiments one can arrive at a limit to the possible extent of errors from these sources in answer to the question, "If the process were set up ab initio, how large a difference in their limiting means would be reasonable?"

A bound to a number of factors can be determined as part of regular measurement. For example, the effect of elevation on sound level measurements could be evaluated by occasionally duplicating a measurement at a different height and taking an appropriate fraction of the observed difference as the limit to the possible offset due to any error in setting elevation. Figure 3 shows some results from sound level meters at two heights with the source at a constant height.

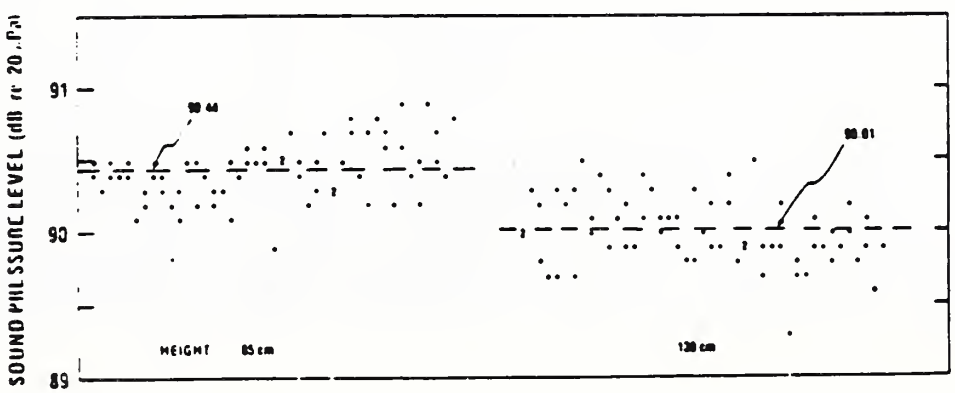


FIGURE 3: DIFFERENCE BETWEEN METER VALUES WITH CHANGE IN HEIGHT

Even if one has a functional relation, $y = f(h)$, expressing the dependence of the result, y, on height, h, one still has to carry out these measurements. The usual propagation of error approach involving partial derivatives, etc., implies that all instruments are equally dependent on the parameter under study, that there are no effects related to the factor except that contained in the formula. This can be verified for a particular instrument by actually measuring its response.

A similar comparison was made for a different orientation of the instrument with respect to this signal source and is shown in Figure 4. The effect of orientation is negligible and one would not be justified in adding an allowance for possible systematic error from this source based on a theoretical calculation.

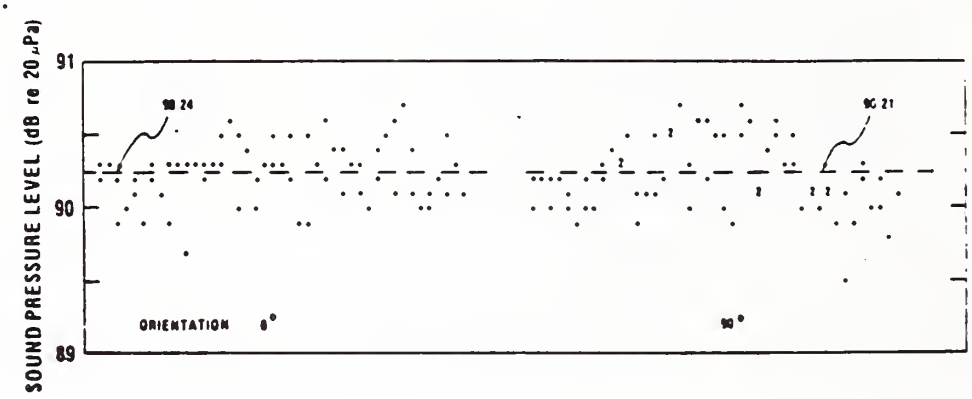


FIGURE 4: DIFFERENCE BETWEEN METER VALUES WITH A CHANGE IN ORIENTATION

From these measurements, one will have a set of bounds E_1, E_2, E_3, ... to the possible offset or systematic error from the various factors. The question as to how to combine these to a single bound to the possible offset depends on knowledge of the joint effects of two or more factors and on the physical model assumed for the process. For example, if the bounds E_i and E_j arise from independent random error

bounds, then it would be appropriate to combine them in quadrature, i.e., $\sqrt{E_i^2 + E_j^2}$. An error in the model e.g., assumed linearity even when nonlinearity exists) would act as an additive error. The properties of any combination rule can be evaluated and a selection made of the most appropriate. The result will be an overall value, E, for the possible offset for the limiting mean of the process from that of the nationally accepted process.

Uncertainty

What can one say about the uncertainty of a measurement made by a process that may be offset from the nationally accepted process by some amount $\pm E$, and is subject to random errors bounded by $\pm R$? How should these values be combined? To begin with, one could raise the question, "If the random error could be made negligible, what uncertainty would one attach to a value from the process?" Clearly the answer is $\pm E$. The next question, "If, in addition, a random error of size R is possible, what do we now say about the uncertainty?" The answer seems obvious--E and R are added to give an uncertainty of $\pm[E + R]$.

But what if E were itself the result of only random errors? The answer depends on what one calls a repetition. By the way E is defined, it is the bound for the systematic offset of the process and although it may be arrived at from consideration of random errors, the factor involved keeps the same (unknown) value throughout. Our ignorance does not make it a random variable.

Consider the case of a mass standard. NBS' certificate states that the uncertainty is based entirely on random variation, the effects from systematic errors being negligible. But unless one recalibrates, the error due to calibration remains fixed in all measurements by the user.

The uncertainty of a measurement--the width of its "shadow of doubt" in a legal proceeding--must therefore be the sum of the random error and systematic error limits.

Measurement Process Control

The essential feature for the validity of the uncertainty statement is that the process remain in a state of statistical control. Once an out-of-control condition occurs, one has lost predictability and the previous uncertainty statements are no longer valid.

To monitor the process some redundancy has to be built into the system. A variety of techniques can be used to give assurance of continued control. For example, one could periodically measure the same reference item or artifact or one could make duplicate measurements on some production items with enough delay to guarantee

independence. The American National Standards Institute Standard N15.18 for mass measurement [10] is an example where this approach is worked out in detail. But one has to verify more than just those parameters related to random variations. One needs to build in tests of the adequacy of the physical model by a variety of tests on the process (e.g., by repeating measurements under different conditions to verify the adequacy of the corrections for such changes) as well as periodic redetermination of the bounds for systematic error. One thus tests that the assumed model is still acceptable and that the parameters assigned to that model have not changed.

An excellent example of the efficacy of this approach is given by the recent announcement [6] of discrepancies of 1 mg in the assignment of mass to aluminum kilogram standards. The mass measurement system has long been shown to be nearly perfect for the usual standards. To check up on the performance of the system at densities nearer to that of most objects involved in practical measurement, an aluminum kilogram was sent to laboratories including several at high elevations. It turns out that the difference between the mass of a stainless steel and an aluminum kilogram is significantly different at different elevations. This unsuspected property of the real measurement system is now the subject of considerable study.

All measurements have some form of measurement assurance program associated with them although, as with quality control, we usually reserve the term for a formal program. In a formal program one treats the whole process--beginning with a study of the need, the development of a measuring process and a procedure for determining and monitoring its performance, and an evaluation of the effectiveness of the whole effort. One needs a criterion of success to be able to determine whether more of one's current measurement activity or perhaps some alternative would contribute most to the overall program, and this is not necessarily provided by the smallness of the uncertainty for a measurement.

For example, when the requirement is for matched sets (e.g., ball bearings) or mated assembly parts, then it is usually cheaper and more accurate to sort into finely divided classes and match for correctness of fit rather than perform direct measurement of each part.

When the measurement requirements are stated in terms of the needs of the system, (number of correctly matching parts, number of correctly measured dosimeters, etc.) one can measure success of the measurement effort in terms of closeness to meeting those goals. Measurement efficiency is thus judged in terms of the output of the organization rather than by the count of the number of significant digits. Also, one needs this measure of performance of the measurement effort to be able to identify those areas which need improvement.

Examples of Measurement Assurance Programs in NBS Measurements

Two easily described measurement assurance programs are those in mass and length. In routine calibration, a check standard is included with each set of weighings and process control is maintained by monitoring the value obtained for the check standard and of the random error from the least squares analysis [8, 9]. Control charts have been maintained since 1963. In the calibration of gage blocks, similar process control has been maintained since 1972 on both the interferometric process by which the assignment of length to the NBS master gage blocks is done and on the comparator process by which length values are transferred to customer gage blocks. [1, 7]

Similar programs are in effect in all divisions, but not all quantities involved in calibration have a formal program worthy of the name, measurement assurance.

Examples of Measurement Assurance Programs At Other Laboratories

Only two examples of measurement assurance programs at other laboratories have ever been reported. One at Autonetics [3] in length and one at Mounds Laboratory in mass. Once the mass measurement system for UF_6 is underway as part of the Safeguards program, NBS will be able to document the efficacy of the approach in practical measurement.

The NBS Measurement Assurance Programs Offered As A Part Of Our Calibration Service

Measurement Assurance Programs are listed as a calibration service in mass, volt, resistence, capacitance, voltage ratio, watthour meters, platinum resistance thermometry, and laser power. These are designed to measure the offset of measurement processes for the calibration of standards by other standards laboratories. These are applicable only to those laboratories who maintain and calibrate standards in the same manner as NBS. [See 11, 5, 13.]

These procedures enable a laboratory to determine the offset between its process of calibrating standards and that of NBS.

Need For Measurement Assurance Program For Practical Measurement

The UF_6 cylinder program for Safeguards [10] is an example of NBS' service in providing a direct method for measuring the offset of practical measurement processes from that accepted as correct, namely mass measurement by NBS. Investigation of the need and possible mechanisms or artifacts for monitoring the offset of practical measurements in quantities such as voltage, resistance, length, radioactivity is underway. (For examples of the application of these principles to sound level meters, see [5].)

In personnel dosimetry procedures are being worked out [14] to monitor the output of firms providing such services. In this case, a table of allowable limits of uncertainty are based on physiological considerations. Process parameters are to be determined by an initial study. Routine monitoring will be used to confirm that the process is "in control" at those levels, otherwise the parameters are redetermined *ab initio*. These "consistency" or "in control" criteria replace the usual one-time round robin approach. The amount of effort needed to establish this predictability is a function of the risk and costs of wrong decisions.

In industrial measurement we could ask

> If some critical measurements on the production line were repeated would the two measurements agree?
>
> How much bad material is passed, or good material rejected because of errors in measurement?

To those who have not properly answered these questions, dollar savings and improved product quality are possible without redesign or changes in production procedures.

Is our faith in instruments justified? Implicit faith in the correctness of instruments means that product variability (as determined by these instruments) is attributed to variability in components, raw materials or even poor design. One wonders how many times this has led to expensive changes in production procedures without apparent improvement because the variability actually arose in the measurements themselves.

How often has the installation and methods of use degraded the output of an instrument capable of much more accuracy than is required when handled properly? Without some surveillance of the actual measurements, one would never know.

One wonders how often a product is redesigned because measurement error has led to the decision that the product does not conform to specifications.

The result of this look at measurement is measurement assurance--the quality control of measurement. If adequate control exists, then one can look elsewhere for improvements in the product line. If it does not, then one has the possibility of savings without changing production procedures.

Some form of redundancy must be built into the process to answer these questions.

References

[1] John S. Beers, "A Gage Block Measurement Process Using Single Wavelength Interferometry," NBS Monograph 152, December 1975

[2] J. M. Cameron, "Measurement Assurance," Review of Standards and Specifications Section of Journal of Quality Technology, Vol. 8, No. 1, pp. 53-55, January 1976

[3] J. A. Hall, "Dimensional Measurement Assurance Programs," Autonetics Report No. X73-1146/031, presentation at 1973 Conference of the National Conference of Standards Laboratories at the National Bureau of Standards, Gaithersburg, Md., November 1973

[4] Edward B. Magrab, "Environmental Effects on Microphones and Type II Sound Level Meters," NBS Technical Note 931, October 1976

[5] N. Michael Oldham, "A Measurement Assurance Program for Electric Energy," NBS Technical Note 930, September 1976

[6] P. E. Pontius, "Mass Measurement: A Study of Anomalies," Science 190, pp. 379-380, October 1975

[7] P. E. Pontius, "Measurement Assurance Program--A Case Study: Length Measurements. Part 1. Long Gage Blocks (5 in. to 20 in.)," NBS Monograph 149, November 1975

[8] P. E. Pontius, "Measurement Philosophy of the Pilot Program for Mass Calibration," NBS Technical Note 288, May 1966

[9] P. E. Pontius, "Notes on the Fundamentals of Measurement and Measurement as a Production Process," NBSIR 74-545, September 1974

[10] "Mass Calibration Techniques for Nuclear Material Control," N15.18-1975, American National Standards Institute, New York, New York

[11] "Measurement Analysis Program in Mass," NBS Technical News Bulletin, September 1970

[12] "Measurement Control Program for Special Nuclear Materials Control and Accounting (10 CFR 70.57)," Federal Register, Vol. 40, No. 155, pp. 33651-33653, August 1975

[13] "New Calibration Service for the Volt," NBS Technical News Bulletin, pp. 46-47, February 1971

[14] "Proposed Standard--Criteria for Testing Personnel-Dosimetry Performance," Health-Physics Society, HPSSC WG/15, October 1976

Measurement Assurance Programs Part II: Development and Implementation

Carroll Croarkin

Statistical Engineering Division
Center for Applied Mathematics
National Engineering Laboratory
National Bureau of Standards
Gaithersburg, MD 20899

Supersedes SP 676-II April 1984

U.S. DEPARTMENT OF COMMERCE, Malcolm Baldrige, Secretary
NATIONAL BUREAU OF STANDARDS, Ernest Ambler, Director

Issued April 1985

Preface

The purpose of this document is to guide the reader through the logical development of a measurement assurance program as it is intended to

i) Tie a measurement process or reference standards to the defined unit of measurement for the quantity in question or to national standards; and

ii) Establish the uncertainty of values reported by the process through the maintenance of statistical control of the measurement process.

The discussion is approached in the context of the assumption that the tie to defined units or national standards is accomplished via a tie to NBS. Participation in a measurement assurance program can satisfy this tie where systematic error is evaluated via measurements made in the participating laboratory on an NBS transfer standard and where it can be shown that the measurement process is continuously in a state of statistical control. This in no way implies that measurement assurance cannot be attained without formal participation in an NBS sponsored program, but the presentation is made more concrete in this context.

The formulation of measurement assurance techniques for all measurement situations is not within the scope of this document. Obviously, such matter is best handled on a subject basis. The dearth of suitable documentation for specific measurement disciplines serves as a motivating factor in the development of this guide which describes statistical procedures and analyses that are generally pertinent to measurement assurance. It is hoped that the reader will be able to adapt the philosophy and techniques contained herein to his own particular measurement needs.

The material in this document is largely statistical in nature because of the measurement assurance approach to quantifying both the random and systematic errors that are generated by a measurement process. It should be recognized, however, that measurement assurance is not achieved by statistical techniques alone but by the totality of procedures such as correct measurement practice, adherence to recommended procedures, control of environmental factors and estimation of process parameters that relate the output of the measurement system to national standards.

This document is the second part of a general treatise, Measurement Assurance Programs, which is divided into Part I: General Introduction and Part II: Development and Implementation. Part I by Dr. Brian C. Belanger is intended as a statement of the goals of measurement assurance from a managerial perspective and advances the basic philosophy of quality in measurement. Part II, which was supported by the NBS Office of Measurement Services, extends the principles so stated to specific measurement situations, drawing extensively on programs that were developed by Mr. Joseph Cameron in consultation with NBS technical divisions. In addition to these examples, measurement control programs with verifiable uncertainty statements are outlined for measurement situations that the author has encountered in consultations with the measurement community outside NBS.

Table I and Table II in this manuscript were compiled using the NBS software package DATAPLOT developed by Dr. James Filliben.

TABLE OF CONTENTS

LIST OF FIGURES

Measurement Assurance Programs
Part II: Development and Implementation

Carroll Croarkin
Statistical Engineering Division
Center for Applied Mathematics
National Bureau of Standards
Gaithersburg, MD 20899

This document is a guide to the logical development of a measurement assurance program in which the tie between a measurement and its reference base is satisfied by measurements on a transfer standard. The uncertainty of values reported by the measurement process is defined; and the validation of this uncertainty for single measurements is developed. Measurement sequences for executing the transfer with NBS and procedures for maintaining statistical control are outlined for eight specific measurement situations with emphasis on characterizing parameters of the measurement process through use of a check standard.

Key Words: Calibration; check standard; measurement assurance; random error; statistical control; statistical methods; systematic error; uncertainty.

1. The Development of a Measurement Assurance Program

1.1 Historical Perspective

The development of measurement assurance at the National Bureau of Standards, over the more than eighty years that the nation's premier measurement laboratory has been in existence, has evolved hand in hand with the NBS central mission of providing quality measurement. We might date this evolution as starting with the early experiments on the velocity of light [1][1]. Since then the principles of measurement assurance have reached realizations of all SI units and numerous derived units of measurement, and even now are influencing innovations in measurement science related to electronics and engineering.

As the reader familiarizes himself with the concepts of measurement assurance, he will come to realize that quality in calibration is dependent upon the inclusion of a check standard in the calibration scheme. The first application of this principle at NBS came in the area of mechanical measurements where a prescribed series of observations known as a weighing design, so called because of the obvious connection to mass weighings, defines the relationship among reference standards, test items and check standards. The first weighing designs published by Hayford in 1893 [2] and Benoit in 1907 [3] had no provision for a check standard, and the creation of suitable designs had to await general progress in the area of experimental design which characterized statistical activity at NBS in the nineteen fifties.

[1] The numbers in brackets refer to references cited at the end of this document.

As early as 1926 an NBS publication by Pienkowsky [4] referred to a standard one gram weight whose mass as "determined in the calibrations just as though it were an unknown weight" was used as a gross check on the calibrations of the other unknown weights. It remained until the nineteen sixties for the concept of measurement as a process to be described by repetitions on a check standard such as the one gram weight described by Pienkowsky. At that time calibrations of mass and length standards were formalized into measurement assurance programs with demonstrable uncertainty of reported values and statistical control of individual calibrations. A compendium of weighing designs for mechanical and electrical quantities with allowance for a check standard in each calibration sequence was published in 1979 (Cameron et al [5]).

Although many experimenters, past and present, have contributed to the quality of measurement science at NBS, the formulation of measurement assurance is the special province of the Statistical Engineering Division. Three members of this group, C. Eisenhart, W. J. Youden and J. M. Cameron, were largely responsible for fruition of the check standard concept, and the advent of electronic computers aided in the rapid application of this concept to NBS calibration programs. In 1962 a paper by Eisenhart [6] laid the groundwork for defining a repetition for a measurement process and assessing the uncertainties associated with such a process. This paper still serves as the primary treatise on the subject. Concurrently, Youden was implementing "ruggedness" testing in physical measurements [7], and at the same time he was introducing experimental design into interlaboratory testing [8].

In 1967 the first documentation of a measurement assurance approach appeared in print as an NBS monograph. The tutorial by Pontius and Cameron [9], treated the entire spectrum of mass measurement as a production process and began the dissemination of measurement assurance outside the NBS community. In the years since then, measurement assurance, both within and outside NBS, has been applied to basic SI units such as length as formulated in reference [10] and complex measurement areas such as dimensional measurements for the integrated circuit industry as formulated in reference [11]. Recently the measurement assurance approach has found its way into an ANSI standard for nuclear material control [12] with the use of "artifact reference mass standards as references for uranium hexafluoride" cylinders reported by Pontius and Doher [13].

1.2 Introduction

The development of a measurement assurance program evolves logically from the specific interpretation that we will give to the term "measurement assurance". The reader is asked to lay aside interpretations given to this term from previous experiences and to concern himself with what it means to have demonstrable scientific assurance about the quality of a measurement. For calibration activities, quality of a measurement is defined by its uncertainty, and the validity of an uncertainty statement for an individual measurement is guaranteed via the measurement assurance program as it is intended to

i) Tie a single measurement to a reference base; and

ii) Establish the uncertainty of the measured value relative to this reference base.

Firstly, in the case of basic SI units, a single measurement of a characteristic embodied in an object or artifact must be related to the defined unit for that quantity; for example, until recently the length of a gage block was defined relative to the wavelength of radiation of krypton 86 as realized through interferometry [14]. Because derived units of measurement can only be indirectly related to basic units, the measurement assurance concept is extended to such quantities by requiring that they be related to a reference base such as artifact standards or a measurement system maintained by the National Bureau of Standards. Secondly, a measurement assurance program must provide a means of maintaining statistical control over the measurement system thereby guaranteeing the validity of the uncertainty for a single measured value relative to its reference base (Cameron [15]).

The definition of measurement assurance is completed by an examination of the properties of measurement. A single measurement is properly related to national standards only if there is agreement between it and a value that would be achieved for the same quantity at NBS--meaning a value that would be arrived at from a sufficiently long history of like measurements at NBS. In actuality it is not possible to estimate the disagreement between a single measurement in a given laboratory and the long-term NBS value. However, if the measurement system of the laboratory is stable or as we say operating in a state of statistical control, the single measurement can be regarded as a random draw from another long history of measurements which also tend to a long-term value. The purpose of calibration is to eliminate or reduce the disagreement, referred to as offset, between a laboratory's long-term value for a measurement and the corresponding NBS long-term value by corrections to the measurement system and/or reference standards.

Where offset cannot be eliminated or reduced by calibration, it is a systematic error accruing to the laboratory's measurement system. Even where there is an accounting for such disagreement, the fact that NBS has imperfect knowledge about the long-term value from its own measurement system, based as it is on a finite though large number of measurements, means that the limits of this knowledge contribute another systematic error to the measurement system of the laboratory. In some special cases systematic and random errors that arise as NBS attempts to tie its measurement system to defined units of measurement may also become part of the systematic error for the laboratory.

The uncertainty that surrounds any single measurement describes the extent to which that single number could disagree with its reference base. The uncertainty includes all systematic errors affecting the measurement system; it also includes limits to random error that define the degree to which the individual laboratory, just as NBS, may be in error in estimating the long-term value for the measurement. Where the calculation of a long-term value for a measurement and limits to random error cannot be done directly, which is the usual case for calibration measurements, the long-term value is referenced to a long-term value of measurements made on an artifact(s) called a check standard.

Measurement assurance is attained when the determination of all sources of systematic error is coupled with statistical control of the measurement process as achieved by adapting quality control techniques to measurements on the check standard. Statistical control consists of comparing current check standard measurements with the value expected for such measurements and making decisions about the condition of the process based on the outcome of this test. The establishment of suitable check standards and implementation of statistical control procedures are discussed in the next two chapters with implementation for specific cases being outlined in chapter 4.

The determination of systematic error is made by intercomparing the laboratory's reference standard(s) or measurement system with national standards or a measurement system maintained by the National Bureau of Standards. This intercomparison can be interfaced with NBS in one of three ways. Firstly, the reference standards can be submitted to the usual calibration exercise wherein values and associated uncertainties are assigned to the reference standards by NBS. The only sources of systematic error that are identifiable in this mode are directly related to the reference standards themselves and to the NBS calibration process. The name "measurement assurance program" is not formally attached to such efforts because the NBS involvement is limited and measurement control is left entirely to the participant, but the goal of measurement assurance is certainly realizable by this route.

Secondly, systematic error can be identified by internal calibration of instrumentation or reference standards through use of a standard reference material distributed by NBS. Thirdly, systematic error can be determined by a formal program in which an NBS calibrated artifact, called a transfer standard, is treated as an unknown in the participant's measurement process. The difference between the participant's assignment for the transfer standard and the NBS assignment determines the offset of the participant's process or reference standards from NBS.

The National Bureau of Standards provides measurement assurance related services that utilize the latter two courses, especially the use of transfer standards, in selected measurement areas [16]. A standard reference material and a transfer standard are comparable in the measurement assurance context. The latter is referred to more frequently in this publication because transfer standards are more publicized in connection with measurement assurance.

The development of a program which satisfies the goals of measurement assurance begins with the measurement problem which must be related to physical reality by a statement called a model. Models covering three aspects of metrology are discussed in this chapter. The first of these, the physical model, relates the realization of the quantity of interest by a measurement process to the fundamental definition for that quantity. Physical models change with changes in fundamental definitions.

For example, until 1960, the standard of length was "the distance between two scratch marks on the platinum-iridium meter bar at the Bureau International des Poids et Mesures" [17]. Models for realizing length related the intercomparison between the international meter bar and the national meter bar and subsequent intercomparison between the national meter bar and gage block standards. In 1960 length was redefined in terms of the wavelength of radiation of krypton-86. The defining wavelength of 86Kr was related to the wavelength of a stabilized laser light[a], thus establishing the relationship of interference fringe patterns observed with the laser interferometer to the length of gage blocks standards. Length has recently been redefined in terms of the velocity of light. This latest change will necessitate another model relating standards to "the length of the path traveled by light in a vacuum during a (given) time interval" [17].

The calibration model describes the relationship among reference standards, items to which the quantity is to be transferred such as secondary or laboratory reference standards, and the instrumentation that is used in the calibration process. For example, calibration of gage blocks by electromechanical intercomparison with gage block standards that have been measured interferometrically includes a correction for the temperature coefficient of blocks longer than 0.35 inches [19]. The calibration model for these intercomparisons assumes a constant instrumental offset that is canceled by the calibration experiment as discussed in section 1.4.

Statistical models further refine the relationship among calibration measurements in terms of the error structure. Section 1.5 describes the type of error structure that is assumed for measurement assurance programs taking as an example the eletromechanical comparison of gage blocks according to the scheme outlined in section 4.

Modeling, usually the responsibility of the national laboratory, is emphasized in this chapter partly to lay the foundation for the remainder of the text and partly so that the reader can form some idea of the degree of success that can be expected from a measurement assurance program. It is an implicit assumption that the validity of any intercomparison, either between transfer standards and reference standards or between reference standards and the workload, depends upon all items responding to test conditions in fundamentally the same way as described by the models.

---

[a] "Direct calibration of the laser wavelength against 86Kr is possible, but is relatively tedius and expensive. The procedure used is a heterodyne comparison of the stabilized He-Ne laser with an iodione stabilized laser" (Pontius [18]).

This logic leads us the next major phase in development--the test of the measurement prescription as a device for transferring a quantity of measurement from the national laboratory to a laboratory participating in a measurement assurance program. The final phase--the application of quality control techniques to the measurement process ensures a continuing tie to the national system of measurement. Several activities can take place during each of these phases. These are listed either in section 1.6 under the role of NBS or in section 1.7 under the role of the participant although it is clear that in practice there is some overlapping of these responsibilities.

In summary, measurement assurance implies that the determination of systematic error and the assignment of values to standards has been done correctly at every step in the measurement chain and, moreover, that this is guaranteed by a statistical control program that is capable of identifying problem measurements at every transfer point in the chain. Accomodation to these principles may require modification of the laboratory's calibration procedures. Where an NBS transfer standard is used to directly calibrate reference standards, the same measurement process and control procedures that are used for this intercomparison should be used for the regular workload. Where the transfer standard is used to calibrate a laboratory's primary standards, a statistical control program should be implemented for this intercomparison, along with similar control programs for the intercomparison of the primary standard with the reference standards and for the intercomparison of the reference standards with the workload. Obviously, the effort required to maintain such a system is greater than is required to maintain a current calibration on the reference standards. Measurement assurance places a substantial portion of the burden of proof on the participant, where it should rightfully be, because it is the quality of his measurements that is of ultimate interest.

1.3 Models for a Measurement System

A measurement system that relies on an artifact demands that the artifact play "two essential roles in the system; it must embody the quantity of interest, and it must produce a signal, (such as the deflection of a pointer on a scale or an electrical impulse) which is unambiguously related to the magnitude or intensity of the specified quantity" (Simpson [20]). The first step that must be undertaken in constructing a measurement system is to reduce the artifact to an idealized model which represents those properties believed to be pertinent to the intended measurement.

This model of the measurement process, based on the laws of physics, in the broadest sense embodies our understanding of the physical universe. It is usually a software model or statement that relates the signal produced by the artifact and all procedures used to produce the desired measured value, called the measurement algorithm, to the realization of the physical quantity of interest taking into account any factors such as environmental conditions that affect this realization.

The integrated circuit industry is a case study of a measurement problem not properly defined in terms of an artifact model. Inability throughout the industry to measure optically the widths of chromium lines to the accuracies needed for producing photomasks for integrated circuits can be traced to misconceptions about the nature of linewidth measurements--misconceptions that led to reliance on a line-scale calibration for making such measurements, in the hope that a correct line-scale for the optical system would guarantee accurate linewidth measurements.

Before attempting to produce a linewidth standard, NBS explored the nature of the systematic errors that are inherent in line-scale and linewidth measurements (Nyyssonen [21]). Line-scale defines the internal ruler of an instrument i.e. it is basically a left-edge to left-edge or a right-edge to right-edge measurement for which any bias in detecting edge location is assumed to cancel out. Linewidth, a more difficult determination, measures the width of a physical object, in this case a chromium line. It is a left-edge to right-edge measurement in which any bias in detecting edge location is assumed to be additive (Jerke [22]).

This theoretical modeling was corroborated by an interlaboratory study which demonstrated that an optical imaging system, although properly calibrated for line-scale, would not necessarily produce linewidth measurements with negligible systematic errors. The study also demonstrated that the same system when properly calibrated using a linewidth artifact would produce linewidth measurements with negligible systematic errors (Jerke et al [23]).

A model is never complete or perfect, and the difference between the model and reality leads to "a particular type of systematic error which exists if the measurement algorithm is flawless. Failure to recognize this fact can lead to major wastes of resources since no improvement in the measurement algorithm can reduce this error" (Simpson [24]).

Thus even though NBS semiconductor research has greatly enhanced linewidth measurement capability, the accuracy of linewidth measurement is still constrained by the difference between the real edge profile of a chromium line and a theoretical profile (Nyyssonen [25]) upon which the model depends. The discrepancy between the edges of chromium lines on production photomasks and the theoretical model is a limiting factor in attaining measurement agreement among photomasks makers, and it will not be reduced by finer tuning of the optical imaging systems or more accurate standards. This points out a problem that exists in going from the calibration laboratory with carefully fabricated artifacts to the production line and prompts us to include a caveat for the claims of measurement assurance programs. This type of systematic error is kept at an acceptable level only if the measured items are close in character to the standards and theoretical model on which their assignments depend. The only strategy which can reduce model ambiguity identically to zero uses objects called "prototypes" and, in effect, takes a particular object and defines it to be its own model. As pointed out by Simpson [26],

> This amounts to saying that this object is the perfect and complete realization of the class of objects to which it belongs, and hence the model ambiguity is, by definition, identically zero. The only SI unit still using this strategy is mass where the Paris[b] Kilogram is the kilogram of mass, and the only objects where mass can be unequivocally defined are one kilogram weights made of platinum.

The comparison of a non-platinum kilogram with the Paris kilogram would produce a systematic error unless the comparison was done in vacuum. High accuracy mass calibrations in air are corrected for air buoyancy — a correction that depends on the material properties of the weight, temperature on the weight at the time of weighing and the local pressure and humidity. Any ambiguity between the model that drives this correction and the Paris kilogram in vacuum contributes a systematic error to the calibration process although admittedly this error is negligible.

1.4 Models for a Calibration Process

1.4.1 The Calibration Experiment

The exploration of the physical and mathematical models that relate a measurement to a quantity of interest leads to a measurement algorithm which defines a reference standard, instrumentation, environmental controls, measurement practices and procedures, and computational techniques for calibrating other artifacts or instruments with respect to the desired property.

> Calibration is a measurement process that assigns values to the response of an instrument or the property of an artifact relative to reference standards or measuring processes. This may involve determining the corrections to the scale (as with direct-reading instruments), determining the response curve of an instrument or artifact as a function of changes in a second variable (as with platinum resistance thermometers), or assigning values to reference objects (as with standards of mass, voltage, etc.) (Cameron [27]).

Calibration consists of comparing an "unknown" or test item which can be an artifact or instrument with reference standards according to the measurement algorithm. The calibration model, which addresses the relationship among measurements of test items and reference standards, must reflect the fact that the individual readings on the test items and reference standards are subject to systematic error that is a function of the measuring system and random error that may be a function of many uncontrollable factors.

---

[b]The international standard of mass resides at the Bureau International des Poids et Mesures in Sèvres, just outside Paris.

There are two common generic types of calibration models, additive models and multiplicative models. Reduction of systematic error by intercomparison with a reference standard involves estimating offset as either an additive factor Δ or a scale factor λ which in turn is used to assign a value to the test item relative to the known value of the reference standard. The choice of an additive or multiplicative model depends on the nature of the relationship among test items and reference standards and properties of the measuring system.

The calibration experiment is designed not only to assign values to test items that will account for systematic error between the requestor and the calibrator but also to estimate the magnitude of random errors in the calibration process. The nature of random error is discussed more fully in section 2.2, but suffice it to say for now that we are talking about small fluctuations that affect every measurement but are unmeasurable themselves for a given measurement. The statistical derivations in this manuscript assume that the random errors are independent and that they affect the measuring process symetrically i.e., that one is not predictable in size or direction from any other one and that the chances are equal of the resulting measurement being either too large or too small. It is also assumed that random errors for a given process conform to a law called a statistical distribution; quite commonly this is assumed to be the normal distribution, and the calibration experiment is designed to estimate a standard deviation which describes the exact shape of this distribution.

In the next three sections we list models that are in common usage in calibration work, and although the list is not exhaustive, it includes those models which form the basis for the calibration schemes in chapter 4. It is noted that the term "reading" or "measurement" in this context does not refer to a raw measurement, but rather to the raw measurement corrected for physical model specifications as discussed in the last section.

1.4.2 Models for Artifact Calibration[c]

In the simplest additive model for a calibration process, a test item x with a value X^*, as yet to be determined, and a reference standard R with a known or assigned value R^* are assumed to be related by:

$$X^* = \Delta + R^* \qquad (1.4.1)$$

where Δ is small but not negligible. The method for estimating the offset Δ between the two artifacts depends upon the response of the calibrating instrument.

If the calibrating instrument is without systematic error, the instrument response x for any item X will attain the value X^* except for the effect of random error; i.e., the instrument responds according to the model

$$x = X^* + \varepsilon$$

[c] The models for artifact calibration are also appropriate for single-point instrument calibration.

where ε represents the random error term. In this case there is no need to compare the test item with a reference standard because the capacity for making the transfer resides in the calibrating instrument. Such is assumed to be the case for direct reading instruments. Normally the calibrating instrument is not invested with such properties, and one calibration approach is to select a reference standard that is almost identical to the test item and compare the two using a comparator type of instrument for which additive instrumental offset is cancelled out in the calibration procedure. Given that the comparator produces a measurement x on the test item and a measurement r on the reference standard, the response is assumed to be of the form:

$$x = \psi + X^* + \varepsilon_x$$

and (1.4.2)

$$r = \psi + R^* + \varepsilon_r$$

where ψ is instrumental offset and the ε_x and ε_r are independent random errors. An estimate[¶] of Δ is gotten by the difference

$$\hat{\Delta} = x - r \ , \tag{1.4.3}$$

and the value of the test item is reported as

$$\hat{X}^* = \hat{\Delta} + R^* \ .$$

An inherent deficiency in relying on a single difference to estimate Δ is that it does not admit a way of assessing the size of the random errors. If the calibration procedure is repeated k times in such a way that the random errors from each repetition can be presumed to be independent, the model for k pairs of readings r_j, x_j $(j=1,\cdots,k)$ becomes

$$x_j = \psi + X^* + \varepsilon_{x_j}$$

$$r_j = \psi + R^* + \varepsilon_{r_j} \tag{1.4.4}$$

and the offset is estimated by

$$\hat{\Delta} = \frac{1}{k} \sum_{i=1}^{k} (x_i - r_i) \ . \tag{1.4.5}$$

Given the further assumption that all the random errors come from the same distribution, the magnitudes of the random errors can be quantified by a standard deviation (see Ku [28] for a clear and concise discussion of standard deviations).

Another less frequently assumed response for a calibrating instrument allows not only for instrumental offset ψ but also for a non-constant error that depends on the item being measured. This type of response is sometimes referred to as non-linear behavior, and in this case two reference standards with known values R_1^* and R_2^* are required to estimate X^*. Given measurements r_1 on the first standard and r_2 on the second standard, the instrument response for the three artifacts is described by:

[¶] The caret (^) over a symbol such as Δ denotes an estimate of the parameter from the data. It is dropped in future chapters where the intent is obvious.

$$x = \psi + \beta X^* + \varepsilon_x$$

$$r_1 = \psi + \beta R_1^* + \varepsilon_{r_1} \qquad (1.4.6)$$

$$\text{and} \quad r_2 = \psi + \beta R_2^* + \varepsilon_{r_2}$$

where the parameter β is non-trivial and different from one, and ε_x, ε_{r_1} and ε_{r_2} are independent random errors.

Then the measured differences $x-r_1$ and r_2-r_1 are used to construct an estimate of Δ, namely,

$$\hat{\Delta} = (R_2^* - R_1^*)\cdot(x - r_1)/(r_2 - r_1). \qquad (1.4.7)$$

The calibrated value of the test item is reported as

$$\hat{X}^* = \hat{\Delta} + R_1^*. \qquad (1.4.8)$$

Equivalently, Δ can be estimated by

$$\hat{\Delta} = (R_1^* - R_2^*)\cdot(x - r_2)/(r_1 - r_2)$$

in which case

$$\hat{X}^* = \Delta + R_2^*.$$

In order to achieve symmetry in the use of the reference standards, before and after readings, x_1 and x_2, can be taken on the test items with the readings in in order x_1, r_1, r_2, and x_2. Then Δ is estimated by

$$\hat{\Delta} = \frac{1}{2}(R_2^* - R_1^*)\cdot(x_1 - r_1 - r_2 + x_2)/(r_2 - r_1), \qquad (1.4.8a)$$

and the value for the test item is given by

$$\hat{X}^* = \hat{\Delta} + \frac{1}{2}(R_1^* + R_2^*).$$

In comparing the models in (1.4.2) and (1.4.6) one sees that the former model amounts to the slope β of the response curve of the instrument being identically one. If this slope is in fact close to one, which is certainly a reasonable assumption for most instruments, any departure from this assumption will contribute only a small systematic error in the assignment to the test item because of the small interval over which the measurements are taken. For this reason (1.4.2) is the commonly accepted model for calibration processes that use a comparator system of measurement.

The model in (1.4.6) amounts to a two-point calibration of the response function of the instrument; it is not dependent on a small calibration interval; and it is commonly used for direct-reading instruments. Notice that for either model a valid calibration for the test item does not depend on the response parameters of the instrument as long as they remain stable.

A multiplicative model for calibration assumes that the test item X and the reference standard R are related by

$$X^* = \gamma R^* \tag{1.4.9}$$

and that the measuring instrument has a response function of the form

$$x = \beta X^* + \varepsilon_x \tag{1.4.10}$$

$$r = \beta R^* + \varepsilon_r$$

where β and ε_x and ε_r are defined as before. The model leads to an estimate of γ; namely,

$$\hat{\gamma} = x/r \ . \tag{1.4.11}$$

The calibrated value of the test item is reported as

$$X^* = \hat{\gamma} R^* . \tag{1.4.12}$$

1.4.3 Models for Instrument Calibration

Models for instrument calibration relate the response of the instrument to a known stimulus called the independent variable. Where non-constant response of the instrument over a range of stimuli can be either theoretically or empirically related to the stimulus, the relationship is called a calibration curve.

The model for a calibration curve assumes that a response X is offset from a known stimulus W by an amount $\Delta(W)$ that depends on W and that the relationship holds over the entire calibration interval within a random error ε. A relationship of the form

$$X = \alpha + \beta W + \varepsilon \tag{1.4.13}$$

where α and β may be unknown parameters is called a linear calibration curve.

Once the parameters of the calibration curve are known or have been estimated by an experiment, future responses can be related back to their corresponding stimuli. In the general case this inversion is not easy nor is the attendant error analysis very tractable because the calibration curve is used in the reverse of the way that the data are fitted by least-squares.

The only case where the solution is straightforward is the linear case where a series of readings $X_j (j=1,\cdots,n)$ at designated points W_j^* $(j=1,\cdots,n)$ are used to obtain estimates $\hat{\alpha}$ and $\hat{\beta}$ of the parameters. The best estimate of offset for the linear case is

$$\hat{\Delta}(W) = \hat{\alpha} + \hat{\beta}(W). \tag{1.4.14}$$

Methods for estimating the parameters and quantifying the random error are discussed by Mandel [29].

1.5 Models for Error Analysis

The models in sections 1.4.2 and 1.4.3 admit random errors that come from a single error distribution whose standard deviation is of interest in quantifying the variability in the calibration process. We now expand this concept to models that include two distinct types of random errors; a random error term for short-term repetitions that is usually attributed to instrument variability and a random error term that allows for changes that are dependent on the conditions of the calibration and as such are assumed to remain constant for a given calibration. These two types of errors give rise to two distinct error distributions with associated standard deviations which can be estimated from the calibration data. The former is usually referred to as a "within" standard deviation and is designated by s_w.

The latter referred to as a "between" standard deviation, meaning between calibrations and designated by s_b, is attributed to changes in the calibration process from day-to-day. These include environmental changes that are not accounted for by modeling, changes in artifact alignment relative to the standard, and other fluctuations that are not reflected in the within standard deviation. For example, the model in (1.4.4) can be rewritten in terms of measured differences d_j $(j=1,\cdots,k)$ as

$$d_j = x_j - r_j = X^* - R^* + \varepsilon_j \qquad (1.5.1)$$

where the subscript j denotes short-term repetition and the ε_j are independent random errors that come from a distribution with standard deviation s_w. When this model is expanded to allow for day-to-day changes, the model becomes

$$d_j = (X^* + \delta_X) - (R^* + \delta_R) + \varepsilon_j \qquad (1.5.2)$$

where δ_X and δ_R are assumed to be independent random errors that come from a distribution with standard deviation s_b.

The quantities s_w and s_b, while of interest in their own right, are components of a "total" standard deviation that includes both "within" and "between" type variations in the measurement process. It is this total standard deviation, whose structure is discussed at length in this and later chapters, that is of primary interest in measurement assurance. The reader can verify that the proposed approach to error modeling is compatible with a components of variance model [30] by considering model (1.5.2) which leads to the estimate of offset given in (1.4.5). In terms of the error structure this offset is

$$\hat{\Delta} = (X^* - R^*) + (\delta_X - \delta_R) + \frac{1}{k}\sum_{j=1}^{k} \varepsilon_j \quad .$$

It can be shown‡ that a reported value based on a single (k=1) measured difference has standard deviation

$$s_r = (2s_b^2 + s_w^2)^{1/2} \quad .$$

‡The methodology for arriving at the standard deviation is not explained in this publication. See Ku [28], pages 312-314, for the computation of standard deviations when several independent errors are involved.

A reported value based on the average of k short-term differences has standard deviation

$$s_r = (2s_b^2 + s_w^2/k)^{1/2}.$$

Notice that the contribution of the component s_b to the standard deviation s_r is not reduced by taking multiple measurements that are closely spaced in time. This is the reason for discouraging short-term repetitions in measurement assurance and insisting that the definition of the total standard deviation encompass a broad range of operating conditions in the laboratory--implications which will be addressed in some detail in later chapters.

In this manuscript the total standard deviation s_c is defined to be the standard deviation of a "check standard" value as estimated from repeated calibration of the check standard. Where the error structure for the check standard value is the same as the error structure for the reported value of the test item, the standard deviation of the reported value which we call s_r, is exactly s_c. Otherwise, s_r must be adjusted accordingly. For example, suppose that a test item X with unknown value X^* is compared with two reference standards R_1 and R_2 with known values R_1^* and R_2^* by consecutive readings x_1, r_1, r_2, x_2 as described in section 4.2.

The error model for the measured differences

$$d_1 = x_1 - r_1$$

and

$$d_2 = x_2 - r_2$$

can be written as

$$\begin{aligned} d_1 &= (X^* + \delta_1) - (R_1^* + \delta_2) + \varepsilon_1 \\ d_2 &= (X^* + \delta_3) - (R_2^* + \delta_4) + \varepsilon_2 \end{aligned} \qquad (1.5.3)$$

where it is assumed that δ_1, δ_2, δ_3 and δ_4 have standard deviation s_b and ε_1 and ε_2 have standard deviation s_w.

The offset is estimated by

$$\hat{\Delta} = \frac{1}{2}(d_1 + d_2) \qquad (1.5.4)$$

and in terms of the error model

$$\hat{\Delta} = X^* - \frac{1}{2}(R_1^* + R_2^*) + \frac{1}{2}(\delta_1 - \delta_2 + \delta_3 - \delta_4 + \varepsilon_1 + \varepsilon_2). \qquad (1.5.5)$$

A check standard defined as the difference between R_1 and R_2 is computed for each calibration by

$$c = (d_2 - d_1). \qquad (1.5.6)$$

In terms of the errors the check standard measurement can be written

$$c = (R_1^* - R_2^*) + (-\delta_1 + \delta_2 + \delta_3 - \delta_4 - \varepsilon_1 + \varepsilon_2) \qquad (1.5.7)$$

The error model (1.5.5) for the reported value

$$X^* = \hat{\Delta} + \frac{1}{2}(R_1^* + R_2^*), \qquad (1.5.8)$$

and the error model (1.5.7) for the check standard measurment c are comprised of the same error terms and differ structurally by a factor of two.

Explicitly, the standard deviation of the reported value X^* is

$$s_r = \frac{1}{2}(4s_b^2 + 2s_w^2)^{1/2} \qquad (1.5.9)$$

and the standard deviation of c is

$$s_c = (4s_b^2 + 2s_w^2)^{1/2}. \qquad (1.5.10)$$

Therefore,

$$s_r = \frac{s_c}{2} \qquad (1.5.11)$$

In practice s_c is estimated by check standard measurements from many calibrations (see chapter 4), and this estimate is used in (1.5.11) to compute s_r.

Where the check standard value is a least-squares estimate from a design or a function of measurements on more than one artifact, the computation of the standard deviation of a reported value is more complicated. In such a case, one must first estimate s_w from a single calibration and compute s_b from an equation for s_c such as (1.5.10). Then the standard deviation of the reported value can be computed from an equation such as (1.5.9).

1.6 NBS Role in the Development of a Measurement Assurance Program

1.6.1 Study of Operations at Participating Laboratories

Before undertaking the development of a measurement assurance program for disseminating a unit of measurement, NBS technical staff familiarize themselves with operations at potential user laboratories so that the program can be structured around the equipment, facilities and personnel available to the laboratories. Suggestions for equipment modifications and additions are made at this time. The range of operating conditions in the participating laboratories is checked for consistency with the model, and in order to determine whether or not the accuracy goals of the measurement assurance program are attainable, NBS is advised of the individual laboratory's measurement requirements and capabilities.

1.6.2 Identification of Factors Capable of Perturbing the System

It is the responsibility of NBS to identify and isolate those factors capable of seriously disrupting the measurement system so that equipment and procedures can be designed to offset the impact of such factors (Youden [31]). This is particularly important if the measurement assurance program is intended for an industrial setting rather than a controlled laboratory setting.

An example of this type of testing, called "ruggedness" testing is found in the NBS flowmeter program for liquids (Mattingly et al [32]). The effects of three types of perturbation on turbine meters were studied experimentally, and it was found that the profile of the flow entering the meter has a significant effect on meter performance. This research led to the development of a flow conditioner which can be inserted in an upstream section of pipe to regulate the profile of the flow entering the meter. Because flow profiles vary from laboratory to laboratory depending on the source of the flow, such a flow conditioner is appended to the turbine meters that are circulated in the industry as NBS transfer standards.

1.6.3 Design of Interlaboratory Exchanges

The purpose of the interlaboratory study or round-robin test that is usually sponsored by NBS at the inception of a measurement assurance program is to determine the extent and size of offsets from NBS that are typical in the target industry. Secondary goals are the evaluation of the adequacy of proposed procedures for resolving the measurement problem, critique of the format and content of directions from NBS, and study of the ease of implementation on the part of participants. Frequently a preliminary interlaboratory test designed to identify significant problem areas is followed by a more comprehensive study which incorporates modifications to artifacts and protocols based on experience gained in the preliminary test.

1.6.4 Development of a Stable Transfer Standard or Standard Reference Material

Either a standard reference material or a transfer standard is developed for each measurement assurance program that is sponsored by NBS. The standard reference material (SRM) is a stable artifact produced either commercially or in-house that is calibrated, certified and sold by NBS in fairly large numbers.[d] Standard reference materials are well known for chemical applications. Recently NBS has certified two separate dimensional artifact standards as SRMs, one a linewidth standard for the integrated circuit industry [NBS SRM-474] and the other a magnification standard for scanning electron microscopes [NBS SRM-484]. An SRM has the unique property that it can be used not only for determining offset from NBS but also as an in-house standard for controlling the measurement process.

[d] A listing of SRM's is contained in the catalog of NBS Standard Reference Materials, NBS Special Publication 260, 1979-80 Edition, available from the Office of Standard Reference Materials, NBS, Gaithersburg, MD.

The transfer standard is a calibrated artifact or instrument standard that is used for disseminating the unit of measurement. It is loaned to the participant to be intercompared with the participant's standards or instrumentation under normal operating conditions in order to determine offset from NBS.

Artifacts that are stable with relation to a physical quantity, such as the mass of an object, do not usually pose any special problems when they are used as transfer standards because they can be shipped from one place to another without a change in the quantity of interest. Transfer standards that are not easily transported are packaged in environmentally controlled containers, but additional redundancy in the form of multiple standards and observations is always included in the measurement assurance program whenever the stability of the transfer standard is in question.

1.6.5 Dissemination of Measurement Technology and Documentation

The participant in a measurement assurance program is entitled to draw upon the expertise and experience that resides in the sponsoring NBS technical group. Technical assistance is disseminated by way of NBS publications, ASTM, standards, ANSI standards, laboratory visits, telephone conversations and NBS sponsored seminars. In conjunction with the advent of a new program a series of seminars is usually offered to the public to explain the philosophy, theory, measurement technology and statistical analyses which form the basis for a measurement assurance program in that discipline.

Documentation for standard reference materials is available through NBS Special Publication Series 260. As part of a long range plan to upgrade its calibration services, the National Bureau of Standards has instituted documentation requirements for all calibration services. Documentation includes theory, laboratory setup and practice, measurement technique, maintenance of standards, specification of measurement sequence, protocol for measurement control and determination of final uncertainty. When these publications become available, they will provide the bulk of the documentation that is needed for implementing a measurement assurance program that is related to an NBS calibration service. Insofar as a measurement assurance program as implemented by the participant may differ from the NBS calibration program in regard to the number of standards, specification of measurement sequence, corrections for environmental conditions, estimation of process parameters, and methods for determining offset and uncertainty, additional user oriented documentation may be made available.

1.6.6 Establishment of Measurement Protocol for Intercomparisons with NBS

Measurement assurance programs currently in existence fall into two categories. The first category contains those services which are highly structured for the participant, with regard to the number of laboratory standards to be employed in the transfer with NBS, the number of repetitions to be made in the exchange, and the protocol to be used for establishing an in-house measurement control program. At this time only the Gage Block Measurement Assurance Program (Croarkin et al [33]) and the Mass Measurement Assurance Program fall into this category.

All other programs allow the participant considerable leeway in regard to the items mentioned above in order to make the service compatible with the unique situation in each laboratory. The advantage of operating within the constraints of equipment and staff resources that are already allocated to the laboratory's normal workload is obvious, especially where accuracy requirements are not difficult to meet. However, there are drawbacks. The data analysis must be tailored to each participant, imposing an additional burden on NBS staff, and responsibility for instituting a rigorous measurement control program is left entirely to the participant.

1.6.7 Data Analyses and Determination of Offset

The determination of offset and associated uncertainty as realized by intercomparison of laboratory reference standards with NBS transfer standards is accomplished in one of two ways:

i) The transfer standard(s) is sent to the participant as a blind sample, and the data from the intercomparison are transmitted to NBS. Based upon the value assigned to the transfer standard by NBS and associated uncertainty from the NBS process, new values with associated uncertainties are assigned to the laboratory standards along with the uncertainty that is appropriate for an item measured by the participant's process.

ii) the transfer standard along with the its assigned value and associated uncertainty are transmitted to the participant, and the analyses and determination of offset become the responsibility of the participant.

Data analyses relating to the regular workload and measurement control procedures in a laboratory are best left to the individual participant. These analyses provide important insights into the pecularities of a measurement process, and, consequently, these analysis are best done internally. Even where much or all of the data analysis is undertaken by NBS, participants are encouraged to develop facility in this area in order to make themselves independent from NBS in the future. Some participants in measurement assurance programs have automated the analysis of calibration data, decisions relating to process control, updating of data files and final determination of uncertainty on minicomputers in their laboratories.

1.7 Participant's Role in a Measurement Assurance Program

1.7.1 Staff Preparation

The success of a properly conceived measurement assurance program depends upon the enthusiasm and dedication of the personnel who are making the measurements and resolving problems that arise in day-to-day operations. The measurement assurance approach is a long-term commitment in terms of evolving a measurement control technique that continually checks on the state of control of the process. Before undertaking such a program, there should be reasonable assurance of continuity of personnel assigned to the project, and steps should be taken to guarantee that new personnel are sufficiently prepared for taking on the assignment before the departure of experienced personnel.

The success of such a program also depends on a certain depth of understanding on the part of the staff. Here we are talking not so much about the intricacies of a particular analysis, but about a basic understanding of scientific methodology, the philosophy of measurement assurance, and the relationship between the control techniques and the validity of the values reported by the measurement process and their associated uncertainties. To this end, NBS offers seminars in which the attendees are instructed in these principles, but some prior staff preparation may be necessary in order to benefit fully from these expositions. Courses at local community colleges are recommended for exploring scientific principles and gaining facility with fundamental mathematical and statistical manipulations.

1.7.2 Selection of a Check Standard

The selection of a check standard must be considered in the preliminary planning for measurement assurance program. In short, its purpose is to provide a continuing thread that characterizes the operation of the measurement process over changing laboratory conditions and over time with regard to both the variability of the process and the long-term average of the process. It is a basic tenet of measurement assurance that the response of the process to the check standard be sufficiently similar to the response of the process to the test items that the performance of the process at all times can be adequately monitored by monitoring the response of the process to the check standard. The value of the check standard at any given time is a decision-making tool, and unexpected behavior on its part is grounds for discontinuing the process until statistical control is resumed.

Careful consideration should be given to the type of artifact that would be suitable for this purpose. It should certainly be of the same character as the items that constitute the workload in the laboratory. For some processes, such as processes dealing with basic units of measurement, the selection is obvious; check standard artifacts are similar to reference standards in design and quality. In general, an artifact that is less stable than the reference standards will not be useful as a check standard if its instability is large enough to mask the properties of the measurement process.

The check standard should be thought of not so much as an artifact but as a data base because it is the measurements that are of interest and not the artifact per se. The check standard data base consists of measurements, properly corrected for environmental factors, or some function of those measurements that have been made on the artifact check standard or on the reference standards. For example, a test item that is compared to two reference standards has its assignment based on the average of the values assigned to the two reference standards. The check standard can be defined to be the difference between the measurements on the reference standards thus eliminating the need for an extraneous measurement or other artifact. Where a calibration involves only one reference standard, an artifact that is similar in response to the test items can be designated as the artifact check standard. This need not be a calibrated artifact, and the properties of the measurement process are ascribed to it as long as it is measured in the same time frame as the other items in the calibration process. Several check standards used separately or in combination may be employed when the stability of the reference standards, such as a bank of standard cells, is cause for concern.

Where reference standards exist at several levels, such as mass standards or length standards, check standards are maintained and monitored at each level. Where the quantity of interest is propagated over several levels from one standard such as a one ohm resistor, which is used to propagate resistances between one and ten ohms, the same check standard artifact may be employed at the different levels, but the data bases for the different levels are regarded as separate check standards.

An SRM makes an ideal check standard if it is not contaminated or otherwise degraded by heavy usage. In any case the artifact or artifacts on which the check standard base is built must be readily available to the measurement process over a long period of time.

The proliferation of check standards involves no small amount of work in maintaining the data base, and serious thought should be given to placement of check standards in the measurement echelon. For a new program, one should start with check standards at a few critical points and gradually increase these as experience is gained with the program.

1.7.3 Initial Experiments to Estimate Process Parameters

The establishment of an initial data base for the laboratory's check standards is the first order of business in a new measurement assurance program. Before one attempts to quantify offset, it must be demonstrated that a measurement process does in fact exist; i.e., that measurements from the process satisfy the requirements for statistical control. This presupposes that the process precision is well known and that this can be documented. If, in fact, the documentation of the process has been lax, or if a substantially new process has been instituted for the measurement assurance program, then measurements taken over as long a time period as practical should be made on the check standard(s) in order to estimate the long-term average of the process and the standard deviation. Procedures for obtaining these initial estimates are discussed in subsequent chapters.

A laboratory planning a transfer with NBS should undertake these experiments well in advance of the arrival of the NBS transfer standard so that any problems encountered in the measuring system can be rectified. This provides a shake-down period for procedures, equipment and software involved in the measurement assurance program. Once the transfer standards are intercompared with the laboratory's reference standards, the resulting measurements involving the check standard are compared with the initial data base to decide if the process is in control at that time, and the transfer between the laboratory process and the NBS process is accomplished only if the process is judged in control. Therefore, participants are urged to make the initial experiments as representative of laboratory conditions as possible and to request help from the sponsoring NBS group if measurement problems or procedural ambiguities exist so that delays with the transfer can be avoided.

1.7.4 Calibration Procedures

Accomodation to measurement assurance principles can mandate a change in calibration procedures within the laboratory. Most often such change will amount to additional redundancy in the design and/or change in the order of measurements. The laboratory should settle upon one calibration design for the

transfer with NBS and the calibration workload. There is considerable advantage in doing this because the uncertainty determined from the transfer with NBS is only valid for that measurement process, and if the uncertainty is to have validity for the workload, the two measurement processes must be identical. There is a further advantage; the same statistical control program will suffice for both processes, and the check standard measurements from both sources can be combined into a single data base.

Another consideration is the manner in which systematic error is handled in the transfer experiment. Some measurement assurance programs are structured so that the determination of systematic error is made relative to the average of two or more reference standards as in section 4.2.4. For example, two reference gage blocks can be calibrated by intercomparison with two NBS transfer blocks by a design that assigns values relative to the average of the two reference blocks called the restraint. Systematic error is estimated as the difference between the restraint and the average computed for the two NBS blocks by the transfer experiment. The laboratory's restraint is then corrected for this offset. Meaningful values cannot be computed for the reference standards individually from the transfer experiment. Thus, the same design that is used for the transfer with NBS is employed in the calibration workload so that all assignments are made relative to the corrected restraint.

1.7.5 Process Control

The measurement assurance concept demands that a value be assigned to an artifact only when the measurement process is in control in order to guarantee the validity of the assignment and associated uncertainty statement. This means that statistical control is employed in the everyday workload of the laboratory as well as during the transfer with NBS. For highest accuracy work, comparable to calibrations at NBS, a check for control is made during every measuring sequence in which an artifact is calibrated by the system. Statistical control procedures based on check standard measurements along with the appropriate statistical tests are discussed in section 3.3.

The choice of a control procedure and its implementation are the responsibility of the participant. Those who are familiar with industrial quality control procedures and Shewhart type control charts should be able to adapt these methodologies to check standard measurements. A general discussion of control charts with examples is contained in chapter 5, and statistical control procedures for specific measurement situations are outlined in chapter 4.

1.7.6 Data Base Maintenance

A record of check standard measurements is kept separately from other laboratory records such as records of past calibrations. This permanent record should include all pertinent information relating to the measurement. For example, it normally includes an identification for the check standard, identification for the instrument, identification for the operator, day, month, year, identification for the type of statistical design used in the intercomparison, observed value of the check standard, environmental conditions that could affect the measurement such as temperature, pressure and relative humidity, standard deviation if applicable, and finally a flag denoting whether or not the check standard was in control on that occasion.

2. Characterization of Error

2.1 Introduction

It is the purpose of this chapter to introduce the reader to the concepts of random error, systematic error and uncertainty. It is the expressed purpose of measurement assurance to identify and quantify all sources of error in the measurement process, because in so doing, the worth of any value reported by the process can be stated in quantitative terms called an uncertainty. In a very real sense, a value assigned to an artifact only has meaning when there is an assessment of how well that number describes the property of interest (be it length, mass or whatever) in terms of its reference base. An uncertainty statement provides that assessment.

Error in measurement is categorized as either systematic, coming from a source that is constant and ever present in the measurement process, or random, coming from a source (or sources) that is continually fluctuating. Systematic error may be known or estimable for any given situation, but random error by its nature is never known for a given measurement situation. The point is that for a single measurement it may be possible to determine the size of the systematic error by intercomparison. On the other hand, the random error that is unique to a single measurement cannot be replicated because conditions of measurement cannot be repeated exactly. Therefore, it is common practice in metrology, as it is in process control [34], to quote limits to random error for all such experiments.

Classification of sources of error as either systematic or random is not always straightforward depending as it does on the way in which the potential source of error enters the measurement process, how it affects the output of that process, and the interpretation of the uncertainty. For example, the maximum observed difference between operators can define a systematic error for a system that is highly operator dependent and for which there are a restricted number of operators or, alternatively, a separate uncertainty statement can be issued for each operator's measurements. Measurement systems that routinely make use of many operators are better served by folding the effect of operator error into the total random error for that system.

At the National Bureau of Standards considerable attention is given to the classification of sources of error. For the participant in a measurement assurance program, systematic error is usually assumed to come from specific sources that are spelled out in this chapter, and remaining sources of error are assumed to be part of the random error of the participant's process and must be estimated as such.

2.2 Process Precision and Random Error

2.2.1 The Standard Deviation

A "measurement process" is said to exist for quantifying a physical attribute of an object, such as its length, only if the process is operating in a state-of-control (Eisenhart [35]). The fact is that, even for such a process, repeated measurements on the same object will not produce identical results. As long as the source of this disagreement is random in nature; i.e., its direction and magnitude not being predictable for any future measurement, the disagreement among measurements is referred to as the process imprecision. A measure of precision, such as the process standard deviation, quantifies this random error or scatter or, more aptly, describes the degree of agreement or closeness among successive measurements of the same object.

The term process precision as used in this publication is not limited to the characterization of the behavior of the particular measuring device per se, but it is intended to describe the total configuration of operator, environmental conditions, instrumentation and whatever other variables go into making any given measurement. As it is rarely possible to measure an item submitted for calibration over a representative set of environmental and working conditions in the laboratory, redundancy is obtained from measurements made on a check standard that is introduced into the measurement sequence on a routine basis. It is assumed that the check standard is similar in response to the test item and that the process precision can be estimated from the measurements made on the check standard.

The simplest measure of process precision is the range--the difference between the largest and smallest measurements in the group. The range is a satisfactory measure of precision when the number of measurements is small, say less than ten. It "does not enjoy the desirable property" (Ku [36]) of tending toward a limiting value as more measurements are taken; it can only increase and not decrease. Therefore, it is desirable to find a measure of precision which takes into account the information in all the measurements and which tends to a limiting value as the sample size increases if we are to use this measure to describe the process behavior as a stable phenomenon.

The standard deviation is such a measure. Small values for the standard deviation are indicative of good agreement and large values are indicative of poor agreement. Because it is necessary to distinguish different kinds of variability that contribute to the total process variability, it is likewise necessary to define different kinds of standard deviations. We routinely identify two levels of standard deviations in calibration work.

These two levels are described briefly in the first chapter where we are dealing with the models covering the error structure among measurements. Reiterating, the first type of standard deviation is a measure of the variability of the measurement process over a short period of time, usually the time necessary to complete one calibration using a particular sequence of measurements called a statistical design. This measure is called the "within standard deviation." Its usage as a check on the internal consistency of an individual calibration experiment is explained in chapter 3 and chapter 4 along with formulas and examples.

The second type of standard deviation that we are dealing with in measurement assurance, and by far the more important of the two, is the total standard deviation s_c. This latter measure includes both the "within" component of variability s_w and a "between" component of variability s_b, which the reader will recall explains incremental changes that can take place from calibration to calibration. The relationship among these quantites is assumed to be of the form

$$s_c = (s_w^2 + s_b^2)^{1/2} .$$

Therefore, the total standard deviation, including as it does both "within" and "between" components of variability, should accurately reflect both the short-term and long-term random errors that are affecting the measurement process.

The limits to random error quoted in the uncertainty statement are computed from the total standard deviation thus assuring that the conditions of a single calibration do not invalidate this measure of the quality of the reported value. As has been noted previously, the total standard deviation, not generally being available from the calibration data, is based on repeated check standard measurements that are structured to include all possible sources of random error. This is accomplished by monitoring the check standard over a long period of time and over the full range of environmental factors for which the uncertainty statement is assumed to be valid.

The total standard deviation depends on the physical model. The most familiar form

$$s_c = \left(\frac{1}{n-1} \sum_{i=1}^{n} (c_i - \bar{c})^2 \right)^{1/2} \qquad (2.2.1)$$

where the arithmetic mean is

$$\bar{c} = \frac{1}{n} \sum_{i=1}^{n} c_i \qquad (2.2.2)$$

assumes that check standard measurements $c_1, \cdots, c_n$ are independent of time and that the effect of other variables is negligible.

The term (n-1), called the degrees of freedom associated with s, is an indication of the amount of information in the standard deviation and is always reported along with the standard deviation.

2.2.2 Pooled Standard Deviation

If several standard deviations with small numbers of degrees of freedom are computed from the same process, they will vary considerably among themselves. It goes without saying that the standard deviation that is quoted in the uncertainty statement must have a sufficient amount of information to guarantee that it is a valid measure of process precision. The question is, "How much

redundancy is sufficient?" As a general rule, fifteen degrees of freedom is a minimum for the initial computation of the standard deviation. As the measurement assurance program progresses, the standard deviation is recomputed to take advantage of the increased data base, and assuming that the process is stable, this will assure a more reliable value of the standard deviation. A standard deviation based on as few as two data points can be combined with other similar estimates that have been obtained on separate occasions for the same process to obtain what is called a "pooled" standard deviation. If the individual standard deviations are $s_1,\cdots,s_k$ with degrees of freedom $\nu_1,\cdots,\nu_k$, respectively, the pooled standard deviation is

$$s_p = \left(\frac{\nu_1 s_1^2 + \cdots + \nu_k s_k^2}{\nu_1 + \cdots + \nu_k} \right)^{1/2} . \qquad (2.2.3)$$

The degrees of freedom associated with s_p is $\nu = \nu_1 + \cdots + \nu_k$.

2.2.3 Limits to Random Error

Limits to random error can be computed with a given probability if the distribution of random errors is known. Limits, so stated, depend upon assumptions concerning the average value and spread of the underlying distribution. For a calibration process it is assumed that random errors of measurement have an equal chance of being negative or positive such that their average value is zero. It is also assumed that the spread of the distribution is adequately estimated by the total process standard deviation.

Limits to random error for a single value from the measurement process are constructed so that the probability is $(1-\alpha)$, for α chosen suitably small, that if the measurement algorithm were to be repeated many times, the average outcome of these experiments would fall within $\pm\, s_c \cdot t_{\alpha/2}(\nu)$ of the reported value, where s_c is the total process standard deviation, ν is the number of degrees of freedom in s_c, and $t_{\alpha/2}(\nu)$ is the $\alpha/2$ percent point of Student's t distribution. (See Ku [37] for a further discussion of Student's t distribution.) Critical values for Student's t are given in Table I for $\alpha = 0.05$ and $\alpha = 0.01$ and degrees of freedom $\nu = 2(2)120$.

Frequently a precise probability interpretation for the limits to error is not needed and, in fact, will not be possible if it cannot be demonstrated that the underlying probability distribution for the data is exactly a normal distribution. In metrology the limits to random error are often taken to be three times the standard deviation. Other technical areas may use two standard deviations. The bounds, plus and minus three standard deviations, are statistically robust (with respect to the coverage of the distribution) in that if the experiment were to be repeated, the chance of reporting a value outside of these bounds would be extremely small. This, of course, assumes that the random errors affecting the experiment come from a distribution that is close in character to the normal distribution and that enough data have been collected to provide a reliable estimate of the standard deviation. The examples given in this chapter use three standard deviation limits.

2.3 Systematic Error

2.3.1 Conventional Calibration

Systematic error takes into account those sources of error, peculiar to the measurement system, that remain constant during the calibration process and explain a difference in results, say, between two different measuring systems trying to realize the same quantity through a large number of measurements. Some obvious examples are: uncertainties in values assumed for reference standards, uncertainties related to the geometry or alignment of instrumentation, differences between operators, differences between comparable systems, etc. The size of the possible discrepancy is estimated, either empirically or theoretically, but its direction is not always known.

In order to define systematic error for a calibration process, it is necessary to define the steps in a calibration echelon that relate the measured value of the quantity of interest back to its basic SI unit or to a national standard. NBS, except in the case of international comparisons, occupies the premier position in the U.S. calibration echelon. Thus the first transfer point in this calibration echelon involves the intercomparison of a laboratory reference standard with the national standard maintained by NBS which may be an artifact standard or an instrument. The second transfer point involves the intercomparison of the laboratory reference standard with an unknown which in turn can be a working standard from the same laboratory or an artifact standard from a lower level calibration laboratory or a finished product. The calibration chain is extended in this way until the final product has been calibrated by an intercomparison involving it and a standard which can be traced back to the National Bureau of Standards.

Systematic error is assessed at every transfer point and passed along to the next lower level in the calibration chain. Thus, the total systematic error for the measurement process that delivers the final product is an aggregate of systematic errors from all transfer points. Systematic error must be defined very specifically for each transfer point in terms of the long-term values for measurements from two systems, and it must also include an estimate of the amount by which the higher level system, such as NBS, may be in error in estimating its long-term value.

The purpose of each transfer point is to reduce or eliminate systematic errors at that level. If we look at an exchange between a laboratory and NBS, a potentially large source of systematic error comes from the values assigned to the laboratory's reference standards. Calibration of the reference standards at NBS can eliminate offset from this source, but the calibration itself is still a source of systematic error whose magnitude depends on how well NBS was able to conduct the calibration as measured by the uncertainty associated with the calibrated values.

The rationalization for assessing a systematic error from this source is that the values for the reference standards remain constant as they are used as a reference for assigning values to other artifacts or instruments. At least they remain constant until they are recalibrated at NBS, and the assignments resulting from their use are all affected in the same way, being either too low or too high, even though the direction and exact magnitude of this error are not known. Thus, uncertainties for values of reference standards are regarded as a systematic error in the laboratory's process (Youden [40]).

Systematic error associated with the uncertainty of a reference standard is assessed proportional to the nominal value of the test item and the nominal value of the reference standard. For example, if a one kilogram standard is used in a weighing design to calibrate a 500g weight, the systematic error from this source is one-half of the uncertainty associated with the assignment for the kilogram standard.

If the value for a test item is reported relative to the average of two reference standards R_1 and R_2, all artifacts being of the same nominal size, and if the assignments for R_1 and R_2 are independent, the systematic error from this source is assessed as

$$U = \frac{1}{2}\left(U_{R1}^{\ 2} + U_{R2}^{\ 2}\right)^{1/2}$$

where U_{R1} and U_{R2} are the uncertainties for R_1 and R_2 respectively. Where the assignments to R_1 and R_2 are not done independently

$$U = (U_{R1} + U_{R2})/2.$$

2.3.2 Measurement Assurance Approach

A laboratory participating in a measurement assurance program measures a transfer standard(s) from NBS as if it were an unknown item using the reference standards and instrumentation that constitute the measurement system in that laboratory. The resulting value for the transfer standard, be it based on one measurement or on several repetitions in the laboratory, is compared with the value assigned the transfer standard by NBS. The relationship between the laboratory's assignment and the NBS assignment for the transfer standard defines an offset which is used to correct the values for the laboratory's reference standards.

This approach has an advantage over the usual calibration route as far as identifying systematic error in the laboratory. Either method suffices for identifying errors related to the values of the reference standards, but given that the reference standards are properly calibrated, the particular conditions of their usage in the laboratory may invite systematic errors that are unsuspected and unidentifiable. The dependence of optical systems on operator was mentioned in an earlier chapter, and systematic error caused by operator effect may be significant for other types of systems as well. Also, instrumentation can differ enough that the reference standards alone are not sufficient for eliminating systematic error. Of course, both of these sources of systematic error might be identifiable by proper experimentation, but it would be difficult to assess the magnitude of such errors without the

measurement assurance program. Other factors that are probably not identifiable within the laboratory itself are systematic errors related to lack of proper environmental control or incorrect measurement of temperature and humidity.

Two sources of systematic error are always present in a measurement assurance program. The uncertainty associated with the value of a transfer standard is one. Because another transfer point has been effectively added to the calibration chain, the limits to random error associated with the transfer measurements in the participating laboratory define another systematic error for the laboratory.

2.3.3 Calibration Curve

A more complex situation arises when the purpose of the program is to calibrate an instrument over a range for all continuous values. In this case transfer artifacts are provided at selected points covering the range of interest, and the intercomparisons are used to establish a functional relationship between the instrument and the NBS system. The assignment of values is based on this functional relationship. For example, systematic errors in linewidth measurements produced by an optical imaging system can be reduced relative to the NBS prototype optical system [38] from measurements made on an NBS dimensional artifact. (This artifact is a glass substrate with a series of chromium lines at spacings spanning the range of interest.)

Measurements made on individual lines on the artifact define a functional relationship between the two systems, and a least-squares technique is used to derive a best fitting curve to the measured values as a function of the NBS values. The empirical fit is called the calibration curve.

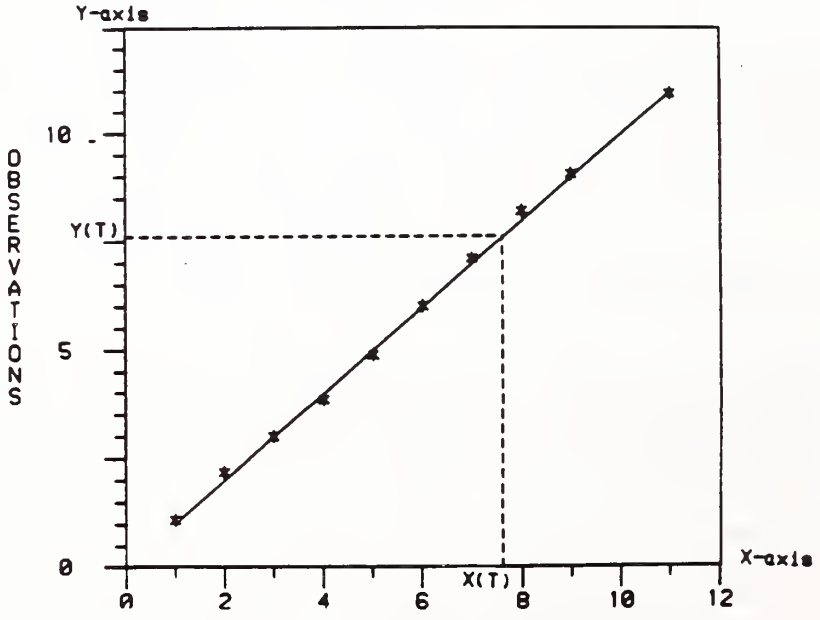


Figure 1
Schematic diagram of a linear calibration curve showing the relationship between an observed value Y(T) and its calibrated value X(T)

In figure 1, each optical measurement is plotted against the corresponding NBS value, and the calibration curve fitted to all the measurements is shown by the solid line. The offset between the user's system and the NBS system is reduced by relating any future measurement back to the NBS value. Schematically, for a future value Y(T) as shown on the Y-axis, a dotted line is drawn through Y(T) parallel to the X-axis. At the point where it intersects the calibration curve another dotted line is drawn parallel to the Y-axis, and its point of intersection on the X-axis, X(T), is the corresponding calibrated value relative to NBS.

Because the functional relationship is not known exactly but is estimated by a series of measurements, the calibration curve can be in error. A discussion of the effect of this error on the final uncertainty of a calibrated value is beyond the scope of this treatise. The reader is referred to Hockersmith and Ku [39] for a discussion relating to quadratic calibration curves and to Croarkin and Varner [40] for a discussion relating to linear calibration curves.

2.4 Uncertainty

2.4.1 Definition

The uncertainty statement assigns credible limits to the accuracy of the reported value stating to what extent that value may differ from its reference base. In practice it quantifies the magnitude of any possible discrepancy between the value actually obtained in the laboratory and the value which would be obtained at NBS for the same property of an object. An uncertainty provides both a measure of the worth of the values reported by the measurement laboratory and an estimate of the systematic error accruing to any organization that makes use of these values.

The uncertainty statement is composed of i) all sources of sytematic error that contribute to the offset from the reference base and ii) a limit to random error that quantifies the variability that is inherent in the measurement process as it transfers from a "known" or calibrated artifact or measurement system to an "unknown".

2.4.2 Combination of Random and Systematic Error

Once the systematic errors and the limits to random error have been estimated, they are combined into a single number which is called the uncertainty. Much controvery arises over the proper way to combine systematic and random errors in an uncertainty statement. Basic premises concerning measurement and its uncertainty as espoused by Youden [41], Eisenhart et al. [42] and others have long been adopted by NBS calibration services and are recommended for measurement assurance programs. A different philosophy that has recently been advanced by the Bureau International des Poids et Mesures is discussed in reference [43]. Basically the question revolves around whether systematic errors should be added linearly or combined in quadrature and around whether the systematic error and the limit to random error should be added linearly or combined in quadrature. For example, if there are several sources of systematic error $S_1,\cdots,S_k$, adding the systematic errors linearly assumes the worst possible combination of errors and gives a total systematic error of

total systematic error S where

$$S = S_1 + S_2 + \cdots + S_k \ . \qquad (2.4.1)$$

Combining the systematic errors in quadrature produces a total systematic error for those sources of

$$S = (S_1^2 + S_2^2 + \cdots + S_k^2)^{1/2} \ . \qquad (2.4.2)$$

Recommended practice for measurement assurance programs is to combine in quadrature systematic errors that are known to be independent as in (2.4.2), to add linearly systematic errors that may not be independent as in (2.4.1), and to combine systematic and random errors linearly.

2.4.3 Final Statement

Because there is no universal agreement on setting limits to random error, such as two or three standard deviation limits, and also because there is no universal agreement either at NBS or internationally as to how the systematic and random components should be combined, it is recommended that for maximum clarity the composition of the uncertainty statement be fully explained. The explanation should include a statement of the limits to random error, a list of sources of systematic error, and a description of the way in which they have been combined. An example of an uncertainty statement from an NBS calibration process is:

> The apparent mass correction for the nominal 10 gram weight is +0.583mg with an overall uncertainty of ±0.042mg, using three times the standard deviation of the reported value as a limit to the effect of random errors of measurement, the magnitude of systematic errors from all known sources being negligible.

The chain of uncertainty as propagated through a calibration echelon starts with the uncertainty assessed at NBS which consists of all sources of error, both systematic and random, associated with that process including the uncertainty of its reference standards relative to basic units of measurements. If the calibration echelon involves one or more standards laboratories, the total uncertainty as assessed at each echelon becomes a systematic error for the next lower echelon laboratory, and the uncertainties at each level are propagated in like manner. In the next section the propagation of uncertainties for a laboratory that uses an NBS calibrated artifact as a reference standard is compared with the propagation of uncertainties for a laboratory that calibrates its own measuring system through the use of an NBS transfer standard.

2.5 Uncertainty of Reported Values

2.5.1 Uncertainty via Conventional Calibration

The uncertainty associated with a value reported for a test item by a measurement process that is operating in a state of statistical control using

a reference standard calibrated by NBS is

$$U = 3s_r + U_{STD} \cdot \qquad (2.5.1)$$

This assumes that the standard is not changed during transport and that environmental and procedural factors are not different from the conditions of calibration. The standard deviation of the reported value s_r depends on the total standard deviation s_c, the error structure for the reported value as discussed in section 1.5, and the number of measurements made on the test item. The quantity U_{STD} is the uncertainty associated with the reference standard as stated in the NBS calibration report.

Note that where the reported value is an average of p measurements, the usual standard deviation of an average, $s_r/\sqrt{p}$, sometimes called the standard error, will apply to the reported value only if the p repetitions were made over the same set of environmental conditions that were sampled in the calculation of the total standard deviation. In a calibration setting where repetitions are done within a day or two, the standard deviation of a reported value depends upon a between component of variability s_b and a within component s_w as explained in section 1.5.

2.5.2 Uncertainty via a Transfer Standard

Where a laboratory has calibrated its own reference standard using an NBS transfer standard, rather than using a reference standard calibrated at NBS, another echelon has effectively been added to the calibration chain. The uncertainty of that transfer must be assessed, and it contributes another systematic error to the process of subsequently assigning values to test items.

The uncertainty of a transfer involving a single transfer standard compared with a single laboratory standard is

$$U_{tr} = 3s_t + U_T \qquad (2.5.2)$$

and the uncertainty associated with a value reported for a test item is

$$U = 3s_r + 3s_t + U_T = 3s_r + U_{tr} \qquad (2.5.3)$$

where s_r is the standard deviation associated with the reported value of the test item as discussed in the last section; s_t is the standard deviation associated with the value assigned to the laboratory's reference standard via measurements made on the transfer standard; and U_T is the uncertainty assigned to the transfer standard by NBS.

Admittedly there can be some concern about qualifying a laboratory's systematic error by means of an NBS transfer standard because of the additional systematic error that this imposes on the uncertainty statement. This fact is inescapable, but the resulting uncertainty statement is, in fact, a realistic expression of the errors affecting the process whereas the usual calibration route does not provide a way of assessing systematic errors that may be affecting measurements, other than those directly involving the artifact standard.

The uncertainty, U_T, associated with a transfer standard will usually be smaller than U_{STD}, the uncertainty associated with a calibrated artifact. The calibration workload at NBS is at least one step removed from the NBS primary standard, and the size of U_T relative to U_{STD} can be reduced by eliminating this step in assignments to transfer standards. For example, transfer standards for voltage measurements are compared directly to an NBS primary reference bank that is in turn compared on a monthly basis to the Josephson effect, which provides a realization of the volt. The regular calibration workload is compared with a secondary bank of cells that is compared to the primary bank on a daily basis.

Transfer standards that are assigned values at NBS based on secondary standards are calibrated several times over a long time period in order to reduce the contribution of random error to the uncertainty of the assignment. For example, values for gage blocks that comprise the transfer set from NBS are averages of approximately nine electro-mechanical calibrations completed over a two year period. Furthermore, because s_t can be made small by sufficient repetition and careful excution of the transfer, the total uncertainty in (2.5.3) can be kept close to the uncertainty in (2.5.1) or at least small enough to meet the goals of the measurement assurance program. See figure 2 for a graphic comparison of uncertainties via measurement assurance and conventional calibration routes.

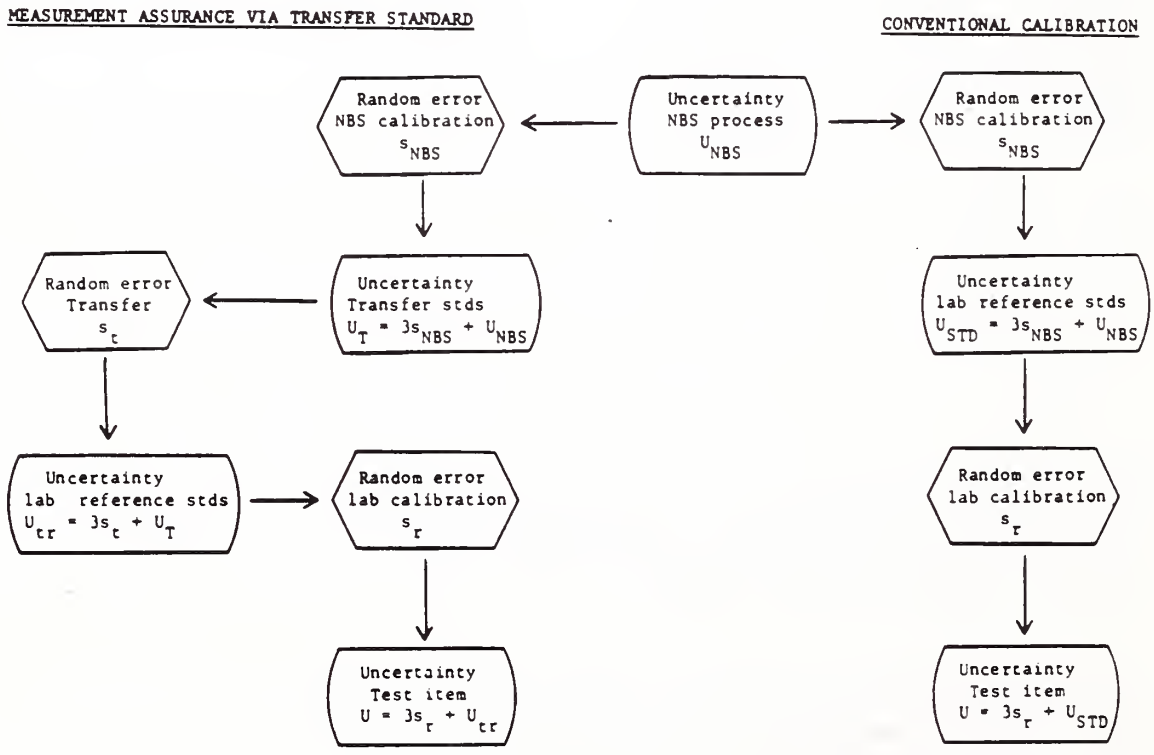


Figure 2
Diagram showing propagation of uncertainties from NBS process to final uncertainty for test item via measurement assurance route compared to the conventional calibration route

2.5.3 Example of an Uncertainty Statement

The principles of this chapter are illustrated by a preliminary experiment at NBS that eventually led to the development of a linewidth standard. Three sources of systematic error were identified in the NBS photometric process that related linewidth measurement to the fundamental definition of length through line-scale interferometry.

The uncertainty from the interferometric process, resulting from random errors associated with making the interferometric determinations and negligible systematic error, translated into a systematic error in the photometric process of 0.01μm. The maximum differences that were observed between the two operators and two instruments that were employed in the NBS system translated into systematic errors of 0.005μm and 0.020μm respectively.

Values assigned to linewidth artifacts were averaged from four photometric readings, and the standard deviation of each assignment was reported as s_r. The limits to random error were taken to be three times the standard deviation of the assignment. An error budget showing the various components contributing to the total uncertainty is shown below.

Components of Uncertainty

Limit to Random Error = $3s_r$			± 0.040μm
Systematic errors:			
	a.	Operator differences	± 0.005μm
	b.	Instrument differences	± 0.020μm
	c.	Uncertainty from interferometry	± 0.010μm
		Total systematic errors	± 0.035μm
Total Uncertainty[§]			± 0.075μm

Based on this analysis NBS assigned a total uncertainty of ± 0.075μm to artifacts that were calibrated by this system. If such an artifact were to be used by a laboratory for calibrating its optical imaging system, this uncertainty would become a systematic error for that process.

[§]It is suggested that uncertainties be stated to no more than two significant figures and that the last decimal place in the reported value of the measured item correspond in place value to the last decimal place in the uncertainty statement.

3. The Check Standard in a Measurement Assurance Program

3.1 Introduction

A check standard provides a means of characterizing the behavior of a measurement process by way of repeated measurements on the same artifact, combination of artifacts, or instrument over a substantial period of time and over fluctuating environmental conditions. It should be thought of as a data base of such measurements rather than as an artifact per se because it is the measurements, or some function of those measurements, corrected according to the model specifications, that actually describe process performance.

The structure of the check standard measurement depends on whether the calibration procedure is based on a single measurement or a calibration design. In some cases the check standard may be a function of readings on two reference standards, thus eliminating the need for an additional artifact. Check standard measurements of the following types form the basis for the measurement assurance programs in the next chapter.

1) Measurements made on a single artifact as close in time as possible to the measurements on the reference standard and the test item.

2) Differences between the observed values of two reference standards whose assigned values are the basis for assigning a value to a test item.

3) Computed value for single artifact from a statistical design involving k intercomparisons of reference standards, test items and artifact check standard.

4) Computed value of difference between two reference standards from a statistical design involving k intercomparisons of reference standards and test items.

5) Measurements made on a calibrated artifact by a direct reading instrument.

6) Calibrated value of a single artifact from a calibration process that uses a ratio technique.

3.2 Process Parameters Defined by the Check Standard

Measurement processes have two properties that are critical to a measurement assurance program. Measurements of a stable quantity tend to a long-term average which may not be the same average that would be achieved if a different laboratory produced the measurements. As discussed in detail in the last chapter, these measurements while tending to an average, will not be identical because of inability to reproduce conditions of measurement exactly, and this latter property is referred to as process variability or imprecision. Process parameters are quantities that describe the long-term value and the process precision from redundant measurements on a check standard.

The statistic for characterizing the long-term value is simply the arithmetic average of the check standard measurements and is referred to as the "accepted

value of the check standard." The check standard measurements supplant the ideal set of measurements that could be made on a test item if it were in the laboratory for a sufficiently long period of time. The average of those hypothetical measurements is, of course, the quantity that is of primary interest, but because such is not at our disposal, we define the process in terms of the accepted value of the check standard. This statistic defines a local base for the measurement process which is intimately related to any discrepancy between the reference base and the average of the measurements that could be made on a test item, and any change in the local base is reason to suspect that this systmatic error has changed.

The statistics for characterizing the process precision are: i) a total standard deviation computed from the same check standard measurements and ii) a within standard deviation computed from each calibration design or group of repetitions for cases where the calibration experiment reports a value based on more than a single measurement on a test item. Within standard deviations are pooled according to (2.2.3) into a single value called the "accepted within standard deviation" which reflects variations that typically take place in the measurement process during the course of a calibration.

If the check standard measurements are properly structured, the accepted total standard deviation reflects the totality of variability in the measurement process. The scatter of check standard measurements will be characteristic of measurements of a test item observed over a period of time in the calibration setting only if both types of measurements are affected by the same sources of error. Then the accepted total standard deviation computed from the check standard measurements can be used to compute the standard deviation for a value reported by the calibration process. Evidently, this computation depends on the type of measurements that are designated as check standard measurements and on the model for the calibration process. Specific examples are discussed in chapter 4.

Before embarking on a full-scale measurement assurance program, the participant conducts a series of experiments to establish a data base of check standard measurements. Accepted values for the process parameters are computed from this data base, and it is emphasized that these experiments should cover several weeks' time and should number at least fifteen to obtain reasonable estimates. The calibration schemes or designs for producing the check standard data must be identical to the procedures for calibrating test items in the workload and measuring transfer standards from NBS.

The importance of the initial check standard measurements dictates that they describe the system in its normal operating mode. Care should be exercised to guarantee that this is indeed the case, so that the standard deviation will be appropriate for an uncertainty statement constructed at any time in the future. This is done by varying the conditions of measurement to cover a representative range of laboratory conditions including operator and environmental variations. These measurements should be scrutinized for outliers because even one significant outlier in a small data set can seriously bias the estimates of the process parameters--perhaps causing an out-of-control condition when the transfer standard is being characterized in the laboratory and invalidating the transfer.

Methods for identifying outliers are highly dependent on underlying distributional assumptions. Several methods for detecting outliers are discussed in ASTM Standard E178[f], but for the foregoing reason, they may not be effective given a limited number of check standard measurements. A plot of the data points is usually satisfactory for detecting outliers. Each check standard measurement should be plotted against a time axis, thus creating a preliminary control chart, and measurements which are obviously aberrant should be deleted from the data set. On the other hand, the data should not be edited in order to achieve seemingly better precision because this will cause failures in the control mechanism at a later time. If a large number of points are suspected as outliers, say more than five percent, the check standard measurements do not constitute a strong data base, and the cause of large variations should be investigated and rectified before proceeding with the measurement assurance program.

3.3 The Check Standard in Process Control

Each check standard measurement is subjected to a statistical test for control, and the outcome of that test is used as a mechanism for accepting or rejecting the results of the measurement process. This presupposes that there is, in fact, a process that is in control, that sufficient data from the process exists to quantify this control, and that the behavior of future measurements is predictable from past behavior of the process. This test is exactly analogous to control chart methodology wherein values that fall inside control limits based on historical data are said to be in control, and values that fall outside the control limits are judged out-of-control.

The technique that is used for control is called a t-test wherein a test statistic is computed from the current check standard measurement, the accepted value of the check standard, and the total standard deviation. This test statistic, when large in absolute value compared to a critical value of Student's t distribution, is indicative of lack of control.

The critical value $t_{\alpha/2}(\nu)$ depends on ν, the number of degrees of freedom in the accepted total standard deviation, and on α, the significance level. The significance level α, the probability of mistakenly flagging a check standard measurement as out-of-control, should be chosen by the participant to be suitably small, say between 0.10 and 0.01, so that the number of remeasurements that must be made because of a chance failure is kept at an acceptable level.

Once the control procedure is installed in the laboratory, the assignments generated by the calibration process are accepted as valid within the stated uncertainty as long as the check standard measurements remain in control. Action is required whenever a check standard measurement is out-of-control. The immediate action is to discard the results of the calibration. Of course, at this point one is faced with a dilemna about what future actions should be taken in regard to the calibration process. Because of the probability of chance failure, exactly α, it is reasonable, while discarding the results of the calibration, to repeat the calibration sequence, hoping that check standard measurements will be in control.

[f] ASTM Standard E178 is available from the American Society for Testing Materials, 1916 Race Street, Philadelphia, Pennsylvania 19103.

In this happy event, one assumes that either something was amiss in the initial calibration, such as insufficient warm-up time for the instrument, or that one was the victim of chance failure. In either case it is permissible to accept the more recent result and proceed as usual. In the event of repeated successive failures or numerous failures over time, one must conclude that a major disruption in the calibration process is affecting the process offset, such as a change in a laboratory reference standard, and the calibration process should be shut down until the problem can be rectified and control reestablished. Each calibrration experiment is intended to reveal the offset of a test item or the client's process relative to NBS, and this offset will not be correctly estimated by the calibrating laboratory if the long-term average for its measurements is not constant relative to the reference base. Therefore, a failure of the check standard test implies that offset has not been eliminated or accounted for by the calibration experiment.

A consideration in choosing α is that the significance level for process control should be the same as the significance level for determining the limits of error in section 2.3. Smaller values of α, the probability of having to remeasure unnecessaarily, that is because of chance failure, correspond to larger associated limits of error. Thus the cost of remeasurement must be weighed against the impact of a larger uncertainty. Values of $\alpha = 0.05$ or $\alpha = 0.01$ are recommended.

An alternative to a critical value based on the t-distribution, as explained in section 2.3, is a factor such as three or two which can be used for computing limits to random error and testing for control. The factor three corresponds approximately to $\alpha = 0.003$ for the normal distribution and is well established in quality control applications. There are no hard and fast rules for picking either a significance level α or a factor such as three for the control procedure, but once it is chosen, it plays a large part in determining the frequency of remeasurement and the magnitude of the uncertainty for the process.

The measurement assurance procedures that are outlined in the next chapter are based upon a critical value of three in almost all cases. Those wishing a more stringent control procedure can substitute the appropriate value of $t_{\alpha/2}$ in the appropriate equations. In calibration work, the purpose of the control procedure is to flag those measurements which are clearly out-of-control, and a critical value of three is suitable for many situations. This approach is the current practice of many calibration services at NBS. Moreover, limits based on the factor three work well, covering a large percentage of the distribution of possible values of the test statistic, even where the test statistic is not strictly distributed as Student's t which is the case for some of the more complicated constructions in the next chapter.

If the measurement sequence allows for a within standard deviation, the ratio of this within standard deviation to the accepted within standard deviation is compared to a critical value based on Snedecor's F distribution (see Ku [44] for a discussion of the F test). A ratio that is large compared to the critical value is indicative of lack of control during the course of the measurement sequence.

The critical value $F_\alpha(\nu_1,\nu_2)$ depends on; ν_1, the number of degrees of freedom in the current within standard deviation; ν_2 the number of degrees of freedom in the accepted within standard deviation; and α, the significance level discussed in preceding paragraphs. Critical values of $F_\alpha(\nu_1,\nu_2)$ are tabulated in Table II for α=0.01, ν_1=1(1)10(2)30(10)120 and ν_2=10(1)20(2)30(5)120.†

The t-test and F test are invoked simultaneously--the failure of either test constituting grounds for discarding the measurement on the test item or transfer standard. The combination of these two tests is a powerful means of detecting shifts in the long-term average of the process as it defines systematic error.

The efficacy of the check standard as a device for guaranteeing that the process is functioning properly and that, therefore, the test items are assigned values with negligible offset relative to NBS, depends on the relationship among the measurements made on the test items, the measurements made on the reference standards and the measurements made on the check standards. The strongest case for measurement assurance exists when all assignments are statistically interrelated as in a statistical design. When the assignments are by nature statistically independent, it is essential that the measurements be temporally related by completing the measurement sequence in as short a time as possible.

There is really no guarantee that a predictable response on the part of the check standard assures a good measurement on the test item if it is possible for the process to change appreciably during the intervening time between the check standard measurement and the other measurements. However, a strong case for confidence in a measurement process exists for a process that is continuously in control. Furthermore, out-of-control findings for the check standard are almost unfailingly indicators of measurement problems because the control limits are specified so that the probability of a single value being out-of-control is extremely small.

The question of how often the process should be checked for control can only be answered in terms of the goals of the measurement program. A criterion based on economic considerations must balance the tradeoff between the cost of making additional measurements to ensure accuracy and the costs incurred when inaccurate measurements are allowed to occur. In order to achieve the highest level of measurement assurance, check standard measurements should be incorporated in every calibration sequence. When this is not possible or not necessary, a check for control should be incorporated in start-up procedures and repeated at intervals thereafter that depend on the level of system control that is desired and on past experiences with the control procedure.

A system that is always in control when checked can be presumed to remain in control between checks, and the time between check standard measurements can be lengthened. Conversely, the same presumption cannot be made for a system that is occasionally out-of-control, and the time between check standard measurements should be shortened if one is to determine how long the system can operate in-control.

†The notation 10(2)30, for example, indicates that the values go in steps of two from ten to thirty.

3.4 The Transfer with NBS

During the transfer between the participating laboratory and NBS, current check standard measurements that result from the transfer experiments are compared with the accepted value of the check standard by a t-test in order to ascertain whether or not there has been a significant change in the long-term average of the process. If the check standard measurements are continually out-of-control while the transfer standard is in the laboratory, the transfer measurements are invalid, and the transfer experiment should be discontinued until the initial check standard measurements are repeated and new accepted values are established. Isolated failures can be treated as they are treated in the calibration workload, and offending measurements that cannot be repeated are deleted from the transfer data.

Similarly, the within standard deviation computed from the transfer measurements is compared with the accepted within standard deviation by an F-test. If possible, sufficient repetitions spaced over a period of time are also included in the procedures for measuring the transfer standards so that the standard deviation for the transfer can be compared to the accepted total standard deviation.

After the completion of the transfer with NBS, the tests for control are continued for the calibration process. When an out-of-control condition is encountered in this mode, the measurement process is discontinued until control is restored which may amount to simply repeating the measurement sequence on the test item and check standard. When it is obvious that the process mean has shifted because of repeated out-of-control findings for the check standard, signifying that the offset from NBS has changed, it is time for another intercomparison with NBS. Theoretically one may be able to analyze the amount of change in the offset, but it seems judicious at this point to reestablish the values of the laboratory's reference standards.

3.5 Updating Process Parameters

After the control procedure has been in place for a year or more, sufficient data should be available so that the process parameters can be updated. The mechanics for doing this depend on the degree of automation that exists in the laboratory and on the computing capability at its disposal. In a sophisticated program one compares the accepted value for the check standard and the accepted total standard deviation with values computed from the check standard data that has been accumulated since the last update. If the two sets of data are essentially in agreement, updated process parameters are computed based on all check standard measurements. In cases where the process has changed significantly in regard to these parameters, the past historical data are discarded, and new process parameters are computed from the most recent data. For computer systems such as micro-computers with limited storage capacity, it may be feasible to retain only a fixed number of check standard measurements. Obviously the number should be sufficient for obtaining reliable estimates. The data file is continually updated by deleting the oldest measurement and adding the newest-thereby always keeping a fixed number of check standard measurements in the data file with which to compute the process parameters.

4. Implementation of Measurement Assurance for Specific Cases

This chapter contains the basic outlines for implementing measurement assurance programs for eight specific measurement situations where the sequence of measurements that constitute an intercomparison depends upon the number of reference standards, the number of test items and the number of redundant measurements to be employed in each intercomparison.

The essential elements that specify the measurement situation for each plan are as follows:

4.1 A comparator process in which one reference standard is compared to a test item and a check standard.

4.2 A comparator process in which a test item is compared to each of two reference standards, and control is maintained on the difference between readings on the two reference standards.

4.3 A comparator process in which three test items are compared to two reference standards in a statistical design, and control is maintained on the difference between the two standards.

4.4 A comparator process for mass calibrations illustrating the use of a 1, 1, 1 design and a 5, 3, 2, 1, 1, 1 design with provision for a check standard for each series.

4.5 A comparator process in which four test items are compared to four reference standards, without direct intercomparison between the test items or reference standards. Control is maintained on the difference between two reference standards.

4.6 Direct reading of the test item with the instrument as the standard. Control is maintained by repetitions on a calibrated artifact.

4.7 Simultaneous measurement of a group of test items relative to a bank of reference standards where a check standard is always included among the test items.

4.8 A ratio technique for one or more test items and one or two reference standards. Control is maintained on calibrated values of an artifact check standard.

Calibration as a process of intercomparing a test item with a reference standard and assigning a value to the test item based on the accepted value of the standard is frequently carried out by a comparator process. For high precision work, the comparator process makes use of an instrument or device which is capable of handling only very small differences between properties of similar objects such as a mechanical comparator for comparing gage blocks of the same nominal length or an electrical bridge for detecting very small differences between resistances. Where individual readings, in scale units, are taken on the unknown and the reference standards and converted to the appropriate units, a value can be assigned to the test item only through the

difference between the reading on the test item and the reading on the reference standard (See section 1.4.2). The calculated difference between the two readings is the "measurement of interest" and the number of such differences determines the redundancy in a measurement scheme.

Where the calibration experiment produces only a difference measurement, such as the difference in emf between two saturated cells as measured by a potentiometer, the term "reading on an unknown" or "reading on a standard" does not have a literal interpretation but refers to the logical intercomparison of the items. In either case, a value is assigned to an unknown relative to the known value of one or more reference standards. This known value is referred to as the restraint.

Where there are a small number of unknowns and reference standards, the calibration experiment may consist of all possible intercomparisons that can be made on the collection of items; this would amount to $k(k-1)/2$ difference measurements for k items being intercompared two at a time. A calibration design consists of a subset of all possible intercomparisons such that, given a restraint or assigned value for the reference standards, the series of intercomparisons can be solved for the unknowns. The method for finding a solution is least-squares, and the resulting values for the unknown items are least-square estimates.

Several factors dictate the choice of intercomparisons that constitute the design. Obviously, it is desirable to keep the number of intercomparisons small. Designs are usually structured so that precision in the assignments to the test items is the same for all items of the same nominal size and so that precision in this sense is optimized for a given number of intercomparisons. Other optimality criteria that are discussed in the statistical literature in references [45] and [46] may be of interest to the reader.

Calibration can also be carried out using a direct reading device or instrument in which case the device is regarded as the standard, and values, already in the appropriate units, are assigned directly to the test items. Such a device, for example an interferometer, can also be used in a comparator mode in which case the difference between a reading on the test item and a reading on the standard is regarded as the measurement of interest.

The eight measurement plans that are discussed in this section have been adapted to both mechanical and electrical measurements. Plan 4.1 is the simplest scheme for a comparator process and may be appropriate when accuracy requirements are moderate. It does not afford a high degree of protection because the linkage between the measurement on the test item and the measurement on the check standard is not as strong as it is for the other comparator schemes. Plan 4.2 affords a higher degree of protection against incorrect measurements by requiring redundant measurements on each test item. This plan is well suited to mechanical measurements and is currently utilized in the Gage Block Measurement Assurance Program. The program is illustrated with data from one participant in section 4.2.7.

Plans 4.3 and 4.5 involve calibration designs that are particularly appropriate for voltage and resistance measurements. The designs have a provision for estimating a so-called left-right effect which is an important

circuit parameter for voltage measurements. The discussion of plan 4.5, which is illustrated with data from the NBS Volt Transfer Program, explains the steps to be followed in process control using a check standard that is either stable or is drifting linearly with time.

Plan 4.4 describes a measurement assurance program for guaranteeing the accuracy of very precise weighings by means of two designs which are routinely used in the NBS mass calibration program. Weighing designs for different combinations of weights along with designs for mechanical and electrical measurements involving more standards and test items are described by Cameron et al [47]. Designs for eliminating temporal effects are described by Cameron and Hailes [48].

Surveillance testing as a means of ensuring the self-consistency of a weight set is described in detail in a recent publication by Jaeger and Davis [49]. The basic idea is to compare a given weight against a collection of other weights in the set whose nominal sum equals the first weight. The authors develop measurement assurance methods for monitoring the difference calculated from the comparison and resolving it with values assigned to the individual weights.

Plan 4.6 is probably the simplest and involves only direct readings on the test items. It is appropriate for large volume workloads that utilize an instrument standard such as interferometer, digital voltmeter, or electronic balance where there is a need to monitor or guarantee the accuracy of the instrument as a matter of course.

Plan 4.7 is appropriate for assigning values to test items or instruments relative to a bank of standards where the calibration consists of subjecting all items including the reference standards to the same stimuli, usually simultaneously. Control is maintained by a check standard which is included as a test item in each measurement sequence. Applications include watthour meter calibration where test meters and reference meters are connected to the same power source and very low pressure calibration where several pressure gages are confined in a vacuum chamber with a reference pressure gage.

By necessity, the analyses are outlined in a straightforward manner, and problems involving drifting reference standards or check standards must be considered separately. It is obviously impossible to anticipate the spectrum of complications that may arise in a given measurement area, and these analyses, offered as a simplistic approach to sometimes difficult problems, are intended to provide a starting point for measurement assurance.

Each measurement assurance program that is presented in this chapter relies upon a check standard concept as discussed at length in the last chapter, and the check standard measurements are crucial to the steps that constitute such a program; namely i) establishment of process parameters; ii) routine process control; iii) evaluation of systematic error by transfer with NBS; iv) determination of uncertainty for test items; v) update of process parameters.

The first four steps are outlined in detail for each program, and the fifth step relating to updating and maintaining the data base was discussed in generality in section 3.5.

4.1 Comparator Process for One Test Item, One Reference Standard, and One Check Standard

4.1.1 Measurement Sequence

This scheme is appropriate for a comparator process where the intercomparison of the test item X with the reference standard R is immediately followed by the intercomparison of an artifact check standard Y with the reference standard R in the sequence X, R, Y, R. The readings are denoted by x, r_1, y, r_2 respectively. This measurement sequence should be followed for all calibrations for which statistical control is to be achieved. The value of the check standard for one such sequence is defined from the reading on the artifact check standard and the duplicate readings on the reference standard as

$$c = y - \frac{1}{2}(r_1 + r_2) . \qquad (4.1.1)$$

All aspects of a measurement assurance program involving this design are explained and illustrated for gage blocks in reference [50].

4.1.2 Process Parameters

Initial values of the process parameters are obtained from n such measurement sequences, where $c_1,\cdots,c_n$ are the observed values of the check standard. The accepted value of the check standard is the mean of the check standard measurements; namely,

$$A_c = \frac{1}{n}\sum_{i=1}^{n} c_i . \qquad (4.1.2)$$

The accepted total standard deviation for the check standard is

$$s_c = \left(\frac{1}{n-1}\sum_{i=1}^{n} (c_i - A_c)^2 \right)^{1/2} \qquad (4.1.3)$$

with $\nu = n-1$ degrees of freedom.

The model assumed for the calibration process is the additive model (1.4.2). Under this model the error structure for the value of the test item and the error structure for the check standard measurement are identical. Thus s_c also estimates the standard deviation of the reported value of the test item which is shown in (4.1.6).

The control limits[h] that are appropriate for future check standard observations are given by

$$\text{Upper control limit} = A_c + 3s_c$$

$$\text{Lower control limit} = A_c - 3s_c \ .$$

4.1.3 Control Procedure

The control procedure applied to each calibration depends on a test statistic t_c that is computed from the observed value of the check standard c for that measurement sequence by

$$t_c = \frac{|c - A_c|}{s_c} \ . \qquad (4.1.4)$$

If

$$t_c < 3 \qquad (4.1.5)$$

the process is in control, and the value of the test item is reported as

$$X^* = x - \frac{1}{2}(r_1 + r_2) + R^* \qquad (4.1.6)$$

where R^* is the value assigned to the reference standard.

If

$$t_c > 3,$$

the calibration of the test item is invalid and must be repeated.

[h]The factor 3 is used in this and all subsequent computations in place of the appropriate percent point of the t distribution; $t_{\alpha/2}(\nu)$.

4.1.4 Transfer with NBS

The transfer with NBS is accomplished by p repetitions of the measurement sequence in which a transfer standard takes the place of the test item in each repetition. Process control as defined by (4.1.5) should be confirmed for each repetition. Any sequence that is out-of-control should be repeated until control is reestablished or else that repetition is deleted from the transfer. If the value assigned to the transfer standard by NBS is T^* with uncertainity U_T, the uncertainty of the transfer is

$$U_{tr} = \frac{3s_c}{\sqrt{p}} + U_T \ . \qquad (4.1.7)$$

The offset Δ of the laboratory process from NBS is

$$\Delta = \frac{1}{p} \sum_{j=1}^{p} X_j^* - T^* \qquad (4.1.8)$$

where $X_1^*, \cdots, X_p^*$ are values calculated according to (4.1.6) for the transfer standard for each of the p repetitions.

This offset is judged significant if

$$\frac{\sqrt{p}\,|\Delta|}{s_c} > 3 \ , \qquad (4.1.9)$$

and in such case the assigned value of the reference standard becomes $R^* - \Delta$.

The assigned value of the reference standard is unchanged if

$$\frac{\sqrt{p}\,|\Delta|}{s_c} < 3 \ .$$

4.1.5 Total Uncertainty

The total uncertainty that is appropriate for a value assigned to a test item by <u>one</u> calibration sequence is

$$U = U_{tr} + 3s_c . \qquad (4.1.10)$$

4.2 Comparator Process for One Test Item and Two Reference Standards

4.2.1 Measurement Sequence

This scheme involving duplicate measurements on the test item is appropriate for a comparator process where the assignment for the test item is made relative to the average of the values assigned to the two reference standards, called the restraint R^*. The intercomparison of the test item X with each of two reference standards, R_1 and R_2, in a trend eliminating design (Croarkin et al [51]) is accomplished by the sequence X, R_1, R_2, X, and the readings are denoted by x_1, r_1 r_2, x_2 respectively. The difference measurements are:

$$d_1 = x_1 - r_1$$

$$d_2 = x_2 - r_2$$

There is no artifact check standard for this design, and a check standard value is defined for each sequence as the calculated difference between the readings on the two reference standards as

$$c = d_2 - d_1 \tag{4.2.1}$$

The value c is structured so as to reflect the maximum variation that occurs in the measurement sequence between the first and the last readings on the test item and not just the variation that occurs between the readings on the two reference standards.

4.2.2 Process Parameters

Initial values of the process parameters are obtained from n such measurement sequences yielding check standard values $c_1, \cdots, c_n$. The accepted value of the check standard is given by the mean of the check standard values; namely,

$$A_c = \frac{1}{n} \sum_{i=1}^{n} c_i \, . \tag{4.2.2}$$

The total standard deviation of the check standard is defined by

$$s_c = \left(\frac{1}{n-1} \sum_{i=1}^{n} (c_i - A_c)^2 \right)^{1/2} \tag{4.2.3}$$

with $\nu = n-1$ degrees of freedom.

The control limits[i] that are appropriate for future observations on the check standard are given by

$$\text{Upper control limit} = A_c + 3s_c$$

$$\text{Lower control limit} = A_c - 3s_c \, .$$

[i] The factor 3 is used in this and all subsequent computations in place of the appropriate percent point of the t distribution; namely, $t_{\alpha/2}(\nu)$.

The model assumed for the process is the additive model (1.4.2). The error structures for the check standard measurement and the reported value of the test item are worked out in detail in section 1.5 where it is shown that the standard deviation for the reported value of the test item is $s_c/2$.

4.2.3 Control Procedure

The control procedure applied to each calibration depends on a statistic t_c that is computed from the observed value of the check standard c for that measurement sequence where

$$t_c = \frac{|c - A_c|}{s_c} . \qquad (4.2.4)$$

If

$$t_c < 3 \qquad (4.2.5)$$

the process is in control, and the value of the test item is reported as

$$X^* = \frac{1}{2}(d_1 + d_2) + R^* \qquad (4.2.6)$$

where the restraint $R^* = \frac{1}{2}(R_1^* + R_2^*)$, and R_1^* and R_2^* are the assigned values of the reference standards.

If

$$t_c > 3,$$

the calibration of the test item is invalid and must be repeated.

4.2.4 Transfer with NBS

The transfer with NBS can be accomplished with two tranfer standards T_1 and T_2. In this mode p_1 repetitions of the measurement sequence are made with T_1 taking the place of the test item and p_2 repetitions of the measurement sequence are made with T_2 taking the place of the test item. This produces a total of $p_1 + p_2$ repetitions for the transfer. Process control as defined by (4.2.5) should be confirmed for each repetition. Any sequence that is out-of-control should be repeated until control is reestablished or else that repetition is deleted from the transfer. If the values assigned to the transfer standards by NBS are T_1^* and T_2^* with uncertainties U_{T1} and U_{T2} respectively, the uncertainty of the transfer is

$$U_{tr} = \frac{3}{4}\left(\frac{p_1+p_2}{p_1 \cdot p_2}\right)^{1/2} s_c + U_T \qquad (4.2.7)$$

where

$$U_T = \frac{1}{2}\left(U_{T1}^{\ 2} + U_{T2}^{\ 2}\right)^{1/2} .$$

The offset Δ of the laboratory process from NBS is defined only in terms of the restraint; i.e., the average of the two reference standards. It is computed from the p_1 values assigned to the first transfer standard according to (4.2.6); namely, $X_1^*,\cdots,X_{p_1}^*$ and the p_2 values assigned to the second transfer standard according to (4.2.6); namely, $X_1^{**},\cdots,X_{p_2}^{**}$.

$$\Delta = \frac{1}{2p_1}\sum_{i=1}^{p_1} X_i^* + \frac{1}{2p_2}\sum_{i=1}^{p_2} X_i^{**} - \frac{1}{2}(T_1^* + T_2^*) \qquad (4.2.8)$$

The offset is judged significant if

$$\tilde{t} > 3, \qquad (4.2.9)$$

where

$$\tilde{t} = \frac{4\sqrt{p_1 \cdot p_2}\,|\Delta|}{\sqrt{p_1+p_2}\,s_c} \qquad (4.2.10)$$

and in such case the assigned value of the restraint is changed to $R^* - \Delta$. The restraint is unchanged if $\tilde{t} < 3$.

4.2.5 Uncertainty

The total uncertainty that is appropriate for a value assigned to a test item by (4.2.6) from one calibration sequence is

$$U = U_{tr} + \frac{3s_c}{2}. \qquad (4.2.11)$$

4.2.6 Example from the Gage Block Measurement Assurance Program

Two sets of eighty-one gage blocks from NBS were sent to industrial participants for the purpose of assigning values to their laboratory reference standards. Before the transfer blocks left NBS, each participant conducted a minimum of six experiments in which his two sets of reference standards were compared to a set of test blocks according to the measurement scheme in section 4.2.1. Because six measurements are not sufficient for estimating a standard deviation, the data were analyzed by groups, with about twenty blocks constituting a group.

In order to check a large data set for outliers, such as the data accumulated on the gage block check standards, it is sometimes possible to make use of the information in the individual standard deviations. Because the measurements are assumed to all come from the same process, a standard deviation that is large compared to the other standard deviations in the group suggests an outlier in the check standard measurements for that nominal size.

If there are k block sizes in a group, the test statistic is the ratio of a single standard deviation s_i to a quantity that has been pooled from the remaining standard deviations in that group; namely, s_j $(j=1,\cdots,k;\ j\neq i)$. The test statistic is

$$F = \left(s_i / s_{p_i}\right)^2$$

where

$$s_{p_i} = \left(\frac{1}{k-1} \sum_{j\neq i} s_j^{\,2}\right)^{1/2}$$

and s_i has ν_1 degrees of freedom and each pooled standard deviation has ν_2 degrees of freedom. If all s_i have the same number of degrees of freedom ν, then $\nu_1 = \nu$ and $\nu_2 = (k-1)\cdot\nu$. If for α chosen suitable small,

$$F > F_\alpha(\nu_1, \nu_2)$$

where $F_\alpha(\nu_1,\nu_2)$ is the upper α percent point of the F distribution with ν_1 and ν_2 degrees of freedom, the standard deviation in question is considered significant, and the individual measurements for that check standard are inspected for an outlier--the outlier being either the largest or the smallest measurement.

Consider the standard deviations in exhibit 4.2.1 which were computed from check standard measurements for nine nominal sizes. The individual measurements are plotted in figure 3 as deviations from the mean for each nominal size as a function of nominal size. Test statistics computed for each nominal size show that the standard deviation for the 0.122000 inch check standard is significantly larger than the others, and figure 3 verifies that the smallest observation is not consistent with the other data for that size and is thus labeled an "outlier."

Exhibit 4.2.1 - Standard deviations from check standard measurements
Values in microinches

Nominal Length (Inches)	Std Devs	Degrees of Freedom	Pooled Std Devs	Degrees of Freedom	Test Statistic
	s_i	ν_1	s_{p_i}	ν_2	F
0.117000	0.445	5	0.723	40	0.38
0.118000	0.288	5	0.733	40	0.15
0.119000	0.952	5	0.659	40	2.09
0.120000	0.382	5	0.727	40	0.28
0.121000	0.616	5	0.707	40	0.76
0.122000	1.303	5	0.579	40	5.06†
0.123000	0.539	5	0.715	40	0.57
0.124000	0.674	5	0.700	40	0.93
0.125000	0.472	5	0.721	40	0.43

† $\left(s_i / s_{p_i}\right)^2 > F_{.01}(5,40)$ where $F_{.01}(5,40) = 3.51$ from Table II.

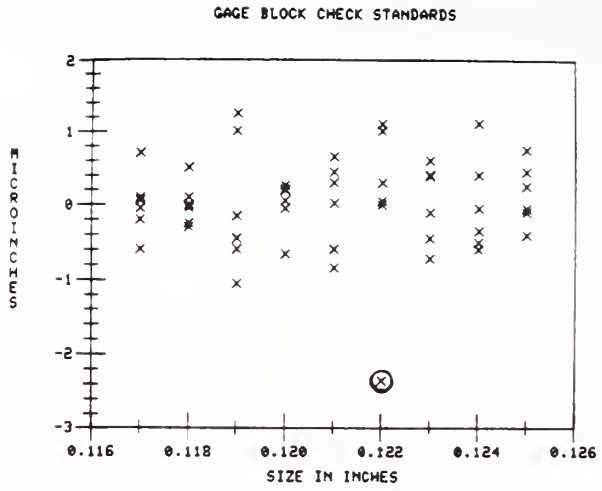


Figure 3
Deviations (microinches) from the mean versus nominal length (inches) for groups of six check standard measurements showing a single outlier

The initial data taken by the participants in the measurement assurance program were inspected for outliers by this method. All outliers were deleted from the data base before calculating the accepted values and standard deviations of the check standard measurements. A subset of the data for one participant is featured in exhibit 4.2.3 with the number of blocks being restricted to five for the purpose of the illustration. The exhibit shows the data from the initial experiments, with a check standard for each repetition computed according to (4.2.1) and initial values for the process parameters A_c and s_c computed using (4.2.2) and (4.2.3) respectively. After the initial data set was edited for outliers, the transfer blocks were sent to the participant. The values assigned to the transfer standards by NBS and the value for the participant's restraint are listed in exhibit 4.2.2.

Exhibit 4.2.2 - Participant's restraint and NBS values for transfer standards
Values in microinches

Nominal	Restraint	Transfer Stds		Uncertainties		Total§
	R^*	T_1^*	T_2^*	U_{T1}	U_{T2}	U_T
0.1006	1.30	-0.63	-0.56	2.17	2.06	2.12
0.1008	0.80	3.21	3.14	2.17	2.06	2.12
0.1010	2.65	2.33	2.52	2.17	2.06	2.12
0.1020	0.45	0.35	0.19	2.17	2.06	2.12
0.1030	-0.05	-2.09	-2.32	2.17	2.06	2.12

§ The systematic errors associated with the transfer standards are added linearly instead of in quadrature because the assignments T_1^* and T_2^* are not independent. Thus $U_T = (U_{T1} + U_{T2})/2$.

Exhibit 4.2.3 - Readings on unknown X and reference standards R_1 and R_2
Corrections to nominal size in microinches

Nominal (inches)	Reps	Readings				Check Standard	Mean	Total S.D.	D.F.
		x_1	r_1	r_2	x_2	c	A_c	s_c	ν
	1	53.9	52.7	46.8	53.9	5.9			
	2	54.8	49.9	45.0	54.5	4.6			
	3	56.0	50.0	44.1	56.1	6.0			
0.1006	4	56.3	50.0	44.1	56.3	5.9	5.80	0.616	5
	5	55.1	49.7	43.7	55.1	6.0			
	6	55.0	50.0	43.9	55.3	6.4			
	1	51.1	51.1	49.2	50.5	1.3			
	2	53.0	49.9	47.9	53.1	2.1			
	3	54.2	50.1	47.5	54.3	2.7			
0.1008	4	54.4	50.2	47.5	54.3	2.6	2.33	0.554	5
	5	53.2	49.9	47.2	53.2	2.7			
	6	53.3	50.0	47.3	53.2	2.6			
	1	52.0	50.1	49.1	52.5	1.5			
	2	54.8	48.8	47.7	54.7	1.0			
	3	55.5	50.0	48.4	55.5	1.6			
0.1010	4	55.4	50.0	48.3	55.4	1.7	1.70	0.593	5
	5	55.5	51.0	48.2	55.5	2.8			
	6	55.8	50.0	48.2	55.6	1.6			
	1	52.1	50.1	48.1	52.2	2.1			
	2	57.3	51.1	49.0	57.2	2.3			
	3	57.0	50.0	48.3	57.0	1.7			
0.1020	4	57.2	50.1	48.4	57.1	1.6	2.07	0.339	5
	5	55.3	50.0	47.6	55.3	2.4			
	6	55.1	49.9	47.6	55.1	2.3			
	1	53.9	49.2	48.5	54.1	0.9			
	2	58.8	50.0	49.0	58.8	1.0			
	3	59.4	50.0	49.1	59.5	1.0			
0.1030	4	59.4	50.0	49.1	59.4	0.9	0.73	0.361	5
	5	59.3	50.0	49.5	58.9	0.1			
	6	59.7	50.2	49.6	59.6	0.5			
							Pooled	0.507	25

Each transfer block was intercompared twice with the participants reference standards by the same scheme used to obtain the initial data, resulting in a total of $p_1 + p_2 = 4$ repetitions. The data for each repetition are shown in exhibit 4.2.4. The readings on the reference standards are designated by r_1 and r_2, and the duplicate readings on a transfer standard are designated by x_1 and x_2. The exhibit also lists the check standard that was computed for each repetition, the test statistic t_c, and the value reported for the NBS transfer standard according to (4.2.6).

Notice that on three occasions the check standard measurement failed the test for control defined by (4.2.5). Because the data were analyzed at NBS after the transfer standards left the participant's laboratory, it was not possible to repeat those sequences, and they were deleted from the transfer data thereby reducing the number of valid repetitions for those block sizes.

Exhibit 4.2.4 - Readings on transfer standards T_1 and T_2
Corrections to nominal size in microinches

Nominal (inches)	Stds	Reps	Readings				Check Std	Test Statistic	Transfer Std
			x_1	r_1	r_2	x_2	c	t_c	X*
	T_1	1	51.2	55.2	48.0	50.8	6.8	2.0	0.70
	T_1	2	51.3	55.2	48.9	51.2	6.2	0.8	0.50
0.1006	T_2	1	50.8	54.9	48.1	51.3	7.3	3.0§	----
	T_2	2	51.2	55.2	48.9	51.3	6.4	1.2	0.50
	T_1	1	56.5	55.3	52.4	56.3	2.7	0.7	3.35
	T_1	2	56.2	55.1	52.5	56.2	2.6	0.5	3.20
0.1008	T_2	1	56.1	55.1	52.3	56.4	3.1	1.5	3.35
	T_2	2	55.7	55.0	52.5	55.8	2.6	0.5	2.80
	T_1	1	54.0	54.9	53.2	54.0	1.7	0	2.60
	T_1	2	53.5	55.0	52.8	53.5	2.2	1.0	2.25
0.1010	T_2	1	53.9	54.9	53.3	53.9	1.6	0.2	2.45
	T_2	2	53.8	55.0	52.8	53.9	2.3	1.2	2.60
	T_1	1	54.9	54.3	52.1	54.7	2.0	0.1	2.05
	T_1	2	55.0	55.1	52.5	55.0	2.6	1.0	1.65
0.1020	T_2	1	54.6	54.3	52.1	54.6	2.2	0.3	1.85
	T_2	2	55.2	55.1	52.5	55.2	2.6	1.0	1.85
	T_1	1	52.9	53.9	52.9	52.8	0.9	0.3	-0.60
	T_1	2	53.9	54.9	52.4	53.9	2.5	3.5§	----
0.1030	T_2	1	52.4	53.9	52.9	52.5	1.1	0.7	-1.00
	T_2	2	53.4	54.9	52.4	53.4	2.5	3.5§	----

§Failed the test for control. The Gage Block Measurement Assuance Program uses a critical value of 3.

Offsets from NBS were computed for each block size by (4.2.8) and were tested for significance by (4.2.9). The participant was advised to change the value of the restraint for those block sizes which showed a significant offset from NBS. The uncertainty of the current transfer with NBS was computed. Results are reported in exhibit 4.2.5. The participant was further advised that the uncertainty appropriate for his process was U = 3.42 microinches as calculated by (4.2.11).

This uncertainty is valid for calibrations conducted according to the measurement scheme in Section 4.1.1 with the value of the restraint as stipulated as long as the process remains in control. Another transfer with NBS will be scheduled in two years to check on the state of the measurement assurance program, and it is anticipated that thereafter transfers with NBS will become increasingly rare. Specific blocks that shows signs of change can be recalibrated or replaced in the interim.

Exhibit 4.2.5 - Offsets from NBS and corrected restraints
Values in microinches

Nominal (inches)	Number of Repetitions	Offset	Test Statistic	Corrected Restraint	Uncertainty of Transfer
	$p_1 + p_2$	Δ	$\tilde{t}$	$R^* - \Delta$	U_{tr}
0.1006	3	1.14	7.3†	0.16	2.59
0.1008	4	0.00	0.0	0.80§	2.50
0.1010	4	0.05	0.4	2.65§	2.50
0.1020	4	1.58	12.5†	-1.13	2.50
0.1030	2	1.40	7.8†	-1.45	2.66

†The test statistic $\tilde{t} > 3$ indicating that the offset from NBS is significant and that the laboratory restraint should be decreased by the amount Δ.

§The restraint is unchanged because the offset is not significant.

4.3 Comparator Process for Three Test Items and Two Reference Standards

4.3.1 Measurement Sequence

In this scheme, which is particularly suitable for electrical measurements, the small difference between two items, such as the difference between the electromotive forces for two saturated cells, constitutes a measurement. The assignments of values to test items are done relative to two reference standards. The statistical design leads not only to equal precision in the assigned value for each test item, but it is also structured so that any position effect in the electrical connection, called left-right effect, is cancelled (Cameron & Eicke [52]). The theory of least-squares estimation which governs the solution of this type of design is explained by Cameron in reference [53].

The design is composed of a subset of all possible difference measurements that could be made on the two standards and three test items. The total number of measurements that could be made in order to achieve left-right balance on such a complement of standards and test items is twenty, and the design is parsimonious in that it requires a subset of ten of the possible measurements while still achieving equal precision for each assignment.

The reference standards are designated by R_1 and R_2, the test items by X, Y, and Z and the corresponding intercomparisons on each by r_1, r_2, x, y, z respectively. The order of measurements is given below:

$$\begin{aligned} d_1 &= r_1 - r_2 \\ d_2 &= r_2 - x \\ d_3 &= x - y \\ d_4 &= y - z \\ d_5 &= z - r_1 \\ d_6 &= y - r_1 \\ d_7 &= r_2 - y \\ d_8 &= z - r_2 \\ d_9 &= x - z \\ d_{10} &= r_1 - x \end{aligned} \qquad (4.3.1)$$

The left-right effect is estimated by

$$\hat{\zeta} = \frac{1}{10} \sum_{i=1}^{10} d_i \, . \qquad (4.3.2)$$

The differences of the reference standards from their average as estimated by least-squares are:

$$\hat{r}_1 = \frac{1}{10}(2d_1 - d_2 - d_5 - d_6 - d_7 + d_8 + d_{10})$$

$$\hat{r}_2 = \frac{1}{10}(-2d_1 + d_2 + d_5 + d_6 + d_7 - d_8 - d_{10})$$

and the corresponding differences for the test items are:

$$\begin{aligned}
\hat{x} &= \frac{1}{10}(-3d_2 + 2d_3 + d_5 + d_6 - d_7 + d_8 + 2d_9 - 3d_{10}) \\
\hat{y} &= \frac{1}{10}(-d_2 - 2d_3 + 2d_4 + d_5 + 3d_6 - 3d_7 + d_8 - d_{10}) \\
\hat{z} &= \frac{1}{10}(-d_2 - 2d_4 + 3d_5 + d_6 - d_7 + 3d_8 - 2d_9 - d_{10}).
\end{aligned} \tag{4.3.3}$$

The within standard deviation for each design is

$$s_w = \left(\frac{1}{\nu} \sum_{i=1}^{10} \xi_i^2 \right)^{1/2} \tag{4.3.4}$$

with degrees of freedom $\nu = 5$.

The individual deviations ξ_i from the least-squares fit are defined by:

$$\begin{aligned}
\xi_1 &= d_1 - \hat{r}_1 + \hat{r}_2 - \hat{\zeta} \\
\xi_2 &= d_2 - \hat{r}_2 + \hat{x} - \hat{\zeta} \\
\xi_3 &= d_3 - \hat{x} + \hat{y} - \hat{\zeta} \\
\xi_4 &= d_4 - \hat{y} + \hat{z} - \hat{\zeta} \\
\xi_5 &= d_5 - \hat{z} + \hat{r}_1 - \hat{\zeta} \\
\xi_6 &= d_6 - \hat{y} + \hat{r}_1 - \hat{\zeta} \\
\xi_7 &= d_7 - \hat{r}_2 + \hat{y} - \hat{\zeta} \\
\xi_8 &= d_8 - \hat{z} + \hat{r}_2 - \hat{\zeta} \\
\xi_9 &= d_9 - \hat{x} + \hat{z} - \hat{\zeta} \\
\xi_{10} &= d_{10} - \hat{r}_1 + \hat{x} - \hat{\zeta}.
\end{aligned} \tag{4.3.5}$$

This design can be used for measurement situations where there is no left-right effect to be estimated. In this case, the equations for the deviations ξ_i do not have the term ζ, and the degrees of freedom associated with s_w is $\nu = 6$. All other computations remain the same.

The value of the check standard for one such sequence is defined as the difference between the estimated values of the two reference standards for the sequence as

$$c = \frac{1}{5}(2d_1 - d_2 - d_5 - d_6 - d_7 + d_8 + d_{10}) \qquad (4.3.6)$$

4.3.2 Process Parameters

Initial values of the process parameters are obtained from n such designs, yielding check standard values $c_1, \cdots, c_n$ and within standard deviations $s_{w_1}, \cdots, s_{w_n}$. The accepted value of the check standard is defined as the mean of the check standard values; namely,

$$A_c = \frac{1}{n}\sum_{i=1}^{n} c_i \; . \qquad (4.3.7)$$

The accepted value of the within standard deviation, describing short-term phenomena that affect the measurements within the design, is the pooled value

$$s_p = \left(\frac{1}{n}\sum_{i=1}^{n} s_{w_i}^2\right)^{1/2} \qquad (4.3.8)$$

with degrees of freedom $\nu_1 = \nu \cdot n$.

The total standard deviation of the check standard is defined as

$$s_c = \left(\frac{1}{n-1}\sum_{i=1}^{n} (c_i - A_c)^2\right)^{1/2} \qquad (4.3.9)$$

with $\nu_2 = n-1$ degrees of freedom.

The model assumed for the process is the additive model (1.4.2). Under this model the error structure for the check standard measurement and the error structure for the reported value of an individual test item are such that the appropriate standard deviation for a value reported for a test item is

$$s_r = \frac{\sqrt{3}}{2} s_c \; .$$

The control limits[k] that are appropriate for future check standard values are given by

$$\text{Upper control limit} = A_c + 3s_c$$

$$\text{Lower control limit} = A_c - 3s_c \ .$$

4.3.3 Control Procedure

A test statistic t_c that depends on the observed value of the check standard c is computed for each design by

$$t_c = \frac{|c - A_c|}{s_c} \ . \qquad (4.3.10)$$

The control procedure depends upon this test statistic and the within standard deviation s_w for that design. A dual control procedure is applied as follows:

If

$$t_c < 3 \qquad (4.3.11)$$

and if

$$s_w < s_p \sqrt{F_\alpha(\nu, \nu_1)} \qquad (4.3.12)$$

for α chosen suitably small, the process is in control and values of the test items are reported as

$$X^* = \hat{x} + R^*$$

$$Y^* = \hat{y} + R^* \qquad (4.3.13)$$

$$Z^* = \hat{z} + R^* \ .$$

The restraint is defined as $R^* = \frac{1}{2}(R_1^* + R_2^*)$ where R_1^* and R_2^* are the assigned values of the reference standards.

If

$$t_c > 3,$$

the calibration of the test items is invalid and must be repeated.

4.3.4 Transfer with NBS

Given three transfer standards T_1, T_2, and T_3, the transfer with NBS could be accomplished in one of several ways such as including only one transfer standard in each design. The most straightforward way is to let the transfer standards take the place of the test items X, Y, and Z in the design. The calibration design is repeated p times, and process control should be confirmed for each repetition as defined by (4.3.11) and (4.3.12).

[k]The factor 3 is used in this and all subsequent computations in place of the appropriate percent point of the t distribution; namely, $t_{\alpha/2}(\nu)$.

Any design that is out-of-control should be repeated until control is reestablished or else that design is deleted from the transfer. If the values assigned to the transfer standards by NBS are T_1^*, T_2^*, and T_3^* with uncertainties U_{T1}, U_{T2}, and U_{T3} respectively, the uncertainty of the transfer is

$$U_{tr} = \frac{3}{2\sqrt{15p}}\, s_c + \frac{1}{3}\left(U_{T1}^2 + U_{T2}^2 + U_{T3}^2\right)^{1/2}. \qquad (4.3.14)$$

A characteristic of the design that is not always recognized is that the offset Δ of the laboratory process from NBS is defined only in terms of the restraint and not in terms of individual reference standards. The reference standards should not be used separately and, if one standard is replaced, the value of the remaining standard and the replacement standard must be reestablished in relationship to NBS.

Given the p values assigned to each transfer standard by (4.3.13); namely,

$$X_1^*, \cdots, X_p^*$$

$$Y_1^*, \cdots, Y_p^*$$

$$Z_1^*, \cdots, Z_p^*,$$

the offset is computed as

$$\Delta = \frac{1}{3p} \sum_{i=1}^{p} (X_i^* + Y_i^* + Z_i^*) - \frac{1}{3}(T_1^* + T_2^* + T_3^*). \qquad (4.3.15)$$

The offset is judged significant if

$$\tilde{t} > 3\,, \qquad (4.3.16)$$

where

$$\tilde{t} = \frac{2\sqrt{15p}\ |\Delta|}{s_c} \qquad (4.3.17)$$

and in such case the assigned value of the restraint R^* is changed to $R^* - \Delta$. The restraint is unchanged if $\tilde{t} < 3$.

4.3.5 Uncertainty

The total uncertainty that is appropriate for a value assigned to a test item by (4.3.13) from one design is

$$U = \frac{3\sqrt{3}}{2}\, s_c + U_{tr}. \qquad (4.3.18)$$

4.4 Comparator Process for Mass Calibrations with One Check Standard for Each Series

4.4.1 Measurement Sequence

High precision mass determination is done by a sequence of intercomparisons that relate the mass of an object to the laboratory's kilogram reference standards which in turn are related to the Paris kilogram. An entire weight set may require several series of intercomparisons in order to assign values to all weights. The weights in each series are intercompared by a statistical design that prescribes the weighings. Each weighing involves a mass difference between two nominally equal weights or groups of weights. Values assigned thereby are least-squares estimates from the design. Provision for a check standard is included with the weights for each series. The reader is referred to Cameron et al. [5] for the statistical theory governing weighing designs; to Jaeger and Davis [54] for the physical theory; to Varner [55] for a description of the NBS software for mass determination; and to Appendix A in this publication for a description of the matrix manipulations needed for a solution to general weighing designs and the propagation of standard deviations and uncertainties through several series.

Normally the first series involves two kilogram reference standards, R_1 and R_2, a test kilogram X_{10}, and a summation Σ_1 of other weights totaling one kilogram nominally. The restraint is the average of the values assigned to R_1 and R_2, and the check standard is defined as the difference between R_1 and R_2 as estimated from the design.

The value assigned to the summation Σ_1 by the first series constitutes the restraint for the second series with the individual weights in the summation being calibrated separately in the second series. For example, if a 500 gram, a 300 gram, and a 200 gram weight make up the summation totaling one kilogram, those weights are assigned values in the second series of intercomparisons. Two series are needed to calibrate a weight set consisting of 1kg, 500g, 300g, 200g, and 100g weights, say. A summation of weights Σ_2 which becomes the restraint for third series is included in the second series if the weight set is to be extended to 50g, 30g, 20g, and 10g weights, and the calibration is extended to lesser weights in like manner.

The weighing designs for two such series are described generically as a 1, 1, 1, 1 design and a 5, 3, 2, 1, 1, 1 design representing the ratios of the weights in the series to each other. A design consists of a subset of all possible intercomparisons that can be made on the group of weights with several factors dictating this choice. A design is always constructed so that the standard deviation of reported values for weights of the same nominal size are equal. The number of intercomparisons is kept small, less than twenty, so that the weighings can be completed with thermal effects being minimized. Furthermore, the number of weights that one is willing to have on the pan at one time and the maximum load of the balance have some bearing on the choice of observations.

Two designs satisfying these criteria are shown below for calibrating the aforementioned weight set. These designs are used routinely in the NBS calibration program. Six observations designated by $d_1,\cdots,d_6$ suffice for the first series. A check standard for the first series is constructed by differencing the values of R_1 and R_2 that were estimated from the design. The second series has eleven observations designated by $d_1,\cdots,d_{11}$. Notice that a 100g weight designated as C is included in this design as a check standard. An observation for a single pan balance is defined as the mass difference between the weights marked by (+) and the weights marked by a (-) as defined by Jaeger and Davis [54].

Design for 1st Series

Obs	1kg	1kg	1kg	1kg
	R_1	R_2	X_{10}	Σ_1
d_1	+	-		
d_2	+		-	
d_3	+			-
d_4		+	-	
d_5		+		-
d_6			+	-

Design for 2nd Series

Obs	500g	300g	200g	100g	100g	100g
	X_5	X_3	X_2	X_1	Σ_2	C
d_1	+	-	-	+	-	
d_2	+	-	-		+	-
d_3	+	-	-	-		+
d_4	+	-	-			
d_5	+		-	-	-	-
d_6		+	-	+	-	-
d_7		+	-	-	+	-
d_8		+	-	-	-	+
d_9			+	-	-	
d_{10}			+	-		-
d_{11}			+		-	-

4.4.2 Process Parameters

The check standard for the first series is defined as

$$c_1 = (1/4)\{2d_1 + d_2 + d_3 - d_4 - d_5\} . \qquad (4.4.1)$$

The check standard for the second series is defined as

$$c_2 = (1/920)\{4d_1 - 111d_2 + 119d_3 + 4d_4 - 108d_5 - 102d_6 - 102d_7 + 128d_8 - 10d_9 - 125d_{10} - 125d_{11}\}. \qquad (4.4.2)$$

The within standard deviation for the first series is

$$s_{w_1} = \left(\frac{1}{4} \sum_{i=1}^{6} \xi_i^2 \right)^{1/2} \tag{4.4.3}$$

with $\nu_1 = 4$ degrees of freedom.

The deviations ξ_i that are needed to compute s_{w_1} are defined by:

$$\xi_1 = d_1 - (1/4)\ [2d_1 - d_2 - d_3 + d_4 + d_5]$$

$$\xi_2 = d_2 - (1/4)\ [-d_1 + 2d_2 - d_3 - d_4 + d_6]$$

$$\xi_3 = d_3 - (1/4)\ [-d_1 - d_2 + 2d_3 - d_5 - d_6]$$

$$\xi_4 = d_4 - (1/4)\ [d_1 - d_2 + 2d_4 - d_5 + d_6]$$

$$\xi_5 = d_5 - (1/4)\ [d_1 - d_3 - d_4 + 2d_5 - d_6]$$

$$\xi_6 = d_6 - (1/4)\ [d_2 - d_3 + d_4 - d_5 + 2d_6]\ .$$

The within standard deviation for the second series is

$$s_{w_2} = \left(\frac{1}{6} \sum_{i=1}^{11} \xi_i^2 \right)^{1/2} \tag{4.4.4}$$

with $\nu_2 = 6$ degrees of freedom.

The deviations needed to compute the within standard deviation s_{w_2} are defined as follows:

$$\begin{aligned}
\xi_1 &= d_1 - \hat{x}_5 + \hat{x}_3 + \hat{x}_2 - \hat{x}_1 + \hat{\Sigma}_2 \\
\xi_2 &= d_2 - \hat{x}_5 + \hat{x}_3 + \hat{x}_2 - \hat{\Sigma}_2 + c_2 \\
\xi_3 &= d_3 - \hat{x}_5 + \hat{x}_3 + \hat{x}_2 + \hat{x}_1 - c_2 \\
\xi_4 &= d_4 - \hat{x}_5 + \hat{x}_3 + \hat{x}_2 \\
\xi_5 &= d_5 - \hat{x}_5 + \hat{x}_2 + \hat{x}_1 + \hat{\Sigma}_2 + c_2 \\
\xi_6 &= d_6 - \hat{x}_3 + \hat{x}_2 - \hat{x}_1 + \hat{\Sigma}_2 + c_2 \\
\xi_7 &= d_7 - \hat{x}_3 + \hat{x}_2 + \hat{x}_1 - \hat{\Sigma}_2 + c_2 \\
\xi_8 &= d_8 - \hat{x}_3 + \hat{x}_2 + \hat{x}_1 + \hat{\Sigma}_2 - c_2 \\
\xi_9 &= d_9 - \hat{x}_2 + \hat{x}_1 + \hat{\Sigma}_2 \\
\xi_{10} &= d_{10} - \hat{x}_2 + \hat{x}_1 + c_2 \\
\xi_{11} &= d_{11} - \hat{x}_2 + \hat{\Sigma}_2 + c_2
\end{aligned} \tag{4.4.5}$$

where

$$\hat{x}_5 = (1/920)\{100(d_1 + d_2 + d_3 + d_4) + 60d_5 - 20(d_6 + d_7 + d_8 + d_9 + d_{10} + d_{11})\}$$

$$\hat{x}_3 = (1/920)\{-68(d_1 + d_2 + d_3 + d_4) - 4d_5 + 124(d_6 + d_7 + d_8) - 60(d_9 + d_{10} + d_{11})\}$$

$$\hat{x}_2 = (1/920)\{-32(d_1 + d_2 + d_3 + d_4) - 56d_5 - 104(d_6 + d_7 + d_8) + 80(d_9 + d_{10} + d_{11})\}$$

$$\hat{x}_1 = (1/920)\{119d_1 + 4d_2 - 111d_3 + 4d_4 - 108d_5 + 128d_6 - 102(d_7 + d_8) - 125(d_9 + d_{10}) - 10d_{11}\} \quad (4.4.6)$$

$$\hat{\Sigma}_2 = (1/920)\{-111d_1 + 119d_2 + 4(d_3 + d_4) - 108d_5 - 125d_6 + 128d_7 - 102d_8 - 125d_9 - 10d_{10} - 125d_{11}\}$$

Accepted values for the check standards, within standard deviations, and total standard deviations are obtained from n initial repetitions of the two series. Check standard values $c_{11},\cdots,c_{1n}$ and $c_{21},\cdots,c_{2n}$ from the respective series are averaged to obtain accepted values,

$$A_{c_1} = \frac{1}{n}\sum_{i=1}^{n} c_{1i}$$

and (4.4.7)

$$A_{c_2} = \frac{1}{n}\sum_{i=1}^{n} c_{2i} \quad .$$

Similarly, within standard deviations $s_{w_{11}},\cdots,s_{w_{1n}}$ from the first series and $s_{w_{21}},\cdots,s_{w_{2n}}$ from the second series are pooled to obtain accepted within standard deviations for the two series:

$$s_{p_1} = \left(\frac{1}{n}\sum_{i=1}^{n} s_{w_{1i}}^2\right)^{1/2}$$

and (4.4.8)

$$s_{p_2} = \left(\frac{1}{n}\sum_{i=1}^{n} s_{w_{2i}}^2\right)^{1/2} \quad .$$

The total standard deviations for the check standards for each series are respectively

$$s_{c_1} = \left(\frac{1}{n-1}\sum_{i=1}^{n} (c_{1i} - A_{c_1})^2\right)^{1/2}$$

and (4.4.9)

$$s_{c_2} = \left(\frac{1}{n-1}\sum_{i=1}^{n} (c_{2i} - A_{c_2})^2\right)^{1/2} \quad .$$

4.4.3 Control Procedure[ℓ]

Statistical control is maintained on the measurements by series. For the first series, test statistics computed from the current check standard value c_1, and the within standard deviation s_{w_1} are used to test for control. Let

$$t_{c_1} = \frac{|A_{c_1} - c_1|}{s_{c_1}} . \tag{4.4.10}$$

If

$$t_{c_1} < 3 \tag{4.4.11a}$$

and if

$$s_{w_1} < s_{p_1} \sqrt{F_\alpha(4,4n)} \tag{4.4.11b}$$

for α chosen suitably small, the measurement process is in control, and the following values are assigned to the the test weight X_{10} and summation Σ_1:

$$X_{10}^* = -(1/8)\{3d_2 + d_3 + 3d_4 + d_5 - 2d_6\} + R^*$$

$$\Sigma_1^* = -(1/8)\{d_2 + 3d_3 + d_4 + 3d_5 + 2d_6\} + R^* \tag{4.4.12}$$

where $R^* = \frac{1}{2}(R_1^* + R_2^*)$ and R_1^* and R_2^* are the corrections to nominal size for the kilogram standards R_1 and R_2.

Statistical control for the second series depends upon the current check standard value c_2 and within standard deviation s_{w_2} for that series. Let

$$t_{c_2} = \frac{|A_{c_2} - c_2|}{s_{c_2}} . \tag{4.4.13}$$

If

$$t_{c_2} < 3 \tag{4.4.14a}$$

and if

$$s_{w_2} < s_{p_2} \sqrt{F_\alpha(6,6n)} \tag{4.4.14b}$$

the measurement process is in control for that series.

Equations (4.4.10) and (4.4.13) are the simplest constructions for testing for offset using a t statistic. The technique for constructing these statistics follows the general method for t statistics; namely, the difference between

[ℓ]The factor 3 is used in this and all subsequent computations in place of the appropriate factor of the t distribution; namely, $t_{\alpha/2}(\nu)$.

the current value of the check standard and its accepted value divided by the standard deviation of the check standard. As such the construction is applicable to any design. In this case the statistic defined by (4.4.10) is precisely correct if the data base for check standard C_1 comes from identical designs with identical restraints, and similarly for the statistic deifined by (4.4.13). In practice a check standard, especially C_2, can be utilized in a variety of designs. This does not affect the interpretation of the accepted value of the check standard, but it does affect the interpretation of the total standard deviation. In such case the test statistics can be computed using the within standard deviations as follows:

$$t_{c_1} = \frac{\sqrt{2}\,|A_{c_1} - c_1|}{s_{w_1}} \qquad (4.4.10a)$$

$$t_{c_2} = \frac{|A_{c_2} - c_2|}{\left(\frac{29}{230}\, s_{w_2}^2 + \frac{1}{100} \cdot \frac{3}{8}\, s_{w_1}^2\right)^{1/2}} \qquad (4.4.13a)$$

These equations are compatible with the documenation in reference [55] where the between component of variance is assumed to be zero --an assumption that is true for the NBS mass calibration process. Notice that the construction of the relevant t statistic becomes increasingly complicated as one moves through the series of weighings depending as it does on the within standard deviations from all prior designs. See Appendix A for the general construction for any design.

Given that (4.4.14a) and (4.4.14b) are satisfied, values are reported for test items and summation for the next series as follows:

Weights	Reported Values	
500g	$X_5^* = \hat{x}_5 + \frac{1}{2}\,\Sigma_1^*$	
300g	$X_3^* = \hat{x}_3 + \frac{276}{920}\,\Sigma_1^*$	
200g	$X_2^* = \hat{x}_2 + \frac{184}{920}\,\Sigma_1^*$	(4.4.15)
100g	$X_1^* = \hat{x}_1 + \frac{92}{920}\,\Sigma_1^*$	
Σ100g	$\Sigma_2^* = \hat{\Sigma}_2 + \frac{92}{920}\,\Sigma_1^*$	

where $\hat{x}_5$, $\hat{x}_3$, $\hat{x}_2$, $\hat{x}_1$ and $\hat{\Sigma}_2$ are defined in (4.4.6) and Σ_1^* is defined in (4.4.12). Whenever a series is out-of-control, the calibration results for the test weights in that series are invalid and must be repeated.

4.4.4 Transfer with NBS

For a mass measurement assurance program the laboratory's starting kilograms are calibrated at NBS and assigned values R_1^* and R_2^* and associated uncertainties U_{R1} and U_{R2}. The transfer is accomplished by relating all weighings to these standards as explained in section 4.4.1.

4.4.5 Uncertainty[m]

The uncertainty associated with the value assigned to any weight is a function of the design and the within standard deviations for that series and all prior series. It also includes as systematic error a proportional part of the uncertainty associated with the starting restraint. For example, the uncertainty for the value assigned to the one kilogram summation Σ_1^* which is the starting restraint for the second series is U_{1000} where

$$U_{1000} = 3\sqrt{k_1}\, s_{w_1} + \frac{1}{2}(U_{R1} + U_{R2}), \qquad k_1 = \frac{3}{8} \quad . \qquad (4.4.16)$$

The uncertainties for the 500g, 300g, 200g, and 100g test weights are respectively:

$$U_{500} = 3\left(k_2\, s_{w_2}^2 + \frac{3}{8}\, m_2^2 s_{w_1}^2\right)^{1/2} + \frac{m_2}{2}(U_{R1} + U_{R2}), \quad k_2 = \frac{50}{920}, \quad m_2 = \frac{1}{2}$$

$$U_{300} = 3\left(k_3\, s_{w_2}^2 + \frac{3}{8}\, m_3^2 s_{w_1}^2\right)^{1/2} + \frac{m_3}{2}(U_{R1} + U_{R2}), \quad k_3 = \frac{82}{920}, \quad m_3 = \frac{3}{10}$$

$$U_{200} = 3\left(k_4\, s_{w_2}^2 + \frac{3}{8}\, m_4^2 s_{w_1}^2\right)^{1/2} + \frac{m_4}{2}(U_{R1} + U_{R2}), \quad k_4 = \frac{64}{920}, \quad m_4 = \frac{1}{5}$$

$$U_{100} = 3\left(k_5\, s_{w_2}^2 + \frac{3}{8}\, m_5^2 s_{w_1}^2\right)^{1/2} + \frac{m_5}{2}(U_{R1} + U_{R2}), \quad k_5 = \frac{116}{920}, \quad m_5 = \frac{1}{10}$$

[m] Uncertainties are computed assuming the between component of variance is zero. See reference [55] for the general construction.

4.5 Comparator Process for Four Reference Standards and Four Test Items

4.5.1 Measurement Sequence

This design for four reference standards and four test items involves the intercomparison of items two at a time where each test item is intercompared with each standard one time, and there is no direct intercomparison among standards or test items. The design is routinely used for voltage measurements where the laboratory's reference standards R_1, R_2, R_3 and R_4 in one temperature controlled box are intercompared with test items W, X, Y and Z or transfer standards in another box, and there are no intercomparisons within a box.

Schematically, the intercomparisons are as shown below where a plus (+) or a minus (-) indicates relative position in the circuit.

Test \ Ref	R_1	R_2	R_3	R_4
W	+	-	+	-
X	-	+	-	+
Y	+	-	+	-
Z	-	+	-	+

Measurements on the laboratory standards R_1, R_2, R_3, and R_4 and the test items W, X, Y and Z are designated by r_1, r_2, r_3 and r_4 and w, x, y, and z respectively. The design consists of the following sequence of difference measurements:

$$\begin{aligned}
d_1 &= r_1 - w \\
d_2 &= r_1 - y \\
d_3 &= r_3 - y \\
d_4 &= r_3 - w \\
d_5 &= r_2 - x \\
d_6 &= r_2 - z \\
d_7 &= r_4 - z \\
d_8 &= r_4 - x \\
d_9 &= x - r_1 \\
d_{10} &= z - r_1 \\
d_{11} &= z - r_3 \\
d_{12} &= x - r_3 \\
d_{13} &= w - r_2 \\
d_{14} &= y - r_2 \\
d_{15} &= y - r_4 \\
d_{16} &= w - r_4
\end{aligned} \qquad (4.5.1)$$

The design has several features that make it particularly suitable for intercomparing saturated standard cells. Let the observations d_i, ordered as in (4.5.1) so as to minimize the number of circuit connections, represent the differences in emf between two cells as measured by a potentiometer. The convention adhered to is, for example, that r_1-w represents the measured difference between R_1 and W with the cells reversed in the circuit relative to their positions for the difference w-r_1.

The design is balanced so as to cancel out any spurious emf that may be present in the circuit [56]. In the presence of such systematic error, called left-right effect, the measurements d_i are assumed to be related to the actual differences D_i in emf between two cells in the following way:

$$d_i = D_i + \zeta + \varepsilon_i \qquad i = 1,\cdots,16$$

where ζ is the left-right effect, and ε_i is random error. For a circuit with negligible left-right effect, one expects that the measurements would sum to zero except for the effect of random error. Any disparity between this expectation and the summation gives an estimate of the magnitude of left-right effect; namely,

$$\hat{\zeta} = \frac{1}{16} \sum_{i=1}^{16} d_i \ . \qquad (4.5.2)$$

A measuring process such as the one described in the foregoing paragraph can be characterized by:

i) a short-term or within standard deviation which describes variability during the time necessary to make the sixteen measurements for one design.

ii) accepted values for check standards which have been specifically chosen for this measurement situation.

iii) a total standard deviation for the process based on the check standard measurements.

The difference of each test item from the average of the reference group is computed by:

$$\hat{w} = -\frac{1}{4}(d_1 + d_4 - d_{13} - d_{16})$$

$$\hat{x} = -\frac{1}{4}(d_5 + d_8 - d_9 - d_{12}) \qquad (4.5.3)$$

$$\hat{y} = -\frac{1}{4}(d_2 + d_3 - d_{14} - d_{15})$$

$$\hat{z} = -\frac{1}{4}(d_6 + d_7 - d_{10} - d_{11})$$

The foregoing quantities in conjunction with the differences of the reference standards from their group average; namely,

$$\hat{r}_1 = \frac{1}{16}(3d_1+3d_2-d_3-d_4-d_5-d_6-d_7-d_8-3d_9-3d_{10}+d_{11}+d_{12}+d_{13}+d_{14}+d_{15}+d_{16})$$

$$\hat{r}_2 = \frac{1}{16}(-d_1-d_2-d_3-d_4+3d_5+3d_6-d_7-d_8+d_9+d_{10}+d_{11}+d_{12}-3d_{13}-3d_{14}+d_{15}+d_{16})$$

$$\hat{r}_3 = \frac{1}{16}(-d_1-d_2+3d_3+3d_4-d_5-d_6-d_7-d_8+d_9+d_{10}-3d_{11}-3d_{12}+d_{13}+d_{14}+d_{15}+d_{16}) \quad (4.5.4)$$

$$\hat{r}_4 = \frac{1}{16}(-d_1-d_2-d_3-d_4-d_5-d_6+3d_7+3d_8+d_9+d_{10}+d_{11}+d_{12}+d_{13}+d_{14}-3d_{15}-3d_{16})$$

and the estimated left-right effect $\hat{\zeta}$ are used to estimate a within standard deviation s_w for each design; namely,

$$s_w = \left(\frac{1}{8} \sum_{i=1}^{16} \xi_i^2 \right)^{1/2} \quad (4.5.5)$$

with $\nu=8$ degrees of freedom. The individual deviations ξ_i are given by:

$$\begin{aligned}
\xi_1 &= d_1 - \hat{r}_1 + \hat{w} - \hat{\zeta} \\
\xi_2 &= d_2 - \hat{r}_1 + \hat{y} - \hat{\zeta} \\
\xi_3 &= d_3 - \hat{r}_3 + \hat{y} - \hat{\zeta} \\
\xi_4 &= d_4 - \hat{r}_3 + \hat{w} - \hat{\zeta} \\
\xi_5 &= d_5 - \hat{r}_2 + \hat{x} - \hat{\zeta} \\
\xi_6 &= d_6 - \hat{r}_2 + \hat{z} - \hat{\zeta} \\
\xi_7 &= d_7 - \hat{r}_4 + \hat{z} - \hat{\zeta} \\
\xi_8 &= d_8 - \hat{r}_4 + \hat{x} - \hat{\zeta} \\
\xi_9 &= d_9 - \hat{x} + \hat{r}_1 - \hat{\zeta} \\
\xi_{10} &= d_{10} - \hat{z} + \hat{r}_1 - \hat{\zeta} \\
\xi_{11} &= d_{11} - \hat{z} + \hat{r}_3 - \hat{\zeta} \\
\xi_{12} &= d_{12} - \hat{x} + \hat{r}_3 - \hat{\zeta} \\
\xi_{13} &= d_{13} - \hat{w} + \hat{r}_2 - \hat{\zeta} \\
\xi_{14} &= d_{14} - \hat{y} + \hat{r}_2 - \hat{\zeta} \\
\xi_{15} &= d_{15} - \hat{y} + \hat{r}_4 - \hat{\zeta} \\
\xi_{16} &= d_{16} - \hat{w} + \hat{r}_4 - \hat{\zeta}
\end{aligned} \quad (4.5.6)$$

Check standards for electrical measurements are not easily defined because of the inherent nature of electrical quantities to drift over time. For this reason, three separate check standards are recommended for measurements on standard cells. The left-right effect reflects many of the sources of error in the measurement system and can be presumed to remain stable over time. For this reason it makes a suitable check standard for process control. Specifically, the value of the first check standard is defined for each design as ζ from (4.5.2).

There is also a need to check on the stability of the reference standards, changes or instabilities in which may not be reflected in the left-right effect. The least-squares estimates for the reference standards from the design (4.5.4) cannot be used to check on the stability of the standards themselves because these estimates are in effect a consequence of the design, subject to the restraint, and are not meaningful separately. For example, if the restraint is changed to exclude one of the reference standards, the least-squares estimates for the remaining reference standards as computed from the same observed differences (4.5.1) can change appreciably.

The information in a design does, however, allow a way of monitoring the change in one reference standard relative to another reference standard. A measured difference between two reference standards that is not subject to the restraint can be computed from each design, and two check standards, each one involving the difference between two reference standards, are recommended for monitoring the stability of the four reference standards.

Check standard C_1 is defined for the difference between R_1 and R_3, and check standard C_2 is defined for the difference between R_2 and R_4. Their respective values c_1 and c_2 are computed for each design as follows:

$$c_1 = \frac{1}{4}(d_1 + d_2 - d_3 - d_4 - d_9 - d_{10} + d_{11} + d_{12}) \qquad (4.5.7)$$

$$c_2 = \frac{1}{4}(d_5 + d_6 - d_7 - d_8 - d_{13} - d_{14} + d_{15} + d_{16}) .$$

Because it is anticipated that the change in one reference standard relative to another may not be stable over time, the method for analyzing check standards C_1 and C_2 is a modified process control technique that allows for linear drift.

4.5.2 Process Parameters for Stable and Drifting Check Standards

Initial values for the process parameters are established from n repetitions of the design in which the four reference standards are compared to any four test items. The resulting check standard measurements are $\zeta_1,\cdots,\zeta_n$; $c_{11},\cdots,c_{1n}$; and $c_{21},\cdots,c_{2n}$. For the left-right effect the n values are averaged to obtain the accepted value

$$A_\zeta = \frac{1}{n}\sum_{i=1}^{n}\hat{\zeta}_i \,. \qquad (4.5.8)$$

A total standard deviation for the left-right effect is also computed from the initial check standard measurements by

$$s_\zeta = \left(\frac{1}{(n-1)}\sum_{i=1}^{n}(\hat{\zeta}_i - A_\zeta)^2\right)^{1/2} \qquad (4.6.9)$$

with $\nu = (n-1)$ degrees of freedom.

The control limits[o] that are appropriate for future measurements on the left-right effect are:

$$\text{Upper Control Limit} = A_\zeta + 3s_\zeta$$

$$\text{Lower Control Limit} = A_\zeta - 3s_\zeta.$$

Similar calculations of accepted values and standard deviations are made for C_1 and C_2 where the check standard measurements $c_{11},\cdots,c_{1n}$ and $c_{21},\cdots,c_{2n}$ are stable over time. More often than not these quantities are not stable over time, and this fact must be taken into account in the analysis. If the check standard values show drift and if the drift is linear with time, check standard values $c_1,\cdots,c_n$ at times $t_1,\cdots,t_n$ can be characterized by

$$c_i = \alpha + \beta t_i \qquad i=1,\cdots,n$$

where the intercept α and the slope β are estimated by

$$\hat{\alpha} = \bar{c} - \hat{\beta}\,\bar{t}$$

and

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(t_i-\bar{t})(c_i-\bar{c})}{\sum_{i=1}^{n}(t_i-\bar{t})^2}$$

[o]The factor 3 is used in this and all subsequent computations in place of the appropriate percent point of the t distribution; namely, $t_{\alpha/2}(\nu)$.

with $$\bar{t} = \frac{1}{n}\sum_{i=1}^{n} t_i \quad \text{and} \quad \bar{c} = \frac{1}{n}\sum_{i=1}^{n} c_i .$$

In the linear case the accepted total standard deviation for each check standard is

$$s_c = \left(\frac{1}{n-2} \sum_{i=1}^{n} (c_i - \hat{\alpha} - \hat{\beta} t_i)^2 \right)^{1/2} \qquad (4.5.10)$$

with $\nu = n-2$ degrees of freedom. See reference [59] for analyses relating to linear regression models.

The parameters of the linear fit and associated standard deviations should be computed for C_1 and C_2 separately resulting in estimates α_1, β_1, s_{c_1} with $\nu_1 = n-2$ degrees of freedom for check standard C_1 and α_2, β_2, s_{c_2} with $\nu_2 = n-2$ degrees of freedom for check standard C_2. The value that a check standard is expected to take on at any given time is thus dependent on the linear fit. Therefore, for a future time t', provided t' is not too far removed from t_n, the accepted values for the check standards are defined by

$$A'_{c_1} = \hat{\alpha}_1 + \hat{\beta}_1 t'$$

and (4.5.11)

$$A'_{c_2} = \hat{\alpha}_2 + \hat{\beta}_2 t' .$$

A total standard deviation for the measurements on C_1 and C_2 can be pooled from s_{c_1} and s_{c_2} by the formula

$$s_c = \left(\frac{1}{2}\left(s_{c_1}^2 + s_{c_2}^2 \right) \right)^{1/2} \qquad (4.5.12)$$

with $\nu = 2(n-2)$ degrees of freedom.

The control procedure assumes that t' is close to t_n because the standard deviation of a predicted value from a linear fit increases dramatically as the linear fit is extrapolated beyond the check standard data. Thus the chance of detecting a real shift in the process diminishes as the tests for control are continued into the future. This fact necessitates frequent updating of the parameters of the linear fit based on recent check standard values.

Furthermore, the control procedure and the assumption of a linear model are interdependent. Because there is no way of separating these two elements, an out-of-control signal can be caused by either lack of process control or a breakdown in the linearity of the check standard measurements. One must recognize this as a short-coming in the control procedure and arrange for other independent checks on the stability of the reference standards.

The control procedure also makes use of the accepted within standard deviation s_p which is not dependent upon model assumptionsfor the check standards . It is computed from the within standard deviations $s_{w_1}, \cdots, s_{w_n}$ for each design as follows:

$$s_p = \left(\frac{1}{n} \sum_{i=1}^{n} s_{w_i}^2 \right)^{1/2} \qquad (4.5.13)$$

with $\nu_3 = 8n$ degrees of freedom.

4.5.3 Process Control

Process control is maintained by monitoring the within standard deviation for each design and the performance of the three designated check standards. If check standard C_1 or C_2 repeatedly fails the test for control, it is likely that one of the two reference standards comprising the check standard has changed in value. In this case it will be necessary to replace one or both of the standards in question or reestablish their values relative to NBS.

Process control should be verified for the within standard deviation s_w as it is calculated for each design and for the current values of the check standards for that design; namely, ζ, c_1, and c_2. For the left-right effect ζ, the test statistic is:

$$t_\zeta = \frac{|\hat{\zeta} - A_\zeta|}{s_\zeta} . \qquad (4.5.14)$$

For check standards C_1 and C_2 that are drifting linearly over time the corresponding test statistics at time t' are:

$$t_{c_1} = \frac{|c_1 - A_{c_1}'|}{\tilde{s}} \qquad (4.5.15)$$

and

$$t_{c_2} = \frac{|c_2 - A_{c_2}'|}{\tilde{s}}$$

where

$$\tilde{s} = s_c \left(\frac{n+1}{n} + \frac{(t' - \bar{t})^2}{\sum_{i=1}^{n} (t_i - \bar{t})^2} \right)^{1/2} .$$

Then the following conditions can be imposed:

$$\text{If } t_\zeta \text{ and } t_{c_1} \text{ and } t_{c_2} \text{ are all } < 3 \qquad (4.5.16a)$$

and if

$$s_w < s_p\sqrt{F_\alpha(8, \nu_3)} \qquad (4.5.16b)$$

for α suitably small, the process is judged in control for that design.

The values of the test items are reported as

$$\begin{aligned} W^* &= \hat{w} + R^* \\ X^* &= \hat{x} + R^* \\ Y^* &= \hat{y} + R^* \\ Z^* &= \hat{z} + R^* \end{aligned} \qquad (4.5.17)$$

where the restraint $R^* = \frac{1}{4}(R_1^* + R_2^* + R_3^* + R_4^*)$, and R_1^*, R_2^*, R_3^*, and R_4^* are the values assigned to the laboratory's reference standards.

If the results of the control procedures along with other experimental evidence indicate instability or other anomalous behavior on the part of one of the reference standards, the entire experiment need not necessarily be discarded. It is possible to delete the reference standard in question from the restraint and obtain new values for the test items if the values of the remaining reference standards are known individually. For example, if one is involved in a transfer with NBS, and if reference standard R_1 shows signs of serious malfunction after several days of intercomparisons between the reference standards and the transfer standards, the values for the transfer standards can be recomputed for each design as follows:

$$\hat{w} = \frac{1}{48}\{-9d_1+3d_2-d_3-13d_4-d_5-d_6-d_7-d_8-3d_9-3d_{10}+d_{11}+d_{12}+13d_{13}+d_{14}+d_{15}+13d_{16}\}$$

$$\hat{x} = \frac{1}{48}\{3d_1+3d_2-d_3-d_4-13d_5-d_6-d_7-13d_8+9d_9-3d_{10}+d_{11}+13d_{12}+d_{13}+d_{14}+d_{15}+d_{16}\} \qquad (4.5.18)$$

$$\hat{y} = \frac{1}{48}\{3d_1-9d_2-13d_3-d_4-d_5-d_6-d_7-d_8-3d_9-3d_{10}+d_{11}+d_{12}+d_{13}+13d_{14}+13d_{15}+d_{16}\}$$

$$\hat{z} = \frac{1}{48}\{3d_1+3d_2-d_3-d_4-d_5-13d_6-13d_7-d_8-3d_9+9d_{10}+13d_{11}+d_{12}+d_{13}+d_{14}+d_{15}+d_{16}\}$$

and W^*, X^*, Y^* and Z^* are computed according to (4.5.17) with the restraint R^* changed to:

$$R^* = \frac{1}{3}(R_2^* + R_3^* + R_4^*).$$

The differences of reference standards R_2, R_3 and R_4 from their average value are recomputed to be:

$$\hat{r}_2 = \frac{1}{12}\{-d_3-d_4+2d_5+2d_6-d_7-d_8+d_{11}+d_{12}-2d_{13}-2d_{14}+d_{15}+d_{16}\}$$

$$\hat{r}_3 = \frac{1}{12}\{2d_3+2d_4-d_5-d_6-d_7-d_8-2d_{11}-2d_{12}+d_{13}+d_{14}+d_{15}+d_{16}\} \quad (4.5.19)$$

$$\hat{r}_4 = \frac{1}{12}\{-d_3-d_4-d_5-d_6+2d_7+2d_8+d_{11}+d_{12}+d_{13}+d_{14}-2d_{15}-2d_{16}\}$$

The within standard deviation for each design (see equations (4.5.5) and (4.5.6)) can be computed using either the original quantities in (4.5.3) and (4.5.4) or the adjusted quantities in (4.5.18) and (4.5.19) with identical results.

4.5.4 Transfer with NBS

Transfer with NBS is accomplished by means of p repetitions of the design in which four transfer standards T_1, T_2, T_3, and T_4 replace the four test items. If one of the tests for control defined by (4.5.16a) and (4.5.16b) is not satisfied, the design should be repeated or else that repetition should be deleted from the transfer.

Given p repetitions of the design in which T_1 replaces W, T_2 replaces X, T_3 replaces Y and T_4 replaces Z, the p values assigned to each transfer standard by the participant's process are computed from (4.5.17); namely,

$$W_1^*,\cdots,W_p^*$$

$$X_1^*,\cdots,X_p^*$$

$$Y_1^*,\cdots,Y_p^*$$

$$Z_1^*,\cdots,Z_p^*.$$

NBS assigns values to electrical transfer standards that take into account their individual and collective behavior both before, during, and after their sojurn in the participant's laboratory. A transfer standard that displays unstable behavior during one of these periods may be excluded from the analysis. Normally the averages for the four transfer standards from the "before and after" NBS determinations are fit by least-squares to a linear function of time; then average values T_j^* are predicted for the times $t_j (j=1,\cdots,p)$ that the transfer standards were in the participant's laboratory by the equation

$$T_j^* = \hat{\alpha}_o + \hat{\beta}_o t_j \qquad j=1,\cdots,p$$

where $\hat{\alpha}_o$ and $\hat{\beta}_o$ are estimated from NBS measurements.

This makes it possible to compute daily offsets $\Delta_j (j=1,\cdots,p)$ for the reference group where

$$\Delta_j = \frac{1}{4}(W_j^* + X_j^* + Y_j^* + Z_j^*) - T_j^* \qquad j=1,\cdots,p \quad (4.5.20)$$

and assuming the reference group is stable, an average offset for the reference group is computed by

$$\bar{\Delta} = \frac{1}{p}\sum_{j=1}^{p} \Delta_j . \quad (4.5.21)$$

The offset is judged significant if

$$\tilde{t} > 3$$

where $$\tilde{t} = \frac{4\sqrt{p}\,|\bar{\Delta}|}{(4s_c^2 - s_p^2)^{1/2}} .$$

In such case the value of the laboratory restraint is changed to $R^* - \Delta$.

Otherwise, the restraint is unchanged.

4.5.5 Uncertainty

The uncertainty of the transfer is

$$U_{tr} = \frac{3(4s_c^2 - s_p^2)^{1/2}}{4\sqrt{p}} + U_T \quad (4.5.22)$$

where U_T is the uncertainty assigned to the transfer standards by NBS.

The uncertainty that is appropriate for the laboratory's process as it assigns a value to a test item based on a single design is

$$U = \frac{3}{4}(10s_c^2 - s_p^2)^{1/2} + U_{tr} . \quad (4.5.23)$$

4.5.6 Example

An example is presented from the Volt Transfer Program where an environmentally controlled box of four standard cells was sent to an industrial participant to be intercompared with the participant's box of four standard cells. After the NBS cells had been in the participant's laboratory for two weeks, thereby giving them a chance to recover from the trip, the laboratory's reference cells were intercompared with the NBS cells each day for 16 days using the design described in sec 4.5.1. The data corrected for the temperature in each box are shown in exhibit 4.5.1.

Exhibit 4.5.1 - Intercomparison of laboratory standard cells with NBS cells
Value in microvolts

Obs \ Day	1	2	3	4	5	6	7	8
d_1	86.70	86.92	86.86	86.98	86.97	86.99	87.07	87.17
d_2	87.28	87.02	86.77	86.69	86.57	86.60	86.63	86.68
d_3	89.29	89.01	88.75	88.64	88.52	88.52	88.54	88.57
d_4	88.84	88.98	88.87	88.94	88.97	88.94	88.98	89.10
d_5	88.06	87.97	87.84	87.89	88.33	88.17	88.03	88.11
d_6	87.43	87.04	86.94	86.94	86.96	86.92	86.92	87.07
d_7	88.96	88.58	88.46	88.47	88.45	88.47	88.50	88.55
d_8	89.63	89.52	89.39	89.43	89.85	89.70	89.63	89.60
d_9	-86.47	-86.60	-86.32	-86.34	-86.71	-86.71	-86.61	-86.71
d_{10}	-85.81	-85.70	-85.40	-85.41	-85.39	-85.49	-85.54	-85.66
d_{11}	-87.82	-87.67	-87.34	-87.37	-87.34	-87.40	-87.42	-87.54
d_{12}	-88.49	-88.62	-88.25	-88.32	-88.72	-88.64	-88.54	-88.59
d_{13}	-88.80	-89.11	-88.84	-88.92	-88.90	-88.88	-88.92	-89.03
d_{14}	-89.15	-89.16	-88.73	-88.64	-88.48	-88.47	-88.48	-88.55
d_{15}	-90.90	-90.69	-90.24	-90.13	-89.93	-89.98	-90.02	-90.08
d_{16}	-90.38	-90.64	-90.33	-90.40	-90.35	-90.41	-90.49	-90.60

Obs \ Day	9	10	11	12	13	14	15	16
d_1	87.25	87.28	87.32	87.45	87.46	87.50	87.53	87.59
d_2	86.72	86.80	86.81	86.87	86.90	86.91	86.92	86.97
d_3	88.60	88.52	88.52	88.62	88.59	88.59	88.59	88.62
d_4	89.13	89.07	89.09	89.16	89.18	89.17	89.21	89.24
d_5	88.09	88.00	87.78	88.12	88.05	88.06	88.04	88.05
d_6	87.07	86.99	86.89	87.17	87.15	87.09	87.07	87.07
d_7	88.58	88.56	88.62	88.69	88.82	88.68	88.69	88.70
d_8	89.60	89.55	89.60	89.68	89.79	89.67	89.65	89.68
d_9	-86.66	-86.66	-86.74	-86.78	-86.78	-86.84	-86.92	-86.89
d_{10}	-85.63	-85.67	-85.79	-85.79	-85.80	-85.88	-85.96	-85.93
d_{11}	-87.52	-87.48	-87.55	-87.58	-87.58	-87.60	-87.59	-87.60
d_{12}	-88.57	-88.47	-88.51	-88.53	-88.57	-88.54	-88.57	-88.57
d_{13}	-89.04	-89.00	-89.01	-89.10	-89.06	-89.10	-89.10	-89.12
d_{14}	-88.53	-88.44	-88.47	-88.55	-88.47	-88.51	-88.55	-88.53
d_{15}	-90.07	-90.00	-90.05	-90.12	-90.14	-90.10	-90.12	-90.14
d_{16}	-90.60	-90.54	-90.58	-90.68	-90.69	-90.69	-90.75	-90.77

Exhibit 4.5.2 - Estimates for transfer standards and reference standards
Values in microvolts

Day	NBS Standard Cells T_1	T_2	T_3	T_4	Laboratory Reference Cells R_1	R_2	R_3	R_4	L–R Effect
	$\hat{w}$	$\hat{x}$	$\hat{y}$	$\hat{z}$	$\hat{r}_1$	$\hat{r}_2$	$\hat{r}_3$	$\hat{r}_4$	$\hat{\zeta}$
1	-88.68	-88.16	-89.15	-87.50	-1.811	-0.016	0.234	1.592	-0.102
2	-88.91	-88.17	-88.96	-87.24	-1.767	-0.007	0.243	1.531	-0.197
3	-88.72	-87.94	-88.62	-87.03	-1.746	+0.004	0.219	1.522	-0.098
4	-88.80	-87.99	-88.50	-87.04	-1.739	+0.003	0.223	1.513	-0.097
5	-88.80	-88.40	-88.38	-87.04	-1.743	+0.015	0.235	1.492	-0.075
6	-88.81	-88.31	-88.40	-87.07	-1.696	-0.033	0.232	1.497	-0.104
7	-88.87	-88.20	-88.42	-87.10	-1.683	-0.058	0.225	1.515	-0.108
8	-88.97	-88.25	-88.46	-87.20	-1.671	-0.036	0.224	1.482	-0.119
9	-89.00	-88.23	-88.48	-87.20	-1.664	-0.047	0.226	1.486	-0.098
10	-88.98	-88.17	-88.44	-87.18	-1.587	-0.082	0.196	1.473	-0.093
11	-89.00	-88.16	-88.46	-87.22	-1.543	-0.171	0.209	1.504	-0.129
12	-89.10	-88.28	-88.54	-87.31	-1.583	-0.071	0.167	1.487	-0.086
13	-89.10	-88.30	-88.53	-87.34	-1.579	-0.132	0.166	1.546	-0.072
14	-89.12	-88.28	-88.53	-87.32	-1.526	-0.118	0.167	1.477	-0.099
15	-89.15	-88.30	-88.55	-87.30	-1.496	-0.139	0.161	1.474	-0.116
16	-89.18	-88.30	-88.57	-87.33	-1.497	-0.149	0.166	1.481	-0.102

Figures 4-7 show the individual behavior of the transfer standards, and figure figure 8 shows the behavior of the transfer group on the average. One might conclude based on these graphs that the cells were not sufficiently stabilized at the beginning of the experiment and that the first two measurements in the participant's laboratory should be deleted from the transfer data.

The differences for the transfer cells and the reference cells from their group means (See equations (4.5.3) and (4.5.4)) are listed in exhibit 4.5.2. The behavior of the reference cells during the transfer with NBS is of interest because the final assignment of offset depends on the assumption that the reference cells are stable. As was noted earlier in this section, the quantities listed in exhibit 4.5.2 do not describe the behavior of the individual reference cells because these quantities are constrained so that their sum is equal to zero.

The only way to observe the individual cells during the transfer is to reverse the way in which the assignments are currently made; i.e., to analyze the data from the intercomparisons using the reference cells as unknowns and the value of the transfer group from NBS as the restraint. This will give individual values for each reference cell and can be done after the fact if the transfer group proves sufficiently stable. The rationalization for computing an offset using the reference cells as the restraint is that one would expect the reference cells, if they are of the same quality as the transfer cells, to be more stable considering they have not recently been in transit.

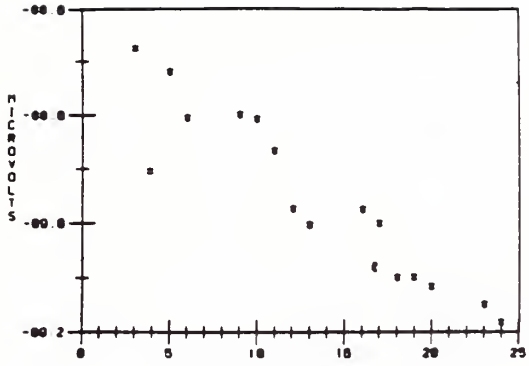


Figure 4
Values (μV) assigned to transfer standard T_1 versus time (days)

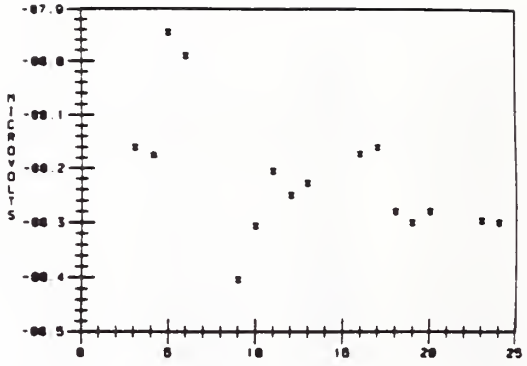


Figure 5
Values (μV) assigned to transfer standard T_2 versus time (days)

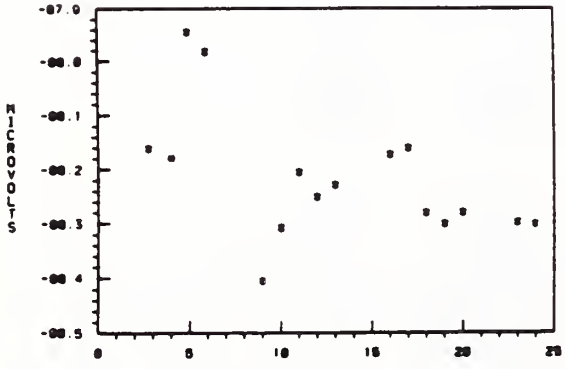


Figure 6
Values (μV) assigned to transfer standard T_3 versus time (days)

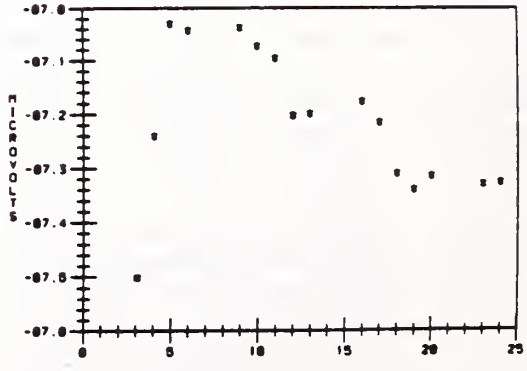


Figure 7
Values (μV) assigned to transfer standard T_4 versus time (days)

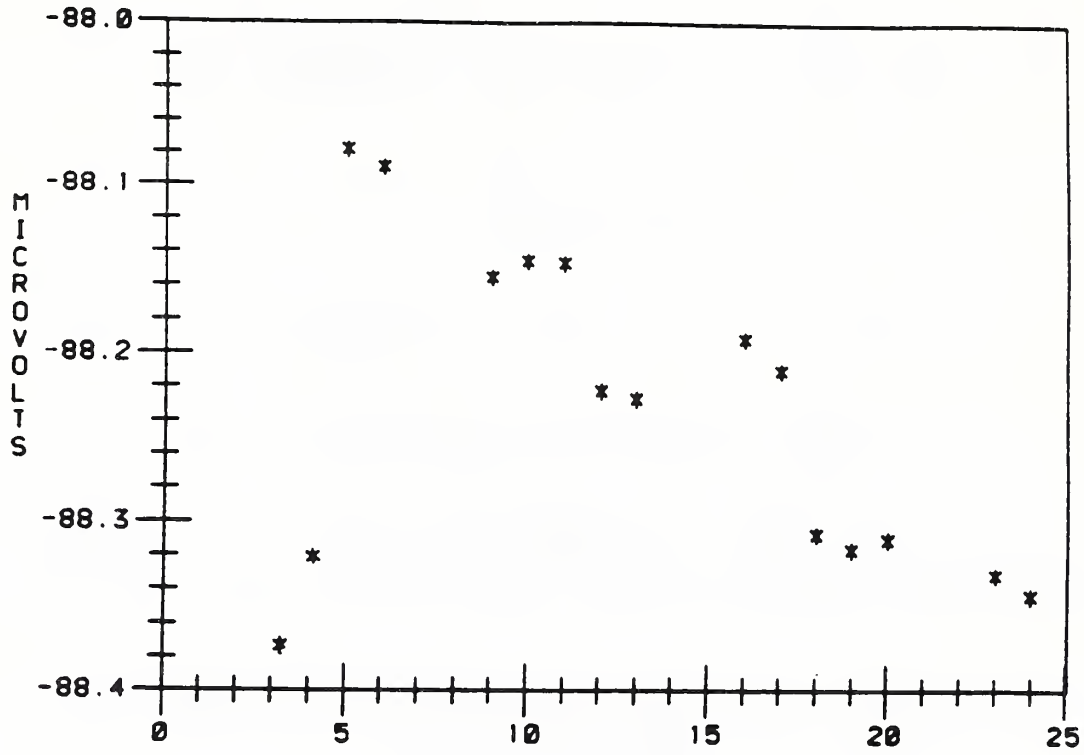


Figure 8
Average values (μV) assigned to four transfer standards versus time (days)

Each day's intercomparisons are analyzed for internal consistency via an F test on the within standard deviation for that day. The stability of the three designated check standards is also tested each day. Results of those designs which show evidence of lack of statistical control or anomalous behavior on the part of one of the check standards are excluded from the transfer experiment. Because we do not have prior history on this measurement process, we rely on hypothetical data to demonstrate to the reader the analysis that should be done for each design.

The left-right effects (4.5.2) are plotted in figure 9. Their respective test statistics (4.5.8) are listed in exhibit 4.5.3. Upper and lower control limits in figure 9 are indicated by dashed lines, and points that fall outside these control limits are equivalent to the corresponding test statistics being significant. These computations assume that prior data on the left-right effect established a standard deviation for the left-right effect of $s_\zeta = 0.02\mu V$ with $\nu_1 = 50$ degrees of freedom and that the accepted value of the left-right effect was established as $A_\zeta = 0.100\mu V$ from the same data.

Check standards C_1 and C_2 as constructed in (4.5.7) are observed differences between two reference cells and do not depend on the restraint or the design. Tracked over a period of time they show the way in which two cells are changing in respect to each other. Their values are listed in exhibit 4.5.3 and plotted as a function of time in figures 10-11.

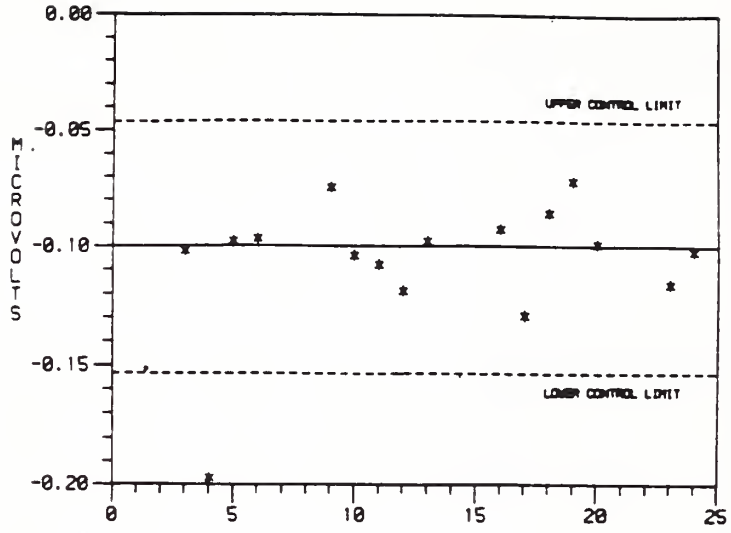


Figure 9
Left-right effect (μV) plotted against time (days) with dashed lines indicating upper and lower control limits at the 1% significance level

Exhibit 4.5.3 - check standards and test statistics[†]
Values in microvolts

Run	Date	Left-right Effect	Test Stat	Check Std	Test Stat	Check Std	Test Stat
#	t'	$\hat{\zeta}$	t_ζ	c_1	t_{c_1}	c_2	t_{c_2}
1	3	-0.102	0.1	-2.0450	0.22	-1.6075[¶]	2.90[¶]
2	4	-0.197[§]	4.8[§]	-2.0100	0.29	-1.5375	0.51
3	5	-0.098	0.1	-1.9650	1.10	-1.5175	0.29
4	6	-0.097	0.2	-1.9625	0.58	-1.5100	0.68
5	9	-0.075	1.2	-1.9775	1.66	-1.4775	2.16
6	10	-0.104	0.2	-1.9275	0.69	-1.5300	0.69
7	11	-0.108	0.4	-1.9075	0.66	-1.5725	0.46
8	12	-0.119	1.0	-1.8950	0.85	-1.5175	1.37
9	13	-0.098	0.1	-1.8900	1.27	-1.5300	1.14
10	16	-0.093	0.4	-1.7825	0.25	-1.5550	0.84
11	17	-0.129	1.4	-1.7525	0.58	-1.6675	2.35
12	18	-0.086	0.7	-1.7500	0.09	-1.5575	1.05
13	19	-0.072	1.4	-1.7450	0.32	-1.6775	2.31
14	20	-0.099	0.0	-1.6925	0.65	-1.5950	0.25
15	23	-0.116	0.8	-1.6575	0.01	-1.6125	0.18
16	24	-0.102	0.1	-1.6625	0.66	-1.6300	0.17

[†]We choose to illustrate the control procedure at the 1% significance level.
[§]Failed test for control at 1% level of significance based on a critical value $t_{.005}(50) = 2.678$ from Table I.
[¶]Failed test for control at 1% level of significance based on a critical value $t_{.005}(100) = 2.626$ from Table I.

For this analysis, it was assumed that data from fifty-one initial designs established a linear relationship with time for each check standard as follows:

$$c_1 = -2.095 + 0.0190t$$
$$c_2 = -1.501 - 0.00513t \qquad (4.5.25)$$

and that standard deviations, s_{c_1} for C_1 and s_{c_2} for C_2, were pooled to form a process standard deviation $s_c = 0.030\mu V$ with $\nu = 100$ degrees of freedom.

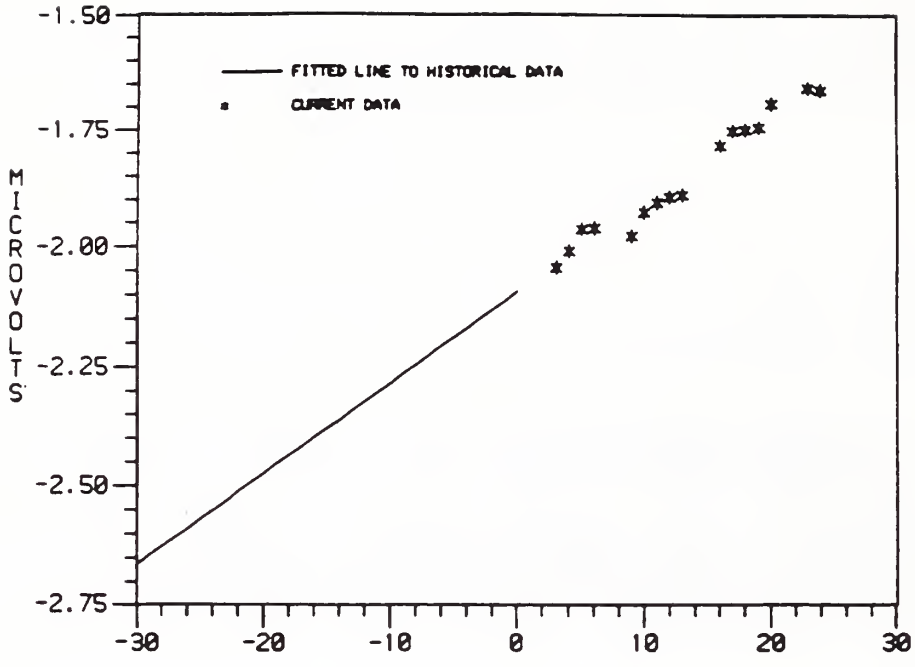
Based on the foregoing assumption, predicted values (4.5.11) for the check standards were computed for each time t' that the transfer standards were measured in the participant's laboratory. Given this information, the check standard measurements on each day were tested for agreement with the extrapolated line by the test statistics listed in exhibit 4.5.3. The test statistics for C_1 and C_2 that are shown in exhibit 4.5.3 were computed from (4.5.15) with n = 31 and values of $t_i(i=1,\cdots,31) = -30(1)0$.

The same analysis is shown graphically in figures 10-11. The upper portion of figure 10 shows the linear fit to the historical data as a solid line, and the values of check standard C_1 for the transfer experience are shown as discrete points, (*) with the convention that the transfer experiment starts at time t = 0.

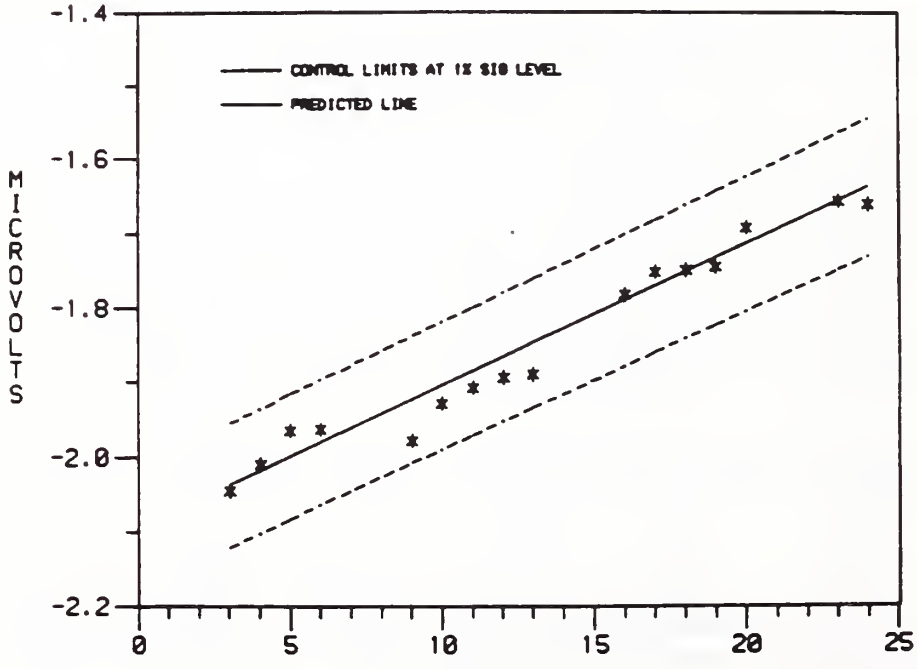
The lower portion of figure 10 shows the analysis of the check standard measurements. The solid line is an extrapolation of the linear fit from the upper portion of the same figure to the time of the transfer experiment. The dashed lines are upper and lower control limits that show the range within which the check standard measurements are expected to deviate from the extrapolated line. A point being outside these control limits is exactly analogous to the corresponding test statistic being significant in exhibit 4.5.3. Although it is not readily apparent from the graph, the control limits become wider as the check standard measurements are further removed in time from their historical data base. Thus, there is a smaller chance of detecting anomalous behavior as the experiments are continued into the future if the data base is not updated frequently.

Figure 11 shows the same analysis for the values of check standard C_2 from the transfer experiment with check standard C_2 out-of-control on the first day.

The within standard deviations are listed in exhibit 4.5.4 and plotted in figure 12. An F test based on an accepted standard deviation $s_p = 0.02\mu V$ with $\nu_3 = 408$ degrees of freedom indicates that there are measurement problems on the first and eleventh days. It is interesting to note that check standard C_2 is low on the eleventh day although it is not actually out-of-control and that the left-right effect is very close to being out-of-control on that same day. Given the responses of the check standards and the transfer standards and the information garnered from the control procedure, it would seem reasonable to delete three measurements from the transfer data; namely, the first, second and eleventh days' measurements.

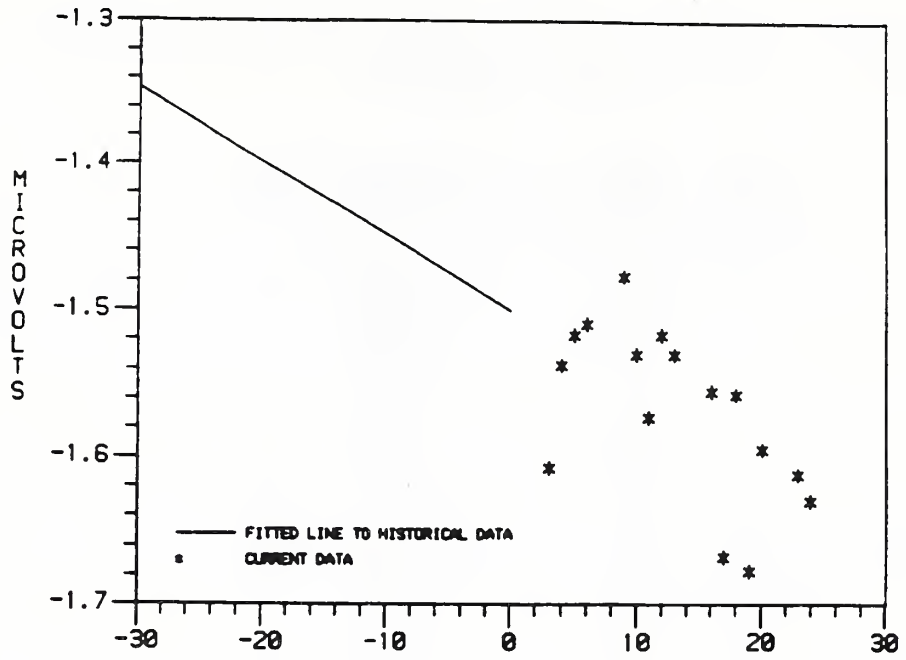


(a)

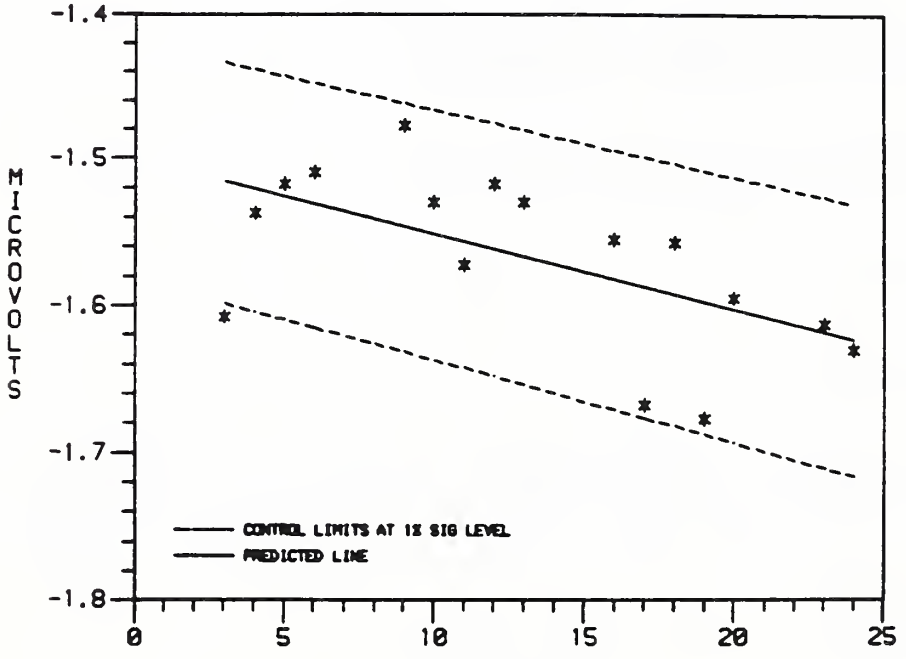


(b)

Figure 10
Check standard C_1 (μV) plotted against time (days)
with a solid line indicating a predicted linear fit and dashed lines
indicating upper and lower control limits at the 1% level of significance



(a)



(b)

Figure 11
Check standard C_2 (μV) plotted against time (days)
with a solid line indicating a predicted linear fit and dashed lines
indicating upper and lower control limits at the 1% level of significance.

Exhibit 4.5.4 - Within sandard deviations and test statistics
Values in microvolts

Run #	Date t	Within SD s_w	DF ν_3
1	3	0.054§	8
2	4	0.018	8
3	5	0.019	8
4	6	0.015	8
5	9	0.022	8
6	10	0.011	8
7	11	0.016	8
8	12	0.021	8
9	13	0.013	8
10	16	0.021	8
11	17	0.054§	8
12	18	0.018	8
13	19	0.031	8
14	20	0.013	8
15	23	0.018	8
16	24	0.011	8

§Failure to satisfy the inequality $s_w < s_p\sqrt{F_{.01}(8,\infty)}$ at the 1% significance level based on $s_p = 0.02\mu V$ and a critical value $F_{.01}(8,\infty) = 2.5$ from Table II.

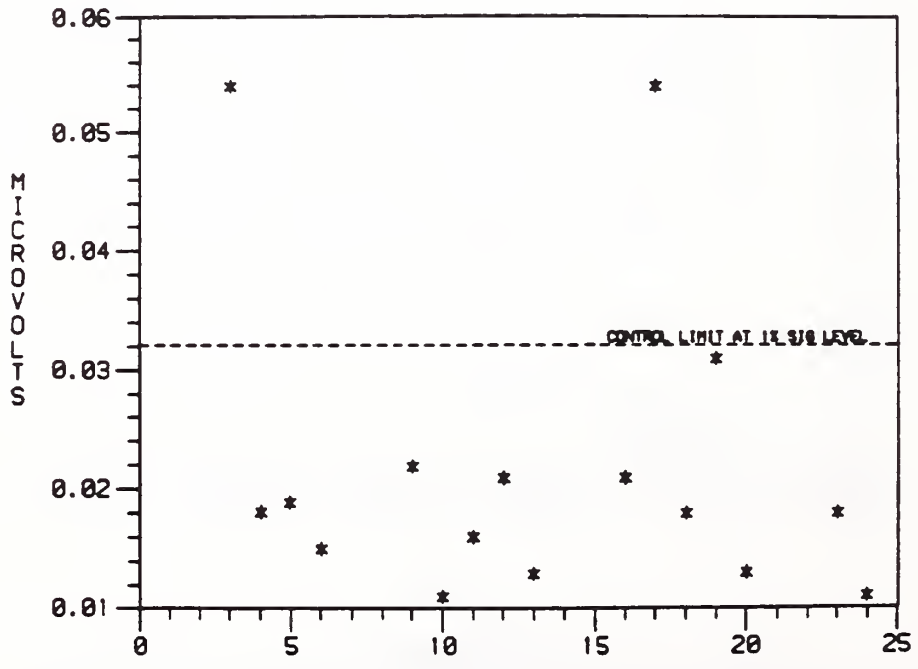


Figure 12
Within standard deviations (μV) plotted against time (days)
with dashed line indicating control limit at 1% level of significance.

4.6 Direct Reading of the Test Item with an Instrument Standard

4.6.1 Measurement Sequence

In this mode of operation a value is directly assigned to a test item X by a calibrated instrument. Observations on a stable artifact that takes on the role of the check standard C are used to establish a base line for the instrument and to maintain and control its variability in what amounts to a surveillance type test. An observation on the test item is denoted by x, and an observation on the check standard is denoted by c.

4.6.2 Process Parameters

Initial values of the process parameters are obtained from n independent measurements on the check standard $c_1, \cdots, c_n$. The accepted value of the check standard is defined by the mean of the check standard measurements; namely,

$$A_c = \frac{1}{n} \sum_{i=1}^{n} c_i \, . \tag{4.6.1}$$

The total standard deviation of the instrument is

$$s_c = \left(\frac{1}{n-1} \sum_{i=1}^{n} (c_i - A_c)^2 \right)^{1/2} \tag{4.6.2}$$

with $\nu = n-1$ degress of freedom. The control limits[n] that are appropriate for future observations on the check standard are given by

$$\text{Upper control limit} = A_c + 3s_c$$

$$\text{Lower control limit} = A_c - 3s_c \, .$$

4.6.3 Control Procedure

The primary purpose of the control procedure is to monitor instrumental drift, and observations on the check standard should be taken frequently enough to ensure that such drift is being contained. A test statistic t_c computed from the most recent check standard measurement c is given by

$$t_c = \frac{|c - A_c|}{s_c} \, .$$

The process is in control at the time of the check standard measurement c if

$$t_c < 3 \, . \tag{4.6.3}$$

[n]The factor 3 is used in this and all subsequent computations in place of the appropriate percent point of the t distribution; namely, $t_{\alpha/2}(\nu)$.

If

$$t_c > 3 ,$$

the process is not in control at the time of the check standard measurement, and measurements should be discontinued until the problem with the instrument is rectified.

4.6.4 Transfer with NBS

Determination of systematic error can be made by making p measurements $r_1,\cdots,r_p$ on a calibrated artifact or transfer standard which has an assigned value T^* and associated uncertainty U_T. Instrumental offset ψ defined by

$$\psi = \frac{1}{p}\sum_{i=1}^{p} r_i - T^* \qquad (4.6.4)$$

is not significant if

$$\frac{\sqrt{p}\,|\bar{\psi}|}{s_c} < 3 . \qquad (4.6.5)$$

It is extremely important to recognize that this approach makes two important assumptions that must be verified experimentally; namely, that the instrument has a constant offset from the NBS process over the regime of interest as in (1.4.2) and that the precision of the instrument is constant over this same regime. The question of constant offset is considered first. A single point is not sufficient for the determination, and the system must be checked using several calibrated artifacts that span the regime of interest. Assume that m transfer standards are sufficient to verify the points of interest and that the transfer standards have assigned values $T_1^*,\cdots,T_m^*$ and associated uncertainties $U_{T1},\cdots,U_{Tm}$. Assume also that m offsets $\psi_1,\cdots,\psi_m$ computed according to (4.6.4) have been determined from measurements made on the transfer standards.

If all ψ_j $(j=1,\cdots,m)$ are insignificant as judged by (4.6.5), no adjustment to the instrument is needed. If the offsets are of varying magnitudes, and if it can be shown that these offsets are functionally related to the assigned values of the transfer standards, it may be possible to calibrate the instrument using a calibration curve based on the offsets (see section 2.3.3). Finally, if the offsets are significant and of the same magnitude, either the instrument is adjusted for the average offset

$$\bar{\psi} = \frac{\sum_{j=1}^{m} p_j \cdot \psi_j}{\sum_{j=1}^{m} p_j} \qquad (4.6.6)$$

where p_j $(j=1,\cdots,m)$ represents the number of measurements on the j^{th} transfer standard or a reading x on a test item x is reported as

$$X^* = x - \bar{\psi} .$$

The uncertainty of the transfer is

$$U_{tr} = \frac{3s_c}{\sqrt{p_1+\cdots+p_m}} + \frac{1}{m}\left(U_{T1}^2 + \cdots + U_{Tm}^2\right)^{1/2} . \qquad (4.6.7)$$

4.6.5 Uncertainty

The total uncertainty that is appropriate for one measurement made on a test item using the calibrated instrument is

$$U = U_{tr} + 3s_c. \qquad (4.6.8)$$

4.6.6 Process Precision.

The question concerning whether or not the precision of an instrument remains constant over a given regime can be addressed by comparing standard deviations from several levels in the regime. A familiar example is an electronic balance that is used over a large range of loads where the precision of the instrument may be load dependent. This assumption can be checked either with calibrated or uncalibrated artifacts.

Standard deviations with their associated degrees of freedom should be tabulated by load and inspected for consistency. It is possible to quote one uncertainty over the entire regime only if the precision is constant over all load levels; i.e., if these standard deviations are all of the same magnitude.

A visual inspection of the values may be sufficient for determining whether or not the standard deviations are of roughly the same magnitude in which case the standard deviations should be pooled using (2.2.3) and the uncertainty computed by replacing s_c in equations (4.6.7) and (4.6.8) with the pooled standard deviation.

If there is some question about the propriety of combining all the standard deviations, the largest standard deviation can be checked for agreement with the others using a test developed by Cochran [57]. A description of the test statistic and tables for deciding whether or not the largest standard deviation in a group is significantly different from the group are tabulated by Eisenhart [58].

If it is logical to assume that the precision of the instrument will vary with the magnitude of the quantity of interest, then a series of check standards should be established, one at each level of interest, with the estimate of process precision (4.6.2), the test for statistical control (4.6.3), and computation of uncertainty (4.6.8) being made at each level independently, thus begging the question of constant variability.

4.7 Simultaneous Measurement of a Group of Test Items and a Group of Reference Standards

4.7.1 Measurement Sequence

This scheme is appropriate for assigning values to individual test items or instruments relative to the average of a bank or group of reference standards, called the restraint R*, when all items including the standards are simultaneously subjected to the same stimuli such as a power source or a vacuum chamber. Assume there are m reference standards $R_1,\cdots,R_m$, and ℓ test items $X_1,\cdots,X_\ell$. One position in the configuration of test items should be reserved for a check standard Y, an artifact similar to the test items, where a reading on Y is always recorded along with the other readings.

Assume that a measurement sequence produces readings $r_1,\cdots,r_m$ on the standards, $x_1,\cdots,x_\ell$ on the test items and y on the check standards. The value that is recorded as the check standard measurement for one sequence is

$$c = y - \frac{1}{m}\sum_{i=1}^{m} r_i \ . \qquad (4.7.1)$$

In other words the measured difference between the artifact check standard and the average of the reference standards is the check standard measurement. In the remainder of this section, the term check standard refers to this recorded difference rather than the measured value y.

4.7.2 Process Parameters

Initial values of the process parameters are obtained from n such measurement sequences where $c_1,\cdots,c_n$ are the check standard measurements.

The accepted value of the check standard is the mean of these values; namely,

$$A_c = \frac{1}{n}\sum_{i=1}^{n} c_i \ . \qquad (4.7.2)$$

The total standard deviation of the check standard is

$$s_c = \left(\frac{1}{n-1}\sum_{i=1}^{n} (c_i - A_c)^2 \right)^{1/2} \ . \qquad (4.7.3)$$

Control limits[q] that are appropriate for future check standard observations are given by

$$\text{Upper Control Limit} = A_c + 3s_c$$

$$\text{Lower Control Limit} = A_c - 3s_c \ .$$

[q]The factor 3 is used in this and all subsequent computations in place of the appropriate percent point of the t distribution; namely, $t_{\alpha/2}(\nu)$.

The control procedure applied to each calibration depends on a test statistic t_c computed from the value of the check standard c for that measurement sequence by

$$t_c = \frac{|c - A_c|}{s_c} . \qquad (4.7.4)$$

If

$$t_c < 3 \qquad (4.7.5)$$

the process in control, and the value of a test item is reported as

$$X_j^* = x_j - \frac{1}{m} \sum_{i=1}^{m} r_i + R^* \qquad j=1,\cdots,\ell \qquad (4.7.6)$$

where $R^* = \frac{1}{m} (R_1^* + \cdots + R_m^*)$ and $R_1^*,\cdots,R_m^*$ are the values assigned to the reference standards. If

$$t_c > 3$$

the calibration of the test items is invalid and must be repeated.

4.7.3 Transfer with NBS

The transfer with NBS is accomplished by p repetitions of the measurement sequence during which a group of ℓ transfer standards $T_1,\cdots,T_\ell$ replaces the group of test items. Process control as defined by (4.7.5) should be confirmed for each repetition. Any sequence that is out-of-control should be repeated until control is restored or else that repetition is deleted from the transfer. The values assigned the transfer standards are $T_1^*,\cdots,T_\ell^*$ with uncertainties $U_{T1},\cdots,U_{T\ell}$.

The offset Δ_i $(i=1,\cdots,p)$ of the laboratory process from NBS for the ith repetition is based on the values assigned to the ℓ transfer standards by (4.7.6); namely, $X_1^*,\cdots,X_\ell^*$ where

$$\Delta_i = \frac{1}{\ell} \sum_{j=1}^{\ell} (X_j^* - T_j^*) \qquad i=1,\cdots,p$$

and the average offset computed for the p repetitions is

$$\bar{\Delta} = \frac{1}{p} \sum_{i=1}^{p} \Delta_i . \qquad (4.7.7)$$

The uncertainty of the transfer is

$$U_{tr} = \frac{3s_c}{\sqrt{p\ell}} + \frac{1}{\ell} \left(U_{T1}^2 + \cdots + U_{T\ell}^2 \right)^{1/2} . \qquad (4.7.8)$$

The offset is judged significant if

$$\frac{\overline{\sqrt{p}\,\ell}\,|\Delta|}{s_c} > 3 \qquad (4.7.8)$$

and in such case the assigned value of the restraint is changed to $R^* - \Delta$. The restraint is unchanged if

$$\frac{\overline{\sqrt{p}\,\ell}\,|\Delta|}{s_c} < 3.$$

4.7.5 Uncertainty

The total uncertainty that is appropriate for a value assigned to a test item by (4.7.6) from one calibration is

$$U = U_{tr} + 3s_c. \qquad (4.7.9)$$

4.8 Ratio Technique for One or More Test Items and One or Two Reference Standards

4.8.1 Measurement Scheme

In this section we describe calibration of a test item X by an instrument such as a scanning electron microscope which has only short-term stability. Consider the case where the test item X and the reference standard R are related by (1.4.9) and the instrument response is of the form (1.4.10). One reference standard R is sufficient to provide a calibrated value X^* for the test item given a single reading x on the test item and a single reading r on the reference standard. The calibrated value is

$$X^* = x \cdot R^*/r \qquad (4.8.1)$$

where R^* is the value assigned to the reference standard.

Where the test item and reference standard are related by (1.4.1) and the instrument response is of the form (1.4.6), two reference standards R_1 and R_2 are needed to calibrate a test item X (Cameron [60]). The artifacts should be measured in the sequence R_1, X, R_2 with the corresponding measurements denoted by r_1, x, r_2. The calibrated value for the test item is

$$X^* = R_1^* + \frac{(R_2^* - R_1^*)\cdot(x - r_1)}{(r_2 - r_1)} \qquad (4.8.2.)$$

where R_1^* and R_2^* are the values assigned to R_1 and R_2 respectively.

If before and after readings are taken on the test item in the sequence X, R_1, R_2, X with the measurements denoted by x_1, r_1, r_2, x_2 respectively, then the calibrated value for the test item is

$$X^* = \frac{1}{2}\left\{ (R_1^* + R_2^*) + \frac{(R_2^* - R_1^*)\cdot(x_1 - r_1 - r_2 + x_2)}{(r_2 - r_1)} \right\} . \quad (4.8.3)$$

More than one unknown can be calibrated from the same pair of readings on R_1 and R_2 only if the sequence of measurements can be arranged so that no test item is too far removed from R_1 and R_2 in the measurement scheme. For example, for test items X, Y, and Z, the sequence X, R_1, Y, R_2, Z minimizes the separation between unknowns and standards, and the calibrated value for each unknown is calculated according to (4.8.2).

In practice, it may be necessary to have several artifact standards that cover the operating range of the instrument. In addition to artifact standards for every level, it is necessary to have one artifact check standard Y for every level. A measurement y on the check standard should be included in the calibration program on a regular basis, and if feasible, with every calibration scheme. The check standard value that is used for controlling the process and for estimating random error is computed in exactly the same way as X^*. For example, for the measurement sequence described by (4.8.2), the check standard value from one calibration is

$$c = R_1^* + \frac{(R_2^* - R_1^*)\cdot(y - r_1)}{(r_2 - r_1)} . \quad (4.8.4)$$

4.8.2 Process Parameters

Initial values of the process parameters are obtained from n such calibration sequences yielding check standard values $c_1, \cdots, c_n$. The accepted value of the check standard is defined as the mean of the check standard values; namely,

$$A_c = \frac{1}{n}\sum_{i=1}^{n} c_i . \quad (4.8.5)$$

The total standard deviation of the check standard is defined by

$$s_c = \left(\frac{1}{n-1} \sum_{i=1}^{n} (c_i - A_c)^2 \right)^{1/2} \quad (4.8.6)$$

with $\nu = n-1$ degrees of freedom.

In this case s_c is the standard deviation of a calibrated value X^* and will reflect not only the imprecision in the measurements x, r_1, and r_2 but also any changes in the response curve for the instrument that are not accounted for by the ratioing device.

The control limits[r] that are appropriate for future check standard values are:

$$\text{Upper control limit} = A_c + 3s_c$$

$$\text{Lower control limit} = A_c - 3s_c \ .$$

4.8.3 Control Procedure

A control procedure is applied to each calibration sequence which includes a check standard measurement. The control procedure is based on a test statistic t_c computed from the check standard value c for that sequence; namely,

$$t_c = \frac{|c - A_c|}{s_c} \ .$$

If

$$t_c < 3 \qquad (4.8.7)$$

the process is in control, and the value of a test item X is reported as X^*.

If

$$t_c > 3,$$

the process is out-of-control, and the calibration of the test item is invalid and must be repeated.

4.8.4 Transfer with NBS

The tie to NBS is via the reference standards which are either standard reference materials from NBS or secondary calibrated artifacts.

4.8.5 Uncertainty

The uncertainty for an artifact calibrated according to (4.8.1) is

$$U = 3s_c + U_R \qquad (4.8.8)$$

where U_R is the uncertainty for R^*. The uncertainty for an artifact calibrated according to (4.8.2) or (4.8.3) is

$$U = 3s_c + \frac{1}{2}\left(U_{R1}{}^2 + U_{R2}{}^2\right)^{1/2} \qquad (4.8.9)$$

where U_{R1} and U_{R2} are the uncertainties for $R_1{}^*$ and $R_2{}^*$ respectively.

[r]The factor 3 is used in this and all subsquent computations in place of the appropriate percent point of the t distribution; namely, $t_{\alpha/2}(\nu)$.

5. Control Charts

5.1 Introduction

The industrial application of control charts involves a production process that yields product that is assumed to be homogeneous with respect to a particular property that is measurable. The control chart is devised to detect any variation in the production process that is not random in nature and which, therefore, can be assigned a cause. Guaranteeing that all variation in the production process is random in nature guarantees that the process is operating in an optimal fashion, and if, given these circumstances, the product is not within specifications, major adjustments to the process are required in order to substantially affect its output.

Once a base line and control limits have been defined for the process, based on prior data from the same process, the control chart is set up with a solid horizontal line representing the base line and dashed lines above and below the base line representing the control limits. Samples drawn at random from the production process are measured for the property of interest, and the resulting values are plotted on the control chart as a function of time. Values that fall within the control limits are referred to as being "in statistical control" and values that fall outside the control limits are referred to as being "out of control". Values outside the control limits are a sufficient indication that the "process should be investigated and corrected" (Bicking & Gryna [61]).

The Shewhart control chart discussed above is appropriate for individual measurements or averages of "natural" groups. This type of control chart, used in conjunction with a control chart for standard deviations, is a powerful means of detecting changes in the measurement process. Other types of control procedures include a cusum chart (Duncan[62]) which is particularly useful for detecting gradual drifts in a continuous process as compared with abrupt shifts. Methods for detecting changes in both the base line of the process and in the variability of the process on a single control chart are discussed by Reynolds and Ghosh in reference [63].

Statistical control as originated by Shewhart [64] assumes that repeated measurements of a reproducible property are available and that these measurements constitute a random sample of all such possible measurements from a known distribution such as the normal distribution. The term random sample implies two important properties of the measurements; namely, that they are independent and that they all come from the same distribution. The average value and standard deviation calculated from a random sample in conjunction with known properties of the distribution are used to calculate limits within which a certain percentage of all measurements should fall. In other words, a series of initial measurements are made to characterize the distribution of all possible measurements, and future measurements are checked for conformity with this distribution.

Notice that one is not concerned with whether or not the product is within certain specification limits, but rather with whether or not the production process is behaving properly. The control procedure for a measurement process is similar in many respects to industrial control. In the measurement

assurance context the measurement algorithm including instrumentation, reference standards and operator interactions is the process that is to be controlled, and its direct product is measurement per se. The measurements are assumed to be valid if the measurement algorithm is operating in a state of control; i.e., if the variations in that process are due to random causes which can be quantified, thus assuring that a value reported by the process will have negligible offset from national standards within predictable limits. This will be the case if the control chart shows that the base line for the process is not changing.

Statistical control in the measurement assurance context can conversely be predicated on the assumption that the measurement process is stable and that lack of control indicates a change in the artifact being measured. There are circumstances where this type of control is needed--that is, when it is necessary to know whether or not an artifact has changed with respect to the property being measured. For example, a transfer standard that is being circulated to several laboratories must be checked periodically at NBS. Similarly, intercomparisons between working standards and primary standards can be subjected to a control procedure to ensure that the working standards have not changed appreciably. In these instances, lack of control will result in either replacing the artifact in question or in reassigning its accepted value.

Calibration control is perhaps dissimilar to industrial control in that although artifacts submitted for measurement are of the same general type, their properties must be quantified individually. Thus, there is an inherent problem in controlling the values assigned to individual artifacts or instruments because the measurement is rarely repeated, let alone repeated sufficiently often to characterize the distribution of possible values. Without a historical data base there is no way of determining whether or not the current calibration is in control or is, in fact, a proper assignment for the item. For this reason a check standard is introduced into the measurement sequence in such a way that it can be assumed that the measurement algorithm acts on the check standard in much the same way as it acts on the item being calibrated. The redundant measurements on the check standard are the basis for both characterizing the distribution of measurements and deciding if the measurement process is in control on a given occasion.

The control limits are chosen so that the probability is 100α percent that future measurements will fall outside the control limits strictly by chance. Therefore, α if always chosen small, say $\alpha = .01$ or $\alpha = .05$ so that very few measurements will be discarded unnecessarily. Smaller values of α correspond to wider control limits which result in the measurement almost always being accepted unless there is a serious shift in the process. The converse is also true--larger values of α correspond to narrower control limits which result in tighter control of the measurement process with more frequent remeasurement. Obviously, the success that can be expected in detecting changes in the process which is referred to as the power of the control procedure is linked to the choice of α.

The reader may have already noted that the procedure for determining control or lack thereof is exactly analogous to a statistical t-test for deciding whether or not a single observation comes from a process with known mean and unknown standard deviation.

5.2 Control Charts for Single Measurements

The measurements for initiating the control chart must be collected over a sufficiently wide range of operating conditions to ensure a correct characterization of the distribution and over a sufficently long period of time to ensure independence. Grant and Leavenworth state that ideally twenty-five measurements should be spread over several months time [65]. As few as ten or fifteen measurements can suffice if this data base is updated when more measurements are available. The measurements are plotted as a function of time without imposing a base line or control limits on the plot in order to track the measurement process and verify that it produces stable measurements whose variability is random in nature. Such a plot also allows one to check specification limits, but specification limits do not constitute statistical control because they do not have a probabilistic interpretation.

When one is satisfied that the initial measurements are adequate for representing the distribution and that process variability is tolerable, a base line and control limits are computed from this data base.

For single measurements the base line is taken to be the average of initial measurements $x_1, \cdots, x_n$; namely

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{5.2.1}$$

and the control limits are taken to be

$$\begin{aligned} &\bar{x} + s \cdot t_{\alpha/2}(\nu) \\ &\bar{x} - s \cdot t_{\alpha/2}(\nu) \end{aligned} \tag{5.2.2}$$

where s, the total standard deviation computed from the initial measurements is

$$s = \left(\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \right)^{1/2} \tag{5.2.3}$$

with $\nu = n-1$ degrees of freedom. The number $t_{\alpha/2}(\nu)$ is the $\alpha/2$ percentage point of Student's t distribution with ν degrees of freedom.

Once the average value and the control limits have been established, future measurements are tested for control. One concludes that measurements that fall within the control limits come from the hypothesized distribution, and that, therefore, the measurement process is acting in an acceptable and predictable manner. The converse is also true. Measurements that fall outside the control limits infer a significant change in the process. Where such a change is noted, one must determine whether the change is permanent or transitory.

In a measurement assurance context, every violation of the control limits requires a remedial action. It may be sufficient to simply repeat the offending measurement in order to reestablish control, but all measurements since the last successful test for control are discarded once an out-of-control condition occurs.

As an example, consider how repeated measurements on a calibrated weight can be used to demonstrate that an electronic balance is, indeed, weighing accurately at all times. Accuracy in this context means that values delivered by the balance are in agreement with national standards (prototype kilogram) as maintained by NBS within the stated uncertainty. Parobeck et al [66] describe a measurement assurance program for large volume weighings on electronic balances where redundancy and control are achieved by repeating weighings of selected test items on different days.

A program to control a weighing process is begun by making n initial measurements on the calibrated weight, being sure to allow enough time between successive measurements to cover a range of operating conditions in the laboratory, and using these initial measurements as a historical base for computing the average $\bar{x}$ and the standard deviation s of the balance.

Given a calibrated value A with uncertainty U_A for the weight, the balance is accurate within the uncertainty $U_A \pm s \cdot t_{\alpha/2}(n-1)$ if

$$A - \frac{s}{\sqrt{n}} \cdot t_{\alpha/2}(n-1) - U_A < \bar{x} < A + \frac{s}{\sqrt{n}} \cdot t_{\alpha/2}(n-1) + U_A .$$

Notice that this test takes into account both the limits to random error for the measurement process, $\pm s \cdot t_{\alpha/2}(n-1)/\sqrt{n}$, and the uncertainty associated with the calibrated value of the weight, U_A.

Once the accuracy has been verified, the control phase of the program is pursued by remeasuring the weight from time to time. The resulting values are plotted on a control chart having base line and control limits as defined in equations (5.2.1) and (5.2.3), and it is presumed that the balance continues to be accurate as long as

$$\bar{x} - s \cdot t_{\alpha/2}(n-1) < y_i < \bar{x} + s \cdot t_{\alpha/2}(n-1)$$

for all future measurements y_i.

There is always a question, in this type of application, of how often one should check for control. It seems obvious, particularly if one is dealing with electronic instrumentation, that there should always be a check for control as part of any start-up procedures. After that, the frequency is dictated by the past performance of the system and by the amount of inconvenience and expense that is generated when an out-of-control condition is encountered--keeping in mind that when the balance is found to be out-of-control, it is necessary to recall all the measurements that were made on that balance since the previous successful check for control.

5.3 Control Charts for Averages or Predicted Values

Thus far, the discussion has centered on control charts for individual measurements, and it is easily extended to include control charts for averages that are completely analagous to the control charts for individual measurements. When the reported value of a measurement sequence, be it an average or a predicted value from a least-squares analysis, is computed from k intercomparisons that were made over a relatively short period of time, the "measurement of interest" is the corresponding average or predicted value of the check standard. This quantity is treated analogously to a single measurement with base line and control limits for the control chart determined from n such initial quantities. That is, given check standard values $x_1,\cdots,x_n$ each of which is an average or predicted value from k intercomparisons, the grand mean $\bar{x}$ computed from (5.2.1) represents the base line of the process and control limits as in (5.2.2) can be calculated using the total standard deviation s from (5.2.3). In this case the quantity s is the standard deviation of an average or predicted value and not the standard deviation of a single measurement from the process.

5.4 Control Charts for Within Standard Deviations

For a measurement scheme involving k intercomparisons, it is possible to generate a control chart for what is called the "within" or short-term variability of the process.

Assume that each check standard value x_i $(i=1,\cdots,n)$ is the result of k intercomparisons; namely, $x_{i1},\cdots,x_{ik}$ where the quantity x_i is the average of these intercomparisons,

$$x_i = \frac{1}{k} \sum_{j=1}^{k} x_{ij} \quad .$$

The within standard deviations are estimated by

$$s_{w_i} = \left(\frac{1}{k-1} \sum_{j=1}^{k} (x_{ij} - x_i)^2 \right)^{1/2} \qquad (5.4.1)$$

with degrees of freedom $\nu_i = k-1$. Where the intercomparisons form a statistical design, the quantity x_i and the within standard deviation are computed from a least-squares analysis.

The base line and limits for controlling short-term process variability make use of the same intercomparisons that were used to establish the control chart for averages. The base line is the pooled within standard deviation

$$s_p = \left(\frac{\nu_1 s_{w_1}^2 + \cdots + \nu_n s_{w_n}^2}{\nu_1 + \cdots + \nu_n} \right)^{1/2} \qquad (5.4.2)$$

The degrees of freedom $\nu = \nu_1 + \cdots + \nu_n$ allow for a different number of degrees of freedom in each estimate of the within standard deviation in (5.4.1). If all measurement schemes contain the same number of intercomparisons, say k, then $\nu = n(k-1)$.

Because a standard deviation is a positive quantity, it is only necessary to test against an upper limit in order to test the short-term variability. Thus for any future measurement sequence involving k intercomparisons, the within standard deviation s_w is computed as in (5.4.1) and is said to be in-control if

$$s_w < s_p \sqrt{F_\alpha(k-1, \nu)} \qquad (5.4.3)$$

where $F_\alpha(k-1, \nu)$ is the upper α percent point of the F distribution with k-1 degrees of freedom in the numerator and ν degrees of freedom in the denominator.

The control chart for averages used in conjunction with the control chart for within standard deviations is a powerful means of detecting changes in the process. The two control procedures are evoked simultaneously, and if an out-of-control condition is encountered for either test, the process is assumed to be out-of-control and the measurement sequence is repeated.

5.5 Alternative Control Limits

The reader may be familiar with control charts with control limits computed as the product of the total standard deviation and a fixed multiplicative factor, such as two or three, instead of the appropriate percentage point of the F or t-distribution. Control charts for within standard deviations should always be based on the F distribution because the critical values of the F distribution change rapidly with changes in degrees of freedom.

The consideration of whether a control chart for averages should be based on the percentage points of Student's t distribution or on a fixed multiplicative factor, such as three or two, is really a matter of choice depending on the level of control that one is hoping to achieve and on the type of measurements that are in question. The use of Student's t distribution is the most rigorous test if the measurements truly represent a random sample from a normal distribution. It allows a strict probability interpretation of the control procedure.

It cannot always be shown, and indeed is not always the case, that measurements come from an idealized distribution such as the normal distribution. If one looked at a large number of measurements on the same item, they might come from a distribution that is slightly skewed; i.e., for example, extreme large values may be more likely than extreme small values.

> The problem of deciding whether to use limits based on the normal distribution, those based on some other distribution, or those which involve no assumption about the form of the distribution is one which, though of a kind common in applied statistics, has no satisfactory solution. Limits based on the normal distribution are substantially shorter for a fixed sample size than those based on no assumption about the distribution, but they may be irrelevant if the distribution is too far from normal. (Bowker [67]).

For this reason it is customary in the United States to use plus or minus three standard deviations as the control limits (Duncan [68]). The factor three guarantees that a large proportion of the distribution is covered for measurements coming from any distribution that is close to the normal distribution in character. These limits are robust, and should be used when the intent is to identify measurements that are clearly out-of-control. Because these limits are so wide, an out-of-control finding is almost certainly an indication of a serious malfunction in the measurement process. If a somewhat tighter control is desired, two standard deviation limits can be considered. Very few values will fall between the two and three standard deviation limits, and the price of remeasuring for those few may be worth the added degree of control.

5.6 Control Charts for Drifting Check Standards

Another consideration concerns the problem of drifting check standards and whether or not they can be used for control purposes. The assumption is made in most measurement control programs that the check standard is stable and that any change that is noted by the control procedure is caused by changes in the measurement process itself. Obviously if the check standard is not completely stable, the ability to detect a change in the process is confounded with any possible drift in the check standard.

Unfortunately the situation in reality is that artifacts may not be completely stable, and this instability will be detected when it is large compared to the process precision. Changes in check standards over time can be expected. Of the forty or more check standards that are in continual use in the NBS mass calibration program, only about half of those standards are completely stable or do not show any drift over time. The question is, "Can a drifting check standard be used for control purposes?" Sometimes it can, but a drifting check standard causes complications in the analysis when, depending on the rate of change, the control limits pick up this change.

There are a few ad hoc procedures that can be used in lieu of a rigorous approach to this problem. Probably the simplest approach is to determine the time interval over which the check standard is stable by studying historical data and to enforce the control procedure over this interval. When this time interval has elapsed or when numerous values have been flagged as being out-of-control, the base line and control limits can be adjusted based on more recent measurements on the check standard.

If the check standard is changing steadily, as is the case for many artifact standards at NBS, it is sometimes possible to model the rate of drift and to predict from this model a value for the check standard at a future time that is not too far removed from the present. This involves fitting a regression equation to the measurements as a function of time by the method of least-squares and computing the values of the check standard for future times. Then the control procedure is time dependent; the base value is the predicted value from the regression equation at that time, and the control limits which depend on the standard deviation of this predicted value become wider with time. This approach has been used at NBS for check standards with linear drift rate as a function of time. It can work reasonably well as long as the drift remains linear, but the cause of a breakdown in the linearity assumption cannot be easily identified because it is never really possible to separate the change in the artifact from the change in the process. In such a situation it is imperative that the process be checked frequently for offset by comparison to a national standard or to other stable laboratory standards.

5.7 Synopsis and Examples of Control Charts

Four important ideas that are pertinent to calibration programs should emerge from the disucssion thus far. First, when dealing with statistical control of the properties of an artifact or statistical control of a measurement process, the control parameters are not imposed upon the process externally but are characteristic of the measurement process itself as described by historical data.

Secondly, if the check standard measurement is outside the established control limits, the calibration sequence is presumed to be out-of-control, and the calibrations of the test items are considered invalid. When such a condition is initially encountered, the instrumentation can be checked and the measurement sequence repeated--testing again for control. Any intervening results should be discarded. If control cannot be restored, a significant change has occurred in the process, and this change must be investigated. If a process is repeatedly out-of-control, the base line and control limits should be reestablished based on more recent data.

If the check standard measurement is in-control, this is taken as evidence that the process is behaving as expected in relation to the item submitted for calibration, and its assignment is assumed to be correct. Lack of control is certainly grounds for rejecting the calibration of the test item, but the complimentary argument is not as strong. The relationship between the measurement on the test item and the measurement on the check standard must be interrelated or executed very close together in time in order to be satisfied that the assignment ot the check standard has, indeed, been done properly.

Thirdly, the process precision is very well characterized by a total standard deviation calculated from measurements on the check standard. In some cases, such measurements provide the only way of obtaining a realistic estimate of this source of uncertainty. Fourthly, even though the tests for control can be automated, it is not only advantageous to visually examine the control charts in order to detect anomalies or slight shifts in the process and possible drifting of the check standard over time, but it is essential for understanding the long term behavior of the measurement process.

In order to demonstrate the value of such critical examinations, four examples that have been encountered in NBS measurement assurance programs are discussed.

The National Bureau of Standards maintains control charts on about forty check standards that are used in the mass calibration program. The control chart shown in figure 13 depicts values of the one kilogram check standard as it has been estimated from the measurement sequence used in the calibration workload for one kilogram weights. The three standard deviation limits shown by the dashed lines are the control limits that are used for this program, and if one compares these limits with the two standard deviation limits shown by the dotted lines, it is apparent that very few points fall between the two sets of limits. It can also be noted that the two standard deviation control limits are almost identical with control limits based on student's t distribution at significance level $\alpha = 0.01$ when the number of points is large as in this case.

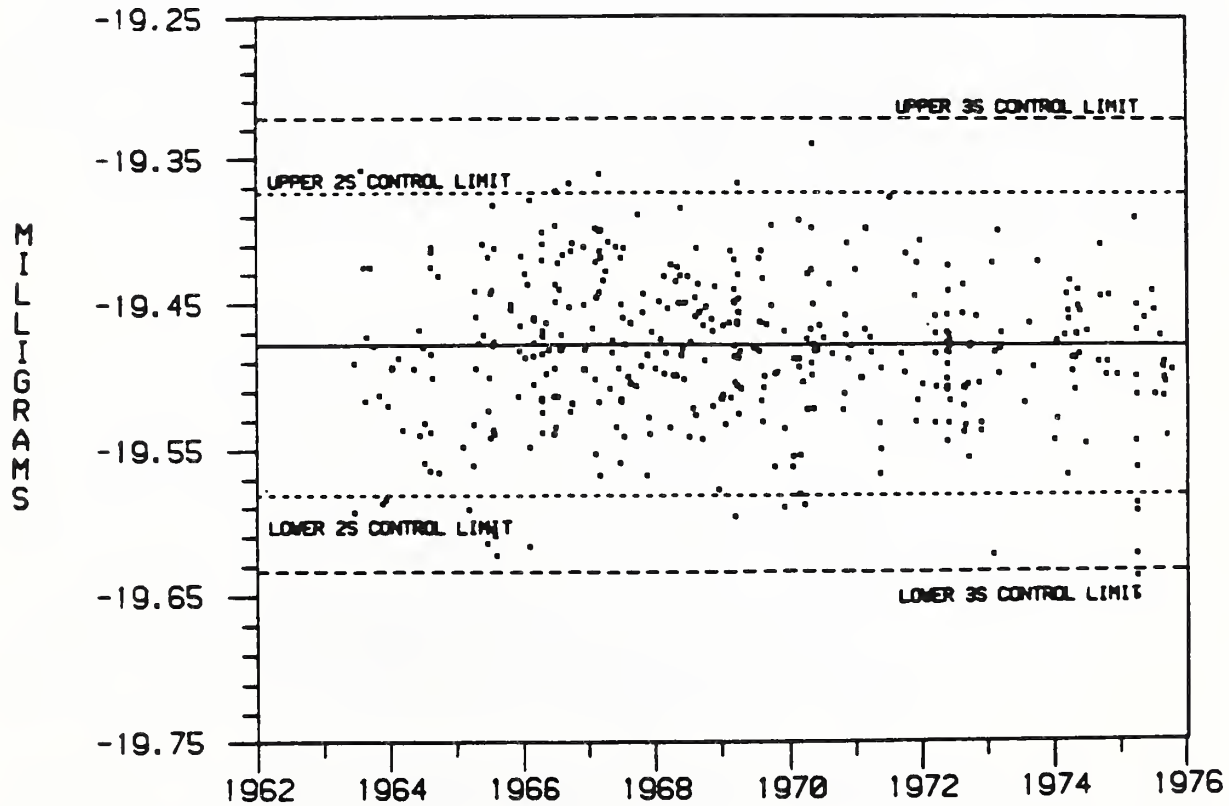


Figure 13
Check standard #41 (mg) as measured on NBS balance #4
plotted against time (years)

At this point the reader should be sufficiently sensitized to this approach to be aware of one shortcoming in this control chart. The chart implies that the process, which is demonstrably in-control, has never been out-of-control. A few points should fall outside of the control limits merely by chance, and as it happens other out-of-control situations have occurred in this program over the years. In fact, the control procedure would serve no useful purpose in the calibration program if there were no out-of-control situations to be detected. Actually this graph represents only the successful tests for control that were made with the one kilogram check standard because the calibration results and the check standard values were automatically discarded whenever the control limits were violated. The software for the NBS mass calibration program has been changed so that all values of the check standard are retained, and each value is flagged as to whether or not it was in control on that occasion. One should know when and how often control limits have been violated, and control charts should contain all findings.

The short-term or within variability of the same process is charted in figure 14 which shows within standard deviations for calibration sequences involving all weights calibrated on NBS balance #4. A calibration sequence typically requires between three and fifteen measurements, and the within standard deviation that is calculated from each sequence reflects the inherent variability of the balance and the effect of any environmental changes that occur during the time needed to make the requisite measurements. The base line for this control procedure, shown by the solid line, is the pooled within standard deviation in (5.4.2). Because the number of degrees of freedom varies with the design, it is not possible to establish a single upper control limit for this process; the control limit for each point is calculated separately, and the control procedure is automated using the control limit based on the F distribution as shown in (5.4.3). Once a year the within standard deviations are plotted to see if any degradation has occurred in the balance over the year.

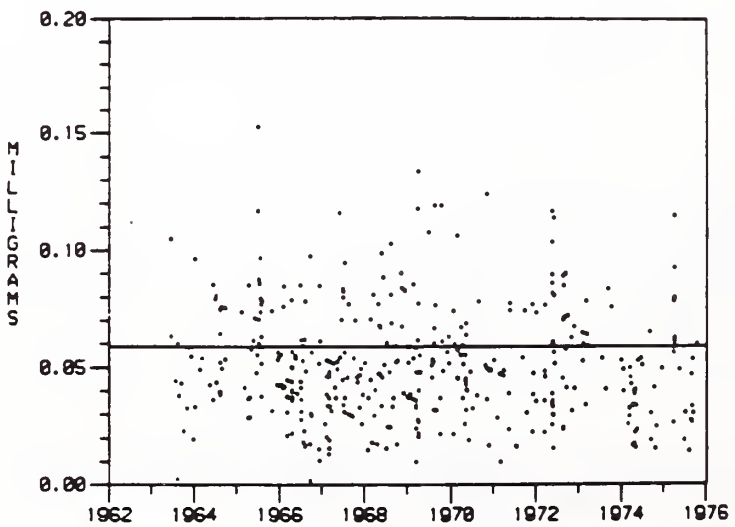


Figure 14
Within standard deviations (mg) for NBS balance #4
plotted against time (years)

The examples cited in figures 13 and 14 are for a process, as was said before, that has been in existence for a long time and that is demonstrably in-control. It may be instructive to examine a few processes, or at least the data from those processes, that have not been carefully monitored and that are not necessarily in-control.

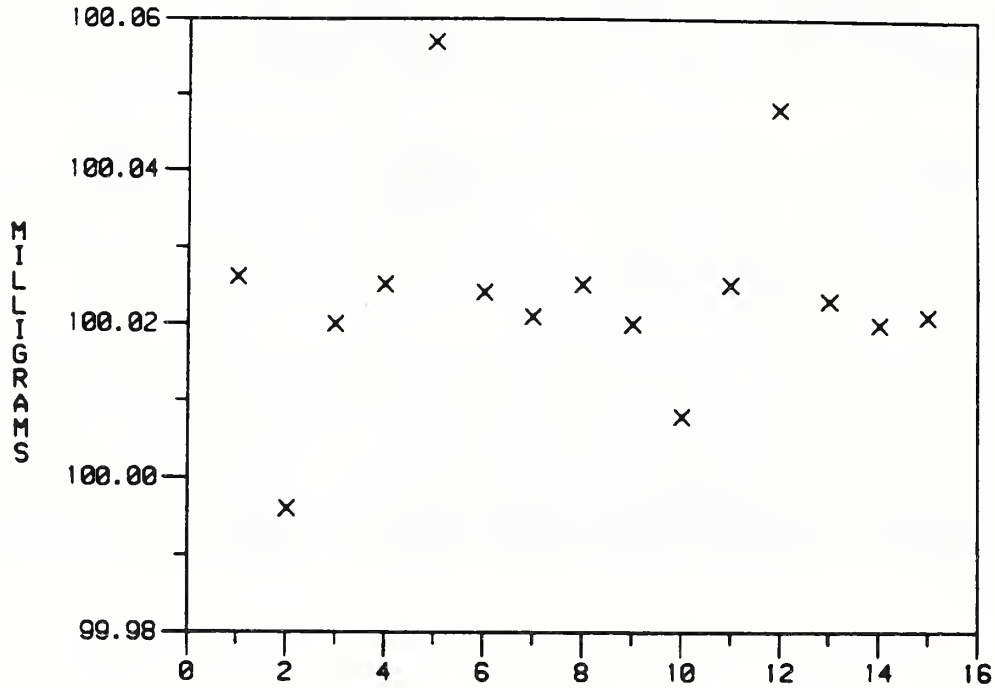


Figure 15
Measurements (mg) on a 100g weight plotted against time (months)

Take, for example, the data in figure 15 which represent repeated weighings made over a fifteen month period on a calibrated weight. Notice that the majority of the values are clustering close together but that there are a relatively large number of extremely discordant values. It is not sensible in this case to ask, "What base line and control limits are appropriate for this process?" In fact, at this point in time, a measurement process does not exist because it is not possible to predict a future value of the process, or in other words, the data as plotted in figure 15 do not represent a random sample from a single error distribution. In this case, a critical deficiency in the measurement process was tracked down; namely, that the elapsed time between two weighings being made on the balance in succession was not sufficient for the balance to come to proper equilibrium.

A control procedure involving a power instrument standard is shown in figure 16. The graph shows assignments made to the power standard as it was intercompared with its primary power source over a two-week period. The sixteen resulting measurements define the base line and control limits for the process.

The results of sixteen additional measurements taken a year later are shown in figure 17, and although they are clearly out-of-control with respect to the initial measurements, they are consistent among themselves raising a question as to whether the power standard itself is changing radically, whether the initial measurements were, in fact, out-of-control and should be discounted, or whether the process is not properly characterized by either set of measurements. Really only one thing is clear at this point -- that the assignment cannot be made with any degree of confidence and that the power standard should not be the basis for a calibration program until the process of assigning a value to the power standard is adequately characterized.

This was accomplished by repeating the intercomparison at three month intervals taking only two or three measurements each time instead of sixteen. The results are shown in figure 18. A large component of variance that did not show up in the initial two-week interval affects the measurement process, and the standard deviation computed from the short-term measurements under-estimates the process variability as it exists over, say, a year's time.

This example demonstrates an extremely important principle of measurement assurance; namely that in general there is little value in closely spaced repetitions. These should be kept to a minimum, and measurements should be taken frequently over a long period of time in order to correctly characterize a process. This practice should be continued until the process parameters are well established and only then should the intervals between intercomparisons be lengthened.

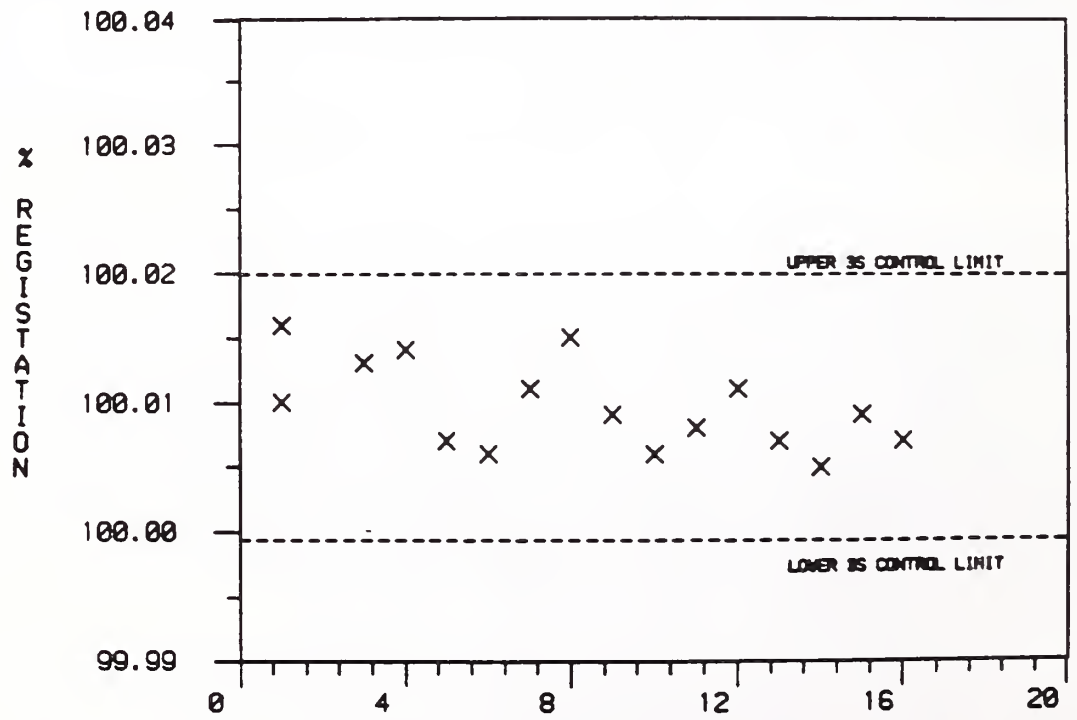


Figure 16
Measurements (% reg) on a power standard plotted against run sequence showing upper and lower three standard deviation limits

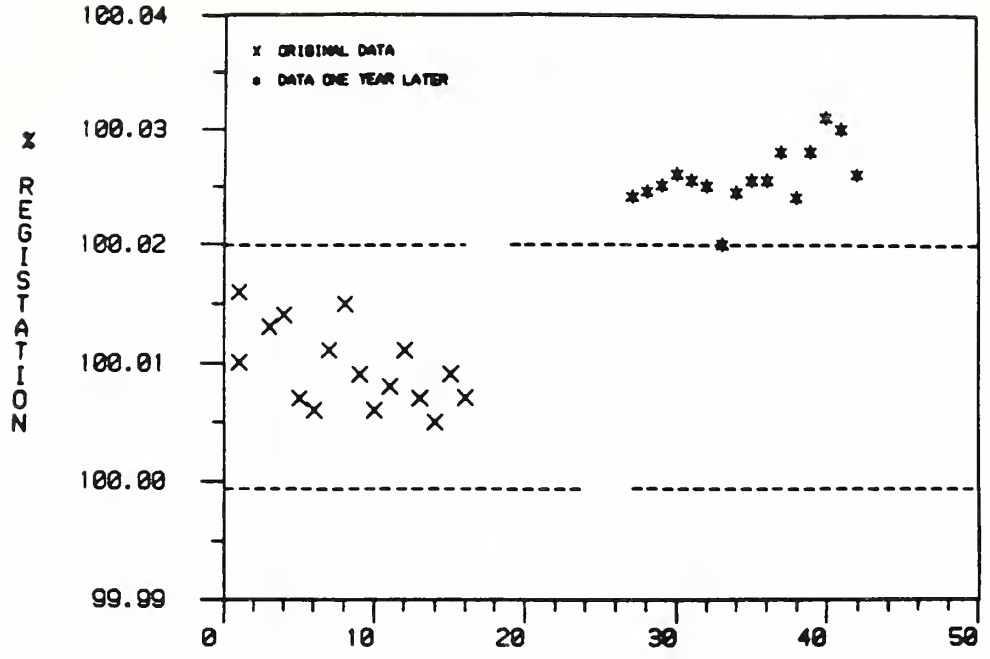


Figure 17
Original measurements (% reg) on power standard and measurements on the same standard a year later with original control limits

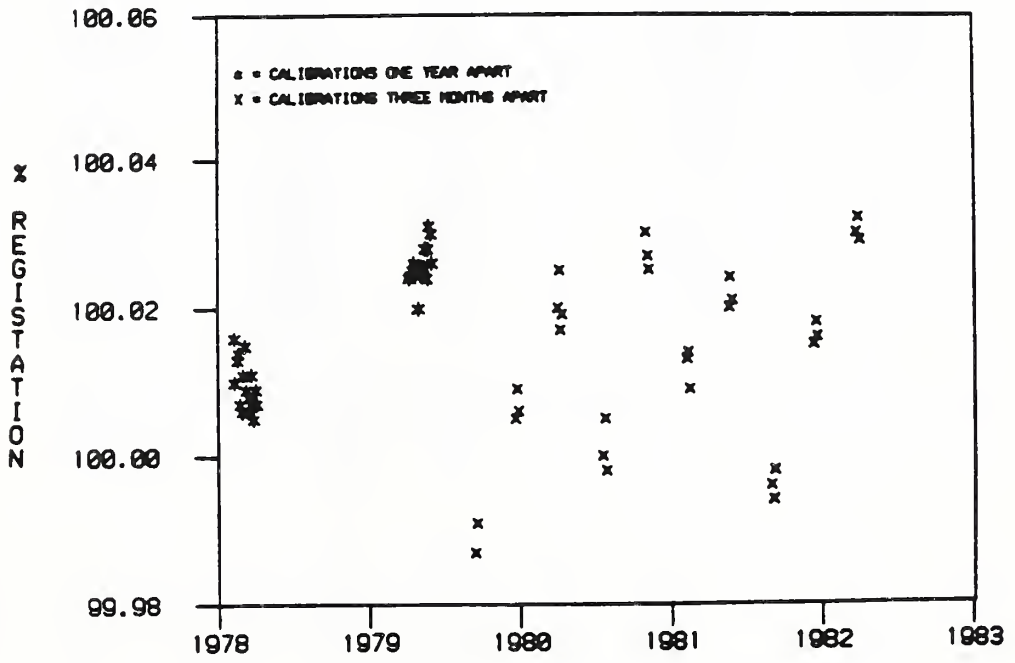


Figure 18
Measurements (% reg) on the power standard at three month intervals over three years

Table I
Critical Values $t_{\alpha/2}(\nu)$ of Student's t Distribution

ν	α=0.05	α=0.01	ν	α=0.05	α=0.01
2	4.303	9.925	62	1.999	2.657
4	2.776	4.604	64	1.998	2.655
6	2.447	3.707	66	1.997	2.652
8	2.306	3.355	68	1.995	2.650
10	2.228	3.169	70	1.994	2.648
12	2.179	3.055	72	1.993	2.646
14	2.145	2.977	74	1.993	2.644
16	2.120	2.921	76	1.992	2.642
18	2.101	2.878	78	1.991	2.640
20	2.086	2.845	80	1.990	2.639
22	2.074	2.819	82	1.989	2.637
24	2.064	2.797	84	1.989	2.636
26	2.056	2.779	86	1.988	2.634
28	2.048	2.763	88	1.987	2.633
30	2.042	2.750	90	1.987	2.632
32	2.037	2.738	92	1.986	2.630
34	2.032	2.728	94	1.985	2.629
36	2.028	2.719	96	1.984	2.628
38	2.024	2.712	98	1.983	2.627
40	2.021	2.704	100	1.983	2.626
42	2.018	2.698	102	1.983	2.625
44	2.015	2.692	104	1.982	2.624
46	2.013	2.687	106	1.982	2.623
48	2.011	2.682	108	1.981	2.622
50	2.009	2.678	110	1.981	2.621
52	2.007	2.674	112	1.981	2.620
54	2.005	2.670	114	1.981	2.620
56	2.003	2.667	116	1.981	2.619
58	2.002	2.663	118	1.980	2.618
60	2.000	2.660	120	1.980	2.617
			∞	1.960	2.576

ν = number of degrees of freedom in the total standard deviation.

Table II

Critical values $F_\alpha(\nu_1, \nu_2)$ of the F Distribution
α=0.01

DF ν_2	Degrees of freedom ν_1									
	1	2	3	4	5	6	7	8	9	10
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98
35	7.42	5.27	4.40	3.91	3.59	3.37	3.20	3.07	2.96	2.88
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80
45	7.23	5.11	4.25	3.77	3.45	3.23	3.07	2.94	2.83	2.74
50	7.17	5.06	4.20	3.72	3.41	3.19	3.02	2.89	2.78	2.70
55	7.12	5.01	4.16	3.68	3.37	3.15	2.98	2.85	2.75	2.66
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63
65	7.04	4.95	4.10	3.62	3.31	3.09	2.93	2.80	2.69	2.61
70	7.01	4.92	4.07	3.60	3.29	3.07	2.91	2.78	2.67	2.59
75	6.99	4.90	4.05	3.58	3.27	3.05	2.89	2.76	2.65	2.57
80	6.96	4.88	4.04	3.56	3.25	3.04	2.87	2.74	2.64	2.55
85	6.94	4.86	4.02	3.55	3.24	3.02	2.86	2.73	2.62	2.54
90	6.93	4.85	4.01	3.53	3.23	3.01	2.84	2.72	2.61	2.52
95	6.91	4.84	3.99	3.52	3.22	3.00	2.83	2.70	2.60	2.51
100	6.90	4.82	3.98	3.51	3.21	2.99	2.82	2.69	2.59	2.50
105	6.88	4.81	3.97	3.50	3.20	2.98	2.81	2.69	2.58	2.49
110	6.87	4.80	3.96	3.49	3.19	2.97	2.81	2.68	2.57	2.49
115	6.86	4.79	3.96	3.49	3.18	2.96	2.80	2.67	2.57	2.48
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47
∞	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32

Table II continued

Critical Values $F_\alpha(\nu_1, \nu_2)$ of the F Distribution
$\alpha = 0.01$

DF	Degrees of freedom ν_1									
ν_2	12	14	16	18	20	22	24	26	28	30
10	4.71	4.60	4.52	4.46	4.41	4.36	4.33	4.30	4.27	4.25
11	4.40	4.29	4.21	4.15	4.10	4.06	4.02	3.99	3.96	3.94
12	4.16	4.05	3.97	3.91	3.86	3.82	3.78	3.75	3.72	3.70
13	3.96	3.86	3.78	3.72	3.66	3.62	3.59	3.56	3.53	3.51
14	3.80	3.70	3.62	3.56	3.51	3.46	3.43	3.40	3.37	3.35
15	3.67	3.56	3.49	3.42	3.37	3.33	3.29	3.26	3.24	3.21
16	3.55	3.45	3.37	3.31	3.26	3.22	3.18	3.15	3.12	3.10
17	3.46	3.35	3.27	3.21	3.16	3.12	3.08	3.05	3.03	3.00
18	3.37	3.27	3.19	3.13	3.08	3.03	3.00	2.97	2.94	2.92
19	3.30	3.19	3.12	3.05	3.00	2.96	2.92	2.89	2.87	2.84
20	3.23	3.13	3.05	2.99	2.94	2.90	2.86	2.83	2.80	2.78
22	3.12	3.02	2.94	2.88	2.83	2.78	2.75	2.72	2.69	2.67
24	3.03	2.93	2.85	2.79	2.74	2.70	2.66	2.63	2.60	2.58
26	2.96	2.86	2.78	2.72	2.66	2.62	2.58	2.55	2.53	2.50
28	2.90	2.79	2.72	2.65	2.60	2.56	2.52	2.49	2.46	2.44
30	2.84	2.74	2.66	2.60	2.55	2.51	2.47	2.44	2.41	2.39
35	2.74	2.64	2.56	2.50	2.44	2.40	2.36	2.33	2.30	2.28
40	2.66	2.56	2.48	2.42	2.37	2.33	2.29	2.26	2.23	2.20
45	2.61	2.51	2.43	2.36	2.31	2.27	2.23	2.20	2.17	2.14
50	2.56	2.46	2.38	2.32	2.27	2.22	2.18	2.15	2.12	2.10
55	2.53	2.42	2.34	2.28	2.23	2.18	2.15	2.11	2.08	2.06
60	2.50	2.39	2.31	2.25	2.20	2.15	2.12	2.08	2.05	2.03
65	2.47	2.37	2.29	2.23	2.17	2.13	2.09	2.06	2.03	2.00
70	2.45	2.35	2.27	2.20	2.15	2.11	2.07	2.03	2.01	1.98
75	2.43	2.33	2.25	2.18	2.13	2.09	2.05	2.02	1.99	1.96
80	2.42	2.31	2.23	2.17	2.12	2.07	2.03	2.00	1.97	1.94
85	2.40	2.30	2.22	2.15	2.10	2.06	2.02	1.98	1.95	1.93
90	2.39	2.29	2.21	2.14	2.09	2.04	2.00	1.97	1.94	1.92
95	2.38	2.28	2.20	2.13	2.08	2.03	1.99	1.96	1.93	1.90
100	2.37	2.27	2.19	2.12	2.07	2.02	1.98	1.95	1.92	1.89
105	2.36	2.26	2.18	2.11	2.06	2.01	1.97	1.94	1.91	1.88
110	2.35	2.25	2.17	2.10	2.05	2.00	1.96	1.93	1.90	1.88
115	2.34	2.24	2.16	2.10	2.04	2.00	1.96	1.92	1.89	1.87
120	2.34	2.23	2.15	2.09	2.03	1.99	1.95	1.92	1.89	1.86
∞	2.19	2.09	2.00	1.94	1.88	1.84	1.79	1.76	1.73	1.70

Table II continued

Critical Values $F_\alpha(\nu_1, \nu_2)$ of the F Distribution
α=0.01

DF ν_2	Degrees of freedom ν_1									
	40	50	60	70	80	90	100	110	120	∞
10	4.17	4.12	4.08	4.06	4.04	4.03	4.01	4.00	4.00	3.91
11	3.86	3.81	3.78	3.75	3.73	3.72	3.71	3.70	3.69	3.60
12	3.62	3.57	3.54	3.51	3.49	3.48	3.47	3.46	3.45	3.36
13	3.43	3.38	3.34	3.32	3.30	3.28	3.27	3.26	3.25	3.17
14	3.27	3.22	3.18	3.16	3.14	3.12	3.11	3.10	3.09	3.01
15	3.13	3.08	3.05	3.02	3.00	2.99	2.98	2.97	2.96	2.87
16	3.02	2.97	2.93	2.91	2.89	2.87	2.86	2.85	2.84	2.76
17	2.92	2.87	2.83	2.81	2.79	2.78	2.76	2.75	2.75	2.65
18	2.84	2.78	2.75	2.72	2.70	2.69	2.68	2.67	2.66	2.57
19	2.76	2.71	2.67	2.65	2.63	2.61	2.60	2.59	2.58	2.49
20	2.69	2.64	2.61	2.58	2.56	2.55	2.54	2.53	2.52	2.42
22	2.58	2.53	2.50	2.47	2.45	2.43	2.42	2.41	2.40	2.31
24	2.49	2.44	2.40	2.38	2.36	2.34	2.33	2.32	2.31	2.21
26	2.42	2.36	2.33	2.30	2.28	2.26	2.25	2.24	2.23	2.13
28	2.35	2.30	2.26	2.24	2.22	2.20	2.19	2.18	2.17	2.07
30	2.30	2.24	2.21	2.18	2.16	2.14	2.13	2.12	2.11	2.01
35	2.19	2.14	2.10	2.07	2.05	2.03	2.02	2.01	2.00	1.89
40	2.11	2.06	2.02	1.99	1.97	1.95	1.94	1.93	1.92	1.81
45	2.05	2.00	1.96	1.93	1.91	1.89	1.88	1.86	1.85	1.74
50	2.01	1.95	1.91	1.88	1.86	1.84	1.82	1.81	1.80	1.69
55	1.97	1.91	1.87	1.84	1.82	1.80	1.78	1.77	1.76	1.64
60	1.94	1.88	1.84	1.81	1.78	1.76	1.75	1.74	1.73	1.60
65	1.91	1.85	1.81	1.78	1.75	1.74	1.72	1.71	1.70	1.57
70	1.89	1.83	1.78	1.75	1.73	1.71	1.70	1.68	1.67	1.54
75	1.87	1.81	1.76	1.73	1.71	1.69	1.67	1.66	1.65	1.52
80	1.85	1.79	1.75	1.71	1.69	1.67	1.65	1.64	1.63	1.50
85	1.83	1.77	1.73	1.70	1.67	1.65	1.64	1.62	1.61	1.48
90	1.82	1.76	1.72	1.68	1.66	1.64	1.62	1.61	1.60	1.46
95	1.81	1.75	1.70	1.67	1.65	1.63	1.61	1.60	1.58	1.45
100	1.80	1.74	1.69	1.66	1.63	1.61	1.60	1.58	1.57	1.43
105	1.79	1.73	1.68	1.65	1.62	1.60	1.59	1.57	1.56	1.42
110	1.78	1.72	1.67	1.64	1.61	1.59	1.58	1.56	1.55	1.41
115	1.77	1.71	1.66	1.63	1.60	1.58	1.57	1.55	1.54	1.40
120	1.76	1.70	1.66	1.62	1.60	1.58	1.56	1.54	1.53	1.39
∞	1.60	1.53	1.48	1.44	1.41	1.38	1.36	1.35	1.33	

REFERENCES

[1] Dorsey, N. E. The velocity of light. Trans. Am. Phil. Soc. XXXIV.; 1944, pp.1-110.

[2] Hayford, J. A. On the least square adjustment of weighings, U.S. Coast and Geodetic Survey Appendix 10, Report for 1892; 1893.

[3] Benoit, M. J. R. L'Etalonnage des Series de Poids Travaux et Memoirs du Bureau International des Poids et Mesures 13(1); 1907.

[4] Pienkowsky, A. T. Short tests for sets of laboratory weights. Scientific Papers of the Bureau of Standards, 21(527); 1926.

[5] Cameron, J. M.; Croarkin, M. C.; Raybold, R. C. Designs for the Calibration of Standards of Mass. Nat. Bur. Stand. (U.S.) Tech Note 952; 1977.

[6]§Eisenhart, C. Realistic evaluation of the precision and accuracy of instrument calibration systems. J. Res. Nat. Bur. Stand. (U.S.) 67C(2); 1962. pp. 161-187.

[7]§Youden, W. J. Experimental design and ASTM committees. Materials Research and Standards, 1(11); 1961. pp. 862-867.

[8]§Youden, W. J. Physical measurements and experiment design. Colloques Internationaux du Centre National de la Recherche Scientifique No. 110, le Plan d'Experiences; 1961. pp. 115-128.

[9]§Pontius, P. E. and Cameron, J. M. Realistic Uncertainties and the Mass Measurement Process. Nat. Bur. Stand. (U.S.) Monogr. 103; 1967.

[10] Croarkin, M. C.; Beers, J. S.; Tucker, C. Measurement Assurance for Gage Blocks. Nat. Bur. Stand. (U.S.) Monogr. 163; 1979.

[11] Croarkin, M. C.; Varner, R. N. Measurement Assurance for Dimensional Measurements on Integrated Circuit Photomasks. Nat. Bur. Stand. (U.S.) Tech. Note 1164; 1982.

[12] American National Standard ANSI N15.18-1975. Mass calibration techniques for nuclear materials control. Available from ANSI, Inc., 1430 Broadway, New York, NY 10018.

[13] Pontius, P. E. and Doher, L. W. The joint ANSI-INMM 8.1-Nuclear Regulatory Commission study of uranium hexafluoride cylinder material accountability bulk measurements. Proc. 18th Ann. Mtg. INMM, VI(III); 1977. p. 480.

§Reprinted in Nat. Bur. Stand. (U.S.) Spec. Publ. 300, Vol I. Precision Methods and Calibration: Statistical Concepts and Procedures. H. H. Ku, editor. 1969.

[14] Beers, J. S. A Gage Block Measurement Process Using Single Wavelength Interferometry. Nat. Bur. Stand. (U.S.) Monogr. 152; 1975.

[15] Cameron, J. M. Measurement Assurance. Nat. Bur. Stand. (U.S.) NBSIR 77-1240; 1977.

[16] Kieffer, L. J., ed. Calibration and Related Measurement Services of the National Bureau of Standards. Nat. Bur. Stand. (U.S.) Spec. Publ. 250; 1982.

[17] Pipkin, F. R., Ritter, R. C. Precision measurements and fundamental constants. Science, 219 (4587); 1983. p. 917.

[18] Pontius, P. E. The Measurement Assurance Program - A Case Study: Length Measurements Part I. Long Gage Blocks (5 in to 20 in). Nat. Bur. Stand. (U.S.) Monogr. 149; 1975.

[19] Beers, J. S., Tucker, C. D. Intercomparison Procedures for Gage Blocks Using Electromechanical Comparators. Nat. Bur. Stand. (U.S.) NBSIR 76-979; 1976. p. 9.

[20] Simpson, J. A. Foundations of metrology. J. Res. Nat. Bur. Stand. (U.S.). 86(3), 1981. p. 282.

[21] Nyyssonen, D. Linewidth measurement with an optical microscope: The effect of operating conditions on the image profile. Appl. Opt. 16(8); Ausust 1977. pp. 2223-2230.

[22] Jerke, J. M., ed. Semiconductor Measurement Technology: Accurate Linewidth Measurements on Integrated Circuit Photomasks. Nat. Bur. Stand. (U.S.) Spec. Publ. 400-43; 1980. pp. 7-15.

[23] Jerke, J. M.; Croarkin, M. C.; Varner, R. N. Semiconductor Measurement Technology: Interlaboratory Study on Linewidth Measurements for Antireflective Chromium Photomasks. Nat. Bur. Stand. (U.S.) Spec. Publ. 400-74; 1982.

[24] See reference 20, p. 283.

[25] See reference 22, p. 50.

[26] See reference 20, p. 283.

[27] Cameron, J. M. Encyclopedia of Statistical Sciences, Vol 1. S. Kotz and N. L. Johnson, ed. New York: John Wiley & Sons, Inc.; 1982. pp. 341-347.

[28] Ku, H. H. Statistical concepts of a measurement process. Precision Methods and Calibration: Statistical Concepts and Procedures. Nat. Bur. Stand. (U.S.) Spec. Publ. 300, Vol I. H. H. Ku, ed. 1969. pp 296-20 to 330-54.

[29] Mandel, J. The Statistical Analysis of Experimental Data. New York: Interscience Publ; 1964. pp. 278-279.

[30] Snedecor, G. W. and Cochran, W. G. Statistical Methods, Sixth ed. Ames, Iowa: The Iowa State University Press; 1976. pp. 279-280.

[31] See reference [7]. pp. 862-863.

[32] Mattingly, G. H.; Pontius, P. E.; Allion, H. H.; Moore, E. F. A laboratory study of turbine meter uncertainty. Proc. Symp. on Flow in Open Channels and Closed Circuits; Nat. Bur. Stand. (U.S.) Spec. Publ. 484; 1977.

[33] See reference [10].

[34] Duncan, A.J. Quality Control and Industrial Statistics, Fourth ed. Homewood: Richard D. Irwin, Inc; 1974. p. 381.

[35] See reference [8], p. 21-22.

[36] See reference [28], p. 299.

[37] See reference [28], p. 305-308.

[38] See reference [23].

[39] Hockersmith, T. E.; Ku, H. H. Uncertainties associated with proving ring calibration. Precision Methods and Calibration: Statistical Concepts and Procedures. Nat. Bur. Stand. (U.S.) Spec. Publ. 300, Vol. 1. H. H. Ku, ed.; 1969. pp. 257-1 to 264-8.

[40] See reference [11], p. 30-33.

[41] Youden, W. J. Uncertainties in calibration. IRE Transactions on Instrumentation, I-11, (3,4); 1962. p. 137.

[42] Eisenhart, C., Ku, H. H. and Colle, R. Expression of the Uncertainties of Final Measurement Results; Reprints. Nat. Bur. Stand. (U.S.) Spec. Pub. 644; 1983.

[43] Giacomo, P. News from BIPM. Metrologia, 17; 1981. pp. 73-74.

[44] See reference [28]. pp. 322-323.

[45] Raghavarao, D. Construction and Combinatorial Problems in Design of Experiments. New York: John Wiley & Sons, Inc; 1971. p. 315.

[46] Mood, A. M. On Hotelling's Weighing Problem. Annals of Mathematical Statistics, 17; 1946. pp.432-446.

[47] See reference [5].

[48] Cameron, J. M.; Hailes, G. E. Designs for the Calibration of Small Groups of Standards in the Presence of Drift. Nat. Bur. Stand. (U.S.) Tech Note 844; 1974. p. 1.

[49] Jaegar, K. B. and Davis, R. S. A Primer for Mass Metrology. Nat. Bur. Stand. (U.S.) Special Publication: Industrial Measurement Series 700-1; 1984.

[50] See reference [10], pp. 13-25.

[51] See reference [10], pp. 27-39.

[52] Cameron, J. M.; Eicke, W. G. Designs for Surveillance of the Volt Maintained by a Small Group of Saturated Standard Cells. Nat. Bur. Stand. (U.S.) Tech Note 430; 1967.

[53] Cameron, J. M. The Use of the Method of Least Squares in Calibration. Nat. Bur. Stand. (U.S.) NBSIR 74-587; 1974.

[54] See reference [49].

[55] Varner, R. N. Mass Calibration Computer Software. Nat. Bur. Stand. (U.S.) Tech. Note 1127; 1980.

[56] See reference [52]. pp. 1-2.

[57] Cochran, W. J. The distribution of the largest of a set of estimated variances as a fraction of their total. Annals of Eugenics, 11; 1941. pp. 47-52.

[58] Eisenhart, C. Significance of the largest of a set of sample estimates of variance. Chapter 15 of Selected Techniques of Statistical Analysis. Eisenhart, C., Hastay, M. W., Wallis, W. A., editors. New York: McGraw Hill Book Co., Inc.; 1947. pp. 383-394.

[59] Draper, N. R.; Smith, H. Applied Regression Analysis. New York: John Wiley & Sons, Inc.; 1966. p. 1-32.

[60] See reference [22], p. 350.

[61] Bicking, C. A.; Gryna, F. M. Jr. Process Control by Statistical Methods. Section 23 of Quality Control Handbook, Third Edition, J. M. Juran, ed. New York: McGraw-Hill Book Co.; 1974. p. 23-2, 23-3.

[62] See reference [34]. pp. 464-484.

[63] Reynolds, J. R. Jr.; Ghosh, B. K. Designing control charts for means and variances. 1081 ASQC Quality Congress Transactions: San Francisco; 1981.

[64] Shewhart, W. A. Statistical Method from the Viewpoint of Quality Control. The Graduate School, U.S. Department of Agriculture, Washington, DC; 1939.

[65] Grant, E. L.; Leavenworth, R. S. Statistical Quality Control, 4th Edition. New York: McGraw Hill Book Co.; 1976. p. 129.

[66] Parobeck, P.; Tomb, T.; Ku, H. H.; Cameron, J. M. Measurement assurance program for weighings of respirable coal mine dust samples. J. Quality Tech, 13(3); 1981. pp. 157-165.

[67] Bowker, A. H. Tolerance limits for normal distributions. Chapter 2 of reference [58]. p. 99.

[68] See reference [63]. p. 381.

APPENDIX A

The purpose of this appendix is to define the matrix manipulations[‡] that produce the least-squares solution to a weighing design along with the propagation of associated standard deviations and uncertainties.[†] The theory is explained by Cameron et al. in reference [5]. It is assumed that a series of weighing designs is required in order to calibrate an entire weight set and that assignments to individual weights depend upon a starting restraint with known value that is invoked in the first design. The starting restraint is usually the known sum of two reference kilograms. It is also assumed that the designs are interconnected in such a way that a value assigned to an individual weight or sum of weights from one design constitutes the restraint for the next design in the series.

Each design in the series involves n intercomparisons among p weights where the p weights include the reference standards composing the restraint, the test weights, and check standard.

The model for the measurement process assumes that these observations are related to the values of the weights by

$$D = AX^* + \varepsilon \tag{A.1}$$

where D is the (nx1) vector of observations; A is an (nxp) design matrix of zeroes and ones such that a plus or minus one in the ijth position indicates that the jth weight is measured by the ith observation, and a zero indicates the converse; X^* is the (px1) vector of unknown values for the p weights; and ε is the (nx1) vector of random errors.

Define

$$D' = (d_1 \cdots d_n) \tag{A.2}$$

$$A = \begin{pmatrix} a_{11} \cdots a_{1p} \\ \vdots \quad\quad \vdots \\ a_{n1} \cdots a_{np} \end{pmatrix} \tag{A.3}$$

$$(X^*)' = (X_1^* \cdots X_p^*) \tag{A.4}$$

and

$$\varepsilon' = (\varepsilon_1 \cdots \varepsilon_n) \tag{A.5}$$

[‡]The matrix notation that is used in this appendix denotes the transpose of the matrix M by M' and the inverse of the matrix M by M^{-1}.

[†]Assuming that there is no significant between component of variance in the measurement process.

In order to define various linear combinations of the weights, we will also define several vectors of size (px1) which have the general form

$$\ell' = (\ell_1 \cdots \ell_p)$$

where each element ℓ_i (i=1,...,p) is either zero, plus one or minus one.

The least-squares estimate for (A.4) depends upon the inverse of the normal equations A'A. The usual case for calibration experiments is that A'A has rank p-1. Where A'A has rank less than p, the inverse does not exist and a solution can be obtained only by imposing a restraint upon the system of equations. Therefore, we let R^* be a scalar with known value called the restraint; and ℓ_R be a (px1) vector of zeroes and ones such that a one in the jth position indicates that the jth weight is in the restraint, and a zero indicates the converse. For example,

$$\ell_R' = (1\ 1\ 0 \cdots 0)$$

indicates that the restraint is over the first two weights.

One approach to finding the least-squares estimate for X^* is via an augmented matrix B where

$$B = \begin{pmatrix} A'A & \ell_R & A'D \\ \ell_R' & 0 & R^* \\ 0 & 0 & -1 \end{pmatrix} \qquad (A.6)$$

is a (p+2)x(p+2) matrix whose inverse

$$B^{-1} = \begin{pmatrix} Q & h & \hat{X}^* \\ h' & 0 & . \\ \cdots & & . \end{pmatrix} \qquad (A.7)$$

can be partitioned as shown above. The (pxp) matrix Q in the upper left hand corner of B^{-1} contains information relating to the variances of the estimates, and the (px1) matrix X^* in the upper right hand corner of B^{-1} contains the least-squares estimates for the p weights. The other quantities in B^{-1} are not of interest for this application. Notice that once the inverse of B has been computed, the estimates are immediately available without further matrix multiplications.[¶]

The individual deviations of the observations from their fitted values are given by the (px1) vector ξ where

$$\xi' = (D - AX^*)', \qquad (A.8)$$

[¶] The caret (^) indicating a least-squares estimate from the data is dropped in future references to X^*.

and the within standard deviation for the design is

$$s_w = \left(\frac{\xi' \xi}{n-p+1} \right)^{1/2} \tag{A.9}$$

with n-p+1 degrees of freedom.

The restraint for the next design in the series can be written in the form

$$\Sigma^* = \ell_\Sigma' X^* \tag{A.10}$$

where ℓ_Σ is a (px1) vector of zeroes and ones where a one in the ith position indicates that the ith weight is to be included in the restraint for the next design, and a zero indicates the converse. The standard deviation for the outgoing restraint is given by

$$s_\Sigma = \left(\ell_\Sigma' Q \ell_\Sigma \, s_w^2 + \left(\frac{\ell_\Sigma' W}{\ell_R' W} \right)^2 s_R^2 \right)^{1/2} \tag{A.11}$$

where s_R is the standard deviation of the incoming restraint R^* as computed from the previous design, and

$$W' = (W_1 \ . \ . \ . \ W_p)$$

where W is a (px1) vector of nominal values for the p weights. If the current design is the first design in the series, then s_R is zero.

Notice that the computation of the standard deviation associated with the check standard as defined in (A.14) and the computation of the standard deviation associated with the values of the test weights as defined in (A.16) are also dependent on s_R. Thus, the standard deviations for each series are dependent on all prior series as they are propagated starting with the first series.

The current value for the check standard from the design can be written in the form

$$c = \ell_c' X^* \tag{A.13}$$

where ℓ_c is a (px1) vector of zeroes and ones such that a plus or minus one in the ith position indicates that the ith weight is in the check standard, and a zero indicates the converse.

The standard deviation of the check standard value is given by

$$s_c = \left(\ell_c' Q \ell_c \, s_w^2 + \left(\frac{\ell_c' W}{\ell_R' W} \right)^2 s_R^2 \right)^{1/2} . \tag{A.14}$$

Then given that the accepted value for the check standard is known from previous experiments to be A_c, a test for control is made by computing the test statistic

$$t_c = \frac{|A_c - c|}{s_c} \tag{A.15}$$

and comparing it to a critical value.

Finally, we are interested in the uncertainty of the value assigned to a single weight or to a collection of weights. For each summation or difference of weights that is of interest, we define a (pxl) vector ℓ_S of zeroes, plus ones and minus ones such that a one in the ith position indicates that the ith weight is involved in the summation or difference, and a zero indicates the converse. The reported value for the summation S is S^* where

$$S^* = \ell_S{}' X^*.$$

The standard deviation for the summation, designated by s_S is

$$s_S = \left(\ell_S{}' Q \ell_S \, s_w{}^2 + \left(\frac{\ell_S{}' W}{\ell_R{}' W} \right)^2 s_R{}^2 \right)^{1/2} \tag{A.16}$$

and the uncertainty associated with the summation is

$$U = 3s_S + \frac{\ell_S{}' W}{\ell_{SR}{}' W} U_{SR} \tag{A.17}$$

where U_{SR} is the uncertainty assigned to the starting restraint in the series, and similarly ℓ_{SR} is the (pxl) vector of zeroes, plus ones and minus ones such that a plus or minus one in the ith position indicates that the ith weight is in the starting restraint.

Notice that if we are talking about a single weight whose value is $X_j{}^*$, then the quantity

$$\ell_S{}' Q \ell_S = q_{jj}$$

where q_{jj} is the jth diagonal element in Q.

For the next design in the series, let the restraint be $R^* = \Sigma^*$ with standard deviation $s_R = s_\Sigma$ and proceed with the calculation starting with equation (A.1).

Expression of the Uncertainties of Final Measurement Results: Reprints

Churchill Eisenhart

Harry H. Ku

R. Collé

National Bureau of Standards
Washington, DC 20234

U.S. DEPARTMENT OF COMMERCE, Malcolm Baldrige, Secretary
NATIONAL BUREAU OF STANDARDS, Ernest Ambler, Director

Issued January 1983

Foreword

The reporting of final measurement results, and the uncertainties associated with the measurement processes used to obtain these results, has always been and continues to be a source of difficulty. The three articles reprinted in this publication are collected here as a convenient reference source for experimenters who must face the difficult task of deciding how to express measurement uncertainties. The philosophical basis, general guidelines, and specific recommendations for expressing uncertainties contained within these articles have evolved at NBS over a period of many years.

The first article originally appeared in *Science* in 1968. This article develops the underlying basis and general guidelines on the forms of expression needed for uncertainty statements, and presents specific recommendations for four distinct cases: (i) when both systematic error and imprecision are negligible; (ii) when systematic error is not negligible, and imprecision is negligible; (iii) when neither systematic error nor imprecision is negligible; and (iv) when systematic error is negligible, and imprecision is not negligible.

The second article, written as a companion to the first, originally appeared in a 1968 issue of *M&D: Measurements and Data*. It gives a condensed summary of the recommendations presented in the first article, and provides tabular guides to commonly used statements of imprecision, systematic error, and uncertainty.

The third article is a postscript to the two preceding articles, and was prepared in 1980 for an internal NBS communications manual. It reinforces the major thrust and content of the earlier articles, but includes more recent thought particularly in regard to overall uncertainty statements.

The first two articles have since been reprinted in several NBS publications including Special Publication 300, Volume 1, *Precision Measurement and Calibration: Statistical Concepts and Procedures* (Harry H. Ku, ed., 1969). The 1980 NBS communications manual incorporated the second and third articles, but did not reprint the first article. Furthermore, this manual is not accessible outside NBS. This special publication, therefore, collects all three articles, for the first time, in one convenient source which is available to the many scientists and engineers throughout the entire measurement community.

Contents

Expression of the Uncertainties of Final Results

Clear statements of the uncertainties of reported values are needed for their critical evaluation.

Churchill Eisenhart

Measurement of some property of a thing in practice always takes the form of a sequence of steps or operations that yield as an end result a number that serves to represent the amount or quantity of some particular property of a thing—a number that indicates how much of this property the thing has, for someone to use for a specific purpose. The end result may be the outcome of a single reading of an instrument, with or without corrections for departures from prescribed conditions. More often it is some kind of average, for example, the arithmetic mean of a number of independent determinations of the same magnitude, or the final result of a least squares "reduction" of measurements of a number of different magnitudes that bear known relations with one another in accordance with a definite experimental plan. In general, the purpose for which the answer is needed determines the precision or accuracy required and ordinarily also the method of measurement employed.

Although the accuracy required of a reported value depends primarily on the *intended* use, or uses, of the value, one should not ignore the requirements of other uses to which it is likely to be put. A reported value whose accuracy is entirely unknown is worthless.

Strictly speaking, the actual *error* of a reported value, that is the magnitude and sign of its deviation from the truth (*1*), is usually unknowable. Limits to this error, however, can usually be inferred—with some risk of being incorrect—from the precision of the measurement process by which the reported value was obtained, and from reasonable limits to the possible bias of the measurement process. The *bias*, or *systematic error*, of a measurement process is the magnitude and direction of its tendency to measure something other than what was intended; its *precision* refers to the typical closeness together of successive independent measurements of a single magnitude generated by repeated applications of the process under specified conditions; and its *accuracy* is determined by the closeness to the true value characteristic of such measurements.

Precision and accuracy are inherent characteristics of the measurement process employed and not of the particular end result obtained. From experience with a particular measurement process and knowledge of its sensitivity to uncontrolled factors, one can often place reasonable bounds on its likely systematic error (bias). It is also necessary to know how well the particular value in hand is likely to agree with other values that the same measurement process might have provided in this instance, or might yield on remeasurement of the same magnitude on another occasion. Such information is provided by the estimated *standard error* (*2*) of the reported value, which measures (or is an index of) the characteristic disagreement of repeated determinations of the same quantity by the same method, and thus serves to indicate the precision (strictly, the imprecision) of the reported value (*3*).

Four Distinct Forms of Expression Needed

The uncertainty of a reported value is indicated by stating credible limits to its likely inaccuracy. No single form of expression for these limits is universally satisfactory. In fact, different forms of expression are recommended, which will depend on the relative magnitudes of the imprecision and likely bias, and their relative importance in relation to the intended use of the reported value, as well as to other possible uses to which it may be put (*4*).

Four distinct cases need to be recognized: (i) both systematic error and imprecision negligible, in relation to the requirements of the intended and likely uses of the result; (ii) systematic error not negligible, imprecision negligible; (iii) neither systematic error nor imprecision negligible; and (iv) systematic error negligible, imprecision not negligible.

Specific recommendations with respect to each of these cases are made below. General guidelines upon which these specific recommendations are based are discussed in the following paragraphs.

Perils of Shorthand Expressions

Final results and their respective uncertainties should be reported in sentence form whenever possible. The shorthand form "$a \pm b$" should be avoided in abstracts and summaries; and never used without explicit explanation of its connotation. If no explanation is given, many persons will take $\pm b$ to signify bounds to the inaccuracy of a. Others may assume that b is the "standard error," or the "probable error," of a, and hence the uncertainty of a is at least $\pm 3b$, or $\pm 4b$, respectively. Still others may take b to be an indication merely of the imprecision of the individual measurements, that is, to be the "standard deviation," or the "average deviation," or the "probable error" of a single observation. Each of these interpretations reflects a practice of which instances can be found in current scientific literature. As a step in the direction of reducing this current confusion, it is recommended that the use of "$a \pm b$" in presenting results be limited to that sanctioned for the case of tabular results in the fourth recommendation of the section below headed "Systematic error not negligible, imprecision negligible."

The author is a senior research fellow and former chief of the Statistical Engineering Laboratory at the National Bureau of Standards, Washington, D.C. 20234. The recommendations presented in this paper have evolved at the Bureau over a period of many years and are made public here for general information, and to educe comments and suggestions.

Imprecision and Systematic Error Require Separate Treatment

Since imprecision and systematic error are distinctly different components of inaccuracy, and are subject to different treatments and interpretations in usage, two numerics respectively expressing the imprecision and bounds to the systematic error of the reported result should be used whenever both of these errors are factors requiring consideration. Such instances are discussed in the section below for the case of "Neither systematic error nor imprecision negligible."

In quoting a reported value and its associated uncertainty from the literature, the interpretation of the uncertainty quoted should be stated if given by the author. If the interpretation is not known, a remark to this effect is in order. This practice may induce authors to use more explicit formulations of their statements of uncertainty.

Standard Deviation and Standard Error

The terms *standard deviation* and *standard error* should be reserved to denote the canonical values for the measurement process, based on considerable recent experience with the measurement process or processes involved. When there is insufficient recent experience, an estimate of the standard error (standard deviation) must of necessity be computed by recognized statistical procedures from the same measurements as the reported value itself. To avoid possible misunderstanding, in such cases, the term "computed (or estimated) standard error" ("computed standard deviation") should be used. A formula for calculating this computed standard error is given in the section below for the case of "Neither systematic error nor imprecision negligible."

Uncertainties of Accepted Values of Fundamental Constants or Primary Standards

If the uncertainty in the accepted value of a national primary standard or of some fundamental constant of nature (for example, in the volt as maintained at the National Bureau of Standards, or in the acceleration of gravity g on the Potsdam basis) is an important source of systematic error affecting the measurement process, no allowance for possible systematic error from this source should be included ordinarily in evaluating overall bounds to the systematic error of the measurement process. Since the error concerned, whatever it is, affects all results obtained by the method of measurement involved, to include an allowance for this error would be to make everybody's results appear unduly inaccurate relative to each other. In such instances one should state: (i) that measurements obtained by the process concerned are expressed in terms of the volt (or the kilogram, or other unit) "as maintained at the National Bureau of Standards," or (ii) that the indicated bounds to the systematic error of the process are exclusive of the uncertainty of the stated value adopted for some particular constant or quantity. An example of the latter form of statement is:

> . . . neglecting the uncertainty of the value 6.6256×10^{-34} joule seconds adopted for Planck's constant.

Systematic Error and Imprecision Both Negligible

In this case the reported result should be given, after rounding, to the number of significant figures consistent with the accuracy requirements of the situation, together with an explicit statement of its accuracy. An example is:

> . . . the wavelengths of the principal visible lines of mercury-198 have been measured relative to the 6057.802106 Å (angstrom units) line of krypton-98, and their values in vacuum are
>
> 5792.2685 Å
> 5771.1984 Å
> 5462.2706 Å
> 4359.5625 Å
> 4047.7146 Å
>
> correct to eight significant figures.

It needs to be emphasized that if no statement of accuracy or precision accompanies a reported number, then, in accordance with the usual conventions governing rounding, this number will ordinarily be interpreted as being accurate within ±½ unit in the last significant figure given; that is, it will be understood that its inaccuracy before rounding was less than ± 5 units in the next place. The statement "correct to eight significant figures" is included explicitly in the foregoing example, rather than left to be understood in order to forestall any concern that an explicit statement of lesser accuracy was inadvertently omitted.

Systematic Error Not Negligible, Imprecision Negligible

When the imprecision of a result is negligible, but the inherent systematic error of the measurement process concerned is not negligible, then the following rules are recommended:

1) Qualification of a reported result should be limited to a single quasi-absolute type of statement that places bounds on its inaccuracy.

2) These bounds should be stated to no more than two significant figures.

3) The reported result itself should be given (that is, rounded) to the last place affected by the stated bounds (unless it is desired to indicate and preserve such relative accuracy or precision of a higher order that it may possess for certain particular uses).

4) Accuracy statements should be given in sentence form in all cases, except when a number of results of different accuracies are presented, for example, in tabular arrangement. If it is necessary or desirable to indicate the respective accuracies of a number of results, the results should be given in the form $a \pm b$ (or $a \pm^{b}_{c}$, if necessary) with an appropriate explanatory remark (as a footnote to the table, or incorporated in the accompanying text) to the effect that the $\pm b$, or $\pm^{b}_{c}$, signify bounds to the systematic errors to which the a's may be subject.

5) The fact that the imprecision is negligible should be stated explicitly.

The particular form of the quasi-absolute type of statement employed in a given instance will depend ordinarily on personal taste, experience, current and past practice in the field of activity concerned, and so forth. Some examples of good practice are:

> . . . is (are) not in error by more than 1 part in (x).
> . . . is (are) accurate within ± (x units) [or ± (x) percent].
> . . . is (are) believed accurate within (. . . .).

Positive wording, as in the first two of these quasi-absolute statements, is appropriate only when the stated bounds to the possible inaccuracy of the reported value are themselves reliably established. However, when the indicated bounds are somewhat conjectural, it is desirable to signify this fact (and put the reader on guard) by inclusion of some modifying expression such as "believed," "considered," "estimated to be," "thought to be," and

so forth, as exemplified by the third of the foregoing examples.

The term *uncertainty* may sometimes be used effectively to achieve a conciseness of expression otherwise difficult or impossible to attain. Thus, one might make a statement such as:

The uncertainties in the above values are not more than ± 0.5°C in the range 0°C to 1100°C, and then increase to ± 2°C at 1450°C,

or

The uncertainty in this value does not exceed . . . excluding (or, including) the uncertainty of . . . in the value . . . adopted for the (reference standard involved).

A statement giving numerical limits of uncertainty as in the above should be followed by a brief discussion telling how the limits were derived.

Finally, the following forms of quasi-absolute statements are considered poor practice, and are to be avoided:

The accuracy of . . . is 5 percent.
The accuracy of . . . is ± 2 percent.

These are presumably intended to mean that the result concerned is not inaccurate, that is, not in error, by more than 5 percent or 2 percent, respectively, but they explicitly state the opposite.

Neither Systematic Error Nor Imprecision Negligible

When neither the imprecision nor the systematic error of a result are negligible, then the following rules are recommended:

1) A reported result should be qualified by a quasi-absolute type of statement that places bounds on its systematic error, and a separate statement of its standard error or its probable error, or of an upper bound thereto, whenever a reliable determination of such value or bound is available. Otherwise a computed value of the standard error, or, probable error, so designated, should be given together with a statement of the number of degrees of freedom on which it is based.

2) The bounds to its systematic error and the measure of its imprecision should be stated to no more than two significant figures.

3) The reported result itself should be stated at most to the last place affected by the finer of the two qualifying statements (unless it is desired to indicate and preserve such relative accuracy or precision of a higher order that it may possess for certain particular uses).

4) The qualification of a reported result with respect to its imprecision and systematic error should be given in sentence form, except when results of different precision or with different bounds to their systematic errors are presented in tabular arrangement. If it is necessary or desirable to indicate their respective imprecisions or bounds to their respective systematic errors, such information may be given in a parallel column or columns, with appropriate identification.

Here, and in the next section, the term *standard error* is to be understood as signifying the standard deviation of the reported value itself, not as signifying the standard deviation of the single determination (unless, of course, the reported value is simply the result of a single determination).

The above recommendations should not be construed to exclude the presentation of a quasi-abolute type of statement placing bounds on the inaccuracy, that is, on the overall uncertainty, of a reported value, provided that separate statements of its imprecision and its possible systematic error are included also. To be in good taste, the bounds indicating the overall uncertainty should not be numerically less than the corresponding bounds placed on the systematic error outwardly increased by at least three times the standard error. The fourth of the following examples of good practice is an instance at point:

The standard errors of these values do not exceed 0.000004 inch, and their systematic errors are not in excess of 0.00002 inch.

The standard errors of these values are less than (x units), and their systematic errors are thought to be less than ± (y units). No additional uncertainty is assigned for the conversion to the chemical scale since the adopted conversion factor is taken as 1.000275 exactly.

. . . with a standard error of (x units), and a systematic error of not more than ± (y units).

. . . with an overall uncertainty of ± 3 percent based on a standard error of 0.5 percent and an allowance of ± 1.5 percent for systematic error.

When a reliably established value for the relevant standard error is available, and the dispersion of the present measurements is in keeping with this experience, then this canonical value of the standard error should be used (*5*). If such experience indicates that the standard error is subject to fluctuations greater than the intrinsic variation of such a measure, then an appropriate upper bound should be given, for example, as in the first two of the above examples, or by changing "a standard error . . ." in the third and fourth examples to "an upper bound to the standard error . . ."

When there is insufficient recent experience with the measurement processes involved, an estimate of the standard error must of necessity be computed by recognized statistical procedures from the same measurements as the reported value itself. It is essential that such computations be carried out according to an agreed-upon standard procedure, and the results thereof presented in sufficient detail to enable the reader to form his own judgment, and make his own allowances for their inherent uncertainties. To avoid possible misunderstanding, in such cases, first, the term *computed standard error* should be used; second, the estimate of the standard error employed should be that obtained from

$$\text{estimate of standard error} = \left(\frac{\text{sum of squared residuals}}{n\nu}\right)^{1/2}$$

where n is the (effective) number of completely independent determinations of which a is the arithmetic mean (or other appropriate least-squares adjusted value) and ν is the number of degrees of freedom involved in the sum of squared residuals (that is, the number of residuals minus the number of fitted constants or other independent constraints on the residuals); and third, the number of degrees of freedom should be explicitly stated. If the reported value a is the arithmetic mean, then:

$$\text{estimate of standard error} = (s^2/n)^{1/2}$$

where

$$s^2 = \sum_{i=1}^{n} (x_i - a)^2/(n-1)$$

and n is the number of completely independent determinations of which a is the arithmetic mean. For example:

. . . which is the arithmetic mean of (n) independent determinations and has a standard error of . . .

. . . with an overall uncertainty of ± 5.2 km/sec based on a standard error of 1.5 km/sec and estimated bounds of ± 0.7 km/sec on the systematic error. (The figure 5.2 is equal to 0.7 plus 3 times 1.5.)

or, if based on a computed standard error,

The computed probable error (or, standard error) of these values is (x units),

based on (ν) degrees of freedom, and the systematic error is estimated to be less than $\pm$ (y units).

. . . with an overall uncertainty of $\pm$ 7 km/sec derived from bounds of $\pm$ 0.7 km/sec on the systematic error and a computed standard error of 1.5 km/sec based on 9 degrees of freedom. [The number 7 is approximately equal to $0.7 + (4.3 \times 1.5)$, where 4.3 is the value of Student's t for 9 degrees of freedom exceeded in absolute value with 0.002 probability. As $\nu \to \infty$, $t_{.002}(\nu) \to 3.090$.]

When the reported value is the result of a complex measurement process and is obtained as a function of several quantities whose standard errors have been computed, these several quantities and their standard errors should usually be reported, together with a description of the method of computation by which the standard errors were combined to provide an overall estimate of imprecision for the reported value.

Systematic Error Negligible, Imprecision Not Negligible

When the systematic error of a result is negligible but its imprecision is not, the following rules are recommended:

1) Qualification of a reported value should be limited to a statement of its standard error or of an upper bound thereto, whenever a reliable determination of such value or bound is available. Otherwise a computed value of the standard error, so designated, should be given together with a statement of the number of degrees of freedom on which it is based.

2) The standard error or upper bound thereto, should be stated to not more than two significant figures.

3) The reported result itself should be stated at most to the last place affected by the stated value or bound to its imprecision (unless it is desired to indicate and preserve such relative precision of a higher order that it may possess for certain particular uses).

4) The qualification of a reported result with respect to its imprecision should be given in sentence form, except when results of different precision are presented in tabular arrangement and it is necessary or desirable to indicate their respective imprecisions in which event such information may be given in a parallel column or columns, with appropriate identification.

5) The fact that the systematic error is negligible should be stated explicitly.

The above recommendations should not be construed to exclude the presentation of a quasi-absolute type of statement placing bounds on its possible inaccuracy, provided that a separate statement of its imprecision is included also. To be in good taste, such bounds to its inaccuracy should be numerically equal to at least three times the stated standard error. The fourth of the following examples of good practice is an instance at point.

The standard errors of these values are less than (x units).

. . . with a standard error of (x units).

. . . with a computed standard error of (x units) based on (ν) degrees of freedom.

. . . with an overall uncertainty of $\pm$ 4.5 km/sec derived from a standard error of 1.5 km/sec. (The figure 4.5 is equal to 3×1.5.)

or, if based on a computed standard error,

. . . with an overall uncertainty of $\pm$ 6.5 km/sec derived from a computed standard error of 1.5 km/sec (based on 9 degrees of freedom). (The number 6.5 is equal to 4.3×1.5, where 4.3 is the value of Student's t for 9 degrees of freedom exceeded in absolute value with 0.002 probability. As $\nu \to \infty$, $t_{.002}(\nu) \to 3.090$.)

The remarks with regard to a computed standard error in the preceding section apply with equal force to the last two examples above.

Conclusion

The foregoing recommendations call for fuller and sharper detail than is general in common pactice. They should be regarded as minimum standards of good practice. Of course, many instances require fuller treatment than that recommended here.

Thus, in the case of determinations of the "fundamental physical constants" and other basic properties of nature, the author or authors should give a detailed account of the various components of imprecision and systematic error, and list their respective individual magnitudes in tabular form, so that (i) the state of the art will be more clearly revealed, (ii) each individual user of the final result may decide for himself which of the indicated components of imprecision or systematic error are, or are not, relevant to his use of the final result, and (iii)—most important—the final result itself or its uncertainty can be modified appropriately in the light of later advances. This is, and has long been, the practice followed in the best reports of fundamental studies, but current efforts to prepare critically evaluated standard reference data have revealed that far too great a fraction of the data in the scientific literature "cannot be critically evaluated because the minimum of essential information is not present" (*6*).

References and Notes

1. The true value defined conceptually by an exemplar measurement process, or the target value intended in a practical measurement process.
2. The standard error is the standard deviation of the probability distribution of estimates (that is, reported values) of the quantity that is being measured. See M. G. Kendall and W. R. Buckland, *A Dictionary of Statistical Terms* (Hafner, New York, 1957).
3. For a comprehensive discussion on precision and accuracy, and a selected bibliography of 80 references, see C. Eisenhart, "Realistic Evaluation of the Precision and Accuracy of Instrument Calibration Systems," *J. Res. Nat. Bur. Std.* **67C**, No. 2, 161–187 (1963). (Reprints are available upon request.)
4. The essential elements of the present recommendations first appeared in a 1955 National Bureau of Standards task group report prepared principally by Malcolm W. Jensen (Office of Weights and Measures), Leroy W. Tilton (Optics and Metrology Division), and Churchill Eisenhart (Applied Mathematics Division), which was based for the most part on detailed recommendations developed some years earlier by Dr. Tilton for the internal guidance of the Optics and Metrology Division. In September 1961, new introductory material was added to the recommendations of the 1955 task group; a few minor changes were made in the illustrative examples, and the resulting revised version was circulated as a working paper of the Subcommittee on Accuracy Statements of the NBS Testing and Calibration Committee. This 1961 version was incorporated without essential change as chapter 23, "Expression of the Uncertainties of Final Results," of NBS Handbook 91, *Experimental Statistics* (U.S. Government Printing Office, Washington, 1963), reprinted with corrections in 1966. (This handbook brought together in a single volume the material on experimental statistics prepared at the National Bureau of Standards for the U.S. Army *Ordnance Engineering Design Handbook*, and printed in 1962 for limited distribution as U.S. Army Ordnance Corps Pamphlets ORDP 20-110 through 20-114. Subsequently, when these five pamphlets became parts of the *AMC Engineering Design Handbook*, they were designated Army Materiel Command Pamphlets AMCP 706-110 through 706-114.)

   In the present version, the content of chapter 23 has been rearranged and, in order to be more appropriate to calibration work, more explicit consideration has been given to the case where the value of the standard deviation σ of the measurement process involved has been well established by recent past experience. A terse summary of the principal recommendations of the present paper in the form of a text figure (Fig. 1) is contained in H. H. Ku, "Expressions of Imprecision, Systematic Error, and Uncertainty Associated with a Reported Value," to be published in *Measurements and Data*. The earlier versions were addressed primarily to the case of isolated experiments or tests, where the relevant value of σ is usually unknown in advance, and the statistical uncertainty of the final results must therefore be expressed entirely in terms of quantities derived from the data of the experiment itself.
5. The control chart is an invaluable tool in providing justification for the use of a canonical value of the standard error. See, for example, ASTM Manual on *Quality Control of Materials* (American Society for Testing and Materials, Philadephia, 1951).
6. L. M. Branscomb, "The misinformation explosion: Is the literature worth reviewing?," a talk presented to the Philosophical Society of Washington, 17 November 1967, and to be published in *Scientific Research*.

EXPRESSIONS OF IMPRECISION, SYSTEMATIC ERROR, AND UNCERTAINTY ASSOCIATED WITH A REPORTED VALUE

HARRY H. KU, National Bureau of Standards

The work of a calibration laboratory may be thought of as a sequence of operations that result in the collection, storage, and transmittal of information. In making a statement of uncertainty of the result of calibration, the calibration laboratory transmits information to its clients on the particular item calibrated.

It is logical, then, to require the transmitted information to be meaningful and unambiguous, and to contain all the relevant information in the possession of the laboratory. *The information content of the statement of uncertainty determines, to a large extent, the worth of the calibrated value.*

A common deficiency in many statements of uncertainty is that they do not convey all the information a calibration laboratory has to offer, information acquired through much ingenuity and hard work. This deficiency usually originates in two ways:

1. Loss of information through oversimplification, and

2. loss of information through the inability of the laboratory to take into account information accumulated from its past experience.

With the increasingly stringent demands for improved precision and accuracy of calibration work, calibration laboratories as a whole just cannot afford such luxury.

Traceability to the national standards, accuracy ratios, and class tolerance requirements are simplified concepts that aim to achieve different degrees of accuracy requirements. These concepts and the resulting statements are useful on certain occasions, but fail whenever the demand is exacting. The general practice of obliterating all the identifiable components of uncertainty, by combining them into an overall uncertainty, just for the sake of simplicity, is another case in point. After all, if the calibration laboratory reports all the pertinent information in separate components, the user can always combine them or use them individually, as he sees fit. On the other hand, if the user is given only one number, he can never disentangle this number into its various components. Since the information buried under these oversimplified statements is available, and may well be useful to sophisticated customers, such practices result in substantial waste of effort and resources.

In calibrating an item by repeating the same calibration procedure, the calibration laboratory gains increments of information about its calibration *system*. These increments of information are quantified and accumulated for the benefit of the calibration laboratory. If the precision of the calibration process remains unchanged, the statistical measure of dispersion (s) – i.e., the standard deviations computed from these sets of data – can be pooled together, weighted by their respective degrees of freedom. When many such increments of information are combined, an accepted or canonical value of standard deviation (σ) is established. This established (canonical) value of standard deviation characterizes the precision of the calibration process, and is treasured information in any calibration laboratory.

Hence, the canonical value of standard deviation is the quantification of information accumulated from past experiences of the calibration laboratory, and is an essential element of the statement of uncertainty. The standard deviation (s) computed from the current calibration is used to check the precision of current work, and to add to the pool of information on the process, but certainly does not represent all the information available in the possession of an established calibration laboratory. Only by passing its accumulated information to the users is the calibration laboratory performing a complete service.

STATEMENT OF UNCERTAINTY

In the preparation of a statement of uncertainty, it is helpful to bear in mind that:

1. The derivation of a statement of uncertainty has as its foundation the work done in the laboratory, and is based on information accumulated from past experience, and

2. In general, information is lost through oversimplification, and demands for improved precision and accuracy cannot be met with simplified statements of uncertainty.

Unless a statement of uncertainty is well formulated and supported, it is difficult to say what is meant by the statement, a difficulty frequently encountered. Since the evaluation of uncertainty is part and parcel of the high standard of work of a calibration laboratory, the statement of uncertainty deserves all the attention required to make the statement both realistic and useful. To this end, Tables 1, 2 and 3 give terms and expressions compiled as a ready reference for those who are searching for some appropriate format or wording, to carry out the thoughts expressed. They summarize the recommended practices on expression of uncertainties as given in Chapter 23 of NBS Handbook 91. A revised version of this chapter with the title "Expression of Uncertainties of Final Results" by Churchill Eisenhart may be found in NBS Special Publication 300-1. Figure 1 gives a condensed summary of this material. Tables 1, 2, and 3 give details of forms of imprecision, systematic error, and uncertainty statements.

TABLE 1 – IMPRECISION STATEMENTS

Value reported	Index or Measure of Error	Remarks
Precision of a measurement (calibration) process	(a). Standard deviation (σ) of a single determination (observation)	σ (or s with the associated degrees of freedom[1]) is of main interest as an index of precision of the measurement process. If the average of n such measurements is also reported, see (b) below.
Arithmetic mean ($\bar{x}_n$) of n numbers	(b). Standard error ($\sigma/\sqrt{n}$) of the reported value	$\bar{x}_n$ is of main interest; the number n is also essential information; σ assumed known [1]
	(c). 2 sigma limits (d). 3 sigma limits	Commonly used bounds of imprecisions; usually used when σ known, or when n large.
	(e). Confidence interval (indicate one- or two-sided)	Data points assumed to be normally distributed; report confidence coefficient (level) $100(1-\alpha)\%$. [2]
	(f). Half-width of confidence interval (or confidence limits)	Same as (e) above; for symmetrical two-sided intervals; an index to bounds of imprecision. [2]
	(g). Probable error of the reported value	Probable error $= .6745\,\frac{\sigma}{\sqrt{n}}$ for normally distributed data points when σ known. Use of $\sigma/\sqrt{n}$ preferred. Incorrect if σ not known.
	(h). Mean deviation, or average deviation, of a measurement from the mean calculated from the sample	Limiting mean of mean deviation $= \sqrt{\frac{2}{\pi}}\,\sqrt{\frac{n-1}{n}}\,\sigma$ for normally distributed data points when σ known. Use of σ usually preferred.
	(i). Any of the above expressed in percent, or ppm of $\bar{x}_n$.	State what is being expressed in percent, eg., $(\sigma/\sqrt{n})(100/\bar{x}_n)$, $\bar{x}_n$ being a fairly constant value.
m means each computed from n measurements	(j). (b), (c), (d) and (f) above	If the measurements are of equal precision and σ unknown, use $s_p^2 = \frac{1}{m}\sum_{i=1}^{m} s_i^2$ as estimate of σ^2. The no. of degrees of freedom associated with s_p is $m(n-1)$.
	(k). Sample coefficient of variation ($v = \frac{s}{\bar{x}_n}$) or relative percent ($v \times 100$)	Appropriate when the m means cover a wide range and where the v's computed for the m sets are about the same magnitude. Give range of v's for the m sets. The means must be positive and bounded away from zero.
Weighted mean $\bar{\bar{x}} = \frac{w_1\bar{x}_1 + w_2\bar{x}_2}{w_1 + w_2}$	(l). Standard error ($\sigma_{\bar{\bar{x}}}^2$) of the weighted mean	If $w_1 = 1/\sigma_{\bar{x}_1}^2$ and $w_2 = 1/\sigma_{\bar{x}_2}^2$, then $\sigma_{\bar{\bar{x}}}^2 = \frac{1}{w_1+w_2}$. Not recommended when the σ's are not known and are estimated by s computed from small number of measurements.
An equation (theoretical or empirical) fitted to data points by the method of least squares	(m). Standard deviation computed from the deviations (residuals) of data points from the fitted curve	Report n, the number of data points, and k, the number of constants fitted, $s^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2/(n-k)$, where $\hat{y}_i$ is the value on the fitted curve for the particular x_i.[3] Value of s usually given in computer print-out.
Constants (coefficients) in the equation fitted to the data points by the method of least squares	(n). Standard errors of the coefficients based on the standard deviation computed under (m)	Standard errors usually given in computer print-out. Report n and k as above.[3]

TABLE 1 – IMPRECISION STATEMENTS – (Continued)

Value reported	Index or Measure of Error	Remarks
A predicted point on the curve $\hat{y}$ for a particular x_o	(o). Standard error ($s_{\hat{y}}$) of the predicted point	For the straight line case, the computer print-out gives the variance-covariance matrix $\begin{pmatrix} s_{11} & \\ s_{12} & s_{22} \end{pmatrix}$. $s_{\hat{y}}^2 = s_{11} + 2s_{12}x_o + s_{22}x_o^2$.[2] Report n and k.
A predicted observed value for a particular x_o	(p). Standard error of the predicted value of y	For the straight line case, $s_y^2 = s_{\hat{y}}^2 + s^2$ where $s_{\hat{y}}^2$ and s^2 are that given in (o) and (m) respectively.[3] Report n and k
Value of function of the arithmetic means of several measured variables	(q). Standard error calculated by the use of propagation of error formulas	Appropriate when errors of measurements are small compared to the values of variables measured. Use standard error of the means of the variables in the formulas.[4] Report number of measurements from which these standard errors are computed.
Percentage or proportion (r/ n), r and n being counts	(r). Confidence limits of the true proportion P	Procedures for obtaining exact and approximate confidence limits are discussed in Chapter 7, NBS Handbook 91. State one-sided or two-sided.

TABLE 2 – SYSTEMATIC ERROR[5] (BIAS) STATEMENTS

Value reported	Index or Measure of Error	Remarks
Numerical value resulting from a measurement process	Reasonable bounds ascribed to the value originating from: (i). systematic error reliably established	Detailed discussions of systematic errors are always helpful. Positive wording is appropriate: "... is not in error by more than ..." "... is accurate within ± ..."
	(ii). systematic error estimated from experience or by judgment	Use modifier such as "believed", "estimated", "considered", to signify the conjectural nature of the statement.
	(iii). combination of a number of elemental systematic errors	State explicitly the method of combination such as "the simple sum of the bounds" or "the square root of the sum of squares".
	(iv). uncertainty in some fundamental constant	Give reference to the value of constant used.
	(v). uncertainty in calibrated values	Ascertain the meaning of the systematic and random components of the uncertainty from the calibration laboratory so that decisions on the uses of these components can be made from the correct interpretations.
	(vi). bias in the method of computation	Correct if feasible, or give the magnitude.

TABLE 3 - UNCERTAINTY STATEMENTS

Value reported	Index or Measure of error	Remarks
Numerical value resulting from a measurement process	Bounds to inaccuracy: (1). Systematic error and imprecision both negligible	Explicit expression of correctness to the last significant figure, interpreted as being accurate within $\pm 1/2$ units in the last significant figure given
	(2). Imprecision negligible. Bounds on inaccuracy given to no more than two significant figures.	Sentence form preferred such as given under remark for (i) and (ii). Footnote needed if bounds are given in tabular form.
	(3). Systematic error negligible. Index of precision (b), (g), (h), (i), (k), or (n) stated to no more than two significant figures	State explicitly the index used and give essential information associated with the index. Qualify index calculated by the word "computed". Avoid using expressions of the form $a \pm b$ unless the meaning of b is explained fully immediately following or in footnote.
	(3'). Systematic error negligible. Bounds to imprecision (c), (d), (e), or (f) stated to no more than two significant figures.	Same as under (3).
	(4). Neither systematic error nor imprecision negligible. Two numerics indicating bounds to systematic error and index of imprecision respectively	(2) and (3) above separately stated.
	(4'). Bounds to systematic error and imprecision combined, indicating the likely inaccuracy of the value	(2) and (3') above where the two components either have been previously described, or explained immediately following (or in footnote).
	(5). Quoted from literature	State reference and give author's interpretation of the uncertainty; add remark if meaning unknown or ambiguous.

[1] If σ is not known, use the computed standard deviation s based on k measurements as an estimate of σ, where $s^2 = \frac{1}{k-1} \sum_{i=1}^{k} (x_i - \bar{x}_k)^2$. The number (k-1) is the degrees of freedom associated with s.

[2] For interpretation see Chapter 1, NBS Handbook 91, **Experimental Statistics**, by M. G. Natrella, 1963.

[3] For details see Chapter 5 (straight line), and Chapter 6 (multivariate and polynomial), NBS Handbook 91.

[4] For details see "Notes on the use of propagation of error formulas", by Harry H. Ku, NBS Journal of Research, Vol. 70C, No. 4, October-December, 1966.

[5] See "Realistic Evaluation of the Precision and Accuracy of Instrument Calibration Systems" by Churchill Eisenhart, NBS Journal of Research, Vol. 67C, No. 2, April-June, 1963, and "Systematic Errors in Physical Constants" by W. J. Youden, Physics Today 14, 1961.

FIGURE 1 – SUMMARY OF RECOMMENDATIONS ON EXPRESSIONS OF THE UNCERTAINTIES OF FINAL RESULTS

SYSTEMATIC ERROR AND IMPRECISION BOTH NEGLIGIBLE (CASE 1)

In this case, the reported result should be given correct to the number of significant figures consistent with the accuracy requirements of the situation, together with an explicit statement of its accuracy or correctness.

SYSTEMATIC ERROR NOT NEGLIGIBLE, IMPRECISION NEGLIGIBLE (CASE 2)

(a) Qualification of a reported result should be limited to a single quasi-absolute type of statement that places bounds on its inaccuracy;

(b) These bounds should be stated to no more than two significant figures;

(c) The reported result itself should be given (i.e., rounded) to the last place affected by the stated bounds, unless it is desired to indicate and preserve such relative accuracy or precision of a higher order that the result may possess for certain particular uses;

(d) Accuracy statements should be given in sentence form in all cases, except when a number of results of different accuracies are presented, e.g., in tabular arrangement. If it is necessary or desirable to indicate the respective accuracies of a number of results, the results should be given in the form $a \pm b$ (or $a \pm \frac{b}{c}$, if necessary) with an appropriate explanatory remark (as a footnote to the table, or incorporated in the accompanying test) to the effect that the $\pm b$, or $\pm \frac{b}{c}$, signify bounds to the errors which the a's may be subject.

(e) The fact that the imprecision is negligible should be stated explicity.

NEITHER SYSTEMATIC ERROR NOR IMPRECISION NEGLIGIBLE (CASE 3)

(a) A reported result should be qualified by: (1) a quasi-absolute type of statement that places bounds on its systematic error; and, (2) a separate statement of its standard error or of an upper bound thereto, whenever a reliable determination of such value or bound is available — otherwise, a computed value of the standard error so designated should be given, together with a statement of a number of degrees of freedom on which it is based;

(b) The bounds to its systematic error and the measure of its imprecision should be stated to no more than two significant figures;

(c) The reported result itself should be stated, at most, to the last place affected by the finer of the two qualifying statements, unless it is desired to indicate and preserve such relative accuracy or precision of a higher order that the result may possess for certain particular uses;

(d) The qualification of a reported result, with respect to its imprecision and systematic error, should be given in sentence form, except when results of different precision or with different bounds to their systematic errors are presented in tabular arrangement. If it is necessary or desirable to indicate their respective imprecisions or bounds to their respective systematic errors, such information may be given in a parallel column or columns, with appropriate identification.

SYSTEMATIC ERROR NEGLIGIBLE, IMPRECISION NOT NEGLIGIBLE (CASE 4)

(a) Qualification of a reported value should be limited to a statement of its standard error or of an upper bound thereto, whenever a reliable determination of such value or bound is available. Otherwise, a computed value of the standard error so designated should be given, together with a statement of the number of degrees of freedom on which it is based;

(b) The standard error, or upper bound thereto, should be stated to not more than two significant figures;

(c) The reported result itself should be stated, at most, to the last place affected by the stated value or bound to its imprecision, unless it is desired to indicate and preserve such relative precision of a higher order that the result may possess for certain particular uses;

(d) The qualification of a reported result with respect to its imprecision should be given in sentence form, except when results of different precision are presented in tabular arrangement and it is necessary or desirable to indicate their respective imprecisions, in which event such information may be given in a parallel column or columns, with appropriate identification.

(e) The fact that the systematic error is negligibie should be stated explicitly.

POSTSCRIPT

Over the intervening years since the publication of Eisenhart's and Ku's articles, it has become apparent that a few additional comments may be useful. It is equally apparent that a complete revision is neither necessary nor desirable inasmuch as the major thrust and content of the articles remain as valid and as appropriate as when first written. For this reason, these comments are made as a postscript.

Uncertainty Assessments Must Be Complete

The uncertainty of a reported value is meant to be a credible estimate of the likely limits to its actual *error*, i.e., the magnitude and sign of its deviation from the truth. As such, uncertainty statements must be based on as nearly complete an assessment as possible. This assessment process must consider every conceivable source of inaccuracy in the result.

A measurement process generally consists of a very complicated sequence of many individual unit operations or steps. Virtually every step in this sequence introduces a conceivable source of inaccuracy whose magnitude must be assessed. These sources include:

- Inherent stochastic variability of the measurement process;
- Uncertainties in standards and calibrated apparatus;
- Effects of environmental factors, such as variations in temperature, humidity, atmospheric pressure, and power supply voltage;
- Time-dependent instabilities due to gradual and subtle changes in standards or apparatus;
- Inability to realize physical model because of instrument limitations;
- Methodology procedural errors, such as incorrect logic, or misunderstanding what one is or should be doing;
- Uncertainties arising from interferences, impurities, inhomogeneity, inadequate resolution, incomplete discrimination, etc.;
- Metrologist errors, such as misreading of an instrument;
- Malfunctioning or damaged apparatus;
- Laboratory practice including handling techniques, cleanliness, etc.; and
- Computational uncertainties as well as errors in transcription of data, and other calculational or arithmetical mistakes.

This list should not be interpreted as exhaustive, but rather as illustrative of the most common generic sources of inaccuracy that may be present.

The various sources of inaccuracy are generally classified into sources of *imprecision* (random components) and sources of *bias* (fixed offsets). To which category a particular source should be properly assigned is often difficult and troublesome. In part, this is because many experimental procedures or individual steps in the overall measurement process embody both systematic and stochastic (random) elements. (For an alternative discussion that questions the need for a clear cut distinction between random and systematic components of uncertainty, see [7].) One practical approach is to classify the sources of inaccuracy according to how the uncertainty is estimated. In this way, sources of imprecision are considered to be those components which *can be* and *are* estimated by a statistical analysis of replicate determinations. For completeness, the *systematic uncertainty components* can be considered to be the residual set of conceivable sources of inaccuracy that are biased and not subject to random variability, and those that may be due to random causes but *cannot be* or *are not* assessed by statistical methods. The systematic category includes sources of inaccuracy other than biases in order to obtain a complete accounting of all sources of inaccuracy in the measurement process. Hence, it is meaningful to report a random uncertainty contribution, only if one has a computed statistic for the magnitude of its imprecision or random variation. Many sources of inaccuracy may exist consisting of several components from both the random and systematic categories and can be assessed only after consideration of the more fundamental processes involved. The uncertainty in the calibration of an instrument with a standard reference material, for example, would have not only components from the uncertainty in the standard itself, but also uncertainty components arising from the use of the standard in performing the calibration.

Assessment of Imprecision (Random Uncertainties)

Although the treatment and expressions of reporting the imprecision of measurement results were adequately covered in the original article, a number of points are of sufficient importance to deserve reemphasis.

The only way to assess realistically the overall imprecision is to make direct—or preferably, when possible, indirect—replicate determinations [1] and calculate an appropriate statistic such as the standard error of the mean. It is extremely important to be definite on what constitutes a "replicate determination" because the extent to which conditions are allowed to vary freely over successive "repetitions" of the measurement process determines the scope of the statistical inferences that may be drawn from measurements obtained [2, sec. 4.1]. When measurements of a particular quantity made on a single occasion exhibit closer mutual agreement than measurements made on different occasions so that differences between occasions are indicated, the value of the computed standard error of the mean of all the measurements obtained by lumping all of the measurements together will underestimate the actual standard error of the mean. A more realistic value is given by taking the arithmetic means of the measurements obtained on the respective occasions as the *replicate determinations* and calculating the standard error of their mean in the usual way [3, sec. 3.5].

In many situations, it may not be possible or feasible because of time and cost constraints to perform a sufficient number of completely independent determinations of the measurement result. For results derived from several component quantities, the individual imprecision estimates must be propagated to obtain the imprecision of the final result. It must be emphasized, however, that

these estimates of imprecision should not be based exclusively on the information derived from just the present measurements. Presently derived information should be added to the information accumulated in the past on the imprecision of the measurement process. In this way, more realistic and reliable canonical values of the imprecision statistics may be established over time. Ideally, every major step or component of the measurement process should be independently assessed. This would include not only the variability inherent in the particular measurement of concern, but also the imprecision arising from corrections, calibration factors, and any other quantities that make up the final result.

Assessment of Systematic Uncertainties

Although a general guideline for the approach to the assessment of systematic uncertainties can be formulated, there are, unfortunately, no rules to objectively assign a magnitude to them. For the most part, it is a subjective process. Their magnitudes should preferably be based on experimental verification, but may have to rely on the judgment and experience of the metrologist. In general, each systematic uncertainty contribution is considered as a quasi-absolute upper bound, overall or maximum limit on its inaccuracy. Its magnitude is typically estimated in terms of an interval from plus to minus δ about the mean of the measurement result. By what method then should the magnitude of these maximum limits be assigned? It may be based on comparison to a standard, on experiments designed for the purpose [4], or on verification with two or more independent and reliable measurement methods. Additionally, the limits may be based on judgment, based on experience, based on intuition, or based on other measurements and data. Or the limits may include combinations of some or all of the above factors. Whenever possible, they should be empirically derived or verified. The reliability of the estimate of the systematic uncertainty will largely depend on the resourcefulness and ingenuity of the metrologist.

The Need for an Overall Uncertainty Statement

Without deprecating the perils of shorthand expressions, there is often a need for an overall uncertainty statement which combines the imprecision and systematic uncertainty components. Arguments that it is incorrect from a theoretical point of view to combine the individual components in any fashion are not always practical. First, an approach which retains all details is not amenable for large compilations of results from numerous sources. And second, this approach shifts the burden of evaluating the uncertainties to users. Many users need a single uncertainty value resulting from the combination of all sources of inaccuracy. These users believe, and rightly so, that this overall estimate of inaccuracy can be most appropriately made by the person responsible for the measurement result. It must be emphasized, however, that there is no one clearly superior appropriate method for reporting an overall uncertainty, and that the choice of method is somewhat arbitrary. Several methods are commonly employed [5,6].

One method is to add linearly all components of the systematic uncertainty and linearly add the total to the imprecision estimate. Since the individual systematic uncertainties (δ_j) are considered to be maximum limits, it logically should be added to an imprecision estimate at a similar confidence level. That is, for example, the overall uncertainty u may be given by

$$u = [t_\nu(\alpha)]s + \sum_{j=1}^{q} \delta_j$$

where s is the computed standard error based on ν degrees of freedom, $t_\nu(\alpha)$ is the Student-t value corresponding to a two-tail significance level of $\alpha = 0.05$, 0.01, or 0.001 (depending on the practice in the measurement field concerned), and δ_j is the magnitude of the estimated systematic uncertainty for each of the identified q systematic uncertainty components. This approach probably overestimates the inaccuracy, but can be considered as an estimate of the maximum possible limits. For example, if someone estimated that five contributions of about equal magnitude made up the total systematic error, that person would have to be very unlucky if all five were plus, or all five were minus. Yet, if there was one dominant contributor, it might be a very valid approximation.

Two other approaches have also been widely used. These methods add in quadrature all of the systematic uncertainty components, and either add the resulting quantity *linearly* to the standard error estimate,

$$s + \sqrt{\sum_{j=1}^{q} \delta_j^2} \quad ,$$

or add it *in quadrature* to the standard error estimate,

$$\sqrt{s^2 + \sum_{j=1}^{q} \delta_j^2}$$

These are frequently considered (erroneously) to correspond to a confidence level with $P = 68\%$.

In another method, often termed the PTB approach [6], the component systematic uncertainties are assumed to be independent and distributed such that all values within the estimated limits are equiprobable (rectangular or uniform distribution) [8]. With these assumptions, the rectangular systematic uncertainty distributions can be convoluted to obtain a combined probability distribution for which the variance may be computed. This may then be combined in quadrature with that for the random uncertainty. In its simplest form, the uncertainty components are combined to form an overall uncertainty by

$$u = k\sqrt{s^2 + (1/3)\sum_{j=1}^{q} \delta_j^2} \quad ,$$

where k is customarily taken as 2 or 3. The above simple form is not appropriate when one of the component δ_j's is much larger than the others; in such a case it will be more informative to keep that component separate from the others and add it linearly.

A Concluding Thought

If there is one fundamental proposition for the expression of uncertainties, it is

> The information content of the statement of uncertainty determines, to a large extent, the worth of the final result.

This information content can be maximized by following a few simple principles:

BE EXPLICIT

PROVIDE DETAILS

DON'T OVERSIMPLIFY

When an overall uncertainty is reported, one should explicitly state how the separate components were combined. In addition, for results of primary importance, a detailed discussion and complete specification of all of the separate uncertainty components is still required. In this way, some users will benefit from having the metrologist's estimate of the overall uncertainty, while more sophisticated users will still have access to all of the information necessary for them to evaluate, combine, or use the uncertainties as they see fit.

REFERENCES AND NOTES

[1] Youden, W. J. Statistical aspects of analytical determinations. The Analyst 77: 874–878; 1952, Dec: Reprinted in Journal of Quality Technology 4: 45–49; 1972, Jan: Youden, W. J., Connor, W. S., Making one measurement do the work of two; Chemical and Engineering Progress 49: 549–552; 1953, Oct. Reprinted in Journal of Quality Technology 4: 25–28; 1972, Jan.

[2] Eisenhart, Churchill. Realistic evaluation of the precision and accuracy of instrument calibration systems. J. Res. Nat. Bur. Stand. (U.S.) 67C (2): 161–187; 1963; Reprinted as paper 1.2 in NBS Special Publication 300-1.

[3] Eisenhart, Churchill. Contribution to panel discussion of adjustments of the fundamental physical constants. Langenberg, D. N., Taylor, B. N., eds. Precision Measurement and Fundamental Constants, Nat. Bur. Stand. (U.S.) Spec. Publ. 343: 509–518; 1971.

[4] Youden's burette experiment, Journal of Quality Technology 4: 20-23 1972, Jan. Youden, W. J., Systematic errors in physical constants. Physics Today, 14, 32–34, 36, 38, 40, 43; 1961, Sept. Reprinted as paper 1.4 in NBS Special Publication 300-1. Youden, W. J., Enduring values, Technometrics 14: 1–11; 1972, Feb.

[5] Campion, P. J.; Burns, J. E.; Williams, A. A code of practice for the detailed statement of accuracy. National Physical Laboratory. London: Her Majesty's Stationery Office. 1953. II-57.

[6] Wagner, Siegfried R. On the quantitative characterization of the uncertainty of experimental results in metrology: PTB-Mitteilungen 89: 83–89; 1979, Feb.

[7] Müller, Jörg G. Some second thoughts on error statements. Nuclear Instruments and Methods 163, 241–251; 1979.

[8] A numerical comparison of uncertainty limits resulting from these assumptions with those implied by several alternative distributional assumptions is provided by table 1 on page 184 of [2], and discussed on the same and following page.

Churchill Eisenhart
Ronald Collé

July 1980

NBS Special Publication 700-2
Industrial Measurement Series

Measurement Evaluation

J. Mandel
National Measurement Laboratory
National Bureau of Standards
Gaithersburg, Maryland 20899

and

L. F. Nanni
Federal University
Porto Alegre, Brazil

March 1986

U.S. Department of Commerce
Malcolm Baldrige, Secretary
National Bureau of Standards
Ernest Ambler, Director

Library of Congress
Catalog Card Number: 86-600510
National Bureau of Standards
Special Publication 700-2
Natl. Bur. Stand. (U.S.),
Spec. Publ. 700-2,
70 pages (March 1986)
CODEN: XNBSAV

U.S. Government Printing Office
Washington: 1986

For sale by the Superintendent
of Documents
U.S. Government Printing Office,
Washington, DC 20402

FOREWORD

When the National Bureau of Standards was established more than 80 years ago, it was given the specific mission of aiding manufacturing and commerce. Today, NBS remains the only Federal laboratory with this explicit goal of serving U.S. industry and science. Our mission takes on special significance now as the country is responding to serious challenges to its industry and manufacturing--challenges which call for government to pool its scientific and technical resources with industry and universities.

The links between NBS staff members and our industrial colleagues have always been strong. Publication of this new Industrial Measurement Series, aimed at those responsible for measurement in industry, represents a strengthening of these ties.

The concept for the series stems from the joint efforts of the National Conference of Standards Laboratories and NBS. Each volume will be prepared jointly by a practical specialist and a member of the NBS staff. Each volume will be written within a framework of industrial relevance and need.

This publication is an addition to what we anticipate will be a long series of collaborative ventures that will aid both industry and NBS.

Ernest Ambler, Director

INTRODUCTION

This paper was published originally as a chapter in the book entitled "Quality Assurance Practices for Health Laboratories".* It is for that reason that the examples used as illustrations are taken from health-related fields of research. However, the statistical concepts and methods presented here are entirely general and therefore also applicable to measurements originating in physics, chemistry, engineering, and other technical disciplines. The reader should have no difficulty in applying the material of this paper to the systems of measurement in his particular field of activity.

J. Mandel
January, 1986

---

* J. Mandel and L.F. Nanni, Measurement Evaluation Quality Assurance Practices for Health Laboratories. Washington: American Public Health Association; 1978: 209-272.1244 p.

ABOUT THE AUTHORS

John Mandel

John Mandel holds an M.S. in chemistry from the University of Brussels. He studied mathematical statistics at Columbia University and obtained a Ph.D in statistics from the University of Eindhoven.

Dr. Mandel has been a consultant on statistical design and data analysis at the National Bureau of Standards since 1947. He is the author of a book, "The Statistical Analysis of Experimental Data", and has contributed chapters on statistics to several others. He has written numerous papers on mathematical and applied statistics, dealing more particularly with the application of statistical methodology to the physical sciences.

Mandel has served as a Visiting Professor at Rutgers University and at the Israel Institute of Technology in Haifa. He has contributed to the educational program of the Chemical Division of the American Society for Quality Control through lectures and courses.

A fellow of the American Statistical Association, the American Society for Testing and Materials, the American Society for Quality Control, and the Royal Statistical Society of Great Britain, Mandel, is the recipient of a number of awards, including the U.S. Department of Commerce Silver Medal and Gold Medal, the Shewhart Medal, the Dr. W. Edwards Deming Medal, the Frank Wilcoxon Prize and the Brumbaugh Award of the American Society for Quality Control.

He was Chairman of one of the Gordon Research Conferences on Statistics in Chemistry and Chemical Engineering and has served on several ASTM committees and is, in particular, an active member of Committee E-11 on Statistical Methods.

Luis F. Nanni

Luis F. Nanni holds a Civil Engineering degree from the National University of Tucuman, Argentina and the M.A. from Princeton University. He was a member of the faculty of Rutgers University School of Engineering for many years and served there as Professor of Industrial Engineering. Professor Nanni also has extensive experience as an industrial consultant on statistics in the chemical sciences, the physical sciences and the health sciences. He is a member of several professional societies including the American Statistical Association, the Institute of Mathematical Statistics, the Operations Research Society of America, the American Institute of Industrial Engineers the American Society for Engineering Education.

Professor Nanni's fields of specialization are statistical analysis and operations research; his scholarly contributions include statistical methods, random processes and simulation, computer programming and engineering analysis. At the present time he is Professor of Civil Engineering at the Federal University in Porto Alegre, Brazil.

CONTENTS

Measurement Evaluation

J. Mandel (*principal author*), and L. F. Nanni.

Basic Statistical Concepts

Random variables

This chapter is concerned with the evaluation of measurements *by means of statistical methods*. This qualification is important, for the total evaluation of measurements involves many different points of view. What differentiates the statistical viewpoint from all others is that each measurement is considered as only one realization of a hypothetical infinite population of similar measurements. Although, in general, all members of this population refer to the measurements of the same property on the same sample (e.g., the glucose content of a given sample of serum), they are not expected to be identical. The differences among them are attributable to chance effects, due to unavoidable fluctuations in many of the conditions surrounding the measuring process. Alternatively, the members of the population of measurements may refer to different samples, or different individuals. Thus, one may consider the glucose content of serum of all healthy individuals in a certain age range. In such cases, the observed differences among the measured values include what is referred to as *sampling error*, meaning the differences in the measured property among the members of the population of samples or individuals. A variable whose value is associated with a statistical population is called a *random variable* or *variate*.

Frequency distribution and histograms

A mathematical representation can be made of a statistical population, such as the hypothetical infinite population of measurements just mentioned. To obtain this representation, called a *frequency distribution*, one divides all the measurements in the population into group intervals and counts the number of measurements in each interval. Each interval is defined in terms of its lower and upper limit, in the scale in which the measurement is expressed. Since in practice one is always limited to a statistical sample, i.e., a finite number of measurements, one can at best only approximate the frequency distribution. Such an approximation is called a *histogram*. Figure 4.1 contains a histogram of glucose values in serum measurements on a sample of 2,197 individuals. It is worth noting that the frequency tends to be greatest in the vicinity of the mean and diminishes gradually as the distance from the mean

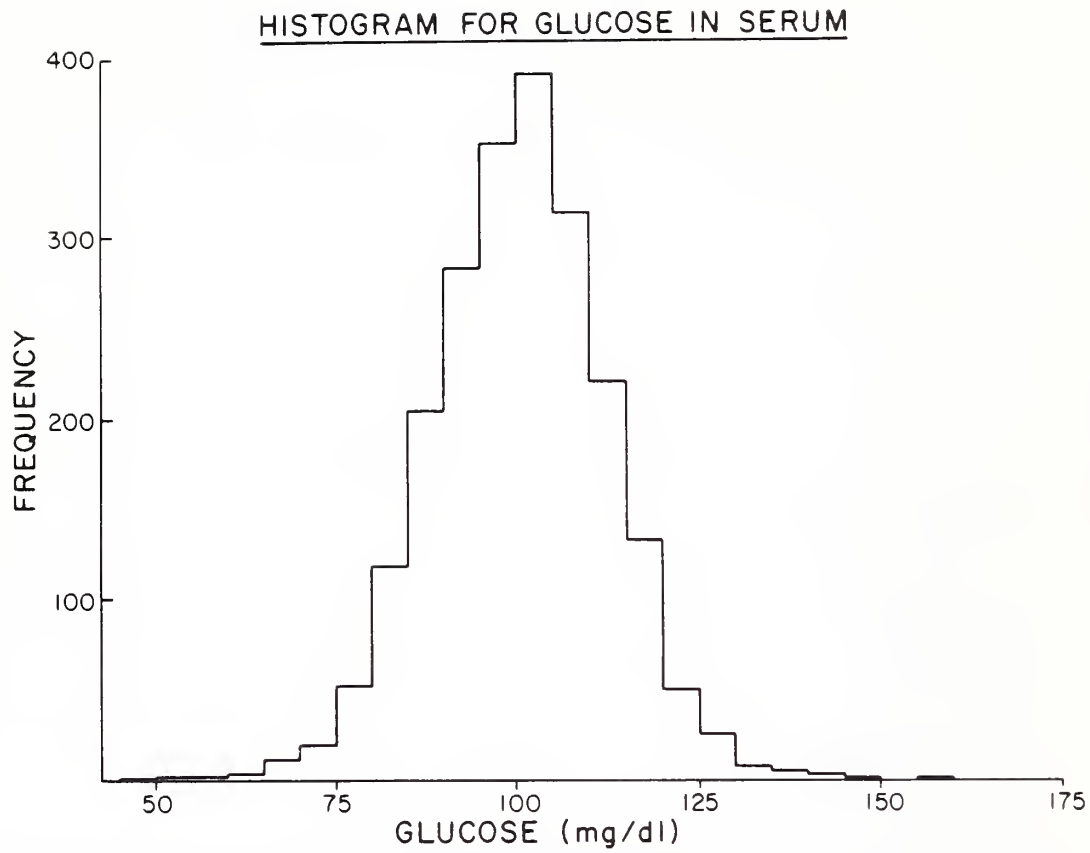


Fig. 4.1. Histogram of glucose serum values on a sample of 2,197 individuals, with a range of 47.5–157.5 mg/dl and a mean of 100.4 mg/dl.

increases. The grouped data on which the histogram is based are given in Table 4.1.

Population parameters and sample estimates

Random samples

The sample of individuals underlying the histogram in Table 4.1 is rather large. A large size, in itself, does not necessarily ensure that the histogram's characteristics will faithfully represent those of the entire population. An additional requirement is that the sample be obtained by a *random selection* from the entire population. A random selection is designed to ensure that each element of the population has an equal chance of being included in the sample. A sample obtained from a random selection is called a *random sample*, although, strictly speaking, it is not the sample but the method of obtaining it that is random. Using the concept of a random sample, it is possible to envisage the population as the limit of a random sample of ever-increasing size. When the sample size N becomes larger and larger, the characteristics of the sample approach those of the entire population. If the random sample is as large as the sample used in this illustration, we may feel confident that its characteristics are quite similar to those of the population.

TABLE 4.1. GROUPED DATA FOR GLUCOSE IN SERUM

Glucose (mg/dl)	Number of individuals	Glucose (mg/dl)	Number of individuals
47.5	1	107.5	313
52.5	2	112.5	220
57.5	2	117.5	132
62.5	3	122.5	50
67.5	12	127.5	26
72.5	20	132.5	8
77.5	52	137.5	6
82.5	118	142.5	4
87.5	204	147.5	1
92.5	281	152.5	0
97.5	351	157.5	1
102.5	390		
		Total number of individuals:	2,197

Thus, upon inspection of Table 4.1, we may feel confident that the mean serum glucose for the entire population is not far from 100.4 mg/dl. We also may feel confident in stating that relatively very few individuals, say about 1 percent of the entire population, will have serum glucose values of less than 70 mg/dl. Our confidence in such conclusions (which, incidentally, can be made more quantitative), however, would have been much less had all of the available data consisted of a small sample, say on the order of five to 50 individuals. Two such sets of data are shown in Table 4.2. Each represents the serum glucose of ten individuals from the population represented in Table 4.1. The mean glucose contents of these samples are 107.57 and 96.37 mg/dl, respectively. If either one of these samples was all the information available

TABLE 4.2. TWO SMALL SAMPLES OF GLUCOSE IN SERUM

Sample I		Sample II	
Individual	Glucose (mg/dl)	Individual	Glucose (mg/dl)
1	134.2	1	88.2
2	119.6	2	82.0
3	91.9	3	96.0
4	96.6	4	94.1
5	118.8	5	96.3
6	105.2	6	108.8
7	103.4	7	106.3
8	112.1	8	101.1
9	97.0	9	89.4
10	96.9	10	101.7
Average	107.57	Average	96.37
Variance	179.44	Variance	70.48
Standard deviation	13.40	Standard deviation	8.40

to us, what could we have concluded about the mean serum glucose of the entire population? And, in that case, what could we have stated concerning the percentage of the population having a serum glucose of less than 70 mg/dl?

Population parameters—general considerations

The answer to these and similar questions requires that we first define some basic characteristics of a statistical sample and relate them to the characteristics of the population. Fortunately, most populations can be characterized in terms of very few quantities, called *parameters*. In many cases, only *two* parameters are required, in the sense that these two parameters contain practically all the pertinent information that is required for answering all useful questions about the population. In cases where more than two parameters are needed, it is often possible to perform a mathematical operation, called a *transformation of scale*, on the measured values, which will reduce the required number of parameters to two. The two parameters in question are the *mean* and the *standard deviation*, measuring, respectively, the location of the center of the population and its spread.

Sample estimates

Let $x_1, x_2, \ldots, x_N$ represent a sample of N measurements belonging to a single population. The *sample mean* is generally denoted by $\bar{x}$ and defined by

$$\bar{x} \equiv \frac{x_1 + x_2 + \ldots + x_N}{N} = \frac{\Sigma x_i}{N} \tag{4.1}$$

The *sample variance* is denoted by s_x^2 and defined by

$$s_x^2 \equiv \frac{\Sigma(x_i - \bar{x})^2}{N - 1} \tag{4.2}$$

The *sample standard deviation* is denoted by s_x and defined by

$$s_x \equiv \sqrt{s_x^2} \tag{4.3}$$

Table 4.2 contains, for each of the samples, the numerical values of $\bar{x}$, s_x^2, and s_x.

Population parameters as limiting values of sample estimates

The quantities defined by Equations 4.1, 4.2, and 4.3 are not the population parameters themselves but rather are *sample estimates* of these parameters. This distinction becomes apparent by the fact that they differ from sample to sample, as seen in Table 4.2. However, it is plausible to assume that as the sample size N becomes very large, the sample estimates become more and more stable and eventually approach the corresponding population parameters. We thus define three new quantities: the *population mean*, denoted by the symbol μ; the *population variance*, denoted by the symbol σ_x^2 or by the symbol Var(x); and the *population standard deviation*, denoted by σ_x. Thus:

$$\sigma_x = \sqrt{\sigma_x^2} = \sqrt{\text{Var}(x)} \tag{4.4}$$

It is customary to denote population parameters by Greek letters (e.g., μ, σ) and sample estimates by Latin letters (e.g., $\bar{x}$, s). Another often used convention is to represent sample estimates by Greek letters topped by a *caret* (ˆ); thus s and $\hat{\sigma}$ both denote a sample estimate of σ. It is apparent from the above definitions that the variance and the standard deviation are not two independent parameters, the former being the square of the latter. In practice, the standard deviation is the more useful quantity, since it is expressed in the same units as the measured quantities themselves (mg/dl in our example). The variance, on the other hand, has certain characteristics that make it theoretically desirable as a measure of spread. Thus, the two basic parameters of a population used in laboratory measurement are: (a) its mean, and (b) either its variance or its standard deviation.

Sums of squares, degrees of freedom, and mean squares

Equation 4.2 presents the sample variance as a ratio of the quantities $\Sigma(x_i - \bar{x})^2$ and $(N - 1)$. More generally, we have the relation:

$$MS = \frac{SS}{DF} \tag{4.5}$$

where *MS* stands for *mean square*, *SS* for *sum of squares*, and *DF* for *degrees of freedom*. The term "sum of squares" is short for "sum of squares of deviations from the mean," which is, of course, a literal description of the expression $\Sigma(x_i - \bar{x})^2$, but it is also used to describe a more general concept, which will not be discussed at this point. Thus, Equation 4.2 is a special case of the more general Equation 4.5.

The reason for making the divisor $N - 1$ rather than the more obvious N can be understood by noting that the N quantities

$$x_1 - \bar{x},\ x_2 - \bar{x},\ \ldots,\ x_N - \bar{x}$$

are not completely independent of each other. Indeed, by summing them we obtain:

$$\sum_i (x_i - \bar{x}) = \Sigma x_i - \Sigma \bar{x} = \Sigma x_i - N\bar{x} \tag{4.6}$$

Substituting for $\bar{x}$ the value given by its definition (Equation 4.1), we obtain:

$$\sum_i (x_i - \bar{x}) = \Sigma x_i - N\,\frac{\Sigma x_i}{N} = 0 \tag{4.7}$$

This relation implies that if any $(N - 1)$ of the N quantities $(x_i - \bar{x})$ are given, the remaining one can be calculated without ambiguity. It follows that while there are N independent measurements, there are only $N - 1$ independent deviations from the mean. We express this fact by stating that the sample variance is based on $N - 1$ *degrees of freedom*. This explanation provides at least an intuitive justification for using $N - 1$ as a divisor for the calculation of s^2. When N is very large, the distinction between N and $N - 1$ becomes unimportant, but for reasons of consistency, we always define the

sample variance and the sample standard deviation by Equations 4.2 and 4.3.

Grouped data

When the data in a sample are given in grouped form, such as in Table 4.1, Equations 4.1 and 4.2 cannot be used for the calculation of the mean and the variance. Instead, one must use different formulas that involve the midpoints of the intervals (first column of Table 4.1) and the corresponding frequencies (second column of Table 4.1).

Formulas for grouped data are given below.

To differentiate the regular average (Equation 4.1) of a set of x_i values from their "weighted average" (Equation 4.8), we use the symbol $\tilde{x}$ (x tilde) for the latter.

$$\tilde{x} = \frac{\sum_i f_i x_i}{\sum f_i} \tag{4.8}$$

$$s_x^2 = \frac{\sum_i f_i (x_i - \tilde{x})^2}{\left(\sum f_i\right) - 1} \tag{4.9}$$

$$s_x = \sqrt{s_x^2} \tag{4.10}$$

where f_i (the "frequency") represents the number of individuals in the ith interval, and x_i is the interval midpoint. The calculation of a sum of squares can be simplified by "coding" the data prior to calculations. The coding consists of two operations:

1) Find an approximate central value x_o (e.g., 102.5 for our illustration) and subtract it from each x_i.
2) Divide each difference $x_i - x_o$ by a convenient value c, which is generally the width of the intervals (in our case, $c = 5.0$).

Let the mean

$$u_i = \frac{x_i - x_o}{c} \tag{4.11}$$

The weighted average $\tilde{u}$ is equal to $(\tilde{x} - x_o)/c$. Operation (1) alters neither the variance nor the standard deviation. Operation (2) divides the variance by c^2 and the standard deviation by c. Thus, "uncoding" is accomplished by multiplying the variance of u by c^2 and the standard deviation of u by c. The formulas in Equations 4.8, 4.9, and 4.10 are illustrated in Table 4.3 with the data from Table 4.1.

We now can better appreciate the difference between population parameters and sample estimates. Table 4.4 contains a summary of the values of the mean, the variance, and the standard deviation for the population (in this case, the very large sample $N = 2{,}197$ is assumed to be identical with the population) and for the two samples of size 10.

TABLE 4.3. CALCULATIONS FOR GROUPED DATA

x	u	f	x	u	f
47.5	−11	1	107.5	1	313
52.5	−10	2	112.5	2	220
57.5	−9	2	117.5	3	132
62.5	−8	3	122.5	4	50
67.5	−7	12	127.5	5	26
72.5	−6	20	132.5	6	8
77.5	−5	52	137.5	7	6
82.5	−4	118	142.5	8	4
87.5	−3	204	147.5	9	1
92.5	−2	281	152.5	10	0
97.5	−1	351	157.5	11	1
102.5	0	390			
	$\bar{u} = -0.4156$		$\bar{x} = 102.5 + 5\bar{u} = 100.42$		
	$s_u^2 = 5.9078$		$s_x^2 = 25 s_u^2 = 147.70$		
	$s_u = 2.4306$		$s_x = 5 s_u = 12.15$		

We first deal with the question: "How reliable is a sample mean as an estimate of the population mean?" The answer requires the introduction of two important concepts—the *standard error of the mean* and the method of *confidence intervals*. Before introducing the latter, however, it is necessary to discuss *normal distribution*.

Standard error of the mean

The widely held, intuitive notion that the average of several measurements is "better" than a single measurement can be given a precise meaning by elementary statistical theory.

Let $x_1, x_2, \ldots, x_N$ represent a sample of size N taken from a population of mean μ and standard deviation σ.

Let $\bar{x}_1$ represent the average of the N measurements. We can visualize a repetition of the entire process of obtaining the N results, yielding a new average $\bar{x}_2$. Continued repetition would thus yield a series of averages $\bar{x}_1, \bar{x}_2, \ldots$. (Two such averages are given by the sets shown in Table 4.2). These averages generate, in turn, a new population. It is intuitively clear, and can readily be proved, that the mean of the population of averages is the same as that of the population of single measurements, i.e., μ. On the other hand, the

TABLE 4.4. POPULATION PARAMETER AND SAMPLE ESTIMATES (DATA OF TABLES 4.1 AND 4.2)

Source	Mean (mg/dl)	Variance (mg/dl)^2	Standard Deviation (mg/dl)
Population[a]	100.42	147.70	12.15
Sample I	107.57	179.55	13.40
Sample II	96.37	70.56	8.40

[a]We consider the sample of Table 4.1 as identical to the population.

variance of the population of averages can be shown to be smaller than that of the population of single values, and, in fact, it can be proved mathematically that the following relation holds:

$$\mathrm{Var}(\bar{x}) = \frac{\mathrm{Var}(x)}{N} \tag{4.12}$$

From Equation 4.12 it follows that

$$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{N}} \tag{4.13}$$

This relation is known as the law of the *standard error of the mean*, an expression simply denoting the quantity $\sigma_{\bar{x}}$. The term *standard error* refers to the variability of *derived* quantities (in contrast to original measurements). Examples are: the mean of N individual measurements and the intercept or the slope of a fitted line (see section on straight line fitting). In each case, the derived quantity is considered a random variable with a definite distribution function. The standard error is simply the standard deviation of this distribution.

Improving precision through replication

Equation 4.13 justifies the above-mentioned, intuitive concept that averages are "better" than single values. More rigorously, the equation shows that the *precision* of experimental results can be improved, in the sense that the *spread* of values is reduced, by taking the average of a number of replicate measurements. It should be noted that the improvement of precision through averaging is a rather inefficient process; thus, the reduction in the standard deviation obtained by averaging ten measurements is only $\sqrt{10}$, or about 3, and it takes 16 measurements to obtain a reduction in the standard deviation to one-fourth of its value for single measurements.

Systematic errors

A second observation concerns the important assumption of *randomness* required for the validity of the law of the standard error of the mean. The N values must represent a *random* sample from the original population. If, for example, *systematic* errors arise when going from one set of N measurements to the next, these errors are not reduced by the averaging process. An important example of this is found in the evaluation of results from different laboratories. If each laboratory makes N measurements, and if the within-laboratory replication error has a standard deviation of σ, the standard deviation between the *averages* of the various laboratories will generally be *larger* than $\sigma/\sqrt{N}$, because additional variability is generally found between laboratories.

The normal distribution

Symmetry and skewness

The mean and standard deviation of a population provide, in general, a great deal of information about the population, by giving its central location

and its spread. They fail to inform us, however, as to the exact way in which the values are distributed around the mean. In particular, they do not tell us whether the frequency or occurrence of values smaller than the mean is the same as that of values larger than the mean, which would be the case for a *symmetrical* distribution. A nonsymmetrical distribution is said to be *skew*, and it is possible to define a parameter of skewness for any population. As in the case of the mean and the variance, we can calculate a sample estimate of the population parameter of skewness. We will not discuss this matter further at this point, except to state that even the set of three parameters, mean, variance, and skewness, is not always sufficient to completely describe a population of measurements.

The central limit theorem

Among the infinite variety of frequency distributions, there is one class of distributions that is of particular importance, particularly for measurement data. This is the class of *normal*, also known as Gaussian, distributions. All normal distributions are symmetrical, and furthermore they can be reduced by means of a simple algebraic transformation to a single distribution, known as the *reduced normal distribution*. The practical importance of the class of normal distributions is related to two circumstances: (a) many sets of data conform fairly closely to the normal distribution; and (b) there exists a mathematical theorem, known as the *central limit theorem*, which asserts that under certain very general conditions the process of averaging data leads to normal distributions (or very closely so), regardless of the shape of the original distribution, provided that the values that are averaged are independent random drawings from the same population.

The reduced form of a distribution

Any *normal* distribution is completely specified by two parameters, its mean and its variance (or, alternatively, its mean and its standard deviation).

Let x be the result of some measuring process. Unlimited repetition of the process would generate a population of values $x_1, x_2, x_3, \ldots$. If the frequency distribution of this population of values has a mean μ and a standard deviation of σ, then the change of scale effected by the formula

$$z = \frac{x - \mu}{\sigma} \tag{4.14}$$

will result in a new frequency distribution of a mean value of *zero* and a standard deviation of *unity*. The z distribution is called the *reduced* form of the original x distribution.

If, in particular, x is normal, then z will be normal too, and is referred to as the *reduced normal distribution*.

To understand the meaning of Equation 4.14, suppose that a particular measurement x lies at a point situated at k standard deviations above the mean. Thus:

$$x = \mu + k\sigma$$

Then, the corresponding z value will be given by

$$z = \frac{(\mu + k\sigma) - \mu}{\sigma} = k$$

Thus the z value simply expresses the distance from the mean, in units of standard deviations.

Some numerical facts about the normal distribution

The following facts about *normal* distributions are noteworthy and should be memorized for easy appraisal of numerical data:

1) In any normal distribution, the fraction of values whose distance from the mean (in either direction) is more than *one* standard deviation is approximately one-third (one in three).
2) In any normal distribution, the fraction of values whose distance from the mean is more than *two* standard deviations, is approximately 5 percent (one in twenty).
3) In any normal distribution, the fraction of values whose distance from the mean is more than *three* standard deviations is approximately 0.3 percent (three in one thousand).

These facts can be expressed more concisely by using the reduced form of the normal distribution:

1) Probability that $|z| > 1$ is approximately equal to 0.33.
2) Probability that $|z| > 2$ is approximately equal to 0.05.
3) Probability that $|z| > 3$ is equal to 0.003.

The concept of coverage

If we define the *coverage* of an interval from A to B to be the fraction of values of the population falling inside this interval, the three facts (1), (2), and (3) can be expressed as follows (where "sigma" denotes standard deviation):

1) A plus-minus *one*-sigma interval around the mean has a coverage of about 2/3 (67 percent).
2) A plus-minus *two*-sigma interval around the mean has a coverage of about 95 percent.
3) A plus-minus *three*-sigma interval around the mean has a coverage of 99.7 percent.

The coverage corresponding to a $\pm z$-sigma interval around the mean has been tabulated for the normal distribution for values of z extending from 0 to 4 in steps of 0.01, and higher in larger steps. Tabulations of the reduced normal distribution, also known as the "normal curve," or "error curve," can be found in most handbooks of physics and chemistry,[1] and in most textbooks of statistics.[2-5] Since the coverage corresponding to $z = 3.88$ is 99.99 percent, it is hardly ever necessary to consider values of z larger than four.

Confidence intervals

A *confidence interval* aims at bracketing the true value of a population parameter, such as its mean or its standard deviation, by taking into account the uncertainty of the sample estimate of the parameter.

Let $x_1, x_2, \ldots, x_N$ represent a sample of size N from a population of mean μ and standard deviation σ. In general μ and σ are unknown, but can be estimated from the sample in terms of $\bar{x}$ and s, respectively.

Confidence intervals for the mean

A confidence interval for the mean μ is an interval, AB, such that we can state, with a prechosen degree of confidence, that the interval AB brackets the population mean μ.

For example, we see in Table 4.3 that the mean of either of the two samples of size 10 is appreciably different from the (true) population mean (100.42 mg/dl). But suppose that the first of the two small samples is all the information we possess. We then would wish to find two values, A and B, *derived completely from the sample*, such that the interval AB is likely to include the true value (100.42). By making this interval long enough we can always easily fulfill this requirement, depending on what we mean by "likely." Therefore, we first express this qualification in a quantitative way by stipulating the value of a *confidence coefficient*. Thus we may require that the interval shall bracket the population mean "with 95 percent confidence." Such an interval is then called a "95 percent confidence interval."

The case of known σ.—We proceed as follows, assuming for the moment that although μ is unknown, the population standard deviation σ is known. We will subsequently drop this restriction.

We have already seen that the population of averages, $\bar{x}$, has mean μ and standard deviation $\sigma/\sqrt{N}$. The reduced variate corresponding to $\bar{x}$ is therefore:

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{N}} \tag{4.15}$$

By virtue of the central limit theorem, the variable $\bar{x}$ generally may be considered to be normally distributed. The variable z then obeys the reduced normal distribution. We can therefore assert, for example, that the probability that

$$-1.96 < z < 1.96 \tag{4.16}$$

is 95 percent. Equation 4.16 can be written

$$-1.96 < \frac{\bar{x} - \mu}{\sigma/\sqrt{N}} < 1.96$$

or

$$\bar{x} - 1.96 \frac{\sigma}{\sqrt{N}} < \mu < \bar{x} + 1.96 \frac{\sigma}{\sqrt{N}} \tag{4.17}$$

The probability that this double inequality will be fulfilled is 95 percent. Consequently, Equation 4.17 provides a confidence interval for the mean. The *lower limit A* of the confidence interval is $\bar{x} - 1.96\,\sigma/\sqrt{N}$; its *upper limit B* is $\bar{x} + 1.96\,\sigma/\sqrt{N}$. Because of the particular choice of the quantity 1.96, the probability associated with this confidence interval is, in this case, 95

percent. Such a confidence interval is said to be a "95 percent confidence interval," or to have a *confidence coefficient* of 0.95. By changing 1.96 to 3.00 in Equation 4.17, we would obtain a 99.7 percent confidence interval.

General formula for the case of known σ.—More generally, from the table of the reduced normal distribution, we can obtain the proper *critical* value z_c (to replace 1.96 in Equation 4.17) for any desired confidence coefficient. The general formula becomes

$$\bar{x} - z_c \cdot \frac{\sigma}{\sqrt{N}} < \mu < \bar{x} + z_c \cdot \frac{\sigma}{\sqrt{N}} \tag{4.18}$$

Values of z_c for a number of confidence coefficients are listed in tables of the normal distribution.

The *length* L of the confidence interval given by Equation 4.18 is

$$L = \left(\bar{x} + z_c \cdot \frac{\sigma}{\sqrt{N}}\right) - \left(\bar{x} - z_c \cdot \frac{\sigma}{\sqrt{N}}\right) = 2z_c \cdot \frac{\sigma}{\sqrt{N}} \tag{4.19}$$

The larger the confidence coefficient, the larger will be z_c and also L. It is also apparent that L increases with σ, but *decreases* as N becomes larger. This decrease, however, is slow, as it is proportional to only the square root of N. By far the best way to obtain short confidence intervals for an unknown parameter is to choose a measuring process for which the dispersion σ is small—in other words, to choose a measuring process of high precision.

The case of unknown σ. *Student's* t *distribution.*—A basic difficulty associated with the use of Equation 4.18 is that σ is generally unknown. However, the sample of N values provides us with an estimate s of σ. This estimate has $N - 1$ degrees of freedom. Substitution of s for σ in Equation 4.18 is not permissible, since the use of the reduced normal variate z in Equation 4.15 is predicated on a knowledge of σ.

It has been shown, however, that if $\bar{x}$ and s are the sample estimates obtained from a sample of size N, from a normal population of mean μ and standard deviation σ, the quantity, analogous to Equation 4.15, given by

$$t \equiv \frac{\bar{x} - \mu}{s/\sqrt{N}} \tag{4.20}$$

has a well-defined distribution, depending only on the degrees of freedom, $N - 1$, with which s has been estimated. This distribution is known as Student's t distribution with $N - 1$ degrees of freedom.

For σ unknown, it is still possible, therefore, to calculate confidence intervals for the mean μ by substituting in Equation 4.18 s for σ, and t_c for z_c. The confidence interval is now given by

$$\bar{x} - t_c \frac{s}{\sqrt{N}} < \mu < \bar{x} + t_c \frac{s}{\sqrt{N}} \tag{4.21}$$

The critical value t_c, for any desired confidence coefficient, is obtained from a tabulation of Student's t distribution. Tables of Student's t values can

be found in several references.[2-5] The length of the confidence interval based on Student's t distribution is

$$L = 2t_c \frac{s}{\sqrt{N}} \tag{4.22}$$

For any given confidence coefficient, t_c will be larger than z_c, so that the length of the interval given by Equation 4.22 is larger than that given by Equation 4.19. This difference is to be expected, since the interval now must take into account the uncertainty of the estimate s in addition to that of $\bar{x}$.

Applying Equation 4.21 to the two samples shown in Table 4.2, and choosing a 95 percent confidence coefficient (which, for 9 degrees of freedom, gives $t_c = 2.26$), we obtain:

1) For the first sample:

$$107.57 - 2.26 \frac{13.40}{\sqrt{10}} < \mu < 107.57 + 2.26 \frac{13.40}{\sqrt{10}}$$

or

$$98.0 < \mu < 117.2$$

The length of this interval is

$$117.2 - 98.0 = 19.2$$

2) For the second sample:

$$96.37 - 2.26 \frac{8.40}{\sqrt{10}} < \mu < 96.37 + 2.26 \frac{8.40}{\sqrt{10}}$$

or

$$90.4 < \mu < 102.4$$

The length of this interval is

$$102.4 - 90.4 = 12.0$$

Remembering that the population mean is 100.4, we see that the confidence intervals, though very different in length from each other, both bracket the population mean. We also may conclude that the lengths of the intervals, which depend on the sample size, show that a sample of size 10 is quite unsatisfactory when the purpose is to obtain a good estimate of the population mean, unless the measurement process is one of high precision.

Confidence intervals for the standard deviation

The chi-square distribution.—In many statistical investigations, the standard deviation of a population is of as much interest, if not more, than the mean. It is important, therefore, to possess a formula that provides a confidence interval for the unknown population standard deviation σ, given a sample estimate s.

If the number of degrees of freedom with which s is estimated is denoted by ν, a confidence interval for σ is given by the formula:

$$s\sqrt{\frac{\nu}{\chi_U^2}} < \mu < s\sqrt{\frac{\nu}{\chi_L^2}} \tag{4.23}$$

In this formula, the quantities χ_U^2 and χ_L^2 are the appropriate upper and lower percentage points of a statistical distribution known as *chi-square*, for the chosen confidence coefficient. These percentage points are found in several references.[2-5]

This formula can be illustrated by means of the two samples in Table 4.2. To calculate 95 percent confidence intervals for σ (the population standard deviation), we locate the limits at points corresponding to the upper and lower 2.5 percentage points (or the 97.5 percentile and the 2.5 percentile) of chi-square. From the chi-square table we see that, for 9 degrees of freedom, the 97.5 percentile is 19.02, and the 2.5 percentile is 2.70. The 95 percent confidence interval in question is therefore:

1) For the first sample:

$$13.40\sqrt{\frac{9}{19.02}} < \sigma < 13.40\sqrt{\frac{9}{2.70}}$$

or

$$9.2 < \sigma < 24.5$$

2) For the second sample:

$$8.40\sqrt{\frac{9}{19.02}} < \sigma < 8.40\sqrt{\frac{9}{2.70}}$$

or

$$5.8 < \sigma < 15.3$$

Here again, both intervals bracket the population standard deviation 12.15, but again the lengths of the intervals reflect the inadequacy of samples of size 10 for a satisfactory estimation of the population standard deviation.

Tolerance intervals

In introducing the data of Table 4.1, we observed that it was possible to infer that about 1 percent of the population has serum glucose values of less than 70 mg/dl. This inference was reliable because of the large size of our sample (N = 2,197). Can similar inferences be made from small samples, such as those shown in Table 4.2? Before answering this question, let us first see how the inference from a very large sample (such as that of Table 4.1) can be made quantitatively precise.

The reduced variate for our data is

$$z = \frac{x - \mu}{\sigma} = \frac{x - 100.42}{12.15}$$

Making $x = 70$ mg/dl, we obtain for the corresponding reduced variate:

$$z = \frac{70 - 100.42}{12.15} = -2.50$$

If we now assume that the serum glucose data are normally distributed (i.e., follow a Gaussian distribution), we read from the table of the normal distribution that the fraction of the population for which z is less than -2.50 is 0.0062, or 0.62 percent. This is a more precise value than the 1 percent estimate we obtained from a superficial examination of the data.

It is clear that if we attempted to use the same technique for the samples of size 10 shown in Table 4.2, by substituting $\bar{x}$ for μ and s for σ, we may obtain highly unreliable values. Thus, the first sample gives a z value equal to $(70 - 107.57)/13.40$ or -2.80, which corresponds to a fraction of the population equal to 0.25 percent, and the second sample gives $z = (70 - 96.37)/8.40 = -3.14$, which corresponds to a fraction of the population equal to 0.08 percent. It is obvious that this approach cannot be used for small samples. It is possible, however, to solve *related* problems, even for small samples. The statistical procedure used for solving these problems is called the method of *tolerance intervals*.

Tolerance intervals for average coverages

Generally speaking, the method of tolerance intervals is concerned with the estimation of coverages or, conversely, with the determination of intervals that will yield a certain coverage. Let us consider an interval extending from $\bar{x} - ks$ to $\bar{x} + ks$, where k is any given value. The coverage corresponding to this interval will be a random variable, since the end points of the interval are themselves random variables. However, we can find a k value such that, on the average, the coverage for the interval will be equal to any preassigned value, such as, for example, 0.98. These k values, for normal distributions, have been tabulated for various sample sizes and desired average coverages.[5,6] As an illustration, we consider the first sample of size 10 given in Table 4.2, where

$$\bar{x} = 107.57, \; s = 13.40$$

For a coverage of 98 percent and 9 degrees of freedom, the tabulated value is

$$k = 3.053$$

Hence the tolerance interval that, on the average, will include 98 percent of the population is

$$107.57 - (3.053)(13.40) \text{ to } 107.57 + (3.053)(13.40)$$

or

$$66.7 \text{ to } 148.5$$

We can compare this interval to the one derived from the population itself (for all practical purposes, the large sample of 2,197 individuals may be considered as the population). Using the normal table, we obtain for a 98 percent coverage

$$100.42 - (2.326)(12.15) \text{ to } 100.42 + (2.326)(12.15)$$

or

$$72.2 \text{ to } 128.7$$

The fact that the small sample gives an appreciably wider interval is due to the uncertainties associated with the estimates $\bar{x}$ and s.

For a more detailed discussion of tolerance intervals, see Proschan.[6] Tables of coefficients for the calculation of tolerance intervals can be found in Snedecor and Cochran[5] and Proschan.[6]

Non-parametric tolerance intervals—order statistics

The tabulations of the coefficients needed for the computation of tolerance intervals are based on the assumption that the measurements from which the tolerance intervals are calculated follow a normal distribution; the table is inapplicable if this condition is grossly violated. Fortunately, one can solve a number of problems related to tolerance intervals for data from *any* distribution, by using a technique known as *non-parametric* or *distribution-free*. The method always involves an ordering of the data. First one rewrites the observation $x_1, x_2, \ldots, x_N$ in increasing order of magnitude. We will denote the values thus obtained by

$$x_{(1)}, x_{(2)}, \ldots, x_{(N)}$$

For example, Sample I in Table 4.2 is rewritten as:

$x_{(1)} = 91.9$	$x_{(6)} = 105.2$
$x_{(2)} = 96.6$	$x_{(7)} = 112.1$
$x_{(3)} = 96.9$	$x_{(8)} = 118.8$
$x_{(4)} = 97.0$	$x_{(9)} = 119.6$
$x_{(5)} = 103.4$	$x_{(10)} = 134.2$

The values $x_{(1)}, x_{(2)}, \ldots x_{(N)}$ are denoted as the first, second, . . . , Nth *order statistic*. The order statistics can now be used in a number of ways, depending on the problem of interest. Of particular usefulness is the following general theorem.

A general theorem about order statistics.—On the average, the fraction of the population contained between any two *successive* order statistics from a sample of size N is equal to $\frac{1}{N+1}$. The theorem applies to any continuous distribution (not only the Gaussian distribution) and to any sample size N.

Tolerance intervals based on order statistics.—It follows immediately from the above theorem that, *on the average*, the fraction of the population contained between the first and the last order statistics (the smallest and the largest values in the sample) is $\frac{N-1}{N+1}$. For example, on the average, the frac-

tion of the population contained between the smallest and the largest value of a sample of size 10 is $\frac{10-1}{10+1} = \frac{9}{11}$. The meaning of the qualification "on the average" should be properly understood. For any particular sample of size 10, the actual fraction of the population contained in the interval $x_{(N)} - x_{(1)}$ will generally not be equal to $\frac{N-1}{N+1}$. But if the average of those fractions is taken for many samples of size N, it will be close to $\frac{N-1}{N+1}$.

Tolerance intervals involving confidence coefficients

One can formulate more specific questions related to coverages by introducing, in addition to the coverage, the *confidence* of the statement about the coverage. For example, one can propose to find two order statistics such that the confidence is at least 90 percent that the fraction of the population contained between them (the coverage) is 95 percent. For a sample of size 200, these turn out to be the third order statistic from the bottom and the third order statistic from the top (see Table A30 in Natrella[3]). For further discussion of this topic, several references are recommended.[3,5,6]

Non-normal distributions and tests of normality

Reasons for the central role of the normal distribution in statistical theory and practice have been given in the section on the normal distribution. Many situations are encountered in data analysis for which the normal distribution does not apply. Sometimes non-normality is evident from the nature of the problem. Thus, in situations in which it is desired to determine whether a product conforms to a given standard, one often deals with a simple dichotomy: the fraction of the lot that meets the requirements of the standard, and the fraction of the lot that does not meet these requirements. The statistical distribution pertinent to such a problem is the *binomial* (see section on the binomial distribution).

In other situations, there is no a priori reason for non-normality, but the data themselves give indications of a non-normal underlying distribution. Thus, a problem of some importance is to "test for normality."

Tests of normality

Tests of normality should never be performed on small samples, because small samples are inherently incapable of revealing the nature of the underlying distribution. In some situations, a sufficient amount of evidence is gradually built up to detect non-normality and to reveal the general nature of the distribution. In other cases, it is sometimes possible to obtain a truly large sample (such as that shown in Table 4.1) for which normality can be tested by "fitting a normal distribution" to the data and then testing the "goodness of the fit."[5]

Probability plots.—A graphical procedure for testing for normality can be performed using the order statistics of the sample. This test is facilitated through the use of "normal probability paper," a type of graph paper on which the vertical scale is an ordinary arithmetic scale and the horizontal

scale is *labeled* in terms of *coverages* (from 0 to 100 percent), but *graduated* in terms of the reduced z-values corresponding to these coverages (see section on the normal distribution). More specifically, suppose we divide the abscissa of a plot of the normal curve into $N + 1$ segments such that the area under the curve between any two successive division points is $\frac{1}{N+1}$. The division points will be $z_1, z_2, \ldots, z_N$, the values of which can be determined from the normal curve. Table 4.5 lists the values $\frac{1}{N+1}, \frac{2}{N+2}, \ldots, \frac{N}{N+1}$, in percent, in column 1, and the corresponding normal z values in column 2, for $N = 10$. According to the general theorem about order statistics, the order statistics of a sample of size $N = 10$ "attempt" to accomplish just such a division of the area into $N + 1$ equal parts. Consequently, the order statistics tend to be linearly related to the z values. The order statistics for the first sample of Table 4.2 are listed in column 3 of Table 4.5. A plot of column 3 versus column 2 will constitute a "test for normality": if the data are normally distributed, the plot will approximate a straight line. Furthermore, the intercept of this line (see the section on straight line fitting) will be an estimate of the mean, and the slope of the line will be an estimate of the standard deviation.[2] For non-normal data, systematic departures from a straight line should be noted. The use of normal probability paper obviates the calculations involved in obtaining column 2 of Table 4.5, since the horizontal axis is graduated according to z but labeled according to the values $\frac{i}{N+1}$, expressed as percent. Thus, in using the probability paper, the ten order statistics are plotted versus the numbers

$$100\,\frac{1}{11},\ 100\,\frac{2}{11},\ \ldots,\ 100\,\frac{10}{11}$$

or 9.09, 18.18, . . . , 90.91 percent. It is only for illustrative purposes that we have presented the procedure by means of a sample of size 10. One would generally not attempt to use this method for samples of less than 30. Even then, subjective judgment is required to determine whether the points fall along a straight line.

In a subsequent section, we will discuss transformations of scale as a means of achieving normality.

The binomial distribution

Referring to Table 4.1, we may be interested in the fraction of the population for which the serum glucose is greater than, say, 110 mg/dl. A problem of this type involves partitioning the range of values of a continuous variable (serum glucose in our illustration) into two groups, namely: (a) the group of individuals having serum glucose less than 110 mg/dl; and (b) the group of individuals having serum glucose greater than 110 mg/dl. (Those having serum glucose exactly equal to 110 mg/dl can be attached to one or the other group, or their number divided equally among them.)

TABLE 4.5. TEST OF NORMALITY USING ORDER STATISTICS[a]

Expected cumulative areas[b] in percent	Reduced normal variate	Order statistics of sample
9.09	−1.335	91.9
18.18	−0.908	96.6
27.27	−0.604	96.9
36.36	−0.348	97.0
45.45	−0.114	103.4
54.54	0.114	105.2
63.64	0.348	112.1
72.73	0.604	118.8
81.82	0.908	119.6
90.91	1.335	134.2

Straight Line Fit of column 3 versus column 2:

Intercept = 107.6 = $\hat{\mu}$

Slope = 15.5 = $\hat{\sigma}$

[a]The example is merely illustrative of the method. In practice one would never test normality on a sample of size 10.

[b]values of $100\ \frac{i}{N+1}$, where $N = 10$.

Suppose now that we have a random sample of only 100 individuals from the entire population. What fraction of the 100 individuals will be found in either group? It is seen that the binomial distribution has shifted the emphasis from the continuous variable (serum glucose) to the *number of individuals* (or the corresponding *fraction*, or *percentage*) in each of the two groups. There are cases in which no continuous variable was ever involved: for example, in determining the number of times a six appears in throwing a die. However, the theory of the binomial applies equally to both types of situations.

The binomial parameter and its estimation

Let P represent the *fraction* (i.e., a number between zero and one) of individuals in one of the two groups (e.g., serum glucose *greater* than 110 mg/dl) *in the population*. It is customary to represent the fraction for the other group by Q. Then it is obvious that $Q = 1 - P$. (If the fractions are expressed as percentages, we have percent $Q = 100 -$ percent P.) For the data in Table 4.1 and the dividing value 110 mg/dl, we can calculate P by using the normal distribution:

The reduced value corresponding to 110 mg/dl is

$$\frac{110 - 100.42}{12.15} = 0.79$$

From the table of the normal distribution, we then obtain for P:

$$P = 0.215$$

Hence $Q = 1 - 0.215 = 0.785$

Let p represent the fraction of individuals having the stated characteristic (serum glucose greater than 110 mg/dl) in the *sample of size* N; and let $q = 1 - p$. It is clear that for a relatively small, or even a moderately large N, p will generally differ from P. In fact, p is a random variable with a well-defined distribution function, namely the *binomial*.

The mean of the binomial (with parameter P) can be shown to be equal to P. Thus

$$E(p) = P \tag{4.24}$$

where the symbol $E(p)$ represents the "expected value" of p, another name for the population mean. Thus the population mean of the distribution of p is equal to the parameter P. If p is taken as an estimate for P, this *estimate* will therefore be *unbiased*.

Furthermore:

$$\text{Var}(p) = \frac{PQ}{N} \tag{4.25}$$

Hence

$$\sigma_p = \sqrt{\frac{PQ}{N}} \tag{4.26}$$

The normal approximation for the binomial distribution

It is a remarkable fact that for a large N, the distribution of p can be approximated by the normal distribution of the same mean and standard deviation. This enables us to easily solve practical problems that arise in connection with the binomial. For example, returning to our sample of 100 individuals from the population given in Table 4.1, we have:

$$E(p) = 0.215$$

$$\sigma_p = \sqrt{\frac{(0.215)(0.785)}{100}} = 0.0411$$

From these values, one may infer that in a sample of $N = 100$ from the population in question, the chance of obtaining p values of less than 0.13 (two standard deviations below the mean) or of more than 0.30 (two standard deviations above the mean) is about 5 percent. In other words, the chances are approximately 95 percent that in a sample of 100 from the population in question the number of individuals found to have serum glucose of more than 110 mg/dl will be more than 13 and less than 30.

Since, in practice, the value of P is generally unknown, all inferences must then be drawn from the sample itself. Thus, if in a sample of 100 one finds a p value of, say, 0.18 (i.e., 18 individuals with glucose serum greater than 110 mg/dl), one will consider this value as an estimate for P, and consequently one will take the value

$$\sqrt{\frac{(0.18)(1 - 0.18)}{100}} = 0.038$$

as an estimate for σ_p. This would lead to the following approximate 95 percent confidence interval for P:

$$0.18 - (1.96)(.038) < P < 0.18 + (1.96)(.038)$$

or

$$0.10 < P < 0.25$$

The above discussion gives a general idea about the uses and usefulness of the binomial distribution. More detailed discussions will be found in two general references.[4,5]

Precision and accuracy

The concept of control

In some ways, a measuring process is analogous to a manufacturing process. The analogue to the raw product entering the manufacturing process is the system or sample to be measured. The outgoing final product of the manufacturing process corresponds to the numerical result produced by the measuring process. The concept of *control* also applies to both types of processes. In the manufacturing process, control must be exercised to reduce to the minimum any random fluctuations in the conditions of the manufacturing equipment. Similarly, in a measuring process, one aims at reducing to a minimum any random fluctuations in the measuring apparatus and in the environmental conditions. In a manufacturing process, control leads to greater uniformity of outgoing product. In a measuring process, control results in higher *precision*, i.e., in less random scatter in repeated measurements of the same quantity.

Mass production of manufactured goods has led to the necessity of interchangeability of manufactured parts, even when they originate from different plants. Similarly, the need to obtain the same numerical result for a particular measurement, regardless of where and when the measurement was made, implies that *local* control of a measuring process is not enough. Users also require *interlaboratory* control, aimed at assuring a high degree of "interchangeability" of results, even when results are obtained at different times or in different laboratories.

Methods of monitoring a measuring process for the purpose of achieving "local" (i.e., within-laboratory) control will be discussed in the section on quality control of this chapter. In the following sections, we will be concerned with a different problem: estimating the precision and accuracy of a *method* of measurement.

Within- and between-laboratory variability

Consider the data in Table 4.6, taken from a study of the hexokinase method for determining serum glucose. For simplicity of exposition, Table

TABLE 4.6. DETERMINATION OF SERUM GLUCOSE

Laboratory	Serum sample			
	A	B	C	D
1	40.9[a]	76.0	137.8	206.3
	42.3	78.6	137.4	208.5
	42.3	77.5	138.5	204.9
	40.5	77.8	138.5	210.3
2	43.4	78.6	135.2	211.6
	43.8	76.0	131.3	201.2
	43.1	76.8	146.7	201.2
	42.3	75.7	133.4	208.7
3	41.3	75.0	134.5	205.1
	40.2	76.1	134.8	200.3
	40.6	76.4	131.5	206.9
	42.0	76.4	133.4	199.9

[a]All results are expressed in mg glucose/dl.

4.6 contains only a portion of the entire set of data obtained in this study. Each of three laboratories made four replicate determinations on each of four serum samples. It can be observed that, for each sample, the results obtained by *different* laboratories tend to show greater differences than results obtained through replication in the same laboratory. This observation can be made quantitative by calculating, for each sample, two standard deviations: the standard deviation "within" laboratories and the standard deviation "between" laboratories. Within-laboratory precision is often referred to as *repeatability*, and between-laboratory precision as *reproducibility*.[7] We will illustrate the method for serum sample A.

The data for serum A can first be summarized as follows:

Laboratory	Average	Standard Deviation
1	41.50	0.938
2	43.15	0.635
3	41.02	0.793

The three standard deviations could be averaged to obtain an "average within-laboratory"standard deviation. However, if one can assume that these three standard deviations are estimates of one and the same population standard deviation, a better way is to "pool" the variances,[2] and take the square root of the pooled variance. Using this procedure, we obtain for the best estimate of the within-laboratory standard deviation s_w:

$$s_w = \sqrt{\frac{(0.938)^2 + (0.635)^2 + (0.793)^2}{3}} = 0.798$$

Let us now calculate the standard deviation among the three average values 41.50, 43.15, and 41.02. Denoting this standard deviation by $s_{\bar{x}}$, we obtain:

$$s_{\bar{x}} = 1.117$$

If the laboratories displayed no systematic differences, this standard deviation, being calculated from averages of four individual results, should be equal to $s_w/\sqrt{4} = 0.798/\sqrt{4} = 0.399$. The fact that the calculated value, 1.117, is appreciably larger than 0.399 can be explained only through the presence of an additional, *between-laboratory* component of variability. This component, expressed as a standard deviation and denoted by s_L (where L stands for "laboratories"), is calculated by subtracting the "anticipated" variance, $(0.399)^2$, from the "observed" variance, $(1.117)^2$, and taking the square root:

$$s_L = \sqrt{(1.117)^2 - (0.399)^2} = 1.04$$

The calculations for all four serum samples are summarized in Table 4.7, in which standard deviations are rounded to two decimal places.

It may be inferred from Table 4.7 that s_w tends to increase with the glucose content of the sample. The between-laboratory component, s_L, shows no such trend. However, the data are insufficient to establish these facts with reasonable confidence. Since our purpose is to discuss general principles, and the use of these data is only illustrative, we will ignore these shortcomings in the discussion that follows.

Accuracy—comparison with reference values

The two components, s_w and s_L, define the *precision* of the method. To estimate its *accuracy*, one requires *reference values* for all samples. Let us assume that such values have been established and are as follows:

Serum Sample	Reference Value
A	40.8
B	76.0
C	133.4
D	204.1

The values given here as "reference values" are actually only tentative. We will assume, however, in our present discussion, that they can be considered to be free of systematic errors. Our task is to decide whether the values obtained in our study are, *within random experimental error*, equal to these reference values. The grand average value for sample A, 41.89 mg/dl, which

TABLE 4.7. SUMMARY OF ANALYSIS FOR SERUM GLUCOSE DATA

Serum sample	Average (mg/dl)	Standard deviation	
		s_w (mg/dl)	s_L (mg/dl)
A	41.89	0.80	1.04
B	76.74	1.05	0.54
C	136.08	4.08	1.07
D	205.41	3.91	1.08

we denote by the symbol $\bar{\bar{x}}$, involves 12 individual determinations and four laboratories. Its variance, therefore, can be estimated by the formula:

$$s_{\bar{\bar{x}}} = \sqrt{\frac{(0.80)^2}{12} + \frac{(1.04)^2}{4}} = 0.57$$

Now, $\bar{\bar{x}}$ differs from the reference value by the amount:

$$41.89 - 40.8 = 1.09$$

Corresponding values for all four samples are shown in Table 4.8.

It can be seen that, on the one hand, all four grand averages are larger than the corresponding reference values but, on the other hand, the differences D are of the order of only one or two standard errors $s_{\bar{\bar{x}}}$. One would tentatively conclude that the method shows a positive systematic error (bias) but, as has been pointed out above, the data are insufficient to arrive at definite conclusions.

Straight line fitting

The fitting of straight lines to experimental data is a subject of great importance, particularly in analytical laboratories. Many analytical and clinical methods make extensive use of linear calibration curves for the purpose of converting a measured quantity, such as an optical absorbance or a ratio of peaks–heights on a mass-spectrometer scan, into a concentration value for an unknown constituent. Calibration curves are established by subjecting samples of known concentrations to the measuring process and fitting lines to the resulting data. Let x be the known concentration, and y the measurement (e.g., optical absorbance). The data will consist of a set of paired values, as shown for an illustrative example in the columns labeled x and y in Table 4.9.

Inspection of the table shows that there is a "blank": for zero concentration, one finds a nonzero absorbance value. If one "corrected" the subsequent two values for the blank, one would obtain $0.189 - 0.050 = 0.139$, and $0.326 - 0.050 = 0.276$. If the "corrected" absorbance were proportional to concentration (as required by Beer's law), these two corrected absorbances should be proportional to 50 and 100, i.e., in a ratio of 1 to 2. Actually, 0.139 is slightly larger than (0.276/2). We will assume that this is due

TABLE 4.8. STUDY OF ACCURACY OF GLUCOSE DETERMINATION

Serum sample	Reference value (R)	Grand average ($\bar{\bar{x}}$)	D ($\bar{\bar{x}} - R$)	$s_{\bar{\bar{x}}}$
A	40.8	41.89	1.09	0.57
B	76.0	76.74	0.74	0.41
C	133.4	136.08	2.68	1.29
D	204.1	205.41	1.31	1.25

TABLE 4.9. CALIBRATION CURVE FOR GLUCOSE IN SERUM

x	y	$\hat{y}$	d
0	0.050	0.0516	−0.0016
50	0.189	0.1895	−0.0005
100	0.326	0.3273	−0.0013
150	0.467	0.4652	0.0015
200	0.605	0.6030	0.0020
400	1.156	1.1545	0.0015
600	1.704	1.7059	−0.0019
214.29	0.6425	0.6425	0

$\hat{y} = 0.0516 + 0.0027571\,x$
$s_\epsilon = 0.0019$

x = concentration of glucose, in mg/dl
y = absorbance
$\hat{y}$ = "fitted value"
d = residual

solely to experimental error in the measured absorbance values, thus assuming that any errors in the concentration values are negligibly small.

A general model

If α represents the true value of the "blank" and β the absorbance per unit concentration, we have, according to Beer's law:

$$E(y) - \alpha = \beta x \tag{4.27}$$

where $E(y)$ is the expected value for absorbance, i.e., the absorbance value freed of experimental error. Now the actual absorbance, y, is affected by an experimental error, which we will denote by e. Hence:

$$y = E(y) + e \tag{4.28}$$

Combining Equations 4.27 and 4.28 we obtain the "model" equation

$$y = \alpha + \beta x + e \tag{4.29}$$

This equation should hold for all x-values, i.e., $x_1, x_2, \ldots, x_N$, with the same values of α and β. Hence

$$y_i = \alpha + \beta x_i + e_i \tag{4.30}$$

where $i = 1$ to N.
The errors e_i should, on the average, be zero, but each one departs from zero by a random amount. We will assume that these random departures from zero do not increase with the absorbance (in some cases, this assumption is not valid) and that their distribution is Gaussian with standard deviation σ_e.

The object of the analysis is to estimate: (a) α and β, as well as the uncertainties (standard errors) of these estimates; and (b) the standard deviation of e; i.e., σ_e.

Formulas for linear regression

The fitting process is known in the statistical literature as the "linear regression of y on x." We will denote the estimates of α, β, and σ_e by $\hat{\alpha}$, $\hat{\beta}$, and s_e, respectively. The formulas involve the following three quantities:

$$U = \Sigma(x_i - \bar{x})^2 \tag{4.31}$$

$$W = \Sigma(y_i - \bar{y})^2 \tag{4.32}$$

$$P = \Sigma(x_i - \bar{x})(y_i - \bar{y}) \tag{4.33}$$

In terms of these three quantities, we have the formulas:

$$\hat{\beta} = \frac{P}{U} \qquad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} \tag{4.34}$$

$$s_e = \sqrt{\frac{W - (P^2/U)}{N - 2}} \tag{4.35}$$

$$s_{\hat{\beta}} = \frac{s_e}{\sqrt{U}} \qquad s_{\hat{\alpha}} = s_e \sqrt{\frac{1}{N} + \frac{\bar{x}^2}{U}} \tag{4.36}$$

For the data of Table 4.9, the calculations result in the following values: $\hat{\alpha} = 0.0516$, $s_{\hat{\alpha}} = 0.0010$, $\hat{\beta} = 0.0027571$, $s_{\hat{\beta}} = 0.0000036$, $s_e = 0.0019$. Since $\hat{\alpha}$ and $\hat{\beta}$ are now available, we can calculate, for each x, a "calculated" (or "fitted") value, $\hat{y}$, given by the equation $\hat{y} = \hat{\alpha} + \hat{\beta}x$. This is, of course, simply the ordinate of the point on the fitted line for the chosen value of x.

The differences between the observed value y and the calculated value $\hat{y}$ is called a "residual." Table 4.9 also contains the values of $\hat{y}$ and the residuals, denoted by the symbol "d."

It is important to observe that the quantity $(W - P^2/U)$, occurring in Equation 4.35, is simply equal to Σd_i^2. Thus:

$$s_e = \sqrt{\frac{\Sigma d^2}{N - 2}} \tag{4.37}$$

This formula, though mathematically equivalent to Equation 4.35, should be used in preference to Equation 4.35, unless all calculations are carried out with many significant figures. The reason for this is that the quantities d_i are less affected by rounding errors than the quantity $(W - P^2/U)$.

Examination of residuals—weighting

The residuals should behave like a set of random observations with a mean of *zero* and a standard deviation σ_e. It follows that the algebraic signs should exhibit a random pattern similar to the occurrence of heads and tails in the flipping of a coin. In our example, the succession of signs raises some suspicion of nonrandomness, but the series is too short to decide on this matter one way or the other. In any case, the errors are quite small, and the calibration curve is quite satisfactory for the intended purpose.

The assumptions underlying this procedure of fitting a straight line are not always fulfilled. The assumption of homoscedasticity (i.e., all e_i have the same standard deviation), in particular, is often violated. If the standard deviation of the error e_i is nonconstant and depends on x_i, the fitting of the straight line requires the application of "weighted regression analysis." Briefly, assuming a different value of σ_{e_i} for each i, one defines a "weight" w_i equal to the reciprocal of the square of σ_{e_i}. Thus:

$$w_i = 1/\sigma_{e_i}^2 \tag{4.38}$$

The weights w_i are then used in the regression calculations, leading to formulas that are somewhat different from those given in this section. For further details, two references can be consulted.[2,5]

Propagation of errors

It is often necessary to evaluate the uncertainty of a quantity that is not directly measured but is derived, by means of a mathematical formula, from other quantities that *are* directly measured.

An example

As an example, consider the determination of glucose in serum, using an enzymatic reaction sequence. The sequence generates a product, the optical absorbance of which is measured on a spectrophotometer. The procedure consists of three steps: (a) apply the enzyme reaction sequence to a set of glucose solutions of known concentrations, and establish in this way a calibration curve of "absorbance" versus "glucose concentration," (b) by use of the same reaction sequences, measure the absorbance for the "unknown," and (c) using the calibration curve, convert the absorbance for the unknown into a glucose concentration.

It turns out that the calibration curve, for this sequence of reactions, is *linear*. Thus, if y represents absorbance, and x concentration, the calibration curve is represented by the equation:

$$y = \alpha + \beta x \tag{4.39}$$

The calibration curve is established by measuring y for a set of known x values. We will again use the data of Table 4.9 for illustration. Fitting a straight line to these data, we obtain:

$$y = 0.0516 + 0.0027571x \tag{4.40}$$

Let us now suppose that an unknown sample of serum is analyzed m times (for example, $m = 4$), and that the average absorbance found is $y_u = 0.3672$ (where y_u stands for absorbance for the unknown). Using the calibration line, we convert the value y_u into a concentration value, $\hat{x}_u$, by solving the calibration equation for x:

$$x_u = \frac{y_u - \alpha}{\beta} = \frac{0.3672 - 0.0516}{0.0027571} = 114.47 \text{ mg/dl} \tag{4.41}$$

How reliable is this estimate?

Let us assume, at this point, that the uncertainty of the calibration line is negligible. Then the only quantity affected by error is y_u, and it is readily seen from Equation 4.41 that the error of $\hat{x}_u$ is equal to that of y_u, divided by β. If we assume that the standard deviation of a single measured y-value is 0.0019 absorbance units, then the standard error of y_u, the average of four determinations, is

$$0.0019/\sqrt{4} = 0.00095$$

Hence the standard deviation of $\hat{x}_u$ is

$$0.00095/\beta = 0.00095/0.0027571 = 0.34 \text{ mg/dl}$$

A more rigorous treatment would also take account of the uncertainty of the calibration line.

The general case

More generally, a calculated quantity z can be a function of several measured quantities $x_1, x_2, x_3, \ldots$, each of which is affected by experimental error. The problem to be solved is the calculation of the standard deviation of the error of z as a function of the standard deviations of the errors of $x_1, x_2, x_3, \ldots$.

We will only deal with the case of *independent* errors in the quantities $x_1, x_2, x_3, \ldots$; i.e., we assume that the error of any one of the x's is totally unaffected by the errors in the other x's. For independent errors in the measured values $x_1, x_2, x_3, \ldots$, some simple rules can be applied. They are all derived from the application of a general formula known as "the law of propagation of errors," which is valid under very general conditions. The reader is referred to Mandel[2] for a general discussion of this formula.

Linear relations.—For

$$y = a_1x_1 + a_2x_2 + a_3x_3 + \ldots \tag{4.42}$$

the law states:

$$\text{Var}(y) = a_1^2 \text{Var}(x_1) + a_2^2 \text{Var}(x_2) + a_3^2 \text{Var}(x_3) + \ldots \tag{4.43}$$

As an example, suppose that the weight of a sample for chemical analysis has been obtained as the difference between two weights: the weight of an empty crucible, W_1, and the weight of the crucible containing the sample, W_2. Thus the sample weight S is equal to

$$S = W_2 - W_1 \tag{4.44}$$

This is in accordance with Equation 4.42 by writing:

$$S = (1)W_1 + (-1)W_2$$

Hence, according to Equation 4.43,

$$\text{Var}(S) = (1)^2\text{Var}(W_1) + (-1)^2\text{Var}(W_2)$$

or

$$\text{Var}(S) = \text{Var}(W_1) + \text{Var}(W_2)$$

Hence

$$\sigma_S = \sqrt{\sigma_{W_1}^2 + \sigma_{W_2}^2} \tag{4.45}$$

Note that in spite of the negative sign occurring in Equation 4.44, the variances of W_1 and W_2 in Equation 4.45 are *added* (not subtracted from each other).

It is also of great importance to emphasize that Equation 4.43 is valid *only* if the errors in the measurements $x_1, x_2, x_3, \ldots$, are *independent* of each other. Thus, if a particular element in chemical analysis was determined as the difference between 100 percent and the sum of the concentrations found for all other elements, the error in the concentrations for that element would *not* be independent of the errors of the other elements, and Equation 4.43 could *not* be used for any linear combination of the type of Equation 4.42 involving the element in question and the other elements. But in that case, Equations 4.42 and 4.43 could be used to evaluate the error variance for the element in question by considering it as the *dependent* variable y. Thus, in the case of three other elements x_1, x_2, and x_3, we would have:

$$y = 100 - (x_1 + x_2 + x_3)$$

where the errors of x_1, x_2, and x_3 are independent. Hence:

$$\mathrm{Var}(y) = \mathrm{Var}(x_1) + \mathrm{Var}(x_2) + \mathrm{Var}(x_3)$$

since the constant, 100, has zero-variance.

Products and ratios.—For products and ratios, the law of propagation of errors states that the squares of the coefficients of variation are additive. Here again, independence of the errors is a necessary requirement for the validity of this statement. Thus, for

$$y = x_1 \cdot x_2 \tag{4.46}$$

with independent errors for x_1 and x_2, we have:

$$\left(100\,\frac{\sigma_y}{y}\right)^2 = \left(100\,\frac{\sigma_{x_1}}{x_1}\right)^2 + \left(100\,\frac{\sigma_{x_2}}{x_2}\right)^2 \tag{4.47}$$

We can, of course, divide both sides of Equation 4.47 by 100^2, obtaining:

$$\left(\frac{\sigma_y}{y}\right)^2 = \left(\frac{\sigma_{x_1}}{x_1}\right)^2 + \left(\frac{\sigma_{x_2}}{x_2}\right)^2 \tag{4.48}$$

Equation 4.48 states that for products of independent errors, the squares of the *relative* errors are additive.

The same law applies for ratios of quantities with independent errors. Thus, when x_1 and x_2 have independent errors, and

$$y = \frac{x_1}{x_2} \tag{4.49}$$

we have

$$\left(\frac{\sigma_y}{y}\right)^2 = \left(\frac{\sigma_{x_1}}{x_1}\right)^2 + \left(\frac{\sigma_{x_2}}{x_2}\right)^2 \tag{4.50}$$

As an illustration, suppose that in a gravimetric analysis, the sample weight is S, the weight of the precipitate is W, and the "conversion factor" is F. Then:

$$y = 100F\frac{W}{S}$$

The constants 100 and F are known without error. Hence, for this example,

$$\left(\frac{\sigma_y}{y}\right)^2 = \left(\frac{\sigma_W}{W}\right)^2 + \left(\frac{\sigma_S}{S}\right)^2$$

If, for example, the coefficient of variation for S is 0.1 percent, and that for W is 0.5 percent, we have:

$$\frac{\sigma_y}{y} = \sqrt{(0.005)^2 + (0.001)^2} = 0.0051$$

It is seen that in this case, the error of the sample weight S has a negligible effect on the error of the "unknown" y.

Logarithmic functions.—When the calculated quantity y is the natural logarithm of the measured quantity x (we assumed that $x > 0$):

$$y = \ln x \tag{4.51}$$

the law of propagation of error states

$$\sigma_y = \frac{\sigma_x}{x} \tag{4.52}$$

For logarithms to the base 10, a multiplier must be used: for

$$y = \log_{10} x \tag{4.53}$$

the law of propagation of error states:

$$\sigma_y = \frac{1}{2.30} \cdot \frac{\sigma_x}{x} \tag{4.54}$$

Sample sizes and compliance with standards

Once the repeatability and reproducibility of a method of measurement are known, it is a relatively simple matter to estimate the size of a statistical sample that will be required to detect a desired effect, or to determine whether a given specification has been met.

An example

As an illustration, suppose that a standard requires that the mercury content of natural water should not exceed 2μg/l. Suppose, furthermore, that the standard deviation of reproducibility of the test method (see section on precision and accuracy, and Mandel[7]), at the level of 2μg/l, is 0.88μg/l. If subsamples of the water sample are sent to a number of laboratories and

each laboratory performs a single determination, we may wish to determine the number of laboratories that should perform this test to ensure that we can detect noncompliance with the standard. Formulated in this way, the problem has no definite solution. In the first place, it is impossible to guarantee *unqualifiedly* the detection of any noncompliance. After all, the decision will be made on the basis of measurements, and measurements are subject to experimental error. Even assuming, as we do, that the method is *unbiased*, we still have to contend with random errors. Second, we have, so far, failed to give precise meanings to the terms "compliance" and "noncompliance"; while the measurement in one laboratory might give a value less than 2μg/l of mercury, a second laboratory might report a value greater than 2μg/l.

General procedure—acceptance, rejection, risks

To remove all ambiguities regarding sample size, we might proceed in the following manner. We consider two situations, one definitely acceptable and the other definitely unacceptable. For example, the "acceptable" situation might correspond to a *true* mercury content of 1.5μg/l, and the "unacceptable" situation to a mercury content of 2.5μg/l (see Fig. 4.2).

Because of experimental errors, we must consider two *risks*: that of *rejecting* (as noncomplying) a "good" sample (1.5μg/1); and that of *accepting* (as complying) a "bad" sample (2.5μg/1). Suppose that both risks are set at 5 percent.

Let us now denote by N the number of laboratories required for the test. The average of the N measurements, which we denote by $\bar{x}$, will follow a normal distribution whose mean will be the true value of the mercury content of the sample and whose standard deviation will be $\sigma/\sqrt{N}$, or $0.88/\sqrt{N}$. For the "acceptable" situation the mean is 1.5μg/l, and for the "unacceptable" situation it is 2.5μg/1. We now stipulate that we will *accept*

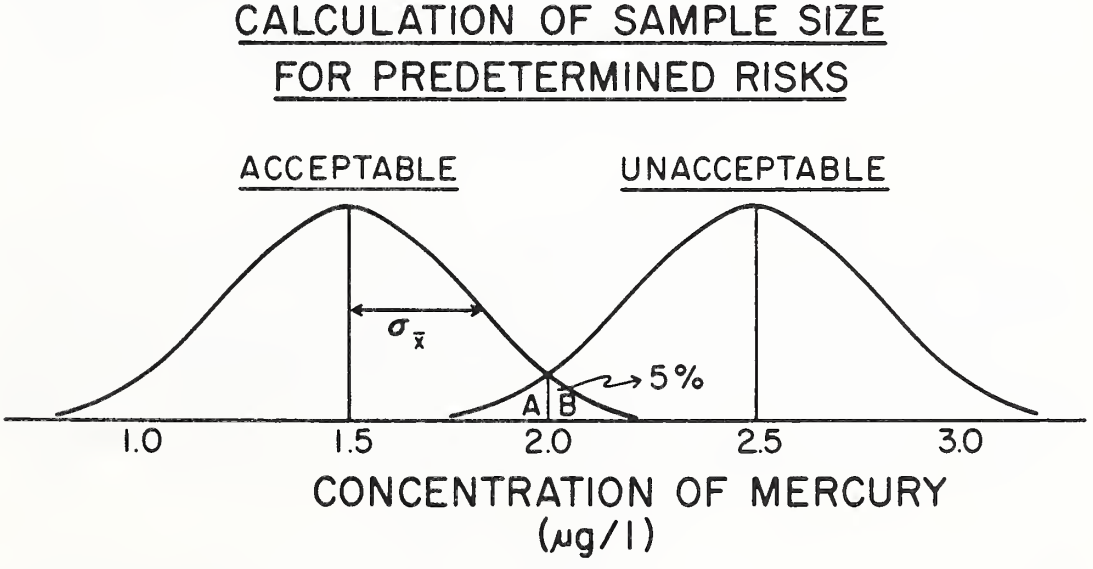


Fig. 4.2. Distribution of measurements of mercury in subsamples of a water sample sent to N laboratories.

the sample, as complying, whenever $\bar{x}$ is less than 2.0, and *reject* it, as noncomplying, whenever $\bar{x}$ is greater than 2.0. As a result of setting our risks at 5 percent, this implies that the areas A and B are each equal to 5 percent (see Fig. 4.2). From the table of the normal distribution, we read that for a 5 percent one-tailed area, the value of the reduced variate is 1.64. Hence:

$$z = \frac{2.0 - 1.5}{0.88/\sqrt{N}} = 1.64$$

(We could also state the requirement that $(2.0 - 2.5)/(0.88/\sqrt{N}) = -1.64$, which is algebraically equivalent to the one above.) Solving for N, we find:

$$N = \left(\frac{1.64 \cdot 0.88}{0.5}\right)^2 = 8.3 \tag{4.55}$$

We conclude that nine laboratories are required to satisfy our requirements. The general formula, for equal risks of accepting a noncomplying sample and rejecting a complying one, is:

$$N = \left(\frac{z_c \cdot \sigma}{d}\right)^2 \tag{4.56}$$

where σ is the appropriate standard deviation, z_c is the value of the reduced normal variate corresponding to the risk probability (5 percent in the above example), and d is the departure (from the specified value) to which the chosen risk probability applies.

Inclusion of between-laboratory variability

If the decision as to whether the sample size meets the requirements of a standard must be made in a single laboratory, we must make our calculations in terms of a different standard deviation. The proper standard deviation, for an average of N determinations in a *single* laboratory, would then be given by:

$$\sigma = \sqrt{\frac{\sigma_w^2}{N} + \sigma_L^2} \tag{4.57}$$

The term σ_L^2 must be included, since the laboratory mean may differ from the true value by a quantity whose standard deviation is σ_L. Since the between-laboratory component σ_L^2 is not divided by N, σ cannot be less than σ_L no matter how many determinations are made in the single laboratory. Therefore, the risks of false acceptance or false rejection of the sample cannot be chosen at will. If in our case, for example, we had $\sigma_w = 0.75\mu$g/l and $\sigma_L = 0.46\mu$g/l, the total σ cannot be less than 0.46. Considering the favorable case, $\mu = 1.5\mu$g/l, the reduced variate (see Fig. 4.2) is:

$$\frac{2.0 - 1.5}{0.46} = 1.09$$

This corresponds to a risk of 13.8 percent of rejecting (as noncomplying) a sample that is actually complying. This is also the risk probability of accept-

ing (as complying) a sample that is actually noncomplying. The conclusion to be drawn from the above argument is that, in some cases, testing error will make it impossible to keep the double risk of accepting a noncomplying product and rejecting a complying product below a certain probability value. If, as in our illustration, the purpose of the standard is to protect health, the proper course of action is to set the specified value at such a level that, even allowing for the between-laboratory component of test error, the risk of declaring a product as complying, when it is actually noncomplying, is low. If, in our illustration, a level of 2.5μg/l is such that the risk of false acceptance of it (as complying) should be kept to 5 percent (and $\sigma_L = 0.46\mu$g/l), then the specification limit should be set at a value x such that:

$$\frac{2.5 - x}{0.46} = 1.64$$

which, solved for x, yields 1.75 μg/l.

Transformation of scale

Some common transformations

Non-normal populations are often *skew* (nonsymmetrical), in the sense that one tail of the distribution is longer than the other. Skewness can often be eliminated by a *transformation of scale*. Consider, for example, the three numbers 1, 10, and 100. The distance between the second and the third is appreciably larger than that between the first and the second, causing a severe asymmetry. If, however, we convert these numbers to logarithms (base 10), we obtain 0, 1, and 2, which constitute a symmetrical set. Thus, if a distribution is *positively skewed* (long-tail on the right), a logarithmic transformation will reduce the skewness. (The simple logarithmic transformation is possible only when all measured values are positive). A transformation of the logarithmic type is not confined to the function $y = \log x$. More generally, one can consider a transformation of the type:

$$y = K \log (A + Bx) \tag{4.58}$$

or even

$$y = C + K \log (A + Bx) \tag{4.59}$$

where C, K, A, and B are properly chosen constants. It is necessary to choose A and B such that $A + Bx$ is positive for all x values. Other common types of transformations are:

$$y = \sqrt{x} \tag{4.60}$$

and

$$y = \arcsin \sqrt{x} \tag{4.61}$$

Robustness

The reason given above for making a transformation of scale is the presence of skewness. Another reason is that certain statistical procedures are

valid only when the data are at least approximately normal. The procedures may become grossly invalid when the data have a severely non-normal distribution.

A statistical procedure that is relatively insensitive to non-normality in the original data (or, more generally, to any set of specific assumptions) is called "robust." Confidence intervals for the mean, for example, are quite robust because, as a result of the central limit theorem, the distribution of the sample mean $\bar{x}$ will generally be close to normality. On the other hand, tolerance intervals are likely to be seriously affected by non-normality. We have seen that nonparametric techniques are available to circumvent this difficulty.

Suppose that, for a particular type of measurement, tests of normality on many sets of data always show evidence of non-normality. Since many statistical techniques are based on the assumption of normality, it would be advantageous to transform these data into new sets that are more nearly normal.

Fortunately, the transformations that reduce skewness also tend, in general, to achieve closer compliance with the requirement of normality. Therefore, transformations of the logarithmic type, as well as the square root and arcsine transformations, are especially useful whenever a nonrobust analysis is to be performed on a set of data that is known to be seriously non-normal. The reader is referred to Mandel[2] for further details regarding transformations of scale.

Transformations and error structure

It is important to realize that any nonlinear transformation changes the *error structure* of the data, and transformations are, in fact, often used for the purpose of making the experimental error more uniform over the entire range of the measurements. Transformations used for this purpose are called "variance-stabilizing" transformations. To understand the principle involved, consider the data in Table 4.10, consisting of five replicate absorbance values at two different concentrations, obtained in the calibration of

TABLE 4.10. ERROR STRUCTURE IN A LOGARITHMIC TRANSFORMATION OF SCALE

	Original data (Absorbance)		Transformed data ($\log_{10}$ Absorbance)	
	Set A[a]	Set B[b]	Set A	Set B
	0.2071	1.6162	−0.6838	0.2085
	0.2079	1.5973	−0.6821	0.2034
	0.1978	1.6091	−0.7038	0.2066
	0.1771	1.7818	−0.7518	0.2509
	0.2036	1.6131	−0.6912	0.2077
Average	0.1987	1.6435	−0.7025	0.2154
Standard deviation	0.0127	0.0776	0.0288	0.0199

[a]Absorbance values for a solution of concentration of 50 mg/dl of glucose.

[b]Absorbance values for a solution of concentration of 600 mg/dl of glucose.

spectrophotometers for the determination of serum glucose. At the higher concentration level, the absorbance values are of course higher, but so is the standard deviation of the replicate absorbance values. The ratio of the average absorbance values is 1.6435/0.1987 = 8.27. The ratio of the standard deviations is 0.0776/0.0127 = 6.11. Thus the standard deviation between replicates tends to increase roughly in proportion to the level of the measurement. We have here an example of "heterogeneity of variance." Let us now examine the two sets of values listed in Table 4.10 under the heading "transformed data." These are simply the logarithms to the base 10 of the original absorbance values. This time, the standard deviations for the two levels are in the proportion 0.0199/0.0288 = 0.69. Thus, the logarithmic transformation essentially has eliminated the heterogeneity of variance. It has, in fact, "stabilized" the variance. The usefulness of variance stabilizing transformations is twofold: (a) a single number will express the standard deviation of error, regardless of the "level" of the measurement; and (b) statistical manipulations whose validity is contingent upon a uniform error variance (homoscedasticity) and which are therefore inapplicable to the original data, can be applied validly to the transformed data.

Presentation of data and significant figures

The law of propagation of errors (see that section) enables one to calculate the number of significant figures in a calculated value. A useful rule of thumb is to report any standard deviation or standard error with two significant figures, and to report a calculated value with as many significant figures as are required to reach the decimal position of the second significant digit of its standard error.

An example

Consider the volumetric determination of manganese in manganous cyclohexanebutyrate by means of a standard solution of sodium arsenite. The formula leading to the desired value of percent Mn is

$$\text{Percent Mn} = 100\frac{v(\text{ml}) \cdot t\left(\frac{\text{mg}}{\text{ml}}\right) \cdot \frac{200\ (\text{ml})}{15(\text{ml})}}{w(\text{mg})}$$

where w is the weight of the sample, v the volume of reagent, and t the titer of the reagent, and the factor 200/15 is derived from taking an aliquot of 15 ml from a total volume of 200 ml.

For a particular titration, the values and their standard errors are found to be:

$v = 23.67$	$\sigma_v = 0.0040$
$t = 0.41122$	$\sigma_t = 0.000015$
200	$\sigma = 0.0040$
15	$\sigma = 0.0040$
$w = 939.77$	$\sigma_w = 0.0060$

The values are reported as they are read on the balance or on the burettes and pipettes; their standard errors are estimated on the basis of previous experience. The calculation gives:

$$\text{Percent Mn} = 13.809872$$

The law of propagation of errors gives:

$$\sigma_{\%Mn} =$$

$$13.8099 \sqrt{\left(\frac{0.0040}{23.67}\right)^2 + \left(\frac{0.000015}{0.41122}\right)^2 + \left(\frac{0.0040}{200}\right)^2 + \left(\frac{0.0040}{15}\right)^2 + \left(\frac{0.0060}{939.77}\right)^2}$$

$$= 0.0044$$

On the basis of this standard deviation, we would report this result as:

$$\text{Percent Mn} = 13.8099;\ \sigma_{\%Mn} = 0.0044$$

It should be well understood that this calculation is based merely on weighing errors, volume reading errors, and the error of the titer of the reagent. In repeating the determination in different laboratories or even in the same laboratory, uncertainties may arise from sources other than just these errors. They would be reflected in the standard deviation calculated from such repeated measurements. In general, this standard deviation will be larger, and often considerably larger, than that calculated from the propagation of weighing and volume reading errors. If such a standard deviation from repeated measurements has been calculated, it may serve as a basis to redetermine the precision with which the reported value should be recorded.

In the example of the manganese determination above, the value given is just the first of a series of repeated determinations. The complete set of data is given in Table 4.11. The average of 20 determinations is 13.8380. The

TABLE 4.11. MANGANESE CONTENT OF MANGANOUS CYCLOHEXANEBUTYRATE

Determination number	Result (Percent Mn)	Determination number	Result (Percent Mn)
1	13.81	11	13.92
2	13.76	12	13.83
3	13.80	13	13.73
4	13.79	14	13.99
5	13.94	15	13.89
6	13.76	16	13.76
7	13.88	17	13.88
8	13.81	18	13.82
9	13.84	19	13.87
10	13.79	20	13.89
	Average = $\bar{x}$ = 13.838		
	$s_x = 0.068$		
	$s_{\bar{x}} = 0.068/\sqrt{20} = 0.015$		

standard deviation of the replicate values is 0.068; therefore, the standard error of the mean is $0.068/\sqrt{20} = 0.015$. The final value reported for this analysis would therefore be:

$$\text{Percent Mn} = \bar{x} = 13.838; s_{\bar{x}} = 0.015$$

This example provides a good illustration of the danger of basing an estimate of the precision of a value solely on the *reading* errors of the quantities from which it is calculated. These errors generally represent only a small portion of the total error. In this example, *the average of 20 values* has a true standard error that is still more than three times larger than the reading error of a *single determination*.

General recommendations

It is good practice to retain, for *individual* measurements, *more* significant figures than would result from calculations based on error propagation, and to use this law only for reporting the final value. This practice enables any interested person to perform whatever statistical calculations he desires on the individually reported measurements. Indeed, the results of statistical manipulations of data, when properly interpreted, are never affected by unnecessary significant figures in the data, but they may be seriously impaired by too much rounding.

The practice of reporting a measured value with a ± symbol followed by its standard error should be avoided at all costs, unless the meaning of the ± symbol is specifically and precisely stated. Some use the ± symbol to indicate a standard error of the value preceding the symbol, others to indicate a 95 percent confidence interval for the mean, others for the standard deviation of a single measurement, and still others use it for an uncertainty interval including an estimate of bias added to the 95 percent confidence interval. These alternatives are by no means exhaustive, and so far no standard practice has been adopted. It is of the utmost importance, therefore, to define the symbol whenever and wherever it is used.

It should also be borne in mind that the same measurement can have, and generally does have, more than one precision index, depending on the framework (statistical population) to which it is referred. For certain purposes, this population is the totality of (hypothetical) measurements that would be generated by repeating the measuring process over and over again on the same sample in the same laboratory. For other purposes, it would be the totality of results obtained by having the sample analyzed in a large number of laboratories. The reader is referred to the discussion in the section on precision and accuracy.

Tests of significance

General considerations

A considerable part of the published statistical literature deals with significance testing. Actually, the usefulness of the body of techniques classified under this title is far smaller than would be inferred from its prominence

in the literature. Moreover, there are numerous instances, both published and unpublished, of serious misinterpretations of these techniques. In many applications of significance testing, a "null-hypothesis" is formulated that consists of a statement that the observed experimental result—for example, the improvement resulting from the use of a drug compared to a placebo—is not "real," but simply the effect of chance. This null-hypothesis is then subjected to a statistical test and, if rejected, leads to the conclusion that the beneficial effect of the drug is "real," i.e., *not* due to chance. A closer examination of the nature of the null-hypothesis, however, raises some serious questions about the validity of the logical argument. In the drug-placebo comparison, the null-hypothesis is a statement of *equality of the means of two populations*, one referring to results obtained with the drug and the other with the placebo. All one infers from the significance test is a probability statement regarding the observed (sample) difference, on the hypothesis that the *true* difference between the population means is *zero*. The real question, of course, is related *not* to the *means* of hypothetical populations but rather to the benefit that any particular subject, selected at random from the relevant population of patients, may be expected to derive from the drug. Viewed from this angle, the usefulness of the significance test is heavily dependent on the *size* of the sample, i.e., on the number of subjects included in the experiment. This size will determine how large the difference between the two populations must be, *as compared to the spread of both populations*, before the statistical procedure will pick it up with a reasonable probability. Such calculations are known as the determination of the "power" of the statistical test of significance. Without indication of power, a test of significance may be very misleading.

Alternative hypotheses and sample size—the concept of power

An example of the use of "power" in statistical thinking is provided by our discussion in the section on sample sizes. Upon rereading this section, the reader will note that *two* situations were considered and that a probability value was associated with each of the two situations, namely, the probability of accepting or rejecting the lot. In order to satisfy these probability requirements, it was necessary to stipulate a value of N, the sample size. Smaller values of N would not have achieved the objectives expressed by the stipulated probabilities.

In testing a drug versus a placebo, one can similarly define two situations: (a) a situation in which the drug is hardly superior to the placebo; and (b) a situation in which the drug is definitely superior to the placebo. More specifically, consider a very large, *hypothetical* experiment in which subjects are paired at random, one subject of each pair receiving the placebo and the other the drug. Situation (a) might then be defined as that in which only 55 percent of all pairs shows better results with the drug than with the placebo; situation (b) might be defined as that in which 90 percent of the pairs shows greater effectiveness of the drug.

If we now perform an *actual* experiment, similar in nature but of moderate size, we must allow for random fluctuations in the percentage of pairs

that show better results with the drug as compared to the placebo. Therefore, our acceptance of the greater effectiveness of the drug on the basis of the data will involve risks of error. If the true situation is (a), we may wish to have only a small probability of declaring the drug superior, say, a probability of 10 percent. On the other hand, if the true situation is (b), we would want this probability to be perhaps as high as 90 percent. These two probabilities then allow us to calculate the required sample size for our experiment. Using this sample size, we will have assurance that the power of our experiment is sufficient to realize the stipulated probability requirements.

An example

An illustration of this class of problems is shown in Table 4.12. The data result from the comparison of two drugs, S (standard) and E (experimental), for the treatment of a severe pulmonary disease. The data represent the reduction in blood pressure in the heart after administration of the drug. The test most commonly used for such a comparison is Student's t test.[2–5] In the present case, the value found for t is 3.78, for DF = 142 (DF = number of degrees of freedom). The probability of obtaining a value of 3.78 or larger by pure chance (i.e., for equal efficacy of the two drugs) is less than 0.0002. The smallness of this probability is of course a strong indication that the hypothesis of equal efficacy of the two drugs is unacceptable. It is then generally concluded that the experiment has demonstrated the superior efficacy of E as compared to S. For example, the conclusion might take the form that "the odds favoring the effectiveness of E over S are better than M to 1" where M is a large number (greater than 100 in the present case). However, both the test and the conclusion are of little value for the solution of the real problem underlying this situation, as the following treatment shows. If we assume, as a first approximation, that the standard deviation 3.85 is the "population parameter" σ, and that the means, 0.10 for S and 2.53 for E, are also population parameters, then the probability of a single patient being better off

TABLE 4.12. TREATMENT OF PULMONARY EMBOLISM—COMPARISON OF TWO DRUGS

Decrease in Right Ventricular Diastolic Blood Pressure (mm Hg)		
	Standard treatment (S)	Experimental treatment (E)
Number of patients	68	76
Average	0.10	2.53
Standard deviation	3.15	4.28
Standard error of average	0.38	0.50
True mean	μ_1	μ_2

t test for H_o: $\mu_1 = \mu_2$, (H_o = null hypothesis)

$s_{pooled} = 3.85$ DF = 67 + 75 = 142, (DF = degrees of freedom)

$$t = \frac{2.53 - 0.10}{3.85\sqrt{\frac{1}{68} + \frac{1}{76}}} = 3.78 \,(P < 0.0002)$$

with E than with S is a function of the quantity Q defined by $Q \equiv (\mu_2 - \mu_1)/\sigma$. In the present case:

$$Q = \frac{2.53 - 0.10}{3.85} = 0.63$$

This can be readily understood by looking at Figure 4.3, in which the means of two populations, S and E, are less than one standard deviation apart, so that the curves show a great deal of overlap. There is no question that the two populations *are* distinct, and this is really all the *t* test shows. But due to the overlap, the probability is far from overwhelming that treatment E will be superior to treatment S for a randomly selected pair of individuals. It can be shown that this probability is that of a random normal deviate exceeding the value $\left(-\frac{Q}{\sqrt{2}}\right)$, or, in our case $\left(-\frac{0.63}{\sqrt{2}}\right) = -0.45$. This probability is 0.67, or about 2/3. Thus, in a large population of patients, two-thirds would derive more benefit from S than from E. Viewed from this perspective, the significance test, with its low "P value" (of 0.0002 in our case) is seen to be thoroughly misleading.

The proper treatment of a problem of this type is to raise the question of interest within a logical framework, derived from the nature of the problem, rather than perform standard tests of significance, which often merely provide correct answers to trivial questions.

Evaluation of diagnostic tests

The concepts of precision and accuracy are appropriate in the evaluation of tests that result in a quantitative measure, such as the glucose level of serum or the fluoride content of water. For medical purposes, different types of tests denoted as "diagnostic tests" are also of great importance. They dif-

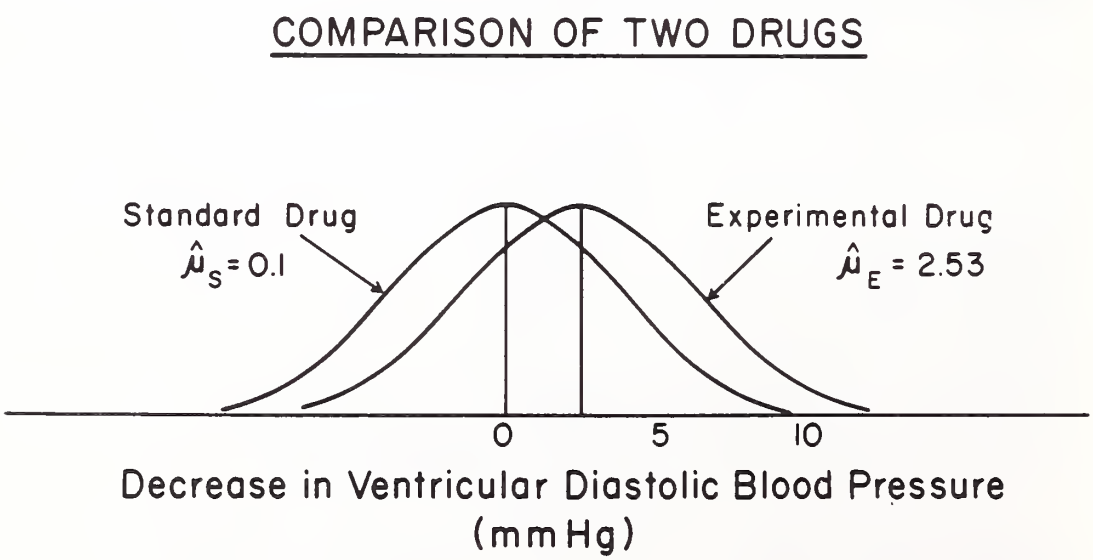


Fig. 4.3. Comparison of two drugs for the treatment of pulmonary disease, as measured by the reduction in right ventricular diastolic blood pressure (mm Hg).

fer from quantitative types of tests in that their outcome is characterized by a simple dichotomy into *positive* or *negative* cases.

As an example, consider Table 4.13, representing data on the alpha-fetoprotein (AFP) test for the diagnosis of hepatocellular carcinoma.[8] What do these data tell us about the value of the AFP test for the diagnosis of this disease?

Sensitivity and specificity

The statistical aspects of this type of problem are best understood by introducing a number of concepts that have been specifically developed for these problems.[8]

Sensitivity is the proportion of positive results among the subjects affected by the disease. Table 4.13 provides as an estimate of sensitivity:

$$\text{Sensitivity} = \frac{90}{107} = 0.8411 = 84.11\%$$

Specificity is the proportion of negative results among the subjects who are free of the disease. From Table 4.13:

$$\text{Specificity} = \frac{2079}{2118} = 0.9816 = 98.16\%$$

The concepts of sensitivity and specificity are useful descriptions of the nature of a diagnostic test, but they are not, in themselves, sufficient for providing the physician with the information required for a rational medical decision.

For example, suppose that a particular subject has a positive AFP test. What is the probability that this subject has hepatocarcinoma? From Table 4.13 we infer that among all subjects for whom the test is positive a proportion of 90/129, or 69.77 percent, are affected by the disease. This proportion is called the *predictive value of a positive test*, or *PV+*.

Predictive values—the concept of prevalence

Predictive value of a positive test.—(PV+) is defined as the proportion of subjects affected by the disease among those showing a positive test. The (PV+) value cannot be derived merely from the sensitivity and the specificity of the test. To demonstrate this, consider Table 4.14, which is fictitious and was derived from Table 4.13 by multiplying the values in the "Present"

TABLE 4.13. RESULTS OF ALPHA-FETOPROTEIN TESTS FOR DIAGNOSIS OF HEPATOCELLULAR CARCINOMA

	Hepatocarcinoma		
Test result	Present	Absent	Total
+	90	39	129
−	17	2,079	2,096
Total	107	2,118	2,225

TABLE 4.14. VALUES FOR ALPHA-FETOPROTEIN TESTS DERIVED FROM TABLE 4.13

Test Result	Hepatocarcinoma		Total
	Present	Absent	
+	900	39	939
−	170	2,079	2,249
Total	1,070	2,118	3,118

column by 10, and by leaving the values in the "Absent" column unchanged. Table 4.14 leads to the same sensitivity and specificity values as Table 4.13. However, the (PV+) value is now 900/939 = 95.85 percent.

It is seen that the (PV+) value depends not only on the sensitivity and the specificity but also on the *prevalence of the disease* in the total population. In Table 4.13, this prevalence is 107/2225 = 4.809 percent, whereas in Table 4.14 it is 1070/3118 = 34.32 percent.

A logical counterpart of the (PV+) value is the *predictive value of a negative test*, or *PV−*.

Predictive value of a negative test.—(PV−) is defined as the proportion of subjects free of the disease among those showing a negative test. For the data of Table 4.13, the (PV−) value is 2079/2096 = 99.19 percent, whereas for Table 4.14, (PV−) = 2079/2249 = 92.44 percent. As is the case for (PV+), the (PV−) value depends on the prevalence of the disease.

The following formulas relate (PV+) and (PV−) to sensitivity, specificity, and prevalence of the disease. We denote sensitivity by the symbol SE, specificity by SP, and prevalence by P; then:

$$(\text{PV}+) = \frac{1}{1 + \dfrac{(1 - SP)\,(1 - P)}{SE \cdot P}} \tag{4.62}$$

$$(\text{PV}-) = \frac{1}{1 + \dfrac{(1 - SE) \cdot P}{SP(1 - P)}} \tag{4.63}$$

As an illustration, the data in Table 4.13 yield:

$$(\text{PV}+) = \frac{1}{1 + \dfrac{(1 - 0.9816)\,(1 - 0.04809)}{(0.8411)\,(0.04809)}} = 0.6978 = 69.78\%$$

$$(\text{PV}-) = \frac{1}{1 + \dfrac{(1 - 0.8411)\,(0.04809)}{(0.9816)\,(1 - 0.04809)}} = 0.9919 = 99.19\%$$

Apart from rounding errors, these values agree with those found by direct inspection of the table.

Interpretation of multiple tests

The practical usefulness of (PV+) and (PV−) is now readily apparent. Suppose that a patient's result by the AFP test is positive and the prevalence

of the disease is 4.809 percent. Then the probability that the patient suffers from hepatocarcinoma is about 70 percent. On the basis of this result, the patient now belongs to a *subgroup* of the total population in which the prevalence of the disease is 70 percent rather than the 4.8 percent applying to the *total* population. Let us assume that a second test is available for the diagnosis of hepatocarcinoma, and that this second test is *independent* of the AFP test. The concept of *independence* of two diagnostic tests is crucial for the correct statistical treatment of this type of problem, but it seems to have received little attention in the literature. Essentially, it means that in the class of patients affected by the disease, the proportion of patients showing a positive result for test B is the same, whether test A was positive or negative. A similar situation must hold for the class of patients free of the disease.

In making inferences from this second test for the patient in question, we can start with a value of *prevalence of the disease* (P) of 70 percent, rather than 4.8 percent, since we know from the result of the AFP test that the patient belongs to the subgroup with this higher prevalence rate. As an illustration, let us assume that the second test has a sensitivity of 65 percent and a specificity of 90 percent and that the second test also is positive for this patient. Then the new (PV+) value is equal to

$$(\mathrm{PV}+) = \frac{1}{1 + \dfrac{(1 - 0.90)\,(1 - 0.70)}{(0.65)\,(0.70)}} = 0.938 = 93.8\%$$

If, on the other hand, the second test turned out to be negative, then the probability that the patient is free of disease would be:

$$(\mathrm{PV}-) = \frac{1}{1 + \dfrac{(1 - 0.65)\,(0.70)}{(0.90)\,(1 - 0.70)}} = 0.524 = 52.4\%$$

In that case, the two tests essentially would have contradicted each other, and no firm diagnosis could be made without further investigations.

A general formula for multiple independent tests

It can easily be shown that the order in which the independent tests are carried out has no effect on the final (PV+) or (PV−) value. In fact, the following general formula can be derived that covers any number of *independent* tests and their possible outcomes.

Denote by $(SE)_i$ and $(SP)_i$ the sensitivity and the specificity of the ith = test, where $i = 1, 2, 3, \ldots, N$. Furthermore, define the symbols A_i and B_i as follows:

$$A_i = \begin{cases} (SE)_i \text{ when the result of test } i \text{ is } + \\ 1 - (SE)_i \text{ when the result of test } i \text{ is } - \end{cases}$$

$$B_i = \begin{cases} 1 - (SP)_i \text{ when the result of test } i \text{ is } + \\ (SP)_i \text{ when the result of test } i \text{ is } - \end{cases}$$

If P is the prevalence rate of the disease *before* administration of any of the tests, and P' is the probability that the subject has the disease *after* administration of the N tests, then:

$$P' = \frac{1}{1 + \frac{(B_1 \cdot B_2 \cdot \ldots \cdot B_N)(1 - P)}{(A_1 \cdot A_2 \cdot \ldots \cdot A_N)P}} \tag{4.64}$$

It is important to keep in mind that Equation 4.64 is valid *only* if all tests are mutually independent in the sense defined above.

Quality Control

The remainder of this chapter deals with the fundamental principles of a quality control and quality assurance program for monitoring and assessing the precision and accuracy of the data being processed within a laboratory.

The definitions of Quality, Quality Assurance, and Quality Control by the American Society for Quality Control (ASQC)[9] apply to either a product or a service, and they are quoted here in their entirety.

1) *Quality.*—"The totality of features and characteristics of a product or service that bear on its ability to satisfy a given need."
2) *Quality assurance.*—"A system of activities whose purpose is to provide assurance that the overall quality-control job is in fact being done effectively. The system involves a continuing evaluation of the adequacy and effectiveness of the overall quality-control program with a view of having corrective measures initiated where necessary. For a specific product or service, this involves verifications, audits, and the evaluation of the quality factors that affect the specification, production, inspection, and use of the product or service."
3) *Quality control.*—"The overall system of activities whose purpose is to provide a quality of product or service that meets the needs of users; also, the use of such a system.

   "The aim of quality control is to provide quality that is satisfactory, adequate, dependable, and economic. The overall system involves integrating the quality aspects of several related steps, including the proper *specification* of what is wanted; *production* to meet the full intent of the specification; *inspection* to determine whether the resulting product or service is in accordance with the specification; and *review of usage* to provide for revision of specification.

   "The term *quality control* is often applied to specific phases in the overall system of activities, as, for example, *process quality control*."

The Control Chart

According to the ASQC,[9] the control chart is "a graphical chart with control limits and plotted values of some statistical measure for a series of samples or subgroups. A central line is commonly shown."

The results of a laboratory test are plotted on the vertical axis, in units of the test results, versus time, in hours, days, etc., plotted on the horizontal axis. Since each laboratory test should be checked at least once a day, the horizontal scale should be wide enough to cover a minimum of one month of data. The control chart should be considered as a tool to provide a "real-time" analysis and feedback for appropriate action. Thus, it should cover a sufficient period of time to provide sufficient data to study trends, "runs" above and below the central line, and any other manifestation of lack of randomness (see section on detection of lack of randomness).

Statistical basis for the control chart

General considerations

W. A. Shewhart, in his pioneering work in 1939,[10] developed the principles of the control chart. They can be summarized, as was done by E. I. Grant,[11] as follows: "The measured quantity of a manufactured product is always subject to a certain amount of variation as a result of chance. Some stable 'System of Chance Causes' is inherent in any particular scheme of production and inspection. Variation within this stable pattern is inevitable. The reasons for variation outside this stable pattern may be discovered and corrected." If the words "manufactured product" are changed to "laboratory test," the above statement is directly applicable to the content of this section.

We can think of the "measured quantity" as the concentration of a particular constituent in a patient's sample (for example, the glucose content of a patient's serum). Under the "system of chance causes," this concentration, when measured many times under the same conditions, will fluctuate in such a way as to generate a statistical distribution that can be represented by a mathematical expression. This expression could be the normal distribution, for those continuous variables that are symmetrically distributed about the mean value, or it could be some other suitable mathematical function applicable to asymmetrically or discretely distributed variables (see section on non-normal distributions). Then, applying the known principles of probability, one can find lower and upper limits, known as *control limits*, that will define the limits of variation within "this stable pattern" for a given acceptable tolerance probability. Values outside these control limits will be considered "unusual," and an investigation may be initiated to ascertain the reasons for this occurrence.

Control limits

According to the ASQC,[9] the control limits are the "limits on a control chart that are used as criteria for action or for judging whether a set of data does or does not indicate lack of control."

Probability limits.—If the distribution of the measured quantity is known, then lower and upper limits can be found so that, on the average, a predetermined percentage of the values (e.g., 95 percent, 99 percent) will fall within these limits if the process is under control. The limits will depend on the nature of the probability distribution. They will differ, depending on

whether the distribution of the measured quantity is symmetric, asymmetric to the left or to the right, unimodal or bimodal, discrete or continuous, etc.

The obvious difficulty of finding the correct distribution function for each measured quantity, and of determining the control limits for this distribution, necessitates the use of procedures that are not overly sensitive to the nature of the distribution function.

Three-sigma limits.—The three-sigma limits, most commonly used in industrial practice, are based on the following expression:

Control limits = Average of the measured quantity ± three standard deviations of the measured quantity

The "measured quantity" could be the mean of two or three replicate determinations for a particular chemical test, the range of a set of replicate tests, a proportion defective, a radioactive count, etc.

The range of three standard deviations around the mean, that is, a width of six standard deviations, usually covers a large percentage of the distribution. For normally distributed variables, this range covers 99.7 percent of the distribution (see section on the normal distribution). For non-normally distributed variables, an indication of the percentage coverage can be obtained by the use of two well-known inequalities:

1) *Tchebycheff's Inequality.* For *any* distribution, (discrete or continuous, symmetric or asymmetric, unimodal or bimodal, etc.) with a finite standard deviation, the interval mean ± $K\sigma$ covers a proportion of the population of at least $1 - \frac{1}{K^2}$. Thus for $K = 3$, the coverage will be at least $1 - \frac{1}{9} = \frac{8}{9}$, or roughly 90 percent of the distribution.

2) *Camp-Meidel Inequality.* If the distribution is unimodal, the interval mean ± $K\sigma$ will cover a proportion of at least $1 - \frac{1}{2.25K^2}$ of the population. Thus, for $K = 3$, the coverage will be at least $1 - \frac{1}{20.25}$ or roughly 95 percent of the population.

From the above discussion, it follows that the three-sigma limits cover a proportion of the population that is at least equal to 90 percent for non-normal distributions and is equal to exactly 99.7 percent when the distribution is normal.

Most control charts are based on the mean of several determinations of the same measured equality. By the Central Limit Theorem, (see section on the normal distribution), the larger the sample size, the closer to normality will be the mean of this measured quantity. However, since most clinical tests are based on single or, at best, duplicate determinations, caution should be used in interpreting the amount of coverage given by the control limits for those distributions that are suspected to be skewed, bimodal, etc.

Warning limits.—The warning limits commonly used in practice are defined as:

Warning limits = Average of the measured quantity ± two standard deviations of the measured quantity.

For interpretation of points falling outside the warning and control limits, see the section on the control chart as a management tool.

Variability between and within subgroups

The hypothesis $\sigma_B = 0$

In control charts for variables, the variability is partitioned into two components: within and between subgroups. To this effect, the sequence of measurements is divided into subgroups of n consecutive values each. The variability within subgroups is estimated by first computing the average of the ranges of all subgroups and dividing this average by a factor that depends on n, which can be found in standard statistical tables. As an example, consider the sequence: 10.2, 10.4, 10.1, 10.7, 10.3, 10.3, 10.5, 10.4, 10.0, 9.8, 10.4, 10.9. When divided into subgroups of four, we obtain the arrangement:

Subgroup	Average	Range
10.2, 10.4, 10.1, 10.7	10.350	0.6
10.3, 10.3, 10.5, 10.4	10.375	0.2
10.0, 9.8, 10.4, 10.9	10.275	1.1
Average	10.333	0.63

In this case $n = 4$, and the average range is $\bar{R} = 0.63$.

Generally, n is a small number, often between 2 and 5. Its choice is sometimes arbitrary, dictated only by statistical convenience. More often, and preferably, the choice of n is dictated by the way in which the data were obtained. In the example above, the data may actually consist of three samples, each measured four times. In this case, "within groups" means "within samples," and "between groups" means "between samples."

Another possibility is that there were actually 12 samples, but that the measuring technique requires that they be tested in groups of four. If that is the situation, the relation of between-group to within-group variability depends not only on the sample-to-sample variability but also on the stability of the measuring instrument or technique from one group of four to another group of four. The location of the control limit and the interpretation of the control chart will depend on the nature and the choice of the subgroup.

If the standard deviation *within* subgroups is denoted by σ_W, and the standard deviation *between* subgroups by σ_B, a control chart is sometimes, but by no means always, a test as to whether σ_B exists (is different from zero). If $\sigma_B = 0$, then the variation between the *averages* of subgroups can be *predicted* from σ_W (or, approximately, from $\bar{R}$). The hypothesis $\sigma_B = 0$ can be tested by observing whether the subgroup averages stay within the control limits calculated on the basis of within-subgroup variability. Failure of this event to occur indicates the presence of causes of variability between subgroups. The nature of these causes depends on the criteria used in the selection of the subgroups.

The case $\sigma_B \neq 0$. Baseline data

In many applications, the hypothesis $\sigma_B = 0$ is not justified by the physical reality underlying the data. It may, for example, already be known that the subgroups vary from each other by more than can be accounted for by within-subgroup variability. Thus, each subgroup may represent a different

day of testing, and there may be more variability between days than within days. The initial set of data (*baseline* data) is then used primarily to estimate both the within- and the between-components of variability, and control limits are calculated on the basis of both these components (see section on computation of control limits). Data that are obtained subsequent to the baseline period are then evaluated in terms of these control lines. From time to time, the control lines are recalculated using *all* the data obtained up to that time, eliminating, however, those data for which abnormal causes of variability were found.

Types of control charts

Depending on the characteristics of the measured quantity, control charts can be classified into three main groups:

1) Control charts for variables (the $\bar{X}$, R Chart). These are used for variables such as clinical chemical determinations, some hematological parameters, etc.
2) Control charts for attributes (the P-Chart). These are used for proportion defective, proportion of occurrence of given disease, etc.
3) Control charts for number of defects per unit (the C-Chart). These may be used for counts, such as the number of cells observed in a given area, radioactive counts, etc.

Preparing a control chart

Objective and choice of variable

The general objectives of a control chart are: (a) to obtain initial estimates for the key parameters, particularly means and standard deviations. These are used to compute the central lines and the control lines for the control charts; (b) to ascertain *when* these parameters have undergone a radical change, either for worse or for better. In the former case, modifications in the control process are indicated; and (c) to determine *when* to look for assignable causes of unusual variations so as to take the necessary steps to correct them or, alternatively, to establish when the process should be left alone.

A daily review of the control chart should indicate whether the resulting product or service is in accordance with specifications. For example, in clinical chemistry, if a control chart based on standard samples shows statistical control for the measurement of a given constituent, then one can proceed with confidence with the determination of this constituent in patient samples. If the chart shows lack of control, an investigation should be started immediately to ascertain the reasons for this irregularity.

No general recommendations can be made here about the types of variables to use for quality control purposes, since they will obviously vary according to the various disciplines of the laboratory. Considerations of this type will be found in the respective specialty chapters of this book. The same statements apply to the types of stable pools or reagents that should be

used, and to the methods of handling these materials in normal laboratory practice.

Selecting a rational subgroup

The generally recommended approach for the selection of a subgroup of data for control purposes (using a single pool of homogeneous material) is that conditions *within* subgroups should be as uniform as possible (same instrument, same reagents, etc.), so if some assignable causes of error are present, they will show up *between* subgroups (see Duncan,[12] p. 347, and Grant,[11] Ch. 6, for further discussions).

When tests on patient samples are performed at regular intervals using standard laboratory equipment, the subgroup becomes automatically defined, since control samples are, or should be, included in each run. Otherwise, tests on control samples should be run at regular intervals during the day in order to detect possible changes in environmental conditions, reagents, calibrations, technicians, etc.

Size and frequency of control sample analyses

A minimum of two replicates should be obtained in each run of the control sample. To account for the possible effects of carryover from other samples, and to have a better indication of the capability of an instrument to reproduce itself under normal conditions within a run, the replicate samples should not be tested back-to-back, but should be separated by patient samples.

As indicated before, the frequency of the runs on control materials is generally tied to the frequency of the tests on patient samples. One general rule is to test the control samples as frequently as possible at the beginning of a control procedure, and to reduce this frequency to a minimum of two or three per day when the results of the control chart show a satisfactory state of control.

Maintaining uniform conditions in laboratory practice

A properly prepared control chart will tend to reflect any change in the precision and accuracy of the results obtained. To avoid wasting time in hunting for unnecessary sources of trouble, care should be taken to maintain laboratory conditions and practices as uniform as possible. These include sampling procedures, dilution techniques, aliquoting methods, storage methods, instrumental techniques, calculating procedures, etc.

Initiating a control chart

When meaningful historical data are not available (as is often the case when a quality control procedure is to be initiated), a plan should be set up to collect a minimum amount of data for each variable to be controlled during an initial ***baseline period***.

For a control chart for *variables*, with a minimum of two replicates for each run, data should be collected for a baseline period of at least one month in order to allow sufficient time for the estimation of day-to-day variability.

Means and ranges should be computed for each run and plotted on separate charts. Records should be accurately kept, using standard quality control (QC) forms that are readily available. Any value that appears to be the result of a blunder should be eliminated, and the source of the blunder carefully noted. It is recommended that the number of runs or subgroups be at least 25 for the baseline period.

The same considerations apply to control charts of *proportions and counts*, except that the number of observations for each subgroup is generally larger than the corresponding number used in a control chart of variables. Statistical procedures for determining the sample size, n, for the P-chart or the C-chart can be found in the literature (see Duncan,[12] pp. 345 and 361). In general, n should be large enough to provide a good chance of finding one or more defectives in the sample.

Determining trial control limits

Based on the initial set of data collected during the baseline period, trial control limits can be determined using the procedure outlined in the section on random samples. After plotting these limits in the initial control chart (see section on the case $\sigma_B \neq 0$), points that are outside or very near the limits should be carefully examined, and if some valid reasons are found for their erratic behavior, they should be eliminated and new control limits should be computed. In general, it is better to start with control limits that are relatively narrow in order to better detect future trends, shifts in mean values, and some other types of irregularities. A common experience is that some initial subgroups of data will not be under control but, in general, after some knowledge is gained in the use of the control chart, the process will tend to reach a state of statistical equilibrium. After this time period, one generally has an adequate amount of data to produce realistic estimates of the mean and standard deviations.

Computing control limits

Two variable control charts should be kept, one for the *average value*, and the other for the *range* of individual determinations in each subgroup. In all cases in which a non-zero component for between-subgroups is known to exist, the control limits for the chart of averages will be based on the "total" standard deviation for subgroup averages.

If the subgroups are of size n, and if $\hat{\sigma}^2_W$ and $\hat{\sigma}^2_B$ represent the estimated *components* of variance within subgroups and between subgroups, respectively, then the "total standard deviation" for the averages of subgroups is

$$\hat{\sigma}_T = \sqrt{\hat{\sigma}^2_B + \frac{\hat{\sigma}^2_W}{n}} \tag{4.65}$$

This quantity can also be obtained by directly calculating the standard deviation of the subgroup averages in the baseline period.

The control chart of the ranges will be used to ascertain whether the variability among individual readings within subgroups is consistent from

subgroup to subgroup. The limits for this chart will be based on the within-subgroup standard deviation.

Calculating the standard deviation

Using the available data for k subgroups, each of size n, we will have the layout shown in Table 4.15. The standard deviation within subgroups can be estimated from

$$S_W \approx \frac{\bar{R}}{d_2} \tag{4.66}$$

where

$$\bar{R} = \frac{\Sigma R_i}{k} \tag{4.67}$$

and the value of d_2 can be obtained from standard control chart tables (see Duncan,[12] p. 927). Values of d_2 for typical sample sizes are given in the following table:

n	d_2
2	1.128
3	1.693
4	2.059
5	2.326

The value of s_W can be accurately determined by pooling the variances from each subgroup (see section on precision and accuracy). However, the above estimate, based on the average range, is sufficiently accurate if the number of subgroups is large enough (say, 25 or more).

The standard deviation of the k sample averages is:

$$S_{\bar{x}} = \sqrt{\frac{\Sigma(\bar{X}_i - \bar{\bar{X}})^2}{k - 1}} \tag{4.68}$$

The between-subgroups standard deviation is given by:

$$S_B = \sqrt{S_{\bar{X}}^2 - \frac{S_W^2}{n}} \tag{4.69}$$

TABLE 4.15. LAYOUT FOR $\bar{X}$, R CONTROL CHARTS

Subgroup	Determinations	Mean	Range
1	$X_{11}, X_{12}, \ldots, X_{1n}$	$\bar{X}_1$	R_1
2	$X_{21}, X_{22}, \ldots, X_{2n}$	$\bar{X}_2$	R_2
3	$X_{31}, X_{32}, \ldots, X_{3n}$	$\bar{X}_3$	R_3
•	•	•	•
•	•	•	•
k	$X_{k1}, X_{k2}, \ldots, X_{kn}$	$\bar{X}_k$	R_k
		$\bar{\bar{X}}$	$\bar{R}$

and the total standard deviation for individual determinations is:

$$S_{T_1} = \sqrt{S_B^2 + S_W^2} \tag{4.70}$$

The total standard deviation for averages of n daily determinations is:

$$S_{T_n} = \sqrt{S_B^2 + \frac{S_W^2}{n}} \tag{4.71}$$

Note that S_{T_n} is identically equal to $S_{\bar{x}}$.

Control limits for the chart of averages

The control limits for the chart of averages are given by:

$$UCL_{\bar{x}} = \bar{\bar{x}} + 3S_{\bar{x}} \tag{4.72}$$

and

$$LCL_{\bar{x}} = \bar{\bar{x}} - 3S_{\bar{x}} \tag{4.73}$$

where UCL = upper control limit; LCL = lower control limit.

The warning limits are:

$$UWL_{\bar{x}} = \bar{\bar{x}} + 2S_{\bar{x}} \tag{4.74}$$

and

$$LWL_{\bar{x}} = \bar{\bar{x}} - 2S_{\bar{x}} \tag{4.75}$$

where UWL = upper warning limit; LWL = lower warning limit.

Control limits for the chart of ranges

Based on the three-sigma limits concept (see section on control limits), the control limits for the chart of ranges are given by $\bar{R} \pm 3\sigma_R$. Using standard control chart notation, these limits are:

$$UCL_R = D_4\bar{R} \tag{4.76}$$

and

$$LCL_R = D_3\bar{R} \tag{4.77}$$

where

$$D_4 = 1 + 3\frac{d_3}{d_2} \tag{4.78}$$

and

$$D_3 = 1 - 3\frac{d_3}{d_2} \tag{4.79}$$

and the values of d_2, d_3, D_3, and D_4 are given in Natrella[3] and Duncan.[12] For $n = 2$, those values are $D_4 = 3.267$, and $D_3 = 0$.

The warning limits for $n = 2$ are:

$$UWL_R = 2.512\,\bar{R}$$

$$LWL_R = 0$$

The numerical value 2.512 is obtained as follows:

$$2.512 = 1 + 2\frac{d_3}{d_2} = 1 + 2\,\frac{(0.853)}{1.128}$$

Examples of average and range ($\bar{X}$ and R) charts

Initial data

The data in Table 4.16 represent 25 daily, duplicate determinations of a cholesterol control, run on a single-channel Autoanalyzer I, 40 per hour. It may appear strange that all 50 values are even. This is due to a stipulation in the protocol that the measured values be read to the nearest even number. The data cover a period of two months, with the analyzer run at a frequency

TABLE 4.16. EXAMPLE OF $\bar{X}$, R CHART: CHOLESTEROL CONTROL RUN

Day	Run 1 X_{i1}	Run 2 X_{i2}	Mean $\bar{X}_i$	Range R_i
1	390	392	391	2
2	392	388	390	4
3	392	388	390	4
4	388	388	388	0
5	378	396	387	18
6	392	392	392	0
7	392	390	391	2
8	398	402	400	4
9	404	406	405	2
10	400	400	400	0
11	402	402	402	0
12	392	406	399	14
13	398	396	397	2
14	380	400	390	20
15	398	402	400	4
16	388	386	387	2
17	402	392	397	10
18	386	390	388	4
19	386	382	384	4
20	390	386	388	4
21	396	390	393	6
22	396	394	395	2
23	384	388	386	4
24	388	382	385	6
25	386	384	385	2
			$\Sigma = 9{,}810$	120

of three days per week. The two daily determinations were randomly located within patient samples. The control consisted of 0.5 ml of sample extracted with 9.5 ml of 99 percent reagent-grade isopropyl alcohol.

Computing trial control limits

From the data in Table 4.16:

$$\bar{\bar{X}} = 9810/25 = 392.4$$

$$\bar{R} = 120/25 = 4.8$$

$$S_{\bar{x}}^2 = \left[\Sigma \bar{X}_i^2 - (\Sigma \bar{X}_i)^2/n\right]/(n-1) = \left[3850320 - (9810)^2/25\right]/24 = 36.5$$

$$S_{\bar{x}} = \sqrt{36.5} = 6.04$$

The control limits for $\bar{X}$ can be computed:

$$UCL_{\bar{x}} = 392.4 + 3(6.04) = 410.5$$

$$LCL_{\bar{x}} = 392.4 - 3(6.04) = 374.3$$

The warning limits for $\bar{X}$ are:

$$UWL_{\bar{x}} = 392.4 + 2(6.04) = 404.5$$

$$LWL_{\bar{x}} = 392.4 - 2(6.04) = 380.3$$

The control limits for R are:

$$UCL_R = (3.367)(4.8) = 15.7$$

$$LCL_R = 0$$

The warning limits of R are:

$$UWL_R = (2.512)(4.8) = 12.1$$

Analysis of data

In Figures 4.4 and 4.5, a graphical representation is shown of the control charts for the mean and range of the daily runs, together with their appropriate control limits.

The means of the daily runs appear to be under control. Only one point, day 9, is above the warning limit, and all points appear to be randomly located around the central line.

The control chart of the range shows two points out of control, days 5 and 14, and one point, day 12, on the upper warning limit.

Let us assume, for the purpose of illustration, that a satisfactory reason was found for those two points to be out of control in the range chart, and that it was decided to recompute new limits for both the $\bar{X}$ and the R charts based on only 23 days of data.

The new values are: $\bar{\bar{X}} = 392.7$, $\bar{R} = 3.57$, $S_{\bar{x}} = 6.17$, and $n = 23$.

$$UCL_{\bar{x}} = 392.7 + 3(6.17) = 411.2; UWL_{\bar{x}} = 405.0$$

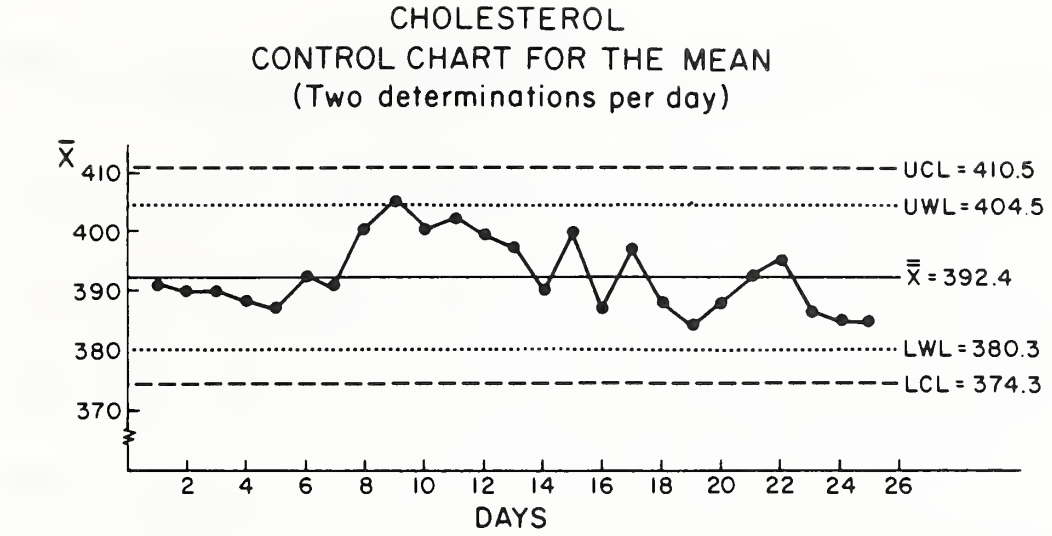


Fig. 4.4. Control chart for the mean, based on 25 daily duplicate determinations of a cholesterol control.

$$LCL_{\bar{x}} = 392.7 - 3(6.17) - 374.2; LWL_{\bar{x}} = 380.4$$

The new limits for the $\bar{X}$ chart are practically the same as the previous limits.

$$UCL_R = (3.267)(3.57) = 11.7; UWL_R = 9.0$$

$$LCL_R = 0 \qquad LWL_R = 0$$

These values establish the final limits, based on the baseline period.

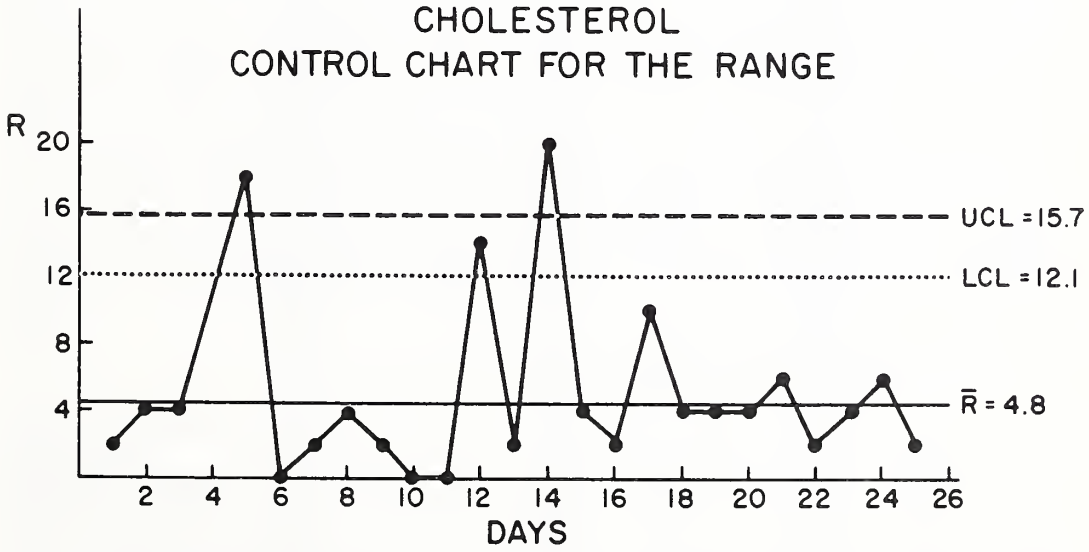


Fig. 4.5. Control chart for the range, based on 25 daily duplicate determinations of a cholesterol control.

Additional data

Nineteen additional points were obtained for days 26 to 44, running through a period of about one-and-a-half months. The values are shown in Table 4.17.

Figures 4.6 and 4.7 show the results of the 19 additional data points plotted against the (corrected) control limits based on the baseline period.

The $\bar{X}$-chart shows two points, days 38 and 39, out of control, about 40 percent of the points near the warning limits, and a definite trend toward large values of $\bar{X}$ after day 30. There is a run of seven points above the central line after day 37 and, in fact, if one considers day 37 to be "above" the central line (the mean of day 37 is 392), the run of points above the central line is of length 12. As indicated in the section on control limits, these considerations are indications of a process out of control.

The R-chart shows one point out of control and two points above the upper warning limit; although the value of $\bar{R}$ based on the 19 additional values, 4.32, is larger than the previous value, $\bar{R} = 3.57$, the difference is not significant.

The new set of points taken by itself produced the following values: $\bar{\bar{X}} = 396.5$, $\bar{R} = 4.32$, and $S_{\bar{x}} = 12.17$, where $n = 19$.

Future control limits

It is generally desirable to have a well-established baseline set so future points can be evaluated with confidence in terms of the baseline central line

TABLE 4.17. ADDITIONAL VALUES FOR CHOLESTEROL CONTROL RUN

Day	Run 1 X_{i1}	Run 2 X_{i2}	Mean $\bar{X}_i$	Range R_i
26	392	392	395	6
27	376	376	376	0
28	390	386	388	4
29	394	384	389	10
30	382	378	380	4
31	384	382	381	2
32	384	388	386	4
33	402	392	397	10
34	390	398	394	8
35	402	402	402	0
36	398	394	396	4
37	390	394	392	4
38	426	428	427	2
39	414	428	421	14
40	402	398	400	4
41	402	400	401	2
42	402	404	403	2
43	400	402	401	2
44	404	404	404	0
			$\Sigma = 7{,}533$	82

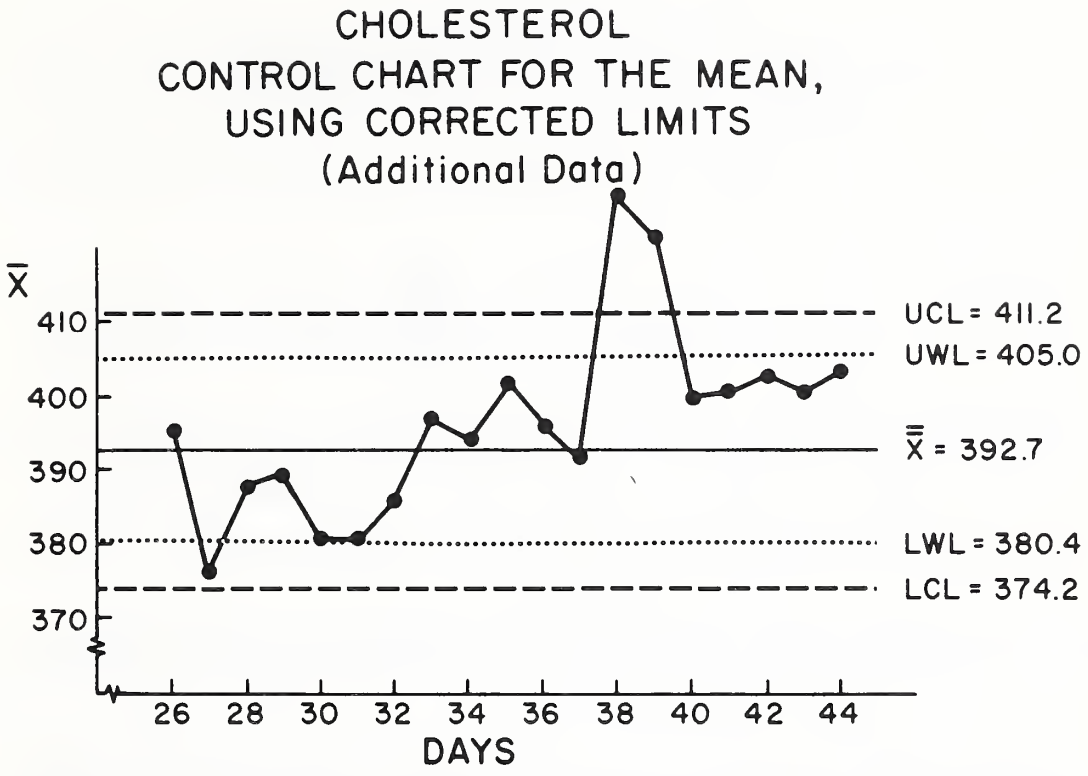


Fig. 4.6. Control chart for the mean, based on 19 additional data points, plotted against the corrected control limits.

and control limits. If, in the example under discussion, the additional set (days 26 to 44) was found to be satisfactorily consistent with the baseline data, then it would be proper to extend the baseline period by this set, i.e., a total of 25 + 19 = 44 points. However, we have already observed a number of shortcomings in the additional set, and the proper action is to search for the causes of these disturbances, i.e., "to bring the process under control." This is of course not a statistical problem.

For the purpose of our discussion, we will assume that an examination of the testing process has revealed faulty procedure starting with day 37. Therefore, we will consider a shortened additional set, of days 26 through 36. The following table gives a comparison of the baseline set (corrected to 23 points as discussed previously) and the shortened additional set (11 points).

	Baseline Set	Additional Set
Number of points, N	23	11
Average, $\bar{\bar{X}}$	392.7	389.5
Average Range, $\bar{R}$	3.57	4.73
Standard deviation, $s_{\bar{x}}$	6.17	8.15

By using the F test, [2-5] it is easily verified that the difference between the two standard deviations is well within the sampling variability that may be expected from estimates derived from samples of 23 and 11 points, respec-

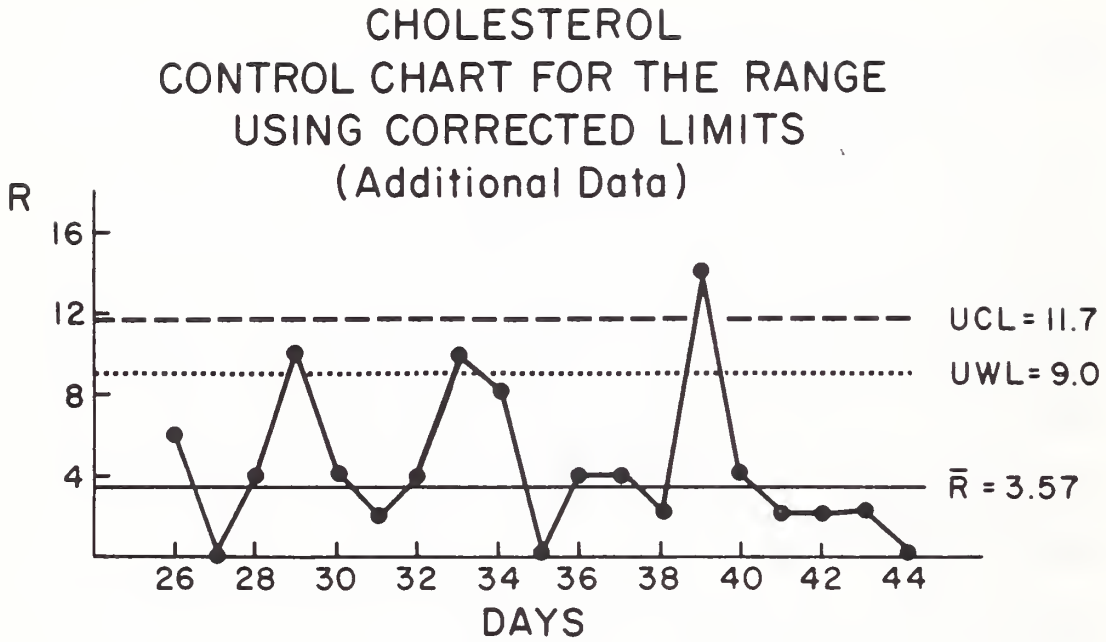


Fig. 4.7. Control chart for the range, based on 19 additional data points, plotted against the corrected control limits.

tively. The difference between the averages, $\bar{\bar{X}}$, is 392.7 − 389.5 = 3.2. A rough test can be made to see whether this difference indicates a real shift between the two sets. The standard error of the difference is approximately $[(6.17)^2/23 + (8.15)^2/11]^{\frac{1}{2}} = 2.77$. Thus the difference, 3.2, is equal to $\frac{3.2}{2.77}$ = 1.15 standard errors, and this is well within sampling errors.

It is therefore not unreasonable in this case to combine the 34 points of both sets to construct a new baseline. This results in the following parameters: $N = 34$, $\bar{\bar{X}} = 391.7$, $\bar{R} = 3.95$, and $s_{\bar{x}} = 6.93$.

The new control limits are:

For $\bar{X}$:	$UCL = 412.5$	$UWL = 405.6$
	$LCL = 370.9$	$LWL = 377.8$
For R:	$UCL = 12.9$	$UWL = 9.9$
	$LCL = 0$	$LWL = 0$

Using these new parameters, it can be noted that the points corresponding to days 37 through 44 may indicate a potential source of trouble in the measuring process.

Control chart for individual determinations

It is possible, although not recommended, to construct charts for individual readings. Extreme caution should be used in the interpretation of points out of control for this type of chart, since individual variations may not follow a normal distribution. When a distribution is fairly skewed, then a transformation (see section on transformation of scale) would be applied before the chart is constructed.

The steps to follow are:

1) Use a moving range of two successive determinations;
2) Compute $\bar{R} = \frac{\Sigma R_i}{k}$;
3) Determine the control limits for $\bar{X}$:

$$\bar{X} \pm 3\frac{\bar{R}}{d_2}$$

For $n = 2$, $d_2 = 1.128$, and hence the control limits are:

$$\bar{X} \pm 2.66\,\bar{R}$$

4) The upper control limit for R is $D_4\bar{R} = 3.267\,\bar{R}$. The lower control limit is equal to zero.

Other types of control charts

Control charts can also be constructed based on the average standard deviation, $\bar{\sigma}$, of several subgroups of sample data, or on "standard" values of σ, called σ' in the quality control literature (See Duncan,[12] Chap. 20).

Control chart for attributes—the P-chart

The fraction defective chart is generally used for quality characteristics that are considered attributes and are not necessarily quantitative in nature. To use this chart, it is only necessary to count the number of entities that have a well-defined property, such as being defective, have a certain type of disease, or have a glucose content greater than a given value, and translate this number into a proportion. The data used in this chart are easy to handle, and the cost of collection is normally not very high. In some instances, the P-chart can do the job of several average and range charts, since the classification of a "defective" element may depend on several quantitative characteristics, each of which would require an individual set of average and range charts for analysis.

The sample size for each subgroup will depend on the value of the proportion P being estimated. A small value of P will require a fairly large sample size in order to have a reasonable probability of finding one or more "defectives" in the sample (See Duncan[12]). In general, a value of n between 25 and 30 is considered adequate for the calculation of a sample proportion.

Control limits and warning limits

Since the standard deviation of a proportion is directly related to the value of the proportion, an estimate p of P is all that is needed for the calculation of the central line and of the control limits.

The central line is located at the value $\bar{p}$. The three-sigma control limits are:

$$UCL_p = \bar{p} + 3\sqrt{\frac{\bar{p}\,\bar{q}}{n}} \qquad (4.80)$$

$$LCL_p = \bar{p} - 3\sqrt{\frac{\bar{p}\,\bar{q}}{n}} \tag{4.81}$$

where $\bar{q} = 1 - \bar{p}$. The estimate $\bar{p}$ is obtained as follows:

Let the data be represented by the table:

Sample Number	Size	Number of Elements Having a Certain Characteristic	Proportion
1	n	X_1	p_1
2	n	X_2	p_2
3	n	X_3	p_3
.	.	.	.
.	.	.	.
.	.	.	.
k	n	X_n	p_k
	Total	ΣX_i	Σp_i

where $p_i = \frac{X_i}{n}$

Average proportion:

$$\bar{p} = \frac{\Sigma p_i}{k} \tag{4.82}$$

The warning limits are:

$$UWL_p = \bar{p} + 2\sqrt{\frac{\bar{p}\,\bar{q}}{n}} \tag{4.83}$$

$$LWL_p = \bar{p} - 2\sqrt{\frac{\bar{p}\,\bar{q}}{n}} \tag{4.84}$$

When the sample size does not remain constant from subgroup to subgroup, the recommended procedure is to compute control limits using the average sample size. However, when a point falls near the control limits thus calculated, then the actual limits for this point, using its own sample size, should be estimated before a conclusion is reached about its state of control.

Control charts for number of defects per unit—the C-chart

In some instances, it is more convenient to maintain control charts for the number of defects per unit, where the unit may be a single article or a subgroup of a given size. The "number of defects" may be, for instance, the number of tumor cells in an area of a specified size, the number of radioactive counts in a specified period of time, etc. In all these instances, the probability of occurrence of a single event (e.g., an individual defect) is very

small, but the unit is large enough to make the average number of occurrences (number of defects) a measurable number.

The Poisson distribution

It can be shown that, when the probability P of an event is very small but the sample size n is large, then the distribution of the number of occurrences c of this event tends to follow a Poisson distribution with parameter $nP = c'$. The mean and standard deviation of c are:

$$E(c) = c' \tag{4.85}$$

$$\sigma_c = \sqrt{c'} \tag{4.86}$$

The random variable c represents the number of defects per unit, the number of radioactive counts in a given period of time, the number of bacteria in a specified volume of liquid, etc.

Control limits.—The upper and lower limits are given by:

$$UCL_c = \bar{c} + 3\sqrt{\bar{c}} \tag{4.87}$$

$$LCL_c = \bar{c} - 3\sqrt{\bar{c}} \tag{4.88}$$

Here $\bar{c}$ is the average number of defects, or counts, obtained using a sufficiently large number, k, of units. $\bar{c}$ is a sample estimate of the unknown, or theoretical value c'.

The warning limits are:

$$UWL_c = \bar{c} + 2\sqrt{\bar{c}} \tag{4.89}$$

$$LWL_c = \bar{c} - 2\sqrt{\bar{c}} \tag{4.90}$$

Detecting lack of randomness

If a process is in a state of statistical control, the observations plotted in the control chart should randomly fall above and below the central line, with most of them falling within the control limits. However, even if all the points fall within the upper and lower control limits, there might still exist patterns of nonrandomness that require action, lest they lead eventually to points outside the control limits. Procedures for detecting such patterns will be discussed.

Rules based on the theory of runs

The most frequent test used to detect a lack of randomness is based on the theory of runs. A run may be defined as a succession of observations of the same type. The length of a run is the number of observations in a given run. For example, if the observations are classified as a or b, depending on whether they fall above or below the mean, then one set of observations may look like:

a a a b a b b b a a b

Here we have six runs, of length 3, 1, 1, 3, 2, 1, respectively.

Another criterion for the definition of a run would be the property of increase or decrease of successive observations. Such runs are called "runs *up* and *down*." For example, the sequence 2, 1.7, 2.2, 2.5, 2.8, 2.0, 1.8, 2.6, 2.5, has three runs *down* and two runs *up*. In order of occurrence, the *lengths* of the runs are 1, 3, 2, 1, 1.

Returning to runs above and below the central value, it is possible through use of the theory of probability, and assuming that the probability is one-half that an observation will fall above the central line (and, consequently, one-half that it will fall below the central line), to determine the probability distribution of the lengths of runs. Tables are available for several of these distributions (See Duncan,[12] Chap. 6). Some rules of thumb based on the theory of runs that are very useful in pointing out some lack of randomness are:

1) A run of length 7 or more. This run may be up or down, above or below the central line in the control chart. (For runs above or below the median, the probability of a run of length 7 is 0.015.)
2) A run of two or three points outside the warning limits.
3) Ten out of 11 successive points on the same side of the central line.

Distribution of points around the central line

When a sufficient number of observations is available, the pattern of distribution of points around the central line should be carefully examined. In particular, if the points tend to cluster near the warning or control limits, or if they show characteristics of bimodality, or if they show a pronounced skewness either to the left or the right, then the assumption of normality will not be satisfied and some transformation of scale may be necessary.

Interpreting patterns of variation in a control chart

Indication of lack of control

A process is out of control when one or more points falls outside the control limits of either the $\bar{x}$ or the R-chart, for control of variables, or outside the limits of the P-chart, for control of attributes.

Points outside the control limits of the R-chart tend to indicate an increase in magnitude of the within-group standard deviation. An increase in variability may be an indication of a faulty instrument which eventually may cause a point to be out of control in the $\bar{x}$-chart.

When two or more points are in the vicinity of the warning limits, more tests should be performed on the control samples to detect any possible reasons for out-of-control conditions.

Various rules are available in the literature about the procedures to follow when control values are outside the limits (see, for example, Haven[13]).

Patterns of variation

By examining the $\bar{x}$ and R-charts over a sufficient period of time, it may be possible to characterize some patterns that will be worth investigating in order to eliminate sources of future troubles.

Some of these patterns are shown in Figure 4.8.

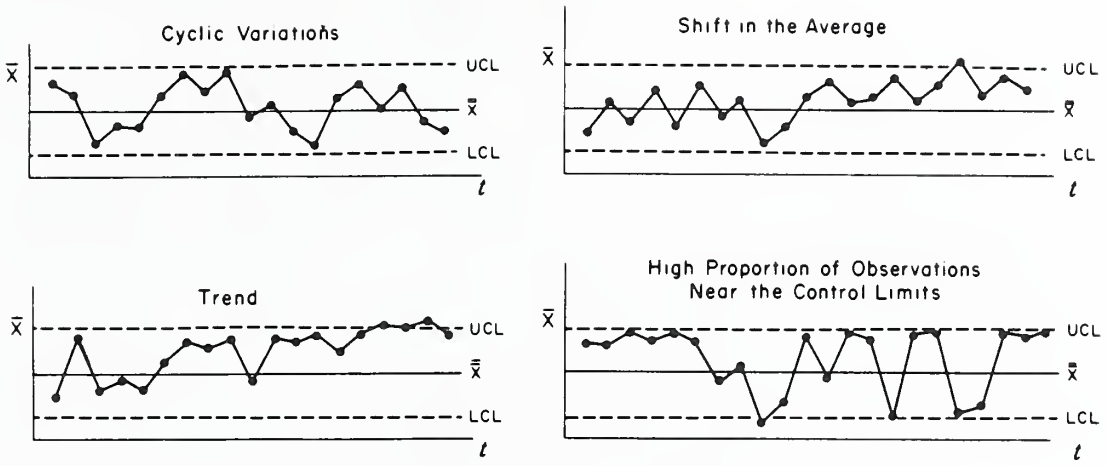


Fig. 4.8. Four patterns of variation in an $\bar{X}$-chart.

The control chart as a management tool

As indicated in the ASQC definition of quality assurance, ". . . The system involves a *continuing evaluation* of the adequacy and effectiveness of the overall quality-control program with a view of *having corrective measures* initiated where necessary . . ."[9]

The key words, "continuing evaluation" and "having corrective measure initiated," indicate the essence of a quality control program. It is important that the results of the control chart be subjected to a daily analysis in order to detect not only the out-of-control points but also any other manifestation of lack of randomness as shown by a time sequence of daily observations. It is always better and more economical to prevent a disaster than to take drastic measures to cure one. Since each test method should be subjected to quality control, the control charts should be prominently displayed at the location where the test is performed, not only to facilitate the logging of results as soon as they are obtained but also to give the technician responsible for the test an easy graphical representation of the time sequence of events. In addition, preprinted forms containing the relevant classification should be available for easy recording of information such as names, dates, time of day, reagent lot number, etc.

When all the pertinent data provided by the control charts are available, the supervisor, or section manager, should have all the meaningful information required to take corrective measures as soon as a source of trouble has been detected. Monthly or periodic review of the results, as performed by a central organization with the aid of existing computer programs, is important to provide the laboratory director with an important management tool, since the output of these programs may include such items as costs, inter- and intra-laboratory averages, historical trends, etc. However, as pointed out by Walter Shewhart[10] and other practitioners of quality control, the most important use of the control chart occurs where the worker is, and it should be continuously evaluated at that location as soon as a new point is displayed on the chart.

References

1. Chemical Rubber Publishing Company. 1974. Handbook of chemistry and physics. 55th ed. Cleveland, Ohio.
2. Mandel, J. 1964. The statistical analysis of experimental data. Interscience–Wiley, New York.
3. Natrella, M. G. 1963. Experimental statistics. Natl. Bur. Stand. Handb. 91, Washington, D.C.
4. Davies, O. L., and P. Goldsmith, eds. 1972. Statistical methods in research and production. Oliver & Boyd, Hafner, New York.
5. Snedecor, G. W., and W. G. Cochran. 1972. Statistical methods. Iowa State Univ. Press, Ames.
6. Proschan, F. 1969. Confidence and tolerance intervals for the normal distribution. *In* H. H. Ku, ed., Precision measurement and calibration, statistical concepts and procedures. Natl. Bur. Stand. Tech. Publ. 300, vol. 1. Washington, D.C.
7. Mandel, J. 1971. Repeatability and reproducibility. Mater. Res. Stand. 11(8): 8–16.
8. Galen, R. S., and S. R. Gambino. 1975. Beyond normality, the predictive value and efficiency of medical diagnoses. Wiley, New York.
9. American Society for Quality Control, Statistical Technical Committee. 1973. Glossary and tables for statistical quality control. Milwaukee, WI.
10. Shewhart, W. A. 1931. Economic control of manufactured product. Van Nostrand, New York.
11. Grant, E. L., and R. S. Leavenworth. 1972. Statistical quality control. McGraw–Hill, New York.
12. Duncan, A. J. 1974. Quality control and industrial statistics. Richard D. Irwin, Homewood, IL.
13. Haven, G. T. 1974. Outline for quality control decisions. The Pathologist 28: 373–378.

U.S. GPO: 1986-491-070/40047

MASS METROLOGY

NIST MEASUREMENT SERVICES:
Mass Calibrations

R. S. Davis

Precision Engineering Division
Center for Manufacturing Engineering
National Engineering Laboratory
National Institute of Standards and Technology
Gaithersburg, MD 20899

January 1989

NOTE: As of 23 August 1988, the National Bureau of Standards (NBS) became the National Institute of Standards and Technology (NIST) when President Reagan signed into law the Omnibus Trade and Competitiveness Act.

U.S. DEPARTMENT OF COMMERCE, C. William Verity, Secretary
Ernest Ambler, Acting Undersecretary for Technology
NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY,
(formerly National Bureau of Standards)
Raymond G. Kammer, Acting Director

Library of Congress Catalog Card Number: 88–600608

National Institute of Standards and Technology Special Publication 250-31
Natl. Inst. Stand. Technol., Spec. Publ. 250-31, 72 pages (Jan. 1989)
CODEN: NSPUE2

U.S. GOVERNMENT PRINTING OFFICE
WASHINGTON: 1989

For sale by the Superintendent of Documents, U.S. Government Printing Office, Washington, DC 20402–9325

NIST Spec. Publ. 250-31

Errata:

p. 16, first paragraph

delete the words "for nominal values 1 kg and below"

p. 22, caption for Figure 5

change 0.4 mg/yr to 0.4 μg/yr

p. A4 (Appendix A)

the first sentence after eq. (11) should read "The array of coefficients, d_{ij}, is given in Appendix B of ref. [8] for some of the designs."

PREFACE

Calibrations and related measurement services of the National Institute of Standards and Technology provide the means for makers and users of measuring tools to achieve levels of measurement accuracy that are necessary to attain quality, productivity and competitiveness. These requirements include the highest levels of accuracy that are possible on the basis of the most modern advances in science and technology as well as the levels of accuracy that are necessary in the routine production of goods and services. More than 450 different calibrations, measurement assurance services and special tests are available from NIST to support the activities of public and private organizations. These services enable users to link their measurements to the reference standards maintained by NIST and, thereby, to the measurement systems of other countries throughout the world. NIST Special Publication 250, NIST Calibration Services Users Guide, describes the calibrations and related services that are offered, provides essential information for placing orders for these services, and identifies expert persons to be contacted for technical assistance.

NIST Special Publication 250 has recently been expanded by the addition of supplementary publications that provide detailed technical descriptions of specific NIST calibration services and, together with the NIST Calibration Services Users Guide, they constitute a topical series. Each technical supplement on a particular calibration service includes:

- specifications for the service
- design philosophy and theory
- description of the NIST measurement system
- NIST operational procedures
- measurement uncertainty assessment
  - error budget
  - systematic errors
  - random errors
- NIST internal quality control procedures

The new publications will present more technical detail than the information that can be included in NIST Reports of Calibration. In general they will also provide more detail than past publications in the scientific and technical literature; such publications, when they exist, tend to focus upon a particular element of the topic and other elements may have been published in different places at different times. The new series will integrate the description of NIST calibration technologies in a form that is more readily accessible and more useful to the technical user.

The present publication, SP 250-31, NIST Measurement Services: Mass Calibration at the National Institute of Standards and Technology, by R. S. Davis, is one of the documents in the new series. It describes calibration technology and

procedures utilized in connection with NIST Service Identification Numbers from 22010C to 22180M listed in the NBS Calibration Services Users Guide 1986-88/Revised (pages 28-31). Inquiries concerning the contents of these documents may be directed to the author(s) or to one of the technical contact persons identified in the Users Guide.

Suggestions for improving the effectiveness and usefulness of the new series would be very much appreciated at NIST. Likewise, suggestions concerning the need for new calibration services, special tests and measurement assurance programs are always welcome.

Joe D. Simmons, Acting Chief
Office of Physical Measurement Services

Contents

1. Description of Service

The National Bureau of Standards maintains the national standard for mass in the form of the prototype kilogram K20 and its companion K4 and provides services to support the segments of the national measurement system which rely directly or indirectly on mass measurements. These services are offered only to those customers whose requirements cannot be met by state laboratories. In order to provide prompt and useful service, the acceptance of the items for calibration or test is based on discussions with each user to determine details necessary to meet measurement and delivery requirements, and on inspection of the item at the Bureau to determine its suitability for the usage intended.

Services are available to enable a user to establish a measurement assurance program for certain measurement processes. This may involve developing procedures for establishing and maintaining a state of statistical control for the measurements, the determination of the offset of the process from the national system, and assisting in the determination of the uncertainty of measurements made by the user's process.

Arrangements for calibration (or test) must be completed before shipping apparatus to the Bureau. While all services are on an actual-cost basis, subject to a $25 minimum charge, a mutual agreement on the work to be performed generally results in substantial savings for the user. Detailed packing and shipping instructions are available on request. Items not accepted for calibration or test will be returned, the cost of inspection or the minimum charge will be applicable.

The results of a calibration or test will be reported in a National Bureau of Standards Report of Calibration Test or of Special Test (which in many cases is prepared by a computer program), a continuation report, or a letter report. In each of these, the values reported are accompanied by an appropriate estimate of uncertainty (allowance for random and systematic errors) as determined by an analysis of the specific measurement process. A continuation report is used for those items submitted for recalibration on which preliminary tests indicate that no significant changes have occurred since the last calibration. Usually a letter report is used to report a test for compliance with a specification which states limits for the departure of the actual value from nominal.

Charges for these services are listed in the NBS SP250 Appendix. Upon receipt of a request for services, an estimated cost will be given along with a firm date for completion. An effort will be made to discuss the measurement requirement with the customer so as to give proper service at minimum cost and delay.

The Bureau's calibration of reference standards of mass provides extensions of the mass unit embodied in the NBS standard of mass. A normal calibration consists of establishing a mass value and the appropriate uncertainty for that value for each weight which has been designated to be a reference standard. It is desirable, but not necessary, that a weight meet the adjustment tolerances

established for NBS Classes A, B, M, or S-1 prior to submission[1].[1] Normally weights are available from manufacturers, many of whom can furnish directly documentation suitable for meeting quality assurance contracts and requirements.

Individual weights or sets of weights in the range of 30 kg to 1 mg or 50 lb to 1 μlb in decimal subdivisions, which are designated as reference standards, must be of design, material, and surface finish comparable to, but not necessarily limited to, present NBS Classes A, B, M, S, or S-1.[2] Design, material, and surface finish of large mass standards (from 50 to 50,000 lb) must be compatible with the intended usage. For these large mass standards an adjustment with reference to a nominal or desired value can be included as a part of the calibration procedure.

The values of true mass (and an apparent mass correction) included in the report will be determined by using computed volumes based on the manufacturer's statement of density of the material, on the density computed from measured volumes, or in the absence of this information, on estimated density values. The apparent mass corrections are computed for 20 °C with reference to Normal Brass (density 8.4 g/cm^3 at 0 °C, volume coefficient of expansion 0.000054/°C) and to stainless steel (density 8.0 g/cm^3 at 20 °C) in an ideal air density of 1.2 mg/cm^3. Apparent mass corrections to any other basis can be furnished if requested.

For periodic recalibrations of reference mass standards, the user need measure only differences between weights or groups of weights within a set and compare them with computed differences. As long as the agreement is within allowable limits, the values can be considered constant within the precision of the comparison process. Mass standards which are submitted to the Bureau for recalibration frequently are tested in this manner. If these tests indicate that no significant changes have occurred, a continuation report so stating and referring to the previous NBS Report of Calibration will be issued.

2. The International System of Units

Virtually all industrialized countries are signatories to a treaty which establishes a consistent set of measurement units. The convention which has been agreed to is called the International System of Units. It is frequently abbreviated as SI (for Systeme International d'Unites, the treaty having been written in French). An international committee, which was established through a provision of the treaty, sees to it that the definitions of units in the SI change to reflect improvements in measurement technology. In the case of the

---

[1]New weights are more likely to be adjusted to ANSI/ASTM or OIML tolerance [2,3]. We will accept ASTM Classes 1, 2, and 3 as well as OIML Classes E1, E2, F1, and F2. (See ASTM E617.)

[2]We will also accept ANSI/ASTM Grades S and 0 as well as OIML classes E1, E2, F1, or F2.

unit of mass, however, there has been no change in the definition for almost 100 years.

The unit of mass in the SI is the kilogram. Its value is defined with reference to an object known as the International Prototype Kilogram (IPK). The definition can be simply stated:

"A kilogram is equal to the mass of the International Prototype of the kilogram."[4]

The IPK is kept and used under the supervision of the International Bureau of Weights and Measures (BIPM) on land provided by the French government in Sèvres, near Paris.[5]

It is then necessary to establish a practical system of mass measurement based on the simple definition.

3. Mass Standards in Practice

3.1 Kilograms

The first step is the easiest to achieve. Countries, such as the United States, possess at least one replica of the IPK. These replicas are made of the same material as the IPK (an alloy of 90 percent platinum/10 percent iridium; density 21.5 g/cm^3), and have the same shape (a cylinder whose height equals its diameter). The replicas are only within 1 milligram of the IPK but differences between the replicas and the IPK can still be measured using the best balances (such balances have the almost incredible precision of 1 microgram in 1 kilogram, or $1x10^{-9}$). Thus each replica must be compared either directly or indirectly with the IPK in order to establish its mass. The U.S. bases its mass measurements on the value of replica no. 20 (sometimes referred to as K20), which is kept at the National Bureau of Standards in Gaithersburg, MD. The difference between the mass of the IPK and K20 was determined in 1890 and again in 1948. In 1984 K20 was compared indirectly with the IPK. A detailed account of these latest measurements and a review of previous measurements involving the replicas can be found in [6]. The mass of K20 is thus known to be 1 kg -0.020 mg with an uncertainty of less than 0.010 mg. Notice that, while the mass of the IPK is 1 kg by definition and thus has no uncertainty, the mass value assigned to K20 is not exactly equal to its nominal value and does have a finite uncertainty.

The IPK and its replicas, such as K20, are made of platinum/iridium for a variety of reasons chief among which is resistance to chemical attack. The expense of this alloy has precluded and still precludes the widespread use of platinum/iridium weights. In the first half of this century, brass weights, usually plated with nickel or rhodium, were the best quality weights commercially available. More recently, stainless steel has supplanted plated brass as the material with which the highest quality commercial weights are fabricated.

Aside from the fact that plated brass or stainless steel are not as resistant to chemical attack as is platinum/iridium, the major difference between kilograms made of Pt/Ir and those of brass or steel is their size, more specifically their volume. One kilogram of Pt/Ir has a volume of about 46.5 cm^3; but a typical stainless steel kilogram has a volume of about 125 cm^3 and one of brass has a volume of about 119 cm^3. The difference in volume between a Pt/Ir kilogram on the one hand and a brass or steel kilogram on the other is so great that the buoyant effect of the air which surrounds the kilograms cannot be neglected during weighing. If, for instance, one constructed a stainless steel kilogram weight which exactly balanced the IPK when the two were compared on the most precise balance available, the stainless weight would actually have a mass of about 1 kg + 94 mg. The "extra mass" came about because the measurements were done in air. Air is a fluid which, like any fluid, produces a buoyant effect on objects it surrounds. The effect can be as large and dramatic as the Goodyear blimp. However, when it comes to mass standards and weighing in general, the effect is small but extremely nettlesome. In the case of the Pt/Ir and steel kilograms just mentioned, the buoyant force on the steel exceeded that on the Pt/Ir so that almost 0.1 g extra steel was needed to balance the two weights. Had the weighing been done in the absence of air, i.e. in vacuum, there would have been no problems due to air buoyancy: the mass of stainless steel which would exactly balance the IPK would have been exactly 1 kilogram. The buoyancy effects are illustrated in figure 1.

Our concept of mass requires that the mass of an object be the same in vacuum or air (or any other fluid) providing that the total amount of material comprising the object has not changed (i.e, the weight does not dissolve, evaporate, react chemically, etc. with the fluid which surrounds it). Thus a correction must be applied to mass measurements made in air between standards of different volumes. Let us return to the imaginary measurement in air which showed that a stainless steel kilogram exactly balanced the IPK. The results of this measurement can be stated in several different ways:

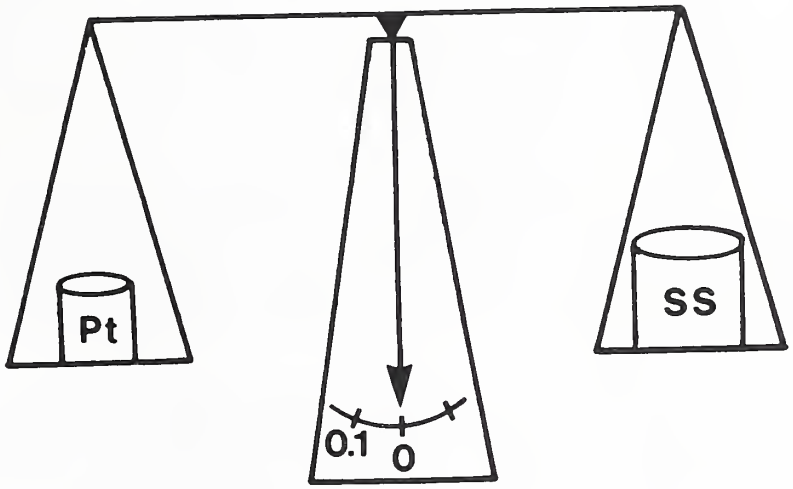
MASS: The mass of the stainless steel object is about 1 kg + 94 mg.

TRUE MASS: The true mass of the stainless steel object is about 1 kg + 94 mg (true mass = mass)

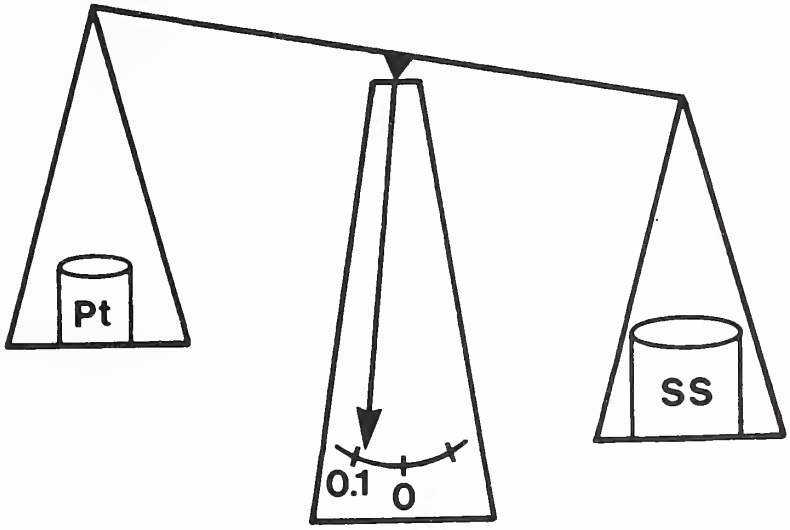
VACUUM MASS: The vacuum mass of the stainless steel object is about 1 kg + 94 mg. (vacuum mass = mass).

APPARENT MASS: The apparent mass of the stainless steel kilogram is about 1 kg + 0.0 mg when measured against Pt/Ir standards in air of density 1.2 g/cm^3 at a temperature of 20 ^{o}C.

Note that when specifying apparent mass, one must define the specific weighing conditions. Whereas the mass of an object is a fundamental attribute, its apparent mass will depend on its density, the density of the standard to which it was compared, and the density of air at the time the measurements were made. (The temperature at which the measurements were made must be specified also because the density of stainless steel, Pt/Ir, and other materials is a slight function of temperature.) Because apparent mass is defined through a particular convention, it is sometimes referred to as "conventional mass."



A.



B.

Figure 1. A. A stainless-steel kilogram (density 8000 kg/m^3) balances a platinum-iridium kilogram (density 21500 kg/m^3) at normal atmospheric conditions. B. Under vacuum conditions one can see that the mass of the stainless-steel kilogram actually exceeds that of the platinum kilogram by about 0.1 g.

The two conventions in widest use are: density 8.0 g/cm^3 at 20 °C in air of density 1.2 g/cm^3; and density 8.39094 g/cm^3 at 20 °C in air of density 1.2 g/cm^3 [7].

Mass in SI units is truly mass (or "vacuum" mass or "true" mass). When the National Bureau of Standards (NBS) calibrates a stainless steel kilogram in terms of its Pt/Ir national standard, a correction of order 0.1 grams must be made to the raw data. This correction is large compared to the precision of the mass comparison. Making the correction requires a great deal of effort. Even doing the best one can, this buoyancy correction can be the major contributor to the uncertainty in the calibration of the stainless steel kilogram. If the steel kilogram is then used to calibrate other steel or brass kilograms, the buoyancy corrections will be small and relatively easy to apply. The use of stainless steel or other nickel-chrome alloys as "working standards" also means that K20 can be used very infrequently, thereby minimizing wear and the chance of accident.

The strategy of NBS was to put great effort into the calibration of two standards, N1 and N2, having a density of 8.35 g/cm^3. These two weights, made of a nickel-chrome alloy, are then used to calibrate other weights of similar density on a more routine basis. At the present time, errors in assigning calibration value of N1 and N2 with respect to the IPK are not included in error budgets. This practice is now in the process of being revised.

At the time N1 and N2 were fabricated, the majority of high-quality weights were made of plated brass; hence the choice of alloy density. Now virtually all high-quality weights are made of stainless steel. For this reason, NBS is in the midst of changing its working standards from N1 and N2 to standards of stainless steel density of 8.0 g/cm^3. This will be a relatively slow process because the long-term stability of the new standards must first be established. In addition, the tie to the SI unit is being carefully established so that a meaningful uncertainty can be assigned.

3.2 Other Denominations

So far, we have described how the SI mass unit is transferred from the Pt/Ir prototype in Sèvres to a nickel-chrome or stainless steel working standard at NBS. Obviously, it is essential to calibrate weights at other nominal values above and below 1 kg.

The concept of how this is accomplished is simple. Once one has a weight which embodies an SI mass value, it is only necessary to find the ratio of the mass of the known weight to the mass of the unknown weight. Since a ratio is dimensionless, these measurements can, in principle, be accomplished by any laboratory with sufficiently sensitive balances. For instance, if we have a kilogram weight calibrated in SI units and we need to know the mass value of a 500-g weight, the following simple approach might be taken: The known 1-kg weight could be used to calibrate a digital electronic balance of 1 kilogram capacity. The unknown 500-g weight could then be placed on the balance and its mass value read directly. (It would also be wise to check the balance linearity and make the necessary corrections for air buoyancy.) Such

calibrations were, of course, done before the days of electronic balances. Even today, higher precision can usually be obtained with mechanical balances although with a great loss in convenience. Using an equal-arm balance, for instance, one would need an additional 500-g weight. The mass difference between the sum of both 500-g weights could be found with respect to the calibrated 1-kg weight. This is enough information to assign mass values to the two unknown weights. Similar, though more sophisticated, procedures are used to calibrate weights of all denominations beginning with calibrated kilogram standards.

Calibration of a set of weights consists of assigning values for the unknown weights in terms of the known mass of one or more standards. For high precision work, this involves the use of the balance as a comparator which measures the difference between two objects (or two groups of objects) which must have nominally the same mass because of the small "on-scale" range of the comparator. In deriving units which are subdivisions of the basic unit or multiples thereof, a variety of different weighing sets have been used because of convenience or other practical considerations. A typical set is the 5 3 2 1 series which bridges the range from 10 to 1. In most cases, the calibration algorithm provides for a check standard, treated as an additional unknown weight, to be used for monitoring the performance of the measuring process.[8]

Precision weighing is usually done by some form of transposition weighing on a two-pan balance or by substitution methods on a one-pan balance [7]. For simplicity, it will be assumed that a well behaved comparator is available and that measurements of differences in the mass of two objects or groups of objects are corrected for air buoyancy effects and other environmental or procedural factors [7]. It is further assumed that the measurements are uncorrelated in the statistical sense and all are of equal precision. (These latter two assumptions are non-trivial and special care has to be taken to insure their validity so that the random error component of the uncertainty is properly evaluated.)

The schedule of measurements for calibration should include provision for a check standard and also for within-run redundancy. The decision as to which one of a number of possible schedules or designs to use for intercomparison of a set of weights depends on items such as the variance associated with individual weights or combinations thereof. The least squares analysis from which the values for the weights and their variances are calculated is presented in Appendix A.

4. Density Determination of Single-Piece Kilograms Using a Submersible Balance

The buoyancy correction is important in precision weighing. For most cases, an assumed density (supplied by the manufacturer) will suffice. However, in the case of 1-kg standards, it is desirable to measure the density of individual weights. This measurement is now done routinely at NBS for single-piece kilogram and pound weights sent for calibration. The density measurements we use are a modification developed in our laboratory of the usual hydrostatic weighings. A brief description of the technique which we use follows.[9]

4.1 Apparatus

The balance modified for this work is a Mettler PL1200,[3] the important specifications of which are:

weighing range	0 - 1200 g
reproducibility	<0.005 g
linearity	<0.01 g

Significant mechanical and electronic modifications were introduced to the balance and its enclosure. Specifically, the weighing cell of the balance was separated from the supporting electronics and placed beneath the surface of an inert liquid (see fig. 2).

Clearly, the fluid in which an electronic measuring cell is immersed must have many special properties: it must be electrically insulating, it must be chemically inert, it must be optically transparent (in order for the servo optics to function properly), and it should not evaporate quickly. These characteristics may be found, for example, in FC-75, a fluorinated fluid manufactured by the 3M Company. A comparison of some of the properties of FC-75 with those of water is given in table 1. An additionally noteworthy property of FC-75 is its immense appetite for gases. For example, the fluid is able to dissolve about 0.3 g of air per kilogram of fluid. This ability to dissolve atmospheric gases, greatly inhibits bubble formation on immersed objects--one of the most serious problems in conventional high-precision hydrostatic weighing. Finally, the fluid is 77 percent more dense than water at room temperature thereby increasing the signal to noise in comparison to a normal hydrostatic weighing. The major disadvantages of this fluid as compared to water are its large coefficient of thermal expansion and its cost. However, use of FC-75 instead of water for conventional "hydrostatic" weighing has many advantages.[10] The density of FC-75 is usually not known accurately enough for the liquid to serve as a density standard. Instead, the fluid density is calibrated at the time of use by including a solid object of known mass and volume in the weighing scheme.

---

[3] Certain trade names and company products are identified in order to adequately specify the experimental procedure. In no case does such identification imply recommendation or endorsement by the National Bureau of Standards nor does it imply that the products are necessarily the best available for the purpose.

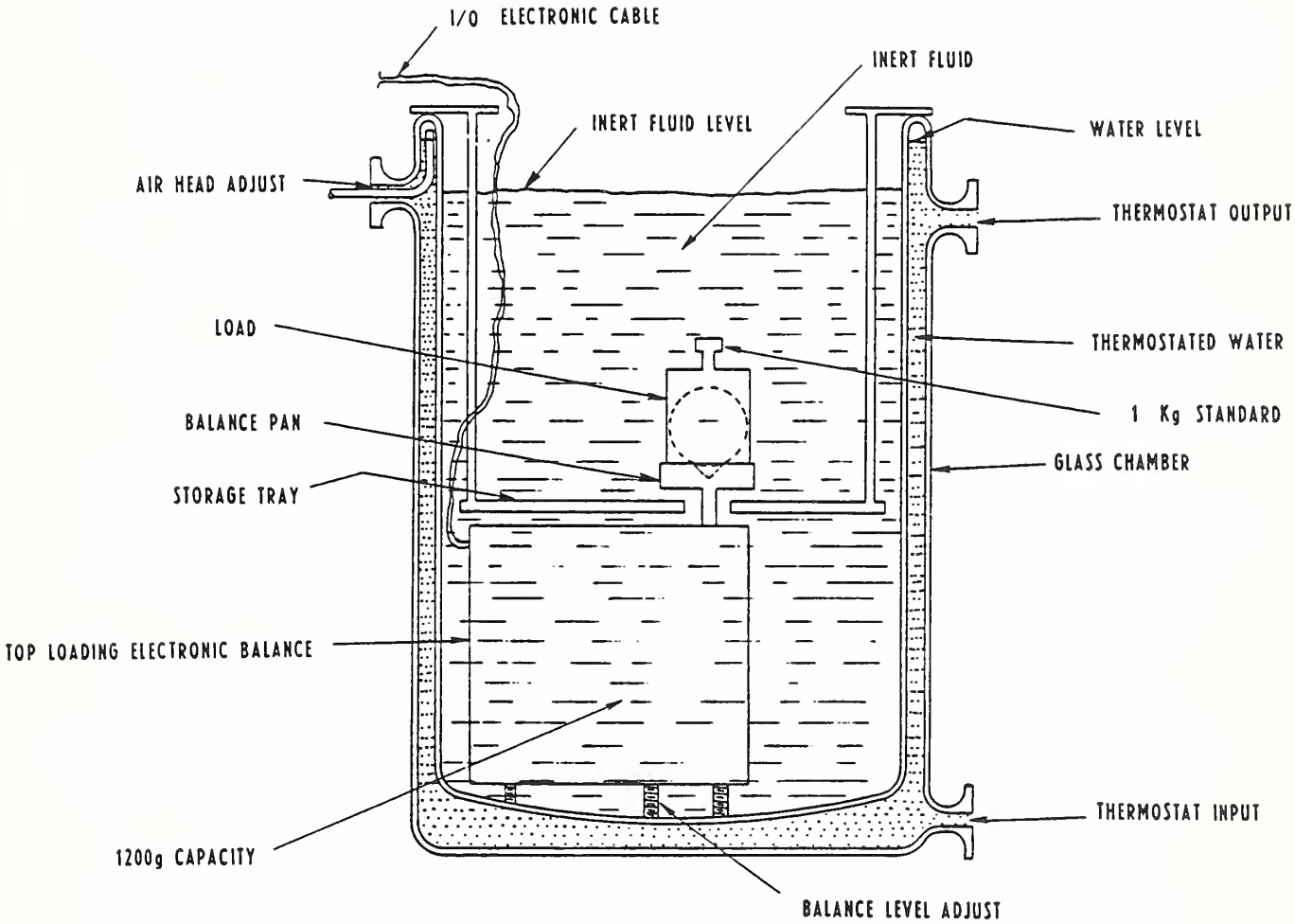


Figure 2. Cross-sectional view of the balance immersed in its thermostated bath. A conventionally-shaped standard weight is shown on the pan. Unconventional loads may also be accommodated by the pan as suggested by the sphere (drawn with dashed lines).

Table 1. Comparison of Properties of FC-75 and Water

Property (at 25 °C)	FC-75[a]	Water
Density (kg/m^3)	1800	1000
Coefficient of expansion ($^oC^{-1}$)	$1.6x10^{-3}$	$2.5x10^{-4}$
Kinematic viscosity (cm^2/s)	$0.82x10^{-2}$	$0.89x10^{-2}$
Vapor pressure (Pa)	4000	3200
Surface tension (N/m)	0.015	0.072
Heat capacity (J/g-°C)	1.0	4.2
Thermal conductivity (W/cm-°C)	$1.4x10^{-3}$	$6.1x10^{-3}$

[a]Values for FC-75 supplied by the manufacturer.

4.2 Principles of Use

We illustrate the use of the submersible balance by finding the volume, V_A, of an object, A, when its mass, M_A, is known. Placing the object on the balance produces a reading, O_A, which is related to the other parameters through the equation

$$O_A = K [M_A - \rho(t)V_A(t)] \qquad (1)$$

where O_A is the difference in reading of the loaded and unloaded balance.

Here ρ is the density of the fluorocarbon and K is a constant scale-factor which may be adjusted by turning a potentiometer controlling the calibration of the balance. Both ρ and V_A are functions of the ambient temperature, t. In succeeding equations the functional dependence on temperature is not shown explicitly.

Normally K is adjusted by the balance manufacturer or user so that the balance will read directly the mass of an object of density $8.0g/cm^3$ in air of density $1.2 \times 10^{-3}g/cm^3$, i.e. K = 1.000150. We found it convenient (though, of course, not essential) to readjust K to exact unity. Thus we can ignore K in the succeeding equations. Hence,

$$V_A = \frac{M_A - O_A}{\rho} \qquad (2)$$

The problem with using eq (2) is that the precision with which V_A can be measured far exceeds the accuracy with which ρ is known. Thus, for best results, one should also measure an object whose mass and volume are known. Placing such an object on the balance essentially calibrates the density of the fluorocarbon at the time of weighing. Let the known object be called S1. Then

$$\rho = \frac{M_{S1} - O_{S1}}{V_{S1}}$$

and

$$V_A = \frac{M_A - O_A}{M_{S1} - O_{S1}} \quad V_{S1} \qquad (3)$$

In practice, S1 is a stainless-steel kilogram whose volume has been determined to better than 1×10^{-5} by classical hydrostatic weighing. We also include a check standard, S2 in the measurements. The check standard is similar to S1 and similarly calibrated.

5. Cleaning of Weights

It is essential that weights being calibrated, as well as the standards used, be clean if the calibration is to be accurate and meaningful. Therefore, a cleaning procedure should be a part of every calibration.[11]

5.1 Categories of Weights

For cleaning purposes, weights may be divided into four categories.

1. One-piece weights.

   This category will include all one-piece weights except lacquered weights, sheet metal weights and small wire weights.

2. Screw-knob weights.

   This category will include all weights with adjusting cavities except lacquered weights.

3. Lacquered weights.

   This category includes all lacquered or painted weights.

4. Sheet metal weights and wire weights.

5.2 Cleaning Procedures

5.2.1 One-Piece Weights

One-piece weights, 1 gram and larger, are generally steam cleaned. The weights are either held or placed in a jet of steam and manipulated so that the entire surface of the weight is subjected to the cleaning action of the steam long enough to clean it. A superficial steaming is not enough. The weight is then dried, either by evaporation or careful wiping with a soft non-abrasive material free from oil and other substances that will leave a residue on the weights, such as high grade cheesecloth. Care must be exercised that no water spots are left on the weights as they dry. Visible particles on the weights should be brushed or wiped off before steam cleaning them. If a steam generator is not available, one-piece weights may be cleaned either by immersing them in a hot or boiling distilled water bath in a non-metallic container or according to the procedures for screw-knob weights.

Occasionally, a weight will have foreign material adhering to it that requires the use of solvents. Ethyl alcohol is a good general solvent.[4] If alcohol does not remove the material, other solvents may be used, such as benzene, 1,1,1- trichloroethane, etc. Alcohol is then used to remove any film left by the other solvents. The weights are then steam cleaned as outlined above. Cleaning and, in particular, steam cleaning, may adversely affect some alloys. For these alloys, only solvent cleaning is used.

5.2.2 Screw-Knob Weights

Weights in this category are usually cleaned by wiping with a soft non-abrasive material, free from oils or other substances that will leave a residue of any kind on the weights, such as high grade cheesecloth. Occasionally, a weight will have foreign matter adhering to it that requires the use of solvents, applied with a cloth. Ethyl alcohol is a good general solvent. If alcohol does not remove the foreign material, other solvents may be used. Alcohol is then used to remove any film left by the other solvents.

To prevent spotting the weights when using solvents, the weights are wiped dry. Care is taken that no liquid gets under the knobs or especially into the adjusting cavity.

A modified steam cleaning procedure may be used on screw-knob weights. The bottoms and sides are steam cleaned, care being taken that no liquid or vapor gets under the knob or into the adjusting cavity.

---

[4] Some solvents are health hazards and should be used in an approved safe manner.

5.2.3 Sheet-Metal Weights

First, the weights are placed in an acetone bath agitated to help loosen any foreign material. A soft brush, such as a camel hair brush, may be used to agitate the weights. The weights are removed from the acetone, allowed to dry and then steam cleaned. For steam cleaning, the weights are held in front of a jet of steam with forceps until the entire surface has been covered with steam. (See Note on next page.) In order that the portion of the surface under the forceps may be steamed, the weight is set down and picked up again with the forceps holding the weight at a different spot than the first time; the weight is again steamed. The weights are not allowed to touch the steam nozzle. Only a low ash filter paper is used for drying the sheet metal weights. A circular disk is folded unsymmetrically. The main body of the weight is placed between the folds of the paper with the turned edge of the weight protruding. The main body of the weight is dried by pressing lightly on the top of the paper. The turned up edge is brushed lightly with a piece of filter paper. In some cases, it may be necessary to brush the body of the weight with filter paper to remove drops of water. Care must be exercised that no water spots are left on the weights as they dry.

Note: The small fractional weights, say smaller than 5 mg, may be placed in a hot or boiling distilled water bath for the final cleaning instead of steam cleaning them. A hot or boiling distilled water bath may also be used for the final cleaning of all sheet metal weights when a steam generator is not available.

5.3 Temperature Equilibrium

Newly cleaned weights are allowed to come to temperature equilibrium before they are calibrated. This may take several hours for the larger weights that have been steam cleaned.

Generally, laboratory weights will come to temperature equilibrium over night.

5.4 Storage

Usually, weights are not placed in the balance immediately after cleaning, but are stored for varying periods. The weights are stored under cover so that they will stay clean. Weights 1 gram and larger may be stored on a tray lined with filter paper and covered with an inverted glass dish. The smaller weights may be stored in a small glass dish covered with a watch glass. In both cases, the container is labeled with the weight identification.

When the weights are placed in the balance, they are carefully brushed to remove any particles that may be on them. A small bulb type rubber syringe is useful in removing lint and other small particles from weights. The particles are blown off the weights. Therefore, neither the nozzle nor any other part of the syringe need touch the weights.

5.5 Brushes

Brushes require special attention because they are easily contaminated and often are the last cleaning instrument used before the weights are calibrated. Only soft brushes, such as camel hair brushes are used on weights.

The brushes are cleaned by washing with soap and water, then rinsing in ethyl alcohol and allowed to dry in air. New brushes are cleaned before using to remove any oil or other matter that might contaminate the weights. Used brushes are cleaned as often as necessary to be sure that the brushes themselves do not contaminate the weights.

6. Method of Calibrating Dead Weights

6.1 Measurement Algorithm

The following method of calibrating dead weights--including pressure gage (piston gage) weights--is routinely used [12]. The method employs simple weighing designs and includes corrections for the standards and the buoyant effect of the atmosphere. Calibrations of dead weights are at a lower accuracy than calibrations of mass standards.

The comparisons are made on several different balances whose precision is adequate for the requirements of the weights being tested.

In general, the "Mass Value" of the standard means its True Mass value and its correction means its True Mass Correction unless otherwise indicated. Measurements are often made on by double-substitution weighing, against the built-in weights of an appropriate single-pan balance. These weights have, in turn, been calibrated against NBS standards. The weighing algorithm is:

1) Read the balance with no load: S_1

2) Read the balance with unknown on the pan: $D_1 + S_2$

3) Read the balance with unknown & sensitivity weight: $D_1 + S_3$

4) Read the balance with no load: S_4

Here S_1 represents a reading of the angle of tilt of the balance relative to an arbitrary zero and D_1 represents the nominal values of the summation of dial weights which was used when the balance is loaded. Tare weights are placed on the balance pan as necessary to ensure that the same dial weights are used for operations 2) and 3) (no dial weights are used if no load is on the balance).

The mass of the unknown is computed to be:

$$M_x = \frac{D_1 - \rho V_1 + \frac{\Delta - \rho V_\Delta}{S_3 - S_2}(S_2 - S_1 + S_3 - S_4)}{1 - \rho/D_x}$$

Where D_1 = the calibrated mass of D_1
V_1 = the volume of D_1
Δ = the mass of the sensitivity weight
V_Δ = the volume of the sensitivity weight
D_x = the density of the unknown (supplied by the customer)
ρ = the computed density of air in the balance,

6.2 Uncertainty of Value Assigned to Piston Weights

It is presumed that the weighings are being carried out by means of a measurement process whose parameters (precision, possible systematic errors, etc.) are known and sufficient evidence is collected to insure that the process is in a state of statistical control.[13] For each method of weighing there is a standard deviation associated with a single measurement of mass difference. This standard deviation is based on considerable history and is used in preference to a standard deviation based on the results of say one day's work. Such a value if available provides the means for judging whether or not to accept that day's measurement as being in control.

The uncertainty of the mass value of the piston weights consists of two parts; the uncertainty due to random errors of measurement and the systematic uncertainty due to the uncertainty in the value of the standard. The limit of the uncertainty due to random errors of measurement may be taken to be three times the standard deviation, σ, where σ is the standard deviation of the process. Therefore:

$$\text{Uncertainty of value} = 3\sigma + \text{uncertainty of standards.}$$

7. Method of Calibrating Standard Weights

7.1 Measurement Algorithm

An unusual aspect of mass calibrations is that, although the standard is defined at one value, one is typically asked to calibrate a set of weights spanning several decades of mass and, often, not even encompassing the nominal defining value (that is, 1 kilogram). We approach this problem in the following way: 1. The weight in the set of unknowns which is closest in value to 1 kg is calibrated against an NBS standard of the same nominal value. 2.

The rest of the weights in the set are calibrated in a self-consistent manner using the weight calibrated in step 1 as the standard.

It is convenient to calibrate a set of weights a decade at a time, using as a comparator the most sensitive balance available which will accommodate the largest weight in the decade. Table 2 shows the set of balances which we use for nominal values of 1 kg and below, along with their present standard deviations. Note that we do not calibrate each individual weight in a weight set against a corresponding NBS standard. This would be inefficient. Instead, weights, or summations of weights within the decade are intercompared at several nominal values. The recipe for choosing weights is referred to as a "design". The designs are chosen so that if one of the weights within the design has a known mass value, the mass values of the other weights can be determined. We always pick designs in which we acquire redundant information and calculate the "least squares" values as the calibration result. The least squares solution minimizes the sum of squares of deviations of the predicted minus observed values in much the same way as fitting a series of points on a graph by the least-squares line minimizes the summation of the distances squared of the points from the line. (See Appendix A.)

7.2 Uncertainty of Value of Standard Weights

The uncertainties which are assigned to weights which we calibrate result from uncertainties in our starting standards and uncertainties in the comparison of the unknown weights with our starting standards. Typically, random uncertainties dominate the comparison operation. These are usually due to the balance which is serving as the mass comparator.

The specific design which is used also enters into the assignment of uncertainty. One can average a set of repeated measurements to find a better estimated value than from one single measurement. So too, using a weight in the set in more than one measurement results in a standard deviation of the value assigned to the weight which is less than that of a single measurement. How much less depends on the design. Table 3 shows typical calibration uncertainties based on one commonly used metric weight set and a typical design for that weight set. Note that large-valued masses are usually in avoirdupois units (50 lb and above). The avoirdupois pound is defined as 0.45359237 kg.

One complication of using designs to assign mass values is that the uncertainties assigned to the weights in a set are correlated. This means that when weights are used in combination, the uncertainty of the combination cannot be inferred directly from the uncertainties assigned to each weight individually.

8. Quality Control

In the previous section we noted that the total uncertainty in the assignment of mass values to unknown weights comes primarily from uncertainties in the starting standards and random errors in the performance of balances used to

Table 2. Capacities and standard deviations of balances used

Balance Capacity	Standard Deviation
3 g	0.0005 mg
20	0.0024
40	0.0039
160	0.014
1 kg	0.032
10	2.5 mg
30	8.2
1 lb	0.031 μlb
6	0.46
50	5.5 mg
2500	0.002 lb
30000	0.017 lb

Table 3. Typical random components of calibration uncertainties

Nom. Val.	Uncertainty	Nom. Value	Uncertainty
1000 g	0.072 mg	1000 g	0.072 mg
500	0.059	2 kg	6.6
300	0.057	3	7.8
200	0.050	5	10.0
100	0.058	10	15.5
50	0.035	20	34
30	0.029	30	55
20	0.024		
10	0.026		
5	0.013		
3	0.0082		
2	0.0061		
1	0.0049		
500 mg	0.0026 mg		
300	0.0017		
200	0.0012		
100	0.0010		
50	0.00089		
30	0.00087		
20	0.00080	1 lb	0.11 μlb
10	0.00090	50	22.
5	0.00083	500	0.0028 lb
3	0.00085	2500	0.0073
2	0.00079	10000	0.054
1	0.00090	30000	0.11

intercompare the unknowns with the standards. To control the quality of this measurement system, we must insure that the mass of the starting standards does not change and that the random error of the balances used has not deteriorated from its accepted value. In addition, we must have a way to detect simple blunders in data entry. We now describe the controls which are presently in place [14] and outline improvements which are underway.

8.1 F-Test [7,13]

Every calibration includes a means of checking the balance performance. We assume we know the balance performance based on a large accumulation of data over a long period of time. Each new calibration provides us with another set of data which can be compared with those previously collected. We check to see whether the scatter in the most recent set of data is statistically consistent with the accepted long-term standard deviation of the balance. Failure of this test indicates either a blunder or a sudden degradation of the balance. Figure 3 shows a control chart of the long-term standard deviation of a kilogram balance used in the calibration service. Control limits vary depending on the number of statistical degrees of freedom in a given design. Similar charts are maintained for all the balances used.

8.2 t-Test [7,13]

Every calibration includes at least two standards along with the unknowns. One standard is used to calibrate both the unknowns and the second standard, which is well-known from many previous measurements. The second standard is called the "check standard" for the following reason: As a result of a calibration, the reference standard is used to assign mass values to the unknowns and the check standard. This mass value of the check standard is then compared to the long-term average of the check standard. A statistically significant difference in the two values indicates a change in the standard, a change in the check standard, or a blunder. Figure 4 is a control chart which shows the long-term variability of the difference in measured mass of our two working kilogram standards. Control charts are also maintained to look for unsuspected variability as a function of ambient temperature, barometric pressure, relative humidity, and air density.

Similar control charts are maintained for all the check standards in use. Occasionally, a check-standard will show a significant change (usually a loss) in mass with time. An example of such behavior is shown in figure 5. The control limits in this case change slowly with time, and are not shown.

Check standards are included in every weighing design. This offers an additional advantage which is best illustrated by an example. Routine calibrations of mass at the 1-kg level begin with starting standards N1 and N2. We actually use the total mass of N1 and N2 as the starting standard and use the difference in mass between N1 and N2 as the check standard. This is a mathematical convenience but is conceptually no different from using N1 as the standard and N2 as the check. While the check standard is adequate for detecting catastrophic changes in N1 or N2, it is obviously insensitive to any

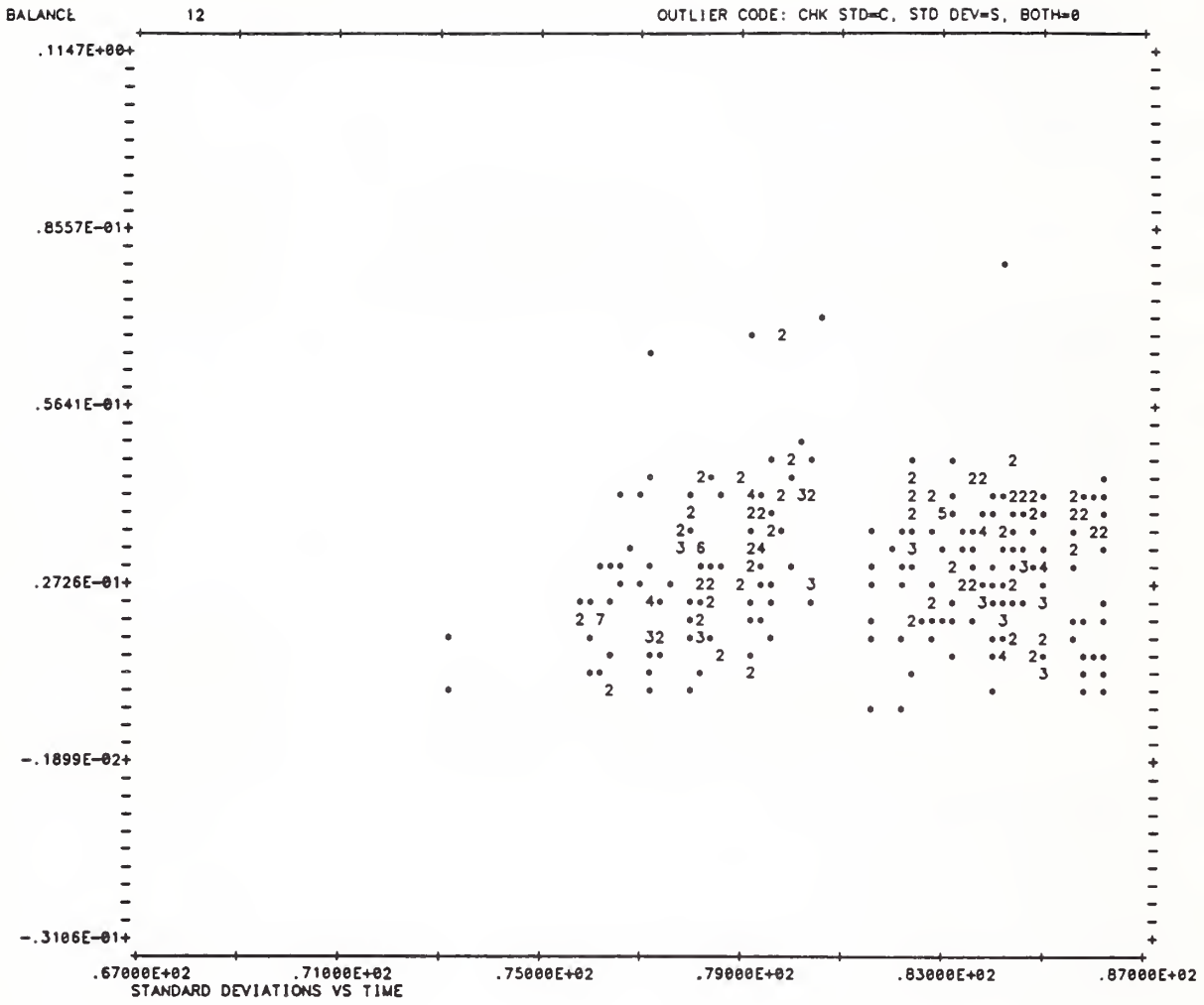


Figure 3. Control chart for the standard deviation of the kilogram balance used in routine mass calibrations. The ordinate gives the measured standard deviation of the balance in milligrams and the abscissa gives the time of the measurement in years since 1900. Control limits are not shown.

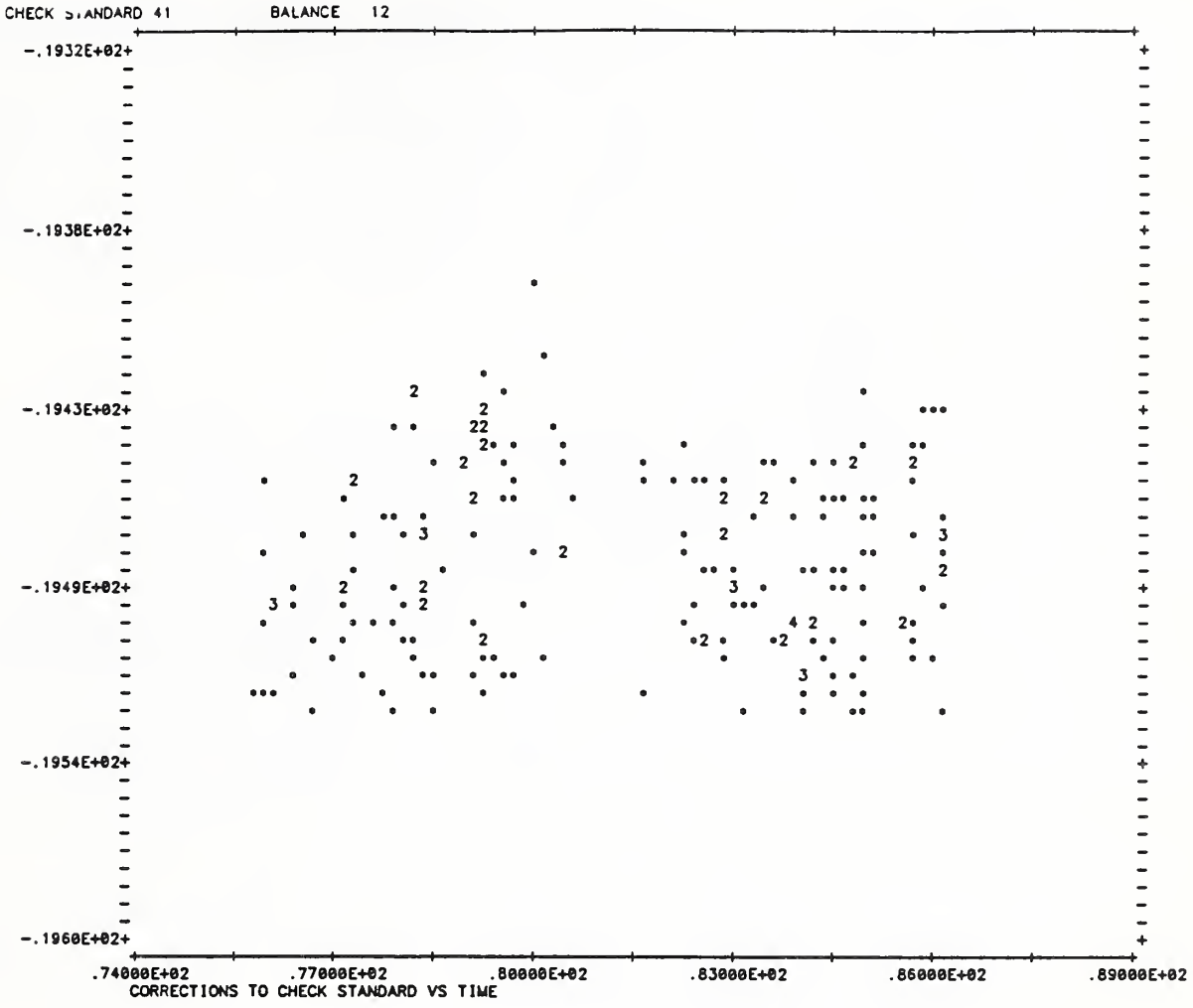


Figure 4. Control chart for the measured difference in mass between kilograms N1 and N2. The ordinate is the difference measured in milligrams and the abscissa is the time of the measurement in years since 1900. Control limits are not shown.

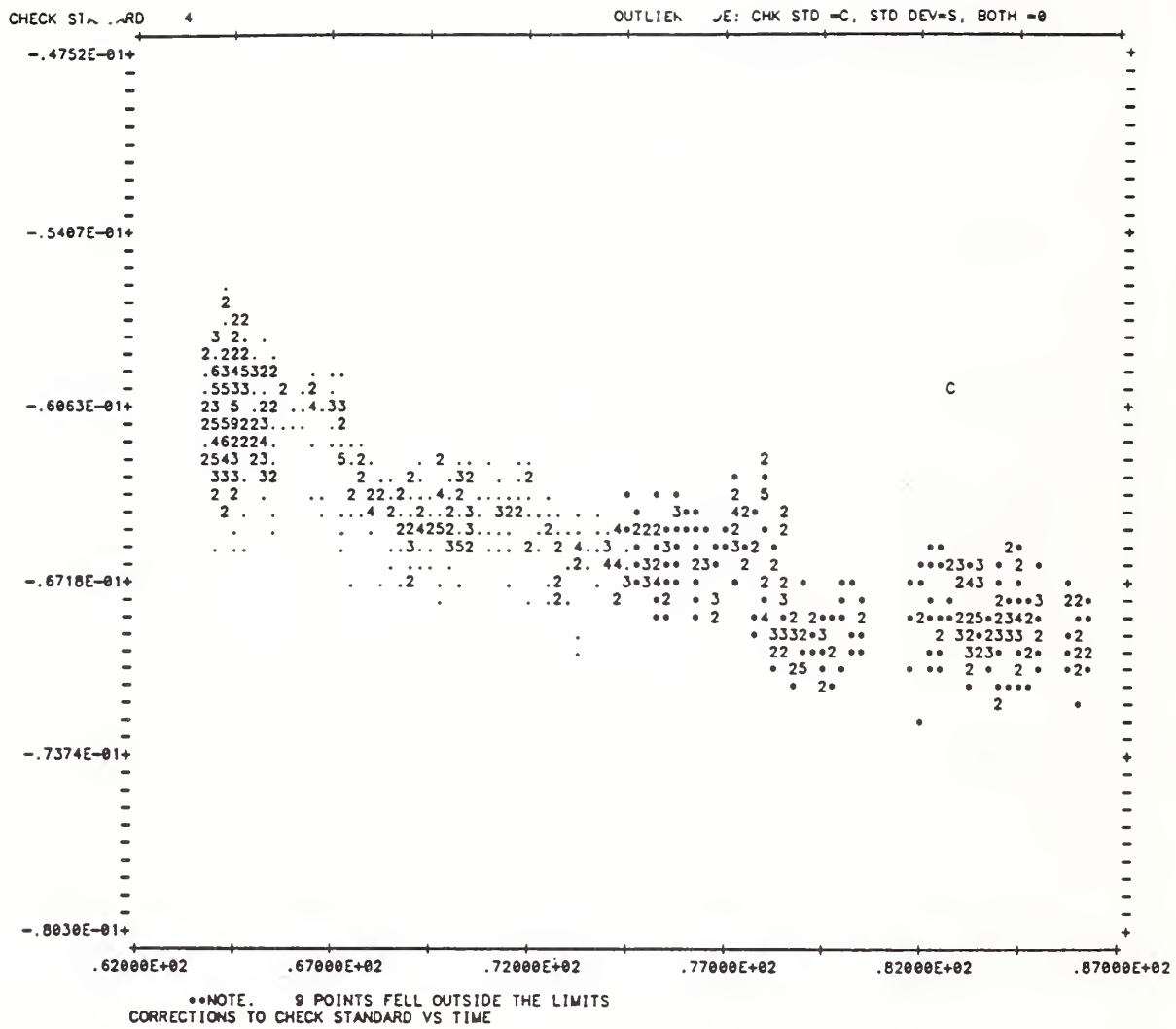


Figure 5. Control chart for the mass of a 1-g check standard. The ordinate is the measured correction to nominal (in milligrams) and the abscissa is the time of the measurement in years since 1900. The check standard is losing mass at the rate of about 0.4 mg/yr. Control limits are not shown.

processes which change N1 and N2 in nearly identical amounts. However, N1 and N2 are used to calibrate weight sets which often include the decade from 1 kg to 100 g. Every time weights in this decade are calibrated, the results are checked with a 100-g check standard. If there were gross changes in N1 and N2 which nevertheless preserved the value of N1 - N2, they would eventually produce failures of the t-test for the 100-g check standards. The problem, then, is to devise a system which will check the constancy of the starting kilogram standards with enough sensitivity to signal a change before its effect is propagated to other masses.

8.3 Between-Times Components

There is a possibility that the standards we are using and measuring have a variability which is a function of time. The shortest time of concern to us is that needed to complete measurements for one design, approximately one hour. We attribute the variability we see in this time scale to the balance (used as comparator) and monitor its constancy by the F-test as described above. A much longer period of interest would be the time between weight calibrations. This is of the order of months or years. The difference between the short-term variability and long-term variability of the measurement process is known as the "between-times" component. We look for the existence of this component by comparing the long-term and short-term standard deviations associated with the check-standards. Ideally, the between-times component will be zero. However, a significant component has been found to exist for calibrations involving the highest relative precision.

For the cases in which we have detected a non-zero between-times component, we have propagated its effect through the calibration process by assuming all weights are subject to the same component as the check standards.

9. Future Plans

There are two major weaknesses in our quality controls at present. The first is that the relationship between the SI unit of mass and the unit of mass as embodied by the values assigned to our working kilogram standards N1 and N2 is not well-enough determined. Second, the use of the difference in mass of N1 and N2 as the check standard at the 1-kilogram level is dangerously insensitive to effects common to both kilograms. This is especially disturbing since N1 and N2 are always stored together and receive identical use. A third but less serious problem is that the alloy of which N1 and N2 are made is closer in density to brass than to most commonly used stainless steels. This makes the usual buoyancy corrections more important than they would need to be if the working standards had a density closer to 8.0 g cm^{-3}.

To improve these three areas, we have completed or are near to completing the following steps:

1. Six new kilograms weights of nominal density 8.0 g cm^{-3} have been purchased. Their densities have been determined to 10 parts per million by hydrostatic weighing.

2. An existing kilogram comparator has been reconditioned and partially automated so that its standard deviation is more than 15 times less than that of the comparator used for routine calibrations. Thus, if there were no between-times components, it would require 225 measurements on the less precise balance to achieve the same uncertainty as one measurement on the more precise balance.

3. The two platinum-iridium prototype kilograms of the United States have recently been recalibrated by the International Bureau of Weights and Measures (BIPM) along with two of our stainless-steel working standards. Good internal consistency was obtained between measurements made at NBS and at BIPM. The results show excellent long-term stability of our platinum-iridium standards with respect to the SI unit as maintained by the BIPM.

4. An exhaustive series of measurements is underway using our best kilogram comparator. These measurements will establish the long-term stability of our new stainless-steel weights as well as determine how reproducibly they can be cleaned. We have also made preliminary measurements of N1 and N2 which can be precisely tied to the SI unit as maintained by the BIPM.

5. When the measurements described in 4. are completed, four of the six new stainless-steel kilograms will be used as working standards. The remaining two will not be used but instead, along with the two stainless-steel weights which have been calibrated at the BIPM, will serve to monitor the constancy of the working standards thus prolonging the times between calibrations against our platinum-iridium prototypes.

10. References

[1] Lashof, T. W. and Macurdy, L. B., Precision Laboratory Standards of Mass and Laboratory Weights. NBS Circular 547, Section 1 (1954 August). Reprinted as NBSIR 78-1476 (1978 October).

[2] ASTM E617-81 Standard Specification for Laboratory Weights and Precision Mass Standards, 1986 Annual Book of ASTM Standards, Vol. 14.02: 480-499.

[3] International Organization for Legal Metrology, OIML Recommendation 20.

[4] Goldman, D. T. and Bell, R. J., eds., The International System of Units (SI): NBS Spec. Publ. 330 (1986 July).

[5] Page, C. H. and Vigoureux, P., eds., The International Bureau of Weights and Measures 1875 - 1975: NBS Spec. Publ. 420 (1975 May).

[6] Davis, R. S., Recalibration of the U.S. National Prototype Kilogram, J. Res. NBS 90 (4), July-August 1985: 263-283.

[7] Jaeger, K. B. and Davis, R. S., A Primer for Mass Metrology, NBS Spec. Publ. 700-1 (1984 November).

[8] Cameron, J. M., Croarkin, M. C., and Raybold, R. C., Designs for the Calibration of Standards of Mass, NBS Tech. Note 952, (1977 June).

[9] Schoonover, R. M. and Davis, R. S., Quick and Accurate Density Determination of Laboratory Weights, Proc. 8th Conf. IMEKO Technical Committee TC3, Krakow Poland, September 9-11, 1980.

[10] Bowman, H. A., Schoonover, R. M. and Carroll, C. L., A Density Scale Based on Solid Objects, J. Res. NBS 78A (1), January-February 1974: 13-40.

[11] Almer, H. E., Weight Cleaning Procedures, NBSIR 74-443 (1973 November).

[12] Almer, H. E., Method of Calibrating Weights for Piston Gages, NBS Tech. Note 577 (1971 May).

[13] Croarkin, M. C., Measurement Assurance Programs Part II: Development and Implementation, NBS Spec. Publ. 676 II; 1984.

[14] Varner, R. N. and Raybold, R. C., National Bureau of Standards Mass Calibration Computer Software, NBS Tech. Note 1127 (1980 July).

Appendix A

LEAST SQUARES ANALYSIS [8]

We begin then with a set of n observations, $y_1, y_2, \ldots y_n$ involving k objects where values, $\beta_1, \beta_2, \ldots \beta_k$ are to be determined. The set of observations can be represented by the equations for their expected values, $E(y_i)$,

$$E(y_1) = x1_1\beta_1 + x_{12}\beta_2 \ldots x_{1k}\beta_k \qquad (1)$$

$$E(y_2) = x_{21}\beta_1 + x_{22}\beta_2 \ldots x_{2k}\beta_k$$

$$\vdots$$

$$E(y_n) = x_{n1}\beta_1 + x_{n2}\beta_2 \ldots x_{nk}\beta_k$$

or in matrix form $E(y) = X\beta$ where the element, x_{ij}, of the X matrix is 0 if the weight is absent, and 1 or -1 depending on the direction of the comparison. In this note we shall adopt the convention of using just the signs so that, for example, all possible comparisons (ignoring direction) of 4 nominally equal objects will have the representation.

β_1	β_2	β_3	β_4
+	-		
+		-	
+			-
	+	-	
	+		-
		+	-

$$X = \begin{pmatrix} 1 & -1 & 0 & 0 \\ 1 & 0 & -1 & 0 \\ 1 & 0 & 0 & -1 \\ 0 & 0 & 1 & -1 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & -1 \end{pmatrix}$$

In the least squares analysis one forms the normal equations

$$X'X\hat{\beta} = X'y$$

where the entries in X'X are merely the sums of squares and sums of cross products of the columns of X. In the above case, one gets

$$\begin{pmatrix} 3 & -1 & -1 & -1 \\ -1 & 3 & -1 & -1 \\ -1 & -1 & 3 & -1 \\ -1 & -1 & -1 & 3 \end{pmatrix} \hat{\beta} = \begin{pmatrix} y_1 + y_2 + y_3 \\ -y_1 + y_4 + y_5 \\ -y_2 + y_4 + y_5 \\ -y_3 + y_4 + y_6 \end{pmatrix}$$

where $\hat{\beta}$ is the column vector with elements $\hat{\beta}1$, $\hat{\beta}2$, $\hat{\beta}_3$, $\hat{\beta}_4$, the caret being used to denote the fact that the values are functions of the observations, and not the sought-after values, β.

It can easily be verified in this case that the system of equations is not of full rank (e.g., the column totals are zero) and this is a property of all designs where only differences are measured. In mass calibration, one has one or more standards whose value can be taken as known and these provide the restraint on the system needed to give a unique set of answers. Usually these involve a starting kilogram or a unique summation such as 5 + 3 + 2 which has been determined in a previous series or is the initial unit value for an ascending series such as the 1, 2, 3, 5 series. One can write the restraint in the form

$$r_1\beta_1 + r_2\beta_2 \ . \ . \ . + r_k\beta_k = m \qquad (2)$$

and use the method of Lagrangian multipliers (with multipliers 2λ) to minimize the function

$$\Phi = \sum(\text{deviations})^2 + 2\Sigma(r_1\beta_1 + \ . \ . \ . \ r_k\beta_k - m) \qquad (3)$$

The normal equations now contain an additional "unknown," namely λ and written out in full are as follows:

$$\begin{aligned} &\sum x_1^2\hat{\beta}_1 + \sum x_1x_2\hat{\beta}_2 \ . \ . \ . \sum x_1x_k\hat{\beta}_k + r_1\lambda = \sum x_1y \qquad (4)\\ &\sum x_2x_1\hat{\beta}_1 + \sum x_2^2\hat{\beta}_2 \ . \ . \ . \sum x_2x_k\hat{\beta}_k + r_2\lambda = \sum x_2y \\ &\ . \\ &\ . \\ &\ . \\ &\sum x_kx_1\hat{\beta}_1 + \sum x_kx_2\hat{\beta}_2 \ . \ . \ . \sum x_k^2\hat{\beta}_k + r_k\lambda = \sum x_ky \\ &r_1\hat{\beta}_1 + r_2\hat{\beta}_2 \ . \ . \ . \ r_k\hat{\beta}_k = m \end{aligned}$$

where

$$\sum x_i x_j = \sum_{k=1}^{n} x_{ik} x_{jk}$$

$$\sum x_i y = \sum_{k=1}^{n} x_{ik} y_k$$

or in matrix notation

$$\begin{bmatrix} X'X & r \\ r' & 0 \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \lambda \end{bmatrix} = \begin{bmatrix} X'y \\ m \end{bmatrix} \qquad (5)$$

The solution may be written out formally as follows:

$$\begin{bmatrix} \hat{\beta} \\ \lambda \end{bmatrix} = \begin{bmatrix} C & h \\ h' & 0 \end{bmatrix} \begin{bmatrix} X'y \\ m \end{bmatrix} = \begin{bmatrix} CX' & h \\ h'X' & 0 \end{bmatrix} \begin{bmatrix} y \\ m \end{bmatrix} \qquad (6)$$

To facilitate computation it is convenient to have the values, $\hat{\beta}$, written out as linear functions of the y's and m, i.e., $\beta = [CX', h]\, y$. This leads to a set of multipliers of the observations of the form m

$$\hat{\beta}_1 = g_{11}y_1 + g_{12}y_2 \cdots g_{1n}y_n + h_1 m \qquad (7)$$

$$\vdots$$

$$\hat{\beta}_k = g_{k1}y_1 + g_{k2}y_2 \cdots g_{kn}y_n + h_k m$$

These multipliers, g_{ij} and h_i, are given in Appendix B in transposed form for some of the designs. The matrix C is important because the variances and covariances of the estimates are given by

$$\text{Variance } (\hat{\beta}_i) = C_{ii}\sigma^2, \quad \text{Covariance } (\hat{\beta}_i, \hat{\beta}_j) = C_{ij}\sigma^2 \qquad (8)$$

The quantity, σ^2, is the variance (square of the long run value of the standard deviation) associated with the process. In a set of n observations on k items and r = 1 restraints one has n - k + r = n - k + 1 degrees of freedom for a standard deviation, s, formed by

$$s^2 = \frac{1}{n - k + 1} \sum (\text{deviations})_i^2 \qquad (9)$$

$$(\text{deviation})_i = y_i - (x_{i1}\hat{\beta}_1 + x_{i2}\hat{\beta}_2 \ . \ . \ . \ x_{ik}\hat{\beta}_k)$$

One can write these deviations as a function of the observations by noting that the predicted values are just $X\beta$ and the deviations are thus

$$\text{dev} = y - X\hat{\beta} = y - X[CX', h] \begin{pmatrix} Y \\ m \end{pmatrix} = y - [XCX', 0] \begin{pmatrix} Y \\ m \end{pmatrix} \qquad (10)$$

$$= [I - XCX']y$$

which can be written as

$$\text{dev}_1 = d_{11}y_1 + d_{12}y_2 \ . \ . \ . \ d_{1n}y_n \qquad (11)$$

.

.

.

$$\text{dev}_n = d_{n1}y_1 + d_{n2}y_2 \ . \ . \ . \ d_{nn}y_n$$

The array of coefficients, d_{ij}, is given in Appendix B for some of the designs. Weights are often used in combination and one needs to know the standard deviation for the various sums. For a sum of two items, β_i and β_j, one has

$$\text{Var}(\hat{\beta}_i + \hat{\beta}_j) = \text{Var}(\hat{\beta}_i) + \text{Var}(\hat{\beta}_j) = 2\text{Cov}(\hat{\beta}_i, \hat{\beta}_j)$$

and for a linear combination

$$L = \ell_1\hat{\beta}_1 + \ell_2\hat{\beta}_2 \ . \ . \ . \ \ell_k\hat{\beta}_k \qquad (12)$$

$$\text{Variance (L)} = \ell' C \ell \sigma^2$$

where $\ell' = (\ell_1 \ . \ . \ . \ \ell_k)$, C comes from the inverse of the matrix of normal equations [see eq (6)].

DESIGNS FOR WEIGHING

The criteria for good weighing designs depend to some extent on the use intended for the resulting values. For example, if the weights are to be used independently of each other, then one would want the standard deviation [$\sigma \ C_{ii}$ from formula (8)] for the value for each unknown weight to be the minimum possible. If the weights are to be used in combination, then one wants the variance of all appropriate linear functions to be as small as possible.

Further, the desirability of a design depends somewhat on the restraint being used. In some cases, one's judgment of a design changes depending on whether one starts with a summation as known (e.g., 5 + 3 + 2) and works down, or with a unit as known and works up (e.g., by use of a 1, 2, 3, 5 series).

Appendix B

Sample Calibration Report

The following is a full calibration report for a set of weights with denominations from 1 g to 1 mg.

U. S. DEPARTMENT OF COMMERCE
NATIONAL BUREAU OF STANDARDS
NATIONAL MEASUREMENT LABORATORY
GAITHERSBURG, MD. 20899

R E P O R T
O F
M A S S V A L U E S

COMPANY X
LOCUS, U.S.A.
SET OF MASS STANDARDS : 500 MG - 1MG
SERIAL NUMBER 00001
MANUFACTURER : COMPANY Y
SEPTEMBER 1, 1986

TEST NUMBER DEMO1

FOR THE DIRECTOR,
NATIONAL MEASUREMENT LABORATORY

JOE D. SIMMONS, CHIEF
LENGTH AND MASS DIVISION
CENTER FOR BASIC STANDARDS

INTRODUCTION

THIS DOCUMENT IS A COMPREHENSIVE REPORT COVERING THE SEQUENCE OF OPERATIONS USED TO ASSIGN MASS VALUES TO THE WEIGHTS IDENTIFIED ABOVE. IT INCLUDES A COMPLETE DESCRIPTION OF THE MEASUREMENT METHODS AND PROCEDURES WHICH WERE USED, ALL OF THE DATA, AND THE ANALYSIS OF THIS DATA. THE RESULTS ARE PRESENTED IN SEVERAL FORMATS. ASSIGNED MASS VALUES, DISPLACEMENT VOLUMES, COEFFICIENTS OF EXPANSION, UNCERTAINTIES, TOGETHER WITH THE SUMMED VALUES FOR LINEAR COMBINATIONS OF THE WEIGHTS IN EACH DECADE ARE PRESENTED AT THE END OF THE APPROPRIATE SERIES. THIS INFORMATION SHOULD BE USEFUL TO THOSE WHO MUST ASSIGN MASS VALUES TO OBJECTS OTHER THAN WEIGHTS. FOR CONVENIENCE, THE VALUES AND UNCERTAINTIES, TOGETHER WITH OTHER APPROPRIATE DATA AND COMMENTS ARE ALSO SUMMARIZED IN TABLES I AND II AT THE END OF THE REPORT. CERTAIN INTERMEDIATE PAGES ARE SUMMARIES OF STATISTICAL DATA WHICH RELATE TO THE MASS MEASUREMENT PROCESS USED TO PERFORM THIS WORK. THESE PAGES HAVE BEEN LEFT IN THE REPORT TO RETAIN CONTINUITY. COPIES OF THESE PAGES BECOME PART OF A COLLECTION OF STATISTICAL DATA WHICH REFLECTS THE MEASUREMENT PROCESS PERFORMANCE OVER A PERIOD OF TIME. SUCH A COLLECTION HAS BEEN USED TO ESTABLISH THE CONTROL LIMITS FOR ACCEPTING THE RESULTS OF THIS MEASUREMENT. THESE COLLECTIONS ARE OPEN FOR INSPECTION AT OUR FACILITY.

THE MASS MEASUREMENT SYSTEM

THE MASS MEASUREMENT SYSTEM WITHIN THIS COUNTRY CONSISTS OF ALL OF THE MEASUREMENT PROCESSES WHICH RELY, DIRECTLY OR INDIRECTLY, ON MASS MEASUREMENTS TO ACCOMPLISH A WIDE VARIETY OF ENDEAVORS. IN ORDER FOR THIS SYSTEM TO FUNCTION PROPERLY, EVERYONE WHO MAKES MEASUREMENTS MUST BE ABLE TO VERIFY THAT HIS MEASUREMENT PROCESS PRODUCES CONSISTENT RESULTS WHICH ARE COMPATIBLE WITH HIS PARTICULAR REQUIREMENTS. THE WEIGHTS COVERED BY THIS REPORT, TOGETHER WITH THE ASSIGNED VALUES AND THE APPROPRIATE UNCERTAINTIES FOR THESE VALUES, PROVIDE IN PART A BASIS FOR CONSISTENT MEASUREMENTS WITHIN THIS SYSTEM OF RELATED MEASUREMENT PROCESSES.

APPROPRIATE CHARACTERIZATION OF ANY MEASUREMENT PROCESS IS FUNDAMENTAL TO VERIFYING THAT RESULTS ARE CONSISTENT WITH THE END REQUIREMENT WITH RESPECT TO CORRECTNESS AND ECONOMY OF THE MEASUREMENT EFFORT. WITHOUT THIS INFORMATION, THE BENEFITS OF OWNERSHIP OF THESE WEIGHTS MAY BE COMPLETELY ILLUSORY. THE ASSIGNED UNCERTAINTIES IN THIS REPORT ARE DESCRIPTIVE OF OUR MASS MEASUREMENT PROCESS. EFFECTIVENESS OF THE TRANSFER OF THE UNIT FROM ONE FACILITY TO ANOTHER SHOULD BE VERIFIED BY AN INDEPENDENT TEST. IT IS PRESUMED THAT THESE WEIGHTS WILL BE USED IN A SIMILARLY WELL-CHARACTERIZED MEASUREMENT PROCESS SO THAT THE STATISTICAL PARAMETERS OF BOTH PROCESSES CAN BE COMBINED TO PROVIDE A REALISTIC ESTIMATE OF THE UNCERTAINTY OF THE MASS UNIT AS ACTUALLY REALIZED IN ANOTHER FACILITY. A COMPREHENSIVE SERVICE DIRECTED TOWARD THE EVALUATION OF A PARTICULAR MASS MEASUREMENT PROCESS IS AVAILABLE THROUGH THE MASS MEASUREMENT ASSURANCE PROGRAM OF THE NATIONAL BUREAU OF STANDARDS.

WEIGHING DESIGN

ONLY DIFFERENCES IN MASS CAN BE MEASURED, THEREFORE THE MASS VALUES FOR THE 'UNKNOWN' WEIGHTS MUST BE DETERMINED BY COMPARISON WITH OTHER WEIGHTS WHICH HAVE ACCEPTED MASS VALUES. THE 'UNKNOWN' WEIGHTS TOGETHER WITH 'CHECK STANDARDS', ARE GROUPED AND INTERCOMPARED ACCORDING TO THE DESIGN SCHEDULE GIVEN AT THE BEGINNING OF EACH SERIES OF WEIGHINGS. THE FIRST SERIES CONTAINS STANDARDS WHICH PROVIDE THE STARTING VALUES FOR THE SERIES OF WEIGHINGS AND PROVIDE THE TIE POINT FOR CONSISTENCY THROUGHOUT THE MEASUREMENT SYSTEM. THE WEIGHING METHOD USED, I.E., DOUBLE SUBSTITUTION, TRANSPOSITION, ETC., IS INDICATED ALONG WITH THE OBSERVED DATA. IN THE COMPUTATIONS, THE DISPLACEMENT VOLUMES ARE TREATED EXPLICITLY, USING THE DATA LISTED IN THE REPORT. IN ALL CASES, A REDUNDANCY IN THE NUMBER OF MEASUREMENTS PROVIDES A MEANS FOR CHECKING ON THE PRECISION OF THE PROCESS.

WHEN THERE ARE MORE EQUATIONS THAN 'UNKNOWNS', NOT ALL OBSERVATIONAL EQUATIONS CAN BE SATISFIED EXACTLY AND THE METHOD OF LEAST SQUARES IS USED TO PROVIDE ESTIMATES OF THE 'UNKNOWN' VALUES. THIS METHOD LEADS TO ESTIMATORS WHICH ARE LINEAR FUNCTIONS OF THE DATA AND WHICH HAVE STANDARD DEVIATIONS READILY CALCULATED FROM THE COEFFICIENTS OF THE LINEAR FUNCTIONS AND THE STANDARD DEVIATION OF AN INDIVIDUAL MEASUREMENT. THE 'CHECK STANDARD' IS ALSO TREATED AS AN UNKNOWN AND THE AGREEMENT OF THE CURRENT RESULT WITH THE ACCEPTED VALUE PROVIDES A TEST OF THE ADEQUACY OF THE CURRENT DATA. THIS SAME CHECK STANDARD IS MEASURED WITH EACH TEST OF UNKNOWNS AND THE COLLECTION OF VALUES OVER TIME IS USED TO EVALUATE THE PERFORMANCE OF THE MEASUREMENT PROCESS.

IN THE CASE OF THE SERIES WHICH INCLUDES THE KNOWN STANDARDS, THE ACCEPTED VALUES OF THESE STANDARDS SERVE AS A RESTRAINT ON THE SOLUTION OF THE EQUATIONS FOR THE VALUES OF ALL OF THE WEIGHTS. THE RESTRAINT FOR THE SOLUTION OF SUBSEQUENT SERIES IS PROVIDED BY THE VALUES ESTABLISHED FOR ONE OR MORE WEIGHTS INCLUDED IN A PREVIOUS SERIES.

ESTIMATED VALUES FOR WEIGHTS WHICH HAVE BEEN GROUPED IN THE SAME SERIES INVOLVE THE SAME OBSERVATIONAL DATA AND ARE, IN ALMOST ALL CASES, CORRELATED. FOR EACH SERIES THERE IS A TABLE OF COMBINATIONS TOGETHER WITH THE APPROPRIATE UNCERTAINTY FOR EACH COMBINATION.

PROCESS CONTROL

THE STANDARD DEVIATION, AS COMPUTED FROM THE LEAST SQUARES SOLUTION, PROVIDES A CHECK ON THE SHORT TERM, OR 'WITHIN-RUN' PROCESS PRECISION. AN AVERAGE OF A NUMBER OF THESE STANDARD DEVIATIONS IS TAKEN AS THE ACCEPTED WITHIN-RUN STANDARD DEVIATION OF THE PROCESS AND IS USED AS A REFERENCE VALUE FOR SURVEILLANCE OF THE PROCESS PRECISION. THE VALUES OBTAINED FOR THE 'CHECK STANDARD' PROVIDE, AS TIME GOES ON, A SEQUENCE OF VALUES THAT REALISTICALLY REFLECTS THE VARIATIONS WHICH BESET PRECISE MEASUREMENTS. COLLECTIONS OF VALUES FOR BOTH THE WITHIN-RUN PRECISION AND THE VALUE OBTAINED FOR THE 'CHECK STANDARD' SHOULD

POSSESS THE PROPERTIES OF RANDOMNESS ASSOCIATED WITH INDEPENDENT MEASUREMENTS FROM A STABLE PROBABILITY DISTRIBUTION. THE REPORTED 'F RATIO' AND 'T VALUE' ARE TESTS OF THE VALUES FROM THE CURRENT RUN FOR CONFORMITY TO THEIR RESPECTIVE DISTRIBUTIONS AND IF SATISFACTORY ARE TAKEN AS EVIDENCE THAT THE PROCESS IS IN CONTROL AND THAT PREDICTIVE STATEMENTS REGARDING UNCERTAINTY ARE VALID.

CONTROL CHARTS ON THE WITHIN-RUN PROCESS PRECISION AND THE VALUES OBTAINED FOR THE CHECK STANDARD ARE KEY ELEMENTS IN MONITORING THE STATE OF CONTROL OF ANY PRECISE MASS MEASUREMENT PROCESS. IN ADDITION TO PROVIDING A BASIS FOR JUDGMENT AS TO THE ADEQUACY OF A GIVEN PROCESS FOR A PARTICULAR REQUIREMENT, THESE DATA PROVIDE A MEANS TO JUDGE THE IMPORTANCE OF LONG TERM, OR 'BETWEEN-RUN' VARIABILITY WHICH CAN BE CHARACTERIZED BY THE STANDARD DEVIATION OF THE VALUES ABOUT THE MEAN. IF THERE IS AN ADDITIONAL COMPONENT OF VARIANCE ENTERING FROM RUN TO RUN, THIS STANDARD DEVIATION WILL BE LARGER THAN CAN BE ACCOUNTED FOR BY THE WITHIN-RUN VARIABILITY. CORRELATION STUDIES, AS WELL AS SUPPLEMENTAL EXPERIMENTS, ARE USED TO DETECT AND REDUCE THE MAGNITUDE OF SIGNIFICANT SYSTEMATIC EFFECTS. APPROPRIATE ACTION, E.G., ADDITIONAL EMPIRICAL CORRECTIONS OR CHANGES IN TECHNIQUE, CAN REDUCE THE EFFECTS FROM KNOWN SOURCES OF SYSTEMATIC VARIABILITY TO A MAGNITUDE WHICH IS NO LONGER IDENTIFIABLE IN THE DATA. IN THE CASES WHERE A SIGNIFICANT LONG TERM, OR BETWEEN-RUN, COMPONENT REMAINS THE UNCERTAINTY HAS BEEN APPROPRIATELY ADJUSTED.

SERIES OF MEASUREMENTS JUDGED AS OUT OF CONTROL RELATIVE TO THE APPROPRIATE PARAMETER ARE CAREFULLY EXAMINED. IF RERUNS WERE NECESSARY IN THE COURSE OF THIS WORK, THE 'OUT OF CONTROL' SERIES, WITH REMARKS AS APPROPRIATE, ARE ATTACHED AT THE END OF THE REPORT FOR YOUR INFORMATION.

UNCERTAINTY

IT IS ASSUMED THAT THE PRESENT 'ACCEPTED VALUES' OF TWO NBS STANDARDS AT THE 1 KILOGRAM LEVEL, DESIGNATED N1 AND N2, ARE WITHOUT ERROR. ESTIMATES OF THE UNCERTAINTY OF THE ACCEPTED VALUES OF THE NBS STANDARDS RELATIVE TO THE INTERNATIONAL PROTOTYPE KILOGRAM CAN BE PROVIDED ON REQUEST. HOWEVER, THESE ESTIMATES HAVE NO REAL MEANING IN EITHER NATIONAL OR INTERNATIONAL COMPARISON. THIS IS BECAUSE OF THE LACK OF SUFFICIENT DATA TO PROVIDE A REALISTIC ESTIMATE OF THE UNCERTAINTY IN THE VALUES ASSIGNED TO THE PROTOTYPE KILOGRAMS K20 AND K4, PARTICULARLY IN REGARD TO LONG TERM, OR BETWEEN-RUN VARIABILITY. CHANGES IN THE ACCEPTED VALUES FOR THE NBS STANDARDS AT THE KILOGRAM LEVEL, AS AND WHEN THEY OCCUR, WILL BE REPORTED IN THE SCIENTIFIC PAPERS OF THE BUREAU AND WILL BE GIVEN WIDE DISTRIBUTION. IN CASES WHERE SUCH CHANGES MAY BE OF IMPORTANCE, OR WHERE CONTINUITY IS DESIRED, INSTRUCTIONS WILL BE INCLUDED FOR UP-DATING PREVIOUSLY REPORTED VALUES. WHEN THE VALUES REPORTED ARE BASED ON THE ACCEPTED VALUES OF STANDARDS OTHER THAN STANDARDS N1 AND N2 MENTIONED ABOVE, THE UNCERTAINTY OF THE ACCEPTED VALUE OF THE STANDARD BECOMES A SYSTEMATIC ERROR IN THE ASSIGNMENT OF VALUES TO OTHER STANDARDS AND IS INCLUDED IN THE REPORT.

A BALANCE UNDER STABLE OPERATING CONDITIONS WILL EXHIBIT A CERTAIN CHARACTERISTIC VARIABILITY WHICH CAN BE DESCRIBED BY THE STANDARD DEVIATION FOR SUCH MEASUREMENTS. THE VALUE FOR A PARTICULAR WEIGHT DETERMINED IN REPEATED TESTS WITH THE SAME WEIGHING DESIGN WILL HAVE ITS OWN STANDARD DEVIATION WHICH WILL BE SOME FUNCTION OF THE BALANCE PRECISION AND (POSSIBLY) OF THE BETWEEN-RUN COMPONENT. AS AN OUTER LIMIT OF THE DISTRIBUTION OF RANDOM ERRORS, THREE TIMES THE STANDARD DEVIATION IS USED. SYSTEMATIC ERRORS DUE TO THE PROCEDURES USED OR TO ENVIRONMENTAL EFFECTS ARE LARGELY BALANCED OUT AND CAN USUALLY BE REGARDED AS NEGLIGIBLE. WHEN A NON-NEGLIGIBLE BOUND TO THE POSSIBLE EFFECT FROM KNOWN SOURCES IS AVAILABLE, IT IS CALCULATED AND REPORTED SEPARATELY, E.G., THE UNCERTAINTY OF ACCEPTED VALUE AT OTHER THAN THE 1 KILOGRAM LEVEL. THE DISTRIBUTION IMPLIED BY THE RANDOM ERRORS MAY THUS BE CENTERED SOMEWHERE IN THE RANGE GIVEN BY THE BOUNDS TO THE SYSTEMATIC ERROR. THE TOTAL UNCERTAINTY IS TAKEN AS THE SUM OF THESE TWO COMPONENTS.

THE UNCERTAINTY ASSOCIATED WITH THE ASSIGNED VALUE CAN BE THOUGHT OF AS A BOUND TO THE DEPARTURE OF THE ASSIGNED VALUE FROM A HYPOTHETICAL AVERAGE VALUE THAT WOULD BE OBTAINED IF IT WERE POSSIBLE TO REPEAT THE MEASUREMENT MANY TIMES OVER A WIDE VARIETY OF CONDITIONS, E.G., SUBSTITUTE THE WEIGHT FOR ONE OF THE CHECK STANDARDS. THIS MEANS THAT THE UNCERTAINTY BAND CENTERED ON THE VALUES OBTAINED FROM EACH OF TWO MEASUREMENTS OF THE SAME OBJECT OVER SOME ARBITRARY TIME INTERVAL SHOULD ALMOST ALWAYS OVERLAP. IN OTHER WORDS, WHILE A SECOND MEASUREMENT WILL PRODUCE A DIFFERENT VALUE, THIS VALUE WILL ONLY RARELY DIFFER FROM THE FIRST VALUE BY MORE THAN THE SUM OF THE TWO UNCERTAINTIES. THE UNCERTAINTY BANDS ARE NOT EXPECTED TO OVERLAP IF SOME EVENT HAS OCCURRED IN THE TIME INTERVAL BETWEEN THE TWO MEASUREMENTS WHICH WILL CHANGE THE MASS OF THE OBJECT, E.G., ABRASIONS, ABUSE, CORROSION, IMPROPER CLEANING AND THE LIKE.

THE UNCERTAINTY IN ASSIGNED VALUE CONTAINED IN THIS REPORT BECOMES A SYSTEMATIC EFFECT FOR THE MEASUREMENT PROCESS IN WHICH THESE WEIGHTS ARE TO BE USED. IN THE ABSENCE OF OTHER SIGNIFICANT SYSTEMATIC EFFECTS IN THE USER'S MEASUREMENT PROCESS (A CONDITION WHICH MUST BE DEMONSTRATED) THE UNCERTAINTY OF THE VALUE ASSIGNED BY THE USER IS AN APPROPRIATE COMBINATION OF THE SYSTEMATIC ERROR IN THE STANDARD AND THE RANDOM COMPONENT ASSOCIATED WITH HIS PROCESS. IF THE MEASUREMENT PROCESSES ARE IN CONTROL AND APPROPRIATE UNCERTAINTIES ARE ASSIGNED, THE VALUES PRODUCED BY DIFFERENT MEASUREMENT FACILITIES WILL HAVE OVERLAPPING UNCERTAINTY BANDS AS DESCRIBED ABOVE. ONE CANNOT DISCUSS DIFFERENCES IN VALUES FOR THE SAME OBJECT OBTAINED BY DIFFERENT FACILITIES WITH ANY DEGREE OF SERIOUSNESS UNLESS EACH VALUE IS ACCOMPANIED BY A REALISTIC UNCERTAINTY STATEMENT.

REFERENCES

THE FOLLOWING REFERENCES ARE SUGGESTED FOR DETAILED DESCRIPTION OF PORTIONS OF THIS REPORT , AND FOR GENERAL INFORMATION CONCERNING THE MASS MEASUREMENT PROCESS:

1. PONTIUS, P. E., AND CAMERON, J. M.
REALISTIC UNCERTAINTIES AND THE MASS MEASUREMENT PROCESS
NAT. BUR. STAND. (U.S.), MONOGR. 103
(AUG. 15, 1967)

2. PONTIUS, P. E.
MEASUREMENT PHILOSOPHY OF THE PILOT PROGRAM FOR MASS CALIBRATION
NAT. BUR. STAND. (U.S.) TECH. NOTE 288
(MAY 6, 1966)

3. BOWMAN, H. A., AND SCHOONOVER, R. M. WITH APPENDIX BY MILDRED JONES
PROCEDURE FOR HIGH PRECISION DENSITY DETERMINATIONS BY HYDROSTATIC WEIGHING
J. RES. NAT. BUR. STAND. (U.S.) 71C. ENGINEERING AND INSTRUMENTATION
NO. 3, 179-198 (JULY-AUG. 1967)

4. NATRELLA, M. B.
EXPERIMENTAL STATISTICS
NAT. BUR. STAND. (U.S.) HANDBOOK 91
(AUGUST 1, 1963)

5. KU, H. H.
PRECISION MEASUREMENT AND CALIBRATION - SELECTED NBS PAPERS ON STATISTICAL CONCEPTS AND PROCEDURES
NAT. BUR. STAND. (U.S.) SPEC. PUBL. 300
VOL. 1 (FEB. 1969)

6. PONTIUS, P. E.
MASS AND MASS VALUES
NAT. BUR. STAND. (U.S.) MONOGR. 133
(JAN. 1974)

7. CAMERON, J. M., CROARKIN, M. C. AND RAYBOLD, R. C.
DESIGNS FOR THE CALIBRATION OF STANDARDS OF MASS
NAT. BUR. STAND. (U.S.) TECH. NOTE 952
(JUNE 1977)

8. VARNER, R. N., AND RAYBOLD, R. C.
NATIONAL BUREAU OF STANDARDS MASS CALIBRATION COMPUTER SOFTWARE
NAT. BUR. STAND. (U.S.) TECH. NOTE 1127
(JULY 1980)

BALANCE 2
OPERATOR 39
ACCEPTED WITHIN STANDARD DEVIATION OF THE PROCESS 0.00240 MG
ACCEPTED BETWEEN STANDARD DEVIATION OF THE PROCESS 0.00000 MG

CALIBRATION DESIGN 41
RESTRAINT VECTOR 1 0 0 0
MASS CORRECTION OF RESTRAINT -0.06971 MG
VOLUME OF WEIGHTS BEING USED IN RESTRAINT AT 22.92 0.11990 CM3
SYSTEMATIC ERROR IN THE RESTRAINT 0.00087 MG
3 STANDARD DEVIATION LIMIT FOR RANDOM ERROR AFFECTING RESTRAINT 0.00000 MG

CHECK STANDARD USED 10
CHECK STANDARD VECTOR 0 1 0 0
ACCEPTED MASS CORRECTION OF CHECK STANDARD -0.00740 MG
REPORT VECTOR 0 0 0 0

TEST CONDITIONS	BEFORE	AFTER	AVERAGE
CORRECTED TEMPERATURE IN DEGREES C	22.85	23.00	22.92
CORRECTED PRESSURE IN MM HG	759.028	758.828	758.928
CORRECTED HUMIDITY IN PERCENT	27.50	27.30	27.40
COMPUTED AIR DENSITY IN MG/CM3	1.1883	1.1874	1.1879
TEMPERATURE CORRECTION	0.00	0.00	
PRESSURE CORRECTION	-0.172	-0.172	
HUMIDITY CORRECTION	0.00	0.00	
OBSERVED TEMPERATURE IN DEGREES C	22.85	23.00	
OBSERVED PRESSURE IN MM HG	759.200	759.000	
OBSERVED HUMIDITY IN PERCENT	27.50	27.30	

	WEIGHTS BEING TESTED	NOMINAL VALUE G	DENSITY G/CM3 AT 20C	COEFFICIENT OF EXPANSION	ACCEPTED CORRECTION MG
NB	1 G	1.0000	8.3406	.000040	-0.06971
AA	1 G	1.0000	7.8704	.000045	-0.00740
	1 G	1.0000	8.4000	.000054	
SUM	1 G	1.0000	16.6000	.000020	

BALANCE 2
OPERATOR 39

CALIBRATION DESIGN 41

	GRAMS 1	1	1	1
A 1	+	-		
A 2	+		-	
A 3	+			-
A 4		+	-	
A 5		+		-
A 6			+	-
R	+			

OBSERVATIONS IN DIVISIONS
DOUBLE SUBSTITUTION ONE PAN BALANCE

A 1	8.1660	8.2200	13.2310	13.1810
A 2	8.1680	8.2030	13.2160	13.1840
A 3	8.1690	8.1500	13.1650	13.1840
A 4	8.2160	8.2060	13.2150	13.2340
A 5	8.2200	8.1500	13.1620	13.2350
A 6	8.2010	8.1460	13.1630	13.2100

BALANCE 2
OPERATOR 39

CALIBRATION DESIGN 41

SENSITIVITY WEIGHT
MASS 4.99198 MG
VOLUME 0.00185 CM3 AT 20 C
COEFFICIENT OF EXPANSION 0.000069
S*=S-PV(S)= 4.98978 MG

	A(I) (MG)	DELTA(I) (MG)	AVERAGE SENSITIVITY (MG/DIV)	DRIFT(I) (MG)	OBSERVED SENSITIVITY (MG/DIV)
A 1	-0.05177	-0.00087	0.99558	0.00199	0.99616
A 2	-0.03335	0.00062	0.99558	0.00149	0.99567
A 3	0.01892	0.00025	0.99558	0.00000	0.99497
A 4	0.01444	-0.00249	0.99558	0.00448	0.99706
A 5	0.07118	0.00162	0.99558	0.00149	0.99587
A 6	0.05077	-0.00187	0.99558	-0.00398	0.99378

ITEM (G)	CORRECTION (MG)	VOLUME (AT T) (CM3)	SYSTEMATIC ERROR (MG)	3 S.D. LIMIT (MG)	UNCERTAINTY LIMIT (MG)
1.0000	-0.06971	0.11990	0.00087	0.00000	0.00087
1.0000	-0.01029	0.12707	0.00087	0.00509	0.00596
1.0000	-0.03673	0.11906	0.00087	0.00509	0.00596
1.0000	-0.15925	0.06023	0.00087	0.00509	0.00596

TEMPERATURE T= 22.92 C

RESTRAINT FOR FOLLOWING SERIES
RESTRAINT VECTOR 0 0 0 1
MASS CORRECTION -0.15925 MG
VOLUME AT 20 C 0.06023 CM3
SYSTEMATIC ERROR 0.00087 MG
3 STANDARD DEVIATION LIMIT 0.00509 MG

BALANCE 2
OPERATOR 39
MAXIMUM LOAD 1.0000 G
STARTING RESTRAINT NUMBER 4

CALIBRATION DESIGN 41

PRECISION CONTROL

OBSERVED STANDARD DEVIATION OF THE PROCESS 0.00212 MG
ACCEPTED STANDARD DEVIATION OF THE PROCESS 0.00240 MG
DEGREES OF FREEDOM 3
F RATIO 0.782

F RATIO IS LESS THAN 3.79 (CRITICAL VALUE FOR PROBABILITY = .01).
THEREFORE THE STANDARD DEVIATION IS IN CONTROL.

CHECK STANDARD VECTOR 0 1 0 0
CHECK STANDARD USED 10
ACCEPTED MASS CORRECTION OF CHECK STANDARD -0.00740 MG
OBSERVED CORRECTION OF CHECK STANDARD -0.01029 MG
STANDARD DEVIATION OF THE OBSERVED CORRECTION 0.00170 MG
T VALUE -1.70

ABSOLUTE VALUE OF T IS LESS THAN 3.
THEREFORE CHECK STANDARD IS IN CONTROL.

TEST CONDITIONS	BEFORE	AFTER	AVERAGE
CORRECTED TEMPERATURE IN DEGREES C	22.85	23.00	22.92
CORRECTED PRESSURE IN MM HG	759.028	758.828	758.928
CORRECTED HUMIDITY IN PERCENT	27.50	27.30	27.40
COMPUTED AIR DENSITY IN MG/CM3	1.1883	1.1874	1.1879
TEMPERATURE CORRECTION	0.00	0.00	
PRESSURE CORRECTION	-0.172	-0.172	
HUMIDITY CORRECTION	0.00	0.00	
OBSERVED TEMPERATURE IN DEGREES C	22.85	23.00	
OBSERVED PRESSURE IN MM HG	759.200	759.000	
OBSERVED HUMIDITY IN PERCENT	27.50	27.30	

BALANCE 13
OPERATOR 39
ACCEPTED WITHIN STANDARD DEVIATION OF THE PROCESS 0.00050 MG
ACCEPTED BETWEEN STANDARD DEVIATION OF THE PROCESS 0.00000 MG

CALIBRATION DESIGN 62
RESTRAINT VECTOR 1 1 1 0 0 0
MASS CORRECTION OF RESTRAINT -0.15925 MG
VOLUME OF WEIGHTS BEING USED IN RESTRAINT AT 23.28 0.06024 CM3
SYSTEMATIC ERROR IN THE RESTRAINT 0.00087 MG
3 STANDARD DEVIATION LIMIT FOR RANDOM ERROR AFFECTING RESTRAINT 0.00509 MG

CHECK STANDARD USED 12
CHECK STANDARD VECTOR 0 0 0 0 1 0
ACCEPTED MASS CORRECTION OF CHECK STANDARD -0.00854 MG
REPORT VECTOR 1 1 1 1 0 0

TEST CONDITIONS	BEFORE	AFTER	AVERAGE
CORRECTED TEMPERATURE IN DEGREES C	23.25	23.30	23.28
CORRECTED PRESSURE IN MM HG	758.828	758.429	758.629
CORRECTED HUMIDITY IN PERCENT	27.30	24.10	25.70
COMPUTED AIR DENSITY IN MG/CM3	1.1863	1.1859	1.1861
TEMPERATURE CORRECTION	0.00	0.00	
PRESSURE CORRECTION	-0.172	-0.171	
HUMIDITY CORRECTION	0.00	0.00	
OBSERVED TEMPERATURE IN DEGREES C	23.25	23.30	
OBSERVED PRESSURE IN MM HG	759.000	758.600	
OBSERVED HUMIDITY IN PERCENT	27.30	24.10	

WEIGHTS BEING TESTED	NOMINAL VALUE G	DENSITY G/CM3 AT 20C	COEFFICIENT OF EXPANSION	ACCEPTED CORRECTION MG
500MG	0.5000	16.6000	.000020	
300MG	0.3000	16.6000	.000020	
200MG	0.2000	16.6000	.000020	
100MG	0.1000	16.6000	.000020	
AN/ 100MG	0.1000	8.4100	.000039	-0.00854
SUM 100MG	0.1000	8.1788	.000049	

BALANCE 13
OPERATOR 39

CALIBRATION DESIGN 62

	MG 500	300	200	100	100	100
A 1	+	-	-	+	-	
A 2	+	-	-		+	-
A 3	+	-	-	-		+
A 4	+	-	-			
A 5	+		-	-	-	-
A 6		+	-	+	-	-
A 7		+	-	-	+	-
A 8		+	-	-	-	+
A 9			+	-	-	
A 10			+	-		-
A 11			+		-	-
R	+	+	+			

OBSERVATIONS IN DIVISIONS
DIRECT READING

A 1	20.4000	10020.4004
A 2	-16.5000	
A 3	6.8500	
A 4	4.1000	
A 5	-28.8500	
A 6	8.8000	
A 7	-26.7000	
A 8	14.2500	
A 9	-25.3000	
A 10	-45.2000	
A 11	-28.3000	9971.7002

BALANCE 13
OPERATOR 39

CALIBRATION DESIGN 62

SENSITIVITY WEIGHT
MASS 10.00000 MG
VOLUME 0.00000 CM3 AT 20 C
COEFFICIENT OF EXPANSION 0.000000
S*=S-PV(S)= 10.00000 MG
ACCEPTED SENSITIVITY = 0.00100 MG/DIV
OBSERVED SENSITIVITY = 0.00100 MG/DIV
T-TEST = 0.000

	A(I) (MG)	DELTA(I) (MG)	OBSERVED SENSITIVITY (MG/DIV)
A 1	0.02040	-0.00081	0.00100
A 2	-0.01650	-0.00016	
A 3	0.00685	-0.00003	
A 4	0.00410	0.00018	
A 5	-0.02885	0.00083	
A 6	0.00880	0.00019	
A 7	-0.02670	-0.00073	
A 8	0.01425	-0.00029	
A 9	-0.02530	-0.00038	
A 10	-0.04520	-0.00003	
A 11	-0.02830	-0.00042	0.00100

ITEM (G)	CORRECTION (MG)	VOLUME (AT T) (CM3)	SYSTEMATIC ERROR (MG)	3 S.D. LIMIT (MG)	UNCERTAINTY LIMIT (MG)
0.5000	-0.07767	0.03012	0.00043	0.00257	0.00300
0.3000	-0.04280	0.01807	0.00026	0.00159	0.00185
0.2000	-0.03879	0.01205	0.00017	0.00109	0.00127
0.1000	0.00171	0.00602	0.00009	0.00074	0.00082
0.1000	-0.00862	0.01189	0.00009	0.00074	0.00082
0.1000	0.01204	0.01223	0.00009	0.00074	0.00082

TEMPERATURE T= 23.28 C

RESTRAINT FOR FOLLOWING SERIES
RESTRAINT VECTOR 0 0 0 0 0 1
MASS CORRECTION 0.01204 MG
VOLUME AT 20 C 0.01223 CM3
SYSTEMATIC ERROR 0.00009 MG
3 STANDARD DEVIATION LIMIT 0.00074 MG

BALANCE 13
OPERATOR 39

CALIBRATION DESIGN 62

SUM (MG)	WEIGHTS USED FOR THE LINEAR COMBINATIONS MG					
	500	300	200	100	100	100
1000	+	+	+			
900	+	+		+		
800	+	+				
700	+		+			
600	+			+		
500		+	+			
400		+		+		
300		+				
200			+			
100				+		

VALUES AND UNCERTAINTIES FOR COMBINATIONS OF WEIGHTS
(UNCERTAINTY IS 3 STANDARD DEVIATION LIMIT PLUS ALLOWANCE FOR SYSTEMATIC ERROR.)

SUM (MG)	CORR (MG)	SYSTEMATIC (MG)	3 S.D. ERROR (MG)	UNCERTAINTY LIMIT (MG)
1000	-0.15925	0.00087	0.00509	0.00596
900	-0.11875	0.00078	0.00463	0.00542
800	-0.12046	0.00070	0.00409	0.00479
700	-0.11646	0.00061	0.00359	0.00420
600	-0.07595	0.00052	0.00312	0.00364
500	-0.08159	0.00043	0.00257	0.00300
400	-0.04108	0.00035	0.00216	0.00251
300	-0.04280	0.00026	0.00159	0.00185
200	-0.03879	0.00017	0.00109	0.00127
100	0.00171	0.00009	0.00074	0.00082

BALANCE 13
OPERATOR 39
MAXIMUM LOAD 0.6000 G
STARTING RESTRAINT NUMBER 4

CALIBRATION DESIGN 62

PRECISION CONTROL

OBSERVED STANDARD DEVIATION OF THE PROCESS 0.00063 MG
ACCEPTED STANDARD DEVIATION OF THE PROCESS 0.00050 MG
DEGREES OF FREEDOM 6
F RATIO 1.584

F RATIO IS LESS THAN 2.81 (CRITICAL VALUE FOR PROBABILITY = .01).
THEREFORE THE STANDARD DEVIATION IS IN CONTROL.

CHECK STANDARD VECTOR 0 0 0 0 1 0
CHECK STANDARD USED 12
ACCEPTED MASS CORRECTION OF CHECK STANDARD -0.00854 MG
OBSERVED CORRECTION OF CHECK STANDARD -0.00862 MG
STANDARD DEVIATION OF THE OBSERVED CORRECTION 0.00025 MG
T VALUE -0.34

ABSOLUTE VALUE OF T IS LESS THAN 3.
THEREFORE CHECK STANDARD IS IN CONTROL.

TEST CONDITIONS	BEFORE	AFTER	AVERAGE
CORRECTED TEMPERATURE IN DEGREES C	23.25	23.30	23.28
CORRECTED PRESSURE IN MM HG	758.828	758.429	758.629
CORRECTED HUMIDITY IN PERCENT	27.30	24.10	25.70
COMPUTED AIR DENSITY IN MG/CM3	1.1863	1.1859	1.1861
TEMPERATURE CORRECTION	0.00	0.00	
PRESSURE CORRECTION	-0.172	-0.171	
HUMIDITY CORRECTION	0.00	0.00	
OBSERVED TEMPERATURE IN DEGREES C	23.25	23.30	
OBSERVED PRESSURE IN MM HG	759.000	758.600	
OBSERVED HUMIDITY IN PERCENT	27.30	24.10	

BALANCE 13
OPERATOR 39
ACCEPTED WITHIN STANDARD DEVIATION OF THE PROCESS 0.00050 MG
ACCEPTED BETWEEN STANDARD DEVIATION OF THE PROCESS 0.00000 MG

CALIBRATION DESIGN 62
RESTRAINT VECTOR 1 1 1 0 0 0
MASS CORRECTION OF RESTRAINT 0.01204 MG
VOLUME OF WEIGHTS BEING USED IN RESTRAINT AT 23.00 0.01223 CM3
SYSTEMATIC ERROR IN THE RESTRAINT 0.00009 MG
3 STANDARD DEVIATION LIMIT FOR RANDOM ERROR AFFECTING RESTRAINT 0.00074 MG

CHECK STANDARD USED 14
CHECK STANDARD VECTOR 0 0 0 0 1 0
ACCEPTED MASS CORRECTION OF CHECK STANDARD -0.00046 MG
REPORT VECTOR 1 1 1 1 0 0

TEST CONDITIONS	BEFORE	AFTER	AVERAGE
CORRECTED TEMPERATURE IN DEGREES C	22.95	23.05	23.00
CORRECTED PRESSURE IN MM HG	760.227	760.027	760.127
CORRECTED HUMIDITY IN PERCENT	30.00	30.70	30.35
COMPUTED AIR DENSITY IN MG/CM3	1.1895	1.1886	1.1891
TEMPERATURE CORRECTION	0.00	0.00	
PRESSURE CORRECTION	-0.173	-0.173	
HUMIDITY CORRECTION	0.00	0.00	
OBSERVED TEMPERATURE IN DEGREES C	22.95	23.05	
OBSERVED PRESSURE IN MM HG	760.400	760.200	
OBSERVED HUMIDITY IN PERCENT	30.00	30.70	

	WEIGHTS BEING TESTED	NOMINAL VALUE G	DENSITY G/CM3 AT 20C	COEFFICIENT OF EXPANSION	ACCEPTED CORRECTION MG
NEW	50MG	0.0500	16.6000	.000020	
	30MG	0.0300	16.6000	.000020	
	20MG	0.0200	2.7000	.000069	
	10MG	0.0100	2.7000	.000069	
AN/	10MG	0.0100	8.4100	.000039	-0.00046
SUM	10MG	0.0100	2.7000	.000069	

BALANCE 13
OPERATOR 39

CALIBRATION DESIGN 62

	MG 50	30	20	10	10	10
A 1	+	-	-	+	-	
A 2	+	-	-		+	-
A 3	+	-	-	-		+
A 4	+	-	-			
A 5	+		-	-	-	-
A 6		+	-	+	-	-
A 7		+	-	-	+	-
A 8		+	-	-	-	+
A 9			+	-	-	
A 10			+	-		-
A 11			+		-	-
R	+	+	+			

OBSERVATIONS IN DIVISIONS
DIRECT READING

A 1	9.1000	10009.1006
A 2	-43.2000	
A 3	-0.1000	
A 4	-11.4000	
A 5	-59.6000	
A 6	-13.8000	
A 7	-55.7000	
A 8	8.4000	
A 9	-11.8000	
A 10	-43.8000	
A 11	-24.0000	9976.0000

BALANCE 13
OPERATOR 39

CALIBRATION DESIGN 62

SENSITIVITY WEIGHT
MASS 10.00000 MG
VOLUME 0.00000 CM3 AT 20 C
COEFFICIENT OF EXPANSION 0.000000
S*=S-PV(S)= 10.00000 MG
ACCEPTED SENSITIVITY = 0.00100 MG/DIV
OBSERVED SENSITIVITY = 0.00100 MG/DIV
T-TEST = 0.000

	A(I) (MG)	DELTA(I) (MG)	OBSERVED SENSITIVITY (MG/DIV)
A 1	0.00910	0.00003	0.00100
A 2	-0.04320	0.00032	
A 3	-0.00010	0.00016	
A 4	-0.01140	0.00017	
A 5	-0.05960	-0.00068	
A 6	-0.01380	0.00058	
A 7	-0.05570	-0.00005	
A 8	0.00840	0.00015	
A 9	-0.01180	0.00054	
A 10	-0.04380	0.00049	
A 11	-0.02400	-0.00035	0.00100

ITEM (G)	CORRECTION (MG)	VOLUME (AT T) (CM3)	SYSTEMATIC ERROR (MG)	3 S.D. LIMIT (MG)	UNCERTAINTY LIMIT (MG)
0.0500	-0.00346	0.00301	0.00004	0.00051	0.00055
0.0300	0.00198	0.00181	0.00003	0.00050	0.00053
0.0200	0.01351	0.00741	0.00002	0.00042	0.00044
0.0100	0.02325	0.00371	0.00001	0.00054	0.00055
0.0100	-0.00039	0.00119	0.00001	0.00054	0.00055
0.0100	0.03457	0.00372	0.00001	0.00054	0.00055

TEMPERATURE T= 23.00 C

RESTRAINT FOR FOLLOWING SERIES
RESTRAINT VECTOR 0 0 0 0 0 1
MASS CORRECTION 0.03457 MG
VOLUME AT 20 C 0.00372 CM3
SYSTEMATIC ERROR 0.00001 MG
3 STANDARD DEVIATION LIMIT 0.00054 MG

BALANCE 13
OPERATOR 39

CALIBRATION DESIGN 62

SUM (MG)	WEIGHTS USED FOR THE LINEAR COMBINATIONS MG					
	50	30	20	10	10	10
100	+	+	+			
90	+	+		+		
80	+	+				
70	+		+			
60	+			+		
50		+	+			
40		+		+		
30		+				
20			+			
10				+		

VALUES AND UNCERTAINTIES FOR COMBINATIONS OF WEIGHTS
(UNCERTAINTY IS 3 STANDARD DEVIATION LIMIT PLUS ALLOWANCE FOR SYSTEMATIC ERROR.)

SUM (MG)	CORR (MG)	SYSTEMATIC (MG)	3 S.D. ERROR (MG)	UNCERTAINTY LIMIT (MG)
100	0.01204	0.00009	0.00074	0.00082
90	0.02178	0.00008	0.00096	0.00104
80	-0.00147	0.00007	0.00071	0.00078
70	0.01006	0.00006	0.00068	0.00074
60	0.01980	0.00005	0.00078	0.00083
50	0.01550	0.00004	0.00051	0.00055
40	0.02523	0.00003	0.00077	0.00081
30	0.00198	0.00003	0.00050	0.00053
20	0.01351	0.00002	0.00042	0.00044
10	0.02325	0.00001	0.00054	0.00055

BALANCE 13
OPERATOR 39
MAXIMUM LOAD 0.0600 G
STARTING RESTRAINT NUMBER 4

CALIBRATION DESIGN 62

PRECISION CONTROL

OBSERVED STANDARD DEVIATION OF THE PROCESS 0.00052 MG
ACCEPTED STANDARD DEVIATION OF THE PROCESS 0.00050 MG
DEGREES OF FREEDOM 6
F RATIO 1.083

F RATIO IS LESS THAN 2.81 (CRITICAL VALUE FOR PROBABILITY = .01).
THEREFORE THE STANDARD DEVIATION IS IN CONTROL.

CHECK STANDARD VECTOR 0 0 0 0 1 0
CHECK STANDARD USED 14
ACCEPTED MASS CORRECTION OF CHECK STANDARD -0.00046 MG
OBSERVED CORRECTION OF CHECK STANDARD -0.00039 MG
STANDARD DEVIATION OF THE OBSERVED CORRECTION 0.00018 MG
T VALUE 0.41

ABSOLUTE VALUE OF T IS LESS THAN 3.
THEREFORE CHECK STANDARD IS IN CONTROL.

TEST CONDITIONS	BEFORE	AFTER	AVERAGE
CORRECTED TEMPERATURE IN DEGREES C	22.95	23.05	23.00
CORRECTED PRESSURE IN MM HG	760.227	760.027	760.127
CORRECTED HUMIDITY IN PERCENT	30.00	30.70	30.35
COMPUTED AIR DENSITY IN MG/CM3	1.1895	1.1886	1.1891
TEMPERATURE CORRECTION	0.00	0.00	
PRESSURE CORRECTION	-0.173	-0.173	
HUMIDITY CORRECTION	0.00	0.00	
OBSERVED TEMPERATURE IN DEGREES C	22.95	23.05	
OBSERVED PRESSURE IN MM HG	760.400	760.200	
OBSERVED HUMIDITY IN PERCENT	30.00	30.70	

BALANCE 13
OPERATOR 39
ACCEPTED WITHIN STANDARD DEVIATION OF THE PROCESS 0.00050 MG
ACCEPTED BETWEEN STANDARD DEVIATION OF THE PROCESS 0.00000 MG

CALIBRATION DESIGN 62
RESTRAINT VECTOR 1 1 1 0 0 0
MASS CORRECTION OF RESTRAINT 0.03457 MG
VOLUME OF WEIGHTS BEING USED IN RESTRAINT AT 22.95 0.00372 CM3
SYSTEMATIC ERROR IN THE RESTRAINT 0.00001 MG
3 STANDARD DEVIATION LIMIT FOR RANDOM ERROR AFFECTING RESTRAINT 0.00054 MG

CHECK STANDARD USED 139
CHECK STANDARD VECTOR 0 0 0 0 1 0
ACCEPTED MASS CORRECTION OF CHECK STANDARD -0.00216 MG
REPORT VECTOR 1 1 1 1 0 0

TEST CONDITIONS	BEFORE	AFTER	AVERAGE
CORRECTED TEMPERATURE IN DEGREES C	22.90	23.00	22.95
CORRECTED PRESSURE IN MM HG	759.527	759.327	759.427
CORRECTED HUMIDITY IN PERCENT	31.60	32.60	32.10
COMPUTED AIR DENSITY IN MG/CM3	1.1884	1.1875	1.1879
TEMPERATURE CORRECTION	0.00	0.00	
PRESSURE CORRECTION	-0.173	-0.173	
HUMIDITY CORRECTION	0.00	0.00	
OBSERVED TEMPERATURE IN DEGREES C	22.90	23.00	
OBSERVED PRESSURE IN MM HG	759.700	759.500	
OBSERVED HUMIDITY IN PERCENT	31.60	32.60	

	WEIGHTS BEING TESTED	NOMINAL VALUE G	DENSITY G/CM3 AT 20C	COEFFICIENT OF EXPANSION	ACCEPTED CORRECTION MG
	5MG	0.0050	2.7000	.000069	
	3MG	0.0030	2.7000	.000069	
	2MG	0.0020	2.7000	.000069	
	1MG	0.0010	2.7000	.000069	
T	1MG	0.0010	8.5000	.000039	-0.00216
SUM	1MG	0.0010	2.7000	.000069	

BALANCE 13
OPERATOR 39

CALIBRATION DESIGN 62

	MG 5	3	2	1	1	1
A 1	+	-	-	+	-	
A 2	+	-	-		+	-
A 3	+	-	-	-		+
A 4	+	-	-			
A 5	+		-	-	-	-
A 6		+	-	+	-	-
A 7		+	-	-	+	-
A 8		+	-	-	-	+
A 9			+	-	-	
A 10			+	-		-
A 11			+		-	-
R	+	+	+			

OBSERVATIONS IN DIVISIONS
DIRECT READING

A 1	8.4000	10008.4004
A 2	2.4000	
A 3	-8.8000	
A 4	1.1000	
A 5	7.2000	
A 6	6.0000	
A 7	-9.1000	
A 8	-11.2000	
A 9	7.6000	
A 10	8.0000	
A 11	17.8000	10017.7998

BALANCE 13
OPERATOR 39

CALIBRATION DESIGN 62

SENSITIVITY WEIGHT
MASS 10.00000 MG
VOLUME 0.00000 CM3 AT 20 C
COEFFICIENT OF EXPANSION 0.000000
S*=S-PV(S)= 10.00000 MG
ACCEPTED SENSITIVITY = 0.00100 MG/DIV
OBSERVED SENSITIVITY = 0.00100 MG/DIV
T-TEST = 0.000

	A(I) (MG)	DELTA(I) (MG)	OBSERVED SENSITIVITY (MG/DIV)
A 1	0.00840	-0.00028	0.00100
A 2	0.00240	0.00039	
A 3	-0.00880	-0.00046	
A 4	0.00110	0.00032	
A 5	0.00720	0.00003	
A 6	0.00600	-0.00059	
A 7	-0.00910	0.00011	
A 8	-0.01120	0.00046	
A 9	0.00760	-0.00009	
A 10	0.00800	-0.00092	
A 11	0.01780	0.00098	0.00100

ITEM (G)	CORRECTION (MG)	VOLUME (AT T) (CM3)	SYSTEMATIC ERROR (MG)	3 S.D. LIMIT (MG)	UNCERTAINTY LIMIT (MG)
0.0050	0.01768	0.00186	0.00000	0.00044	0.00045
0.0030	0.00600	0.00111	0.00000	0.00048	0.00048
0.0020	0.01089	0.00074	0.00000	0.00041	0.00041
0.0010	0.00555	0.00037	0.00000	0.00054	0.00054
0.0010	-0.00265	0.00012	0.00000	0.00054	0.00054
0.0010	-0.00358	0.00037	0.00000	0.00054	0.00054

TEMPERATURE T= 22.95 C

BALANCE 13
OPERATOR 39

CALIBRATION DESIGN 62

SUM (MG)	WEIGHTS USED FOR THE LINEAR COMBINATIONS MG					
	5	3	2	1	1	1
10	+	+	+			
9	+	+		+		
8	+	+				
7	+		+			
6	+			+		
5		+	+			
4		+		+		
3		+				
2			+			
1				+		

VALUES AND UNCERTAINTIES FOR COMBINATIONS OF WEIGHTS
(UNCERTAINTY IS 3 STANDARD DEVIATION LIMIT PLUS ALLOWANCE FOR SYSTEMATIC ERROR.)

SUM (MG)	CORR (MG)	SYSTEMATIC (MG)	3 S.D. ERROR (MG)	UNCERTAINTY LIMIT (MG)
10	0.03457	0.00001	0.00054	0.00055
9	0.02923	0.00001	0.00084	0.00085
8	0.02368	0.00001	0.00058	0.00059
7	0.02857	0.00001	0.00058	0.00059
6	0.02322	0.00001	0.00072	0.00073
5	0.01689	0.00000	0.00044	0.00045
4	0.01155	0.00000	0.00075	0.00075
3	0.00600	0.00000	0.00048	0.00048
2	0.01089	0.00000	0.00041	0.00041
1	0.00555	0.00000	0.00054	0.00054

BALANCE 13
OPERATOR 39
MAXIMUM LOAD 0.0060 G
STARTING RESTRAINT NUMBER 4

CALIBRATION DESIGN 62

PRECISION CONTROL

OBSERVED STANDARD DEVIATION OF THE PROCESS 0.00070 MG
ACCEPTED STANDARD DEVIATION OF THE PROCESS 0.00050 MG
DEGREES OF FREEDOM 6
F RATIO 1.955

F RATIO IS LESS THAN 2.81 (CRITICAL VALUE FOR PROBABILITY = .01).
THEREFORE THE STANDARD DEVIATION IS IN CONTROL.

CHECK STANDARD VECTOR 0 0 0 0 1 0
CHECK STANDARD USED 139
ACCEPTED MASS CORRECTION OF CHECK STANDARD -0.00216 MG
OBSERVED CORRECTION OF CHECK STANDARD -0.00265 MG
STANDARD DEVIATION OF THE OBSERVED CORRECTION 0.00018 MG
T VALUE -2.77

ABSOLUTE VALUE OF T IS LESS THAN 3.
THEREFORE CHECK STANDARD IS IN CONTROL.

TEST CONDITIONS	BEFORE	AFTER	AVERAGE
CORRECTED TEMPERATURE IN DEGREES C	22.90	23.00	22.95
CORRECTED PRESSURE IN MM HG	759.527	759.327	759.427
CORRECTED HUMIDITY IN PERCENT	31.60	32.60	32.10
COMPUTED AIR DENSITY IN MG/CM3	1.1884	1.1875	1.1879
TEMPERATURE CORRECTION	0.00	0.00	
PRESSURE CORRECTION	-0.173	-0.173	
HUMIDITY CORRECTION	0.00	0.00	
OBSERVED TEMPERATURE IN DEGREES C	22.90	23.00	
OBSERVED PRESSURE IN MM HG	759.700	759.500	
OBSERVED HUMIDITY IN PERCENT	31.60	32.60	

SUMMARY

FOR CONVENIENCE, THE RESULTS OF THIS WORK ARE SUMMARIZED IN TABLES I AND II. THE VALUES ASSIGNED ARE WITH REFERENCE TO THE STANDARDS IDENTIFIED ON THE DATA SHEETS. THE UNCERTAINTY FIGURE IS AN EXPRESSION OF THE OVERALL UNCERTAINTY USING THREE STANDARD DEVIATIONS AS A LIMIT TO THE EFFECT OF RANDOM ERRORS OF THE MEASUREMENT ASSOCIATED WITH THE MEASUREMENT PROCESSES. THE MAGNITUDE OF SYSTEMATIC ERRORS FROM SOURCES OTHER THAN THE USE OF ACCEPTED VALUES FOR CERTAIN STARTING STANDARDS ARE CONSIDERED NEGLIGIBLE. IT SHOULD BE NOTED THAT THE MAGNITUDE OF THE UNCERTAINTY REFLECTS THE PERFORMANCE OF THE MEASUREMENT PROCESS USED TO ESTABLISH THESE VALUES. THE MASS UNIT, AS REALIZABLE IN ANOTHER MEASUREMENT PROCESS, WILL BE UNCERTAIN BY AN AMOUNT WHICH IS A COMBINATION OF THE UNCERTAINTY OF THIS PROCESS AND THE PROCESS IN WHICH THESE STANDARDS ARE USED.

THE ESTIMATED MASS VALUES LISTED IN TABLE I ARE BASED ON AN EXPLICIT TREATMENT OF DISPLACEMENT VOLUMES, E.G., 'TRUE MASS', 'MASS IN VACUO', MASS IN THE NEWTONIAN SENSE. THE DISPLACEMENT VOLUME ASSOCIATED WITH EACH VALUE IS LISTED AS WELL AS THE VOLUMETRIC COEFFICIENT OF EXPANSION. THESE VALUES SHOULD BE USED, TOGETHER WITH APPROPRIATE CORRECTION FOR THE BUOYANT EFFECTS OF THE ENVIRONMENT, TO ESTABLISH CONSISTENT MASS VALUES FOR OBJECTS WHICH DIFFER SIGNIFICANTLY IN DENSITY AND/OR FOR MEASUREMENTS WHICH MUST BE MADE IN DIFFERING ENVIRONMENTS. THE RELATION 1LB AVDP=.45359237KG IS USED AS REQUIRED.

THE ESTIMATED MASS VALUES LISTED IN TABLE II ARE BASED ON AN IMPLICIT TREATMENT OF DISPLACEMENT VOLUMES, E.G., 'APPARENT MASS', 'APPARENT MASS VERSUS BRASS', 'APPARENT MASS VERSUS DENSITY 8.0'. THE VALUES ARE LISTED AS CORRECTIONS TO BE APPLIED TO THE LISTED NOMINAL VALUE (A POSITIVE CORRECTION INDICATES THAT THE MASS IS LARGER THAN THE STATED NOMINAL VALUE BY THE AMOUNT OF THE CORRECTION). THESE VALUES ARE COMPUTED FROM THE VALUES BASED ON AN EXPLICIT TREATMENT OF DISPLACEMENT VOLUMES USING THE FOLLOWING DEFINING RELATIONS AND ARE UNCERTAIN BY THE AMOUNT SHOWN IN TABLE I.

THE ADJUSTMENT OF WEIGHTS TO MINIMIZE THE DEVIATION FROM NOMINAL ON THE BASIS OF 'NORMAL BRASS' (IN ACCORDANCE WITH COR. A BELOW) IS WIDESPREAD IN THIS COUNTRY AND IN MANY PARTS OF THE WORLD. VALUES STATED ON EITHER BASIS ARE INTERNALLY CONSISTENT AND DEFINITE. THERE IS, HOWEVER, A SYSTEMATIC DIFFERENCE BETWEEN THE VALUES ASSIGNED ON EACH BASIS, THE VALUE ON THE BASIS OF 'DENSITY 8.0' BEING 7 MICROGRAMS/GRAM LARGER THAN THE VALUE ON THE BASIS OF NORMAL BRASS. THIS SYSTEMATIC DIFFERENCE IS CLEARLY DETECTABLE ON MANY DIRECT READING BALANCES.

CORRECTION A - 'APPARENT MASS VERSUS BRASS' OR 'WEIGHT IN AIR AGAINST BRASS' IS DETERMINED BY A HYPOTHETICAL WEIGHING OF THE WEIGHT AT 20 CELSIUS IN AIR HAVING A DENSITY OF 1.2 MG/CM3, WITH A (NORMAL BRASS) STANDARD HAVING A DENSITY OF 8.4 G/CM3 AT 0 CELSIUS WHOSE COEFFICIENT OF VOLUMETRIC EXPANSION IS 0.000 054 PER DEGREE CELSIUS, AND WHOSE VALUE IS BASED

ON ITS TRUE MASS OR WEIGHT IN VACUO.

CORRECTION B - 'APPARENT MASS VERSUS DENSITY 8.0' IS DETERMINED BY A HYPOTHETICAL WEIGHING OF THE WEIGHT, IN AIR HAVING A DENSITY OF 1.2 MG/CM3, WITH A STANDARD HAVING A DENSITY OF 8.0 G/CM3 AT 20 CELSIUS, AND WHOSE VALUE IS BASED ON ITS TRUE MASS OR WEIGHT IN VACUO.

TABLE I

	ITEM	MASS (G)	UNCERTAINTY (G)	VOL AT 20 (CM3)	COEF OF EXP
	500MG	.49992233	0.00000300	0.03012	0.000020
	300MG	.29995722	0.00000185	0.01807	0.000020
	200MG	.19996121	0.00000127	0.01205	0.000020
	100MG	.10000171	0.00000082	0.00602	0.000020
NEW	50MG	.04999654	0.00000055	0.00301	0.000020
	30MG	.03000198	0.00000053	0.00181	0.000020
	20MG	.02001351	0.00000044	0.00741	0.000069
	10MG	.01002325	0.00000055	0.00371	0.000069
	5MG	.00501768	0.00000045	0.00186	0.000069
	3MG	.00300600	0.00000048	0.00111	0.000069
	2MG	.00201089	0.00000041	0.00074	0.000069
	1MG	.00100555	0.00000054	0.00037	0.000069

TABLE II

ITEM	COR.A (MG)	COR.B (MG)
500MG	-.04231	-.03881
300MG	-.02158	-.01948
200MG	-.02465	-.02325
100MG	.00879	.00948
NEW 50MG	.00008	.00043
30MG	.00410	.00431
20MG	.00748	.00762
10MG	.02023	.02030
5MG	.01616	.01620
3MG	.00510	.00512
2MG	.01028	.01030
1MG	.00525	.00525

Appendix C

Surveillance Test

The following is the report of a surveillance test [7]. Subsequent to a calibration such as that shown in Appendix B, weights may be resubmitted for periodic surveillance. Surveillance is a more rapid and less costly procedure than calibration. The surveillance test can provide assurance that the values of mass previously assigned to a set of weights are still valid.

UNITED STATES DEPARTMENT OF COMMERCE
National Bureau of Standards
Gaithersburg, Maryland 20899

September 18, 1986

In reply refer to:

Subject:

Items:

The above items have been intercompared in sums. The differences as measured have been compared with the differences computed from the values under 225716-B. One or more of the items have been checked against national standards. The results of this test indicate that there is no significant change since the last calibration. This test assures the continuing accuracy of the values under 225716-B.

Sincerely,

JOE D. SIMMONS, Deputy Director
Center for Basic Standards

Attachment

Appendix D

Calibration of Dead Weights

The following is a typical report of calibration for a set of dead weights.

Form NBS-443
(REV. 12-68)

U.S. DEPARMENT OF COMMERCE
NATIONAL BUREAU OF STANDARDS
GAITHERSBURG, MARYLAND 20899

REPORT OF CALIBRATION

NBS Test Number:

For:

Items:

The above items have the mass values shown with reference to the NBS standard of mass.

Item	Mass (g)	Uncertainty (g)	Density (g/cm^3)
1kg-1	999.9968	0.0033	7.92
1kg-2	999.9985	0.0033	7.92
1kg-3	1000.0007	0.0033	7.92
1kg-4	999.9969	0.0033	7.92
1kg-5	999.9960	0.0033	7.92
1kg-6	999.9952	0.0033	7.92

The uncertainty figure is an expression of the overall uncertainty using three standard deviation as a limit to the effect of random errors of measurement plus the systematic errors, assuming the density is correct within 1%. Test conditions: mass computed using air density 1.175mg/cm^3 for all items.

The National Bureau of Standards uses the following relationship between the metric unit of mass and the U.S. customary unit of mass: one pound (avoirdupois) equals 0.45359237 kilogram.

For the Director,
National Measurement Laboratory

Joe D. Simmons, Chief
Length and Mass Division
Center for Basic Standards

Test completed: September 3, 1986

Note: Mass and associated density values listed above are appropriate for M_m and ρ_m in Equation (24) from NBS Monograph 65, "Reduction of Data for Piston Gage Pressure Measurements."

NBS-114A (REV. 2-8C)

U.S. DEPT. OF COMM. **BIBLIOGRAPHIC DATA SHEET** (See instructions)	1. PUBLICATION OR REPORT NO. NIST/SP-250/31	2. Performing Organ. Report No.	3. Publication Date January 1989

4. TITLE AND SUBTITLE

NIST Measurement Services: Mass Calibrations

5. AUTHOR(S)

Richard S. Davis

6. PERFORMING ORGANIZATION (If joint or other than NBS, see instructions)

NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY
(formerly NATIONAL BUREAU OF STANDARDS)
U.S. DEPARTMENT OF COMMERCE
GAITHERSBURG, MD 20899

7. Contract/Grant No.

8. Type of Report & Period Covered

Final

9. SPONSORING ORGANIZATION NAME AND COMPLETE ADDRESS (Street, City, State, ZIP)

Same as Item #6

10. SUPPLEMENTARY NOTES

Library of Congress Catalog Card Number: 88-600608

☐ Document describes a computer program; SF-185, FIPS Software Summary, is attached.

11. ABSTRACT (A 200-word or less factual summary of most significant information. If document includes a significant bibliography or literature survey, mention it here)

The NIST calibration service for standard masses is described. Weights which are accepted for calibration range in nominal values from 1 mg to 13,600 kg (30,000 pounds). We also accept weights used to generate standard pressures in piston gages. Cleaning procedures used on weights prior to calibration are described. The measurement algorithms (including density determinations of single-piece kilogram weights) and the uncertainties assigned to calibrated weights are discussed. We also describe the system now in place to monitor the quality of calibrations. Finally, we assess the limitations of the present controls on measurement quality and outline improvements which are underway.

12. KEY WORDS (Six to twelve entries; alphabetical order; capitalize only proper names; and separate key words by semicolons)

calibration; density; kilogram; least squares; mass; piston weights; uncertainty

13. AVAILABILITY

- [X] Unlimited
- ☐ For Official Distribution. Do Not Release to NTIS
- [X] Order From Superintendent of Documents, U.S. Government Printing Office, Washington, D.C. 20402.
- [X] Order From National Technical Information Service (NTIS), Springfield, VA. 22161

14. NO. OF PRINTED PAGES

72

15. Price

USCOMM-DC 6043-P80

NBS Special Publication 700-1
Industrial Measurement Series

A Primer for Mass Metrology

K.B. Jaeger
Measurement Standards Laboratory
Lockheed Missiles and Space Company, Inc.
Sunnyvale, CA 94086

and

R.S. Davis
Center for Basic Standards
National Measurement Laboratory
National Bureau of Standards
Gaithersburg, MD 20899

November 1984

U.S. Department of Commerce
Malcolm Baldrige, Secretary

National Bureau of Standards
Ernest Ambler, Director

Library of Congress
Catalog Card Number: 84-601090
National Bureau of Standards
Special Publication 700-1
Natl. Bur. Stand. (U.S.),
Spec. Publ. 700-1,
85 pages (Nov 1984)
CODEN: XNBSAV

U.S. Government Printing Office
Washington: 1984

For sale by the Superintendent
of Documents,
U.S. Government Printing Office,
Washington, DC 20402

FOREWORD

When the National Bureau of Standards was established more than 80 years ago, it was given the specific mission of aiding manufacturing and commerce. Today, NBS remains the only Federal laboratory with this explicit goal of serving U.S. industry and science. Our mission takes on special significance now as the country is responding to serious challenges to its industry and manufacturing--challenges which call for government to pool its scientific and technical resources with industry and universities.

The links between NBS staff members and our industrial colleagues have always been strong. Publication of this new Industrial Measurement Series, aimed at those responsible for measurement in industry, represents a strengthening of these ties.

The concept for the series stems from the joint efforts of the National Conference of Standards Laboratories and NBS. Each volume will be prepared jointly by an industrial specialist and a member of the NBS staff. Each volume will be written within a framework of industrial relevance and need.

This publication, A Primer for Mass Metrology, represents the first of what we anticipate will be a long series of collaborative ventures that will aid both industry and NBS.

Ernest Ambler, Director

ABOUT THE AUTHORS

Klaus Jaeger

Dr. Klaus Jaeger was educated at Syracuse University (B.S., Physics 1965; Ph.D., High Energy Physics, 1970). He was employed at Argonne National Laboratory from 1970 to 1980 and Brookhaven National Laboratory from 1980 through 1981. During these periods Dr. Jaeger was engaged in laboratory research in neutrino and meson interaction physics, bubble chamber experiments, particle beam designs and superconducting magnet technology. In addition to high energy physics his experience includes electronics, computer science and cryogenics.

In January 1982 Dr. Jaeger joined Lockheed Missiles and Space Company, Sunnyvale, California as a research specialist responsible for upgrading all primary measurement standards required by the organization. Since October 1983 he has been the supervisor responsible for primary electrical measurement standards. Dr. Jaeger resides in Saratoga, California.

Richard S. Davis

Dr. Richard Davis was educated at Brown University (B.S., Physics, 1967) and the University of Maryland (Ph.D., Solid State Physics, 1972). His fields of study include experimental solid state physics, theoretical and experimental problems in fluid dynamics, properties of materials at very low temperatures, superfluid properties of liquid helium, electronics, absolute electrical measurements and basic mass metrology.

Dr. Davis joined the National Bureau of Standards in 1972. His laboratory research programs at NBS have been concerned with the improvement of the measured value of the Faraday (the fundamental constant of electrochemistry), high voltage capacitor calibrations, the accurate determination of air density effects in mass measurements and improved methods for mass comparison at the highest levels of accuracy. Dr. Davis resides in Washington, D.C.

CONTENTS

A Primer for Mass Metrology

K. B. Jaeger
Measurement Standards Laboratory
Lockheed Missiles and Space Company, Inc.
Sunnyvale, CA 94086

and

R. S. Davis
Center for Basic Standards
National Bureau of Standards
Gaithersburg, MD 20899

1. INTRODUCTION

Much of the mass program at the National Bureau of Standards has been documented during the last two decades [1-8]. What we attempt here is to explain the relevance of these publications to mass metrology conducted at laboratories whose standards are calibrated by NBS. We try to begin with basic physics concepts deriving the essential results with a minimum of rigor. The reader should always consult the appropriate reference for a more sophisticated treatment.

We have also tried to use a consistent notation. Consequently, our notation may vary slightly from that found in the references cited above.

Many examples are provided in the belief that understanding of a general case is much easier if a special case is understood well. Cautionary statements are provided at places in the text where important practical complications may exist.

In general, we have strived to be understandable rather than scholarly where the combination was elusive.

CAUTION: This document is incomplete. Many extremely important topics--manufacture, specification, cleaning, handling, and storage of weights being among them--have been omitted. The first two of these subjects are treated in ref. [14].

Our ultimate goal is to outline the way in which mass measurements can be made to acceptable metrological accuracies. There are three major concepts which must be explored but which are somewhat interrelated - what one means by "mass;" how one can use a measuring device to determine the mass of an unknown object in terms of an accepted unit, for example, the kilogram; and how to assign a realistic uncertainty to the results. It will be useful to introduce some basic equations of physics in order to make our arguments precise. The first major step is to explore the basic properties of a mass-measuring device, or balance, and then see how the device can be put to the service of metrology. Mathematically, this means that the first step is to derive the basic equations of weighing. The remainder of our efforts modify the basic equations in order to make them practical. We also show their relevance to mass metrology.

2. BASIC MASS EQUATIONS

In order to derive the basic equations of weighing, we need a precise notion of mass. For the purpose of this discussion we will introduce mass through the familiar equation:

$$\vec{F} = M\vec{a}$$

(Force = Mass x Acceleration) (1)

That is, each object possesses a property called "mass" which appears in eq. (1) as the constant of proportionality between a force applied to that object and the resulting acceleration of the object. Note that both force and acceleration are vector quantities: they have a direction associated with them. One should also note that mass is always a positive number (that is, for example, the acceleration is always in the same direction as the gravitational force, never opposite to it). These comments about mass are consistent with the qualitative idea of the mass of an object being a measure of the "amount of substance" in the object.

Unfortunately, our intuitive notions of "mass" are often confused with "weight." Such confusion is unacceptable for science and metrology. Therefore we will now see how the notion of weight differs both logically and practically from that of mass.

We can take it as an experimental fact that over a small plane area of the earth's surface (such as the space occupied by a metrology laboratory) the acceleration of gravity, $\vec{g}$, is essentially constant. Since $\vec{g}$ is a vector quantity, constant $\vec{g}$ implies that, over a small area of the earth's surface, the direction of $\vec{g}$ is also practically constant. That is, the acceleration vectors are all parallel and define the direction "down." In this approximation, we can replace $\vec{g}$ by a numerical constant, $\bar{g}$.

The gravitational force on an object of mass M is then:

$$F_1 = M\bar{g} \quad . \tag{2}$$

The weight of an object of mass M is defined as F_1.[1] That is, weight is a force, not a mass.

Using $W = F_1$, we have

$$W = M\bar{g} \quad . \tag{3}$$

From what was said about $\bar{g}$, it is clear that weight is not a constant property of matter, but depends on location. Consider, for example, a body of mass M taken out into deep space such that the gravitational forces are negligible (they are never zero). In this case, we can approximate (3) as

$$W = M\bar{g} \quad 0$$

which holds since

$$\bar{g} \quad 0; \ (M \neq 0) \quad .$$

---

[1] This definition was adopted for international use by the General Conference for Weights and Measures (C.G.P.M.) in 1901.

Hence, the body still has mass M but its weight is zero. It is only because most people live exclusively on the surface of the earth that weight and mass are often confused.

Let us now consider the action of a scale or balance - the device which we will use to measure mass. There are a great many schemes in use for constructing scales and balances. All have their strengths and weaknesses but most are designed to approximate the workings of the following hypothetical device:

This device is a "black box" which has a pan on which to place an object and an indicator which reads zero when the pan is empty and some other number when an object is placed on the pan (fig. 1). Our ideal scale has the following two desirable properties:

1) The indicator always reads the same amount to any desired precision when the same object is put on the pan at the same conditions of temperature, barometric pressure, relative humidity, etc.
2) If any two objects are put on the pan, the indicator reading is the sum of the readings for each object placed individually.

The first property means that the imprecision of the balance readings is zero. The second property means that the balance is perfectly linear. Later we will show how to use real balances which only approximate these two important features.

A final property of our idealized balance, which it shares with actual scales and balances is that the indicator responds to a force on the pan, not a mass. In other words, our balance would give different readings on the moon than on earth (though properties 1) and 2) would be unaffected).

When the balance has reached an equilibrium condition, the sum of all forces acting on the system must be zero. We then have

$$\sum_{i=1}^{n} \vec{F}_i = 0$$

and since the forces acting on the system are known to act in one direction and its exact opposite (i.e. up and down) we can once again ignore the vector notation and write

$$\sum_{i=1}^{n} F_i = 0 \qquad (4)$$

where n = total number of forces acting on the system.

The forces acting on the object placed on the pan of the balance are:

$F_1 = M\bar{g}$ gravitational force

F_2, balance force exerted by the balance via deflection of a beam, stretching of a spring, or some other method. Obviously this force has to be opposite to F_1. Since F_2 acts opposite to F_1, we adopt the convention

$$F_2 = -k \cdot \theta_1 \cdot \bar{g} \quad , \qquad (5)$$

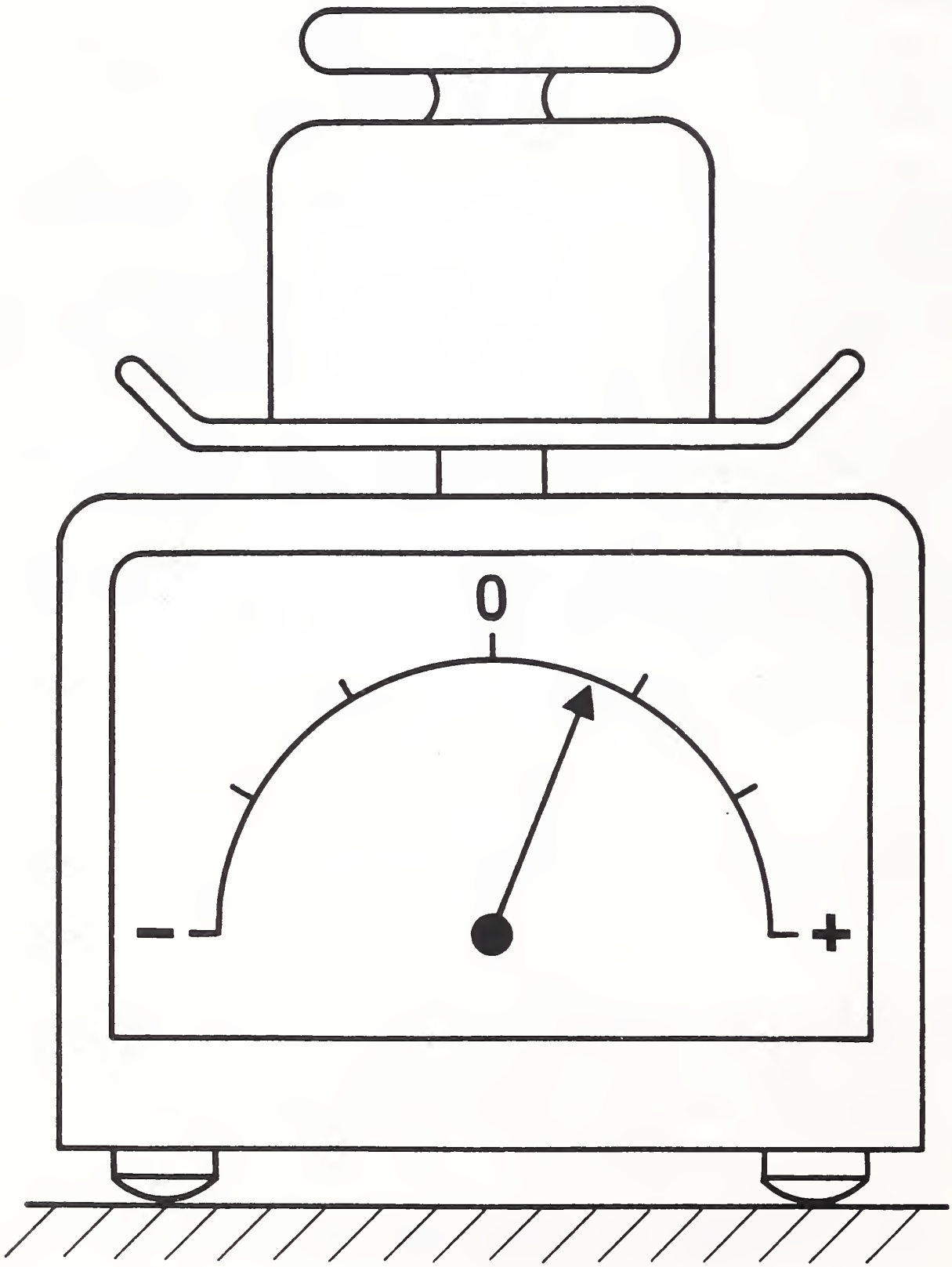


Figure 1. An idealized scale. The indicating dial is marked in equally spaced increments. The number pointed to by the needle can be resolved to any desired precision.

where θ_1 is the balance reading and k is a positive constant (yet to be evaluated) which converts balance readings to mass units.

In addition to these main forces, we have to consider buoyant, electrostatic, and electromagnetic forces on the object. Of these three only the first one plays a significant role in usual metrology and we now look at the effect.

Under the influence of gravity, the force on a body submerged in a fluid is equal in magnitude but opposite in direction to the weight of the displaced medium. This upward force is known as the "buoyant force." From the definition we see that

$$F_3 = -m\bar{g}$$

where m = mass of displaced medium. The buoyant force was first quantified by Archimedes in the third century B.C. We also can write

$$m = V\rho_m$$

so that

$$F_3 = -V\rho_m\bar{g} \qquad (6)$$

where

V = volume of the displaced medium
= volume of object
ρ_m = density of medium $\equiv$ m/V .

With these three forces, we can now write eq. (4) as

$$F_1 + F_2 + F_3 = 0$$

or

$$F_1 + F_3 = -F_2 \ .$$

Substituting from eqs. (2), (5) and (6) yields

$$M\bar{g} - \rho_m V\cdot\bar{g} = k\theta_1\bar{g}$$

$$(M - \rho_m V)\bar{g} = k\theta_1\bar{g} \qquad (7a)$$

or

$$(M - \rho_m V) = k\theta_1 \ . \qquad (7b)$$

It is very important to realize from eq. (7a) that if the weighing is done in a vacuum chamber then the F_3 term drops out and eq. (7a) reduces to

$$M\bar{g} = k\theta_1\bar{g} \ . \qquad (7c)$$

Furthermore, the medium or environment can be any fluid--air, water, oil, etc. For most applications and for practical purposes, the medium is usually air. Even though air is a mixture of gases and the density is quite small compared to densities of objects usually measured, the correction term, F_3, cannot be ignored by metrologists, as we shall see.

Let us now consider the weighing of two objects of mass M_1 and M_2 under the same conditions, i.e. in the same medium. We then have from eq. (7a)

$$\bar{g}(M_1-\rho_m V_1) = k\theta_1\bar{g} \qquad (7d)$$

$$\bar{g}(M_2-\rho_m V_2) = k\theta_2\bar{g} \ . \qquad (7e)$$

The density of an object is its mass devided by its volume. That is,

$$\rho_1 \equiv M_1/V_1$$

where ρ_1 is the density of object 1. We can, therefore, always replace a quantity such as

$$M_1 - \rho_m V_1$$

by the identical quantity

$$M_1(1 - \rho_m/\rho_1) \quad .$$

The choice of representation is completely optional, although the proper choice can often simplify a calculation. For the present, we will continue with the volume representation.

Subtracting eq. (7e) from 7(d) yields

$$\bar{g}(M_1 - \rho_m V_1) - \bar{g}(M_2 - \rho_m V_2) = \bar{g}\, k\theta_1 - \bar{g}\, k\theta_2 \qquad (7f)$$

and by setting

$$\theta_1 - \theta_2 = \Delta\theta$$

we have

$$\bar{g}(M_1 - \rho_m V_1) - \bar{g}(M_2 - \rho_m V_2) = \bar{g}k\Delta\theta \quad . \qquad (8)$$

Cancelling $\bar{g}$ yields

$$M_1 - M_2 = k\Delta\theta + \rho_m\,(V_1 - V_2) \quad . \qquad (9)$$

Equation (9) is one of the fundamental equations in mass metrology and is known as the "Weighing Equation." Its importance will become evident in Section 3. For now we simply point out that, using the best balances available, one can determine $\Delta\theta$ to much higher accuracy than either θ_1 or θ_2. Thus eq. (9) turns out to be more useful than eq. (7b).

In most cases, the laboratory conditions are such that the medium is air, in which case

$$\rho_A = \rho_m = \text{density of air}$$

and eq. (9) becomes

$$M_1 - M_2 = k\Delta\theta + \rho_A(V_1 - V_2) \qquad (10)$$

Note: When we refer to "weighing" in this text, we will always mean the process which determines the product $k \cdot \theta$. It should be clear from the above paragraphs that finding the mass of an object involves more than just weighing the object. Also, we have no future need to refer to "weight" as defined by (3). From this point on, the term "weight" will be reserved solely to designate an object manufactured to have particular nominal mass. For example, a "1-g weight" is an object whose mass is close to 1 g.

From eqs. (9) and (10) it is obvious the condition

$$M_1 - M_2 = k\Delta\theta$$

can exist if and only if

$$\rho_m(V_1 - V_2) \equiv 0 \quad .$$

This can only hold if

1) $\rho_m \equiv 0$ i.e., the measurement is done in vacuum

or

2) $V_1 \equiv V_2$.

The first case is usually not encountered in metrology. The second case can be approximated quite easily. Consider for example two weights of nearly equal mass, such that

$$M_1 \approx M_2 \quad ,$$

made from the same billet of alloy. In this case, the densities are equal so that

$$\rho_1 = \rho_2 \equiv \rho \quad .$$

Since

$$V_1 = \frac{M_1}{\rho_1} \quad \text{and} \quad V_2 = \frac{M_2}{\rho_2}$$

we have from (10)

$$M_1 - M_2 = k\Delta\theta + \rho_A\left(\frac{M_1}{\rho_1} - \frac{M_2}{\rho_2}\right)$$

$$M_1 - M_2 = k\Delta\theta + \rho_A\left(\frac{M_1 - M_2}{\rho}\right)$$

$$(M_1 - M_2)\left(1 - \frac{\rho_A}{\rho}\right) = k\Delta\theta$$

$$M_1 - M_2 = \frac{k\Delta\theta}{1-\rho_A/\rho}$$

$$M_1 - M_2 \simeq k\Delta\theta$$

because $\rho_A << \rho$ and $M_1 - M_2$ is already known to be a small number.

Relations such as that expressed by eq. (1) or (10) are of use to metrology only if all variables are expressed in a consistent set of units. In what follows, we will assume that we are working in the International System of Units (also known as the Système International des Unités, or simply, the S.I.).

The unit of mass in the S.I. is the kilogram. By definition, one kilogram is exactly equal to the mass of an object known as the International Prototype Kilogram, or I.P.K. This object is made of a platinum-iridium alloy and is stored at the International Bureau of Weights and Measures in Sèvres, France. All laboratory mass standards must ultimately be traceable to the I.P.K.

2.1. True Mass and Apparent Mass

So far, the "mass" we have considered is that found in Newton's basic equation (i.e., eq. (1)). For that reason, this concept of mass is sometimes called "true mass". It is a quantity intrinsic to the object alone and not to its situation. We have also seen that, in a vacuum, the force registered by the balance is exactly proportional to the Newtonian mass. For this reason the "true" mass is also known as the "vacuum" mass.[2] The terms "mass", "true mass" and "vacuum mass" are interchangeable.

The need to apply buoyancy corrections to mass measurements has led to the adoption of so-called "apparent" masses. As explained below and in Appendix A, use of apparent mass can simplify the job of both the weight manufacturer and the metrologist. The drawback, however, is that one must learn a new concept (that is, apparent mass).

Of limited scientific significance, the apparent mass approach nevertheless often simplified calculations for the metrologist at a time when exact computations were tedious (either done by hand or through nomographs) and difficult to check. Today, when every metrologist has access, at the very least, to a digital pocket-calculator, the argument for apparent masses is less compelling. Nevertheless, both for completeness and because commercial weights are still manufactured on the basis of apparent mass, we cannot avoid this topic.

The fact is that most mass metrology is carried out: (1) in air; (2) in laboratories near sea level; and (3) at a temperature of $\sim$20 °C.

Air at such conditions has a density nearly equal to 1.2 mg/cm^3. Fluctuations about the mean density are known rarely to exceed 3 percent. Thus 97 percent of the buoyancy correction can be "built into" the calibration of a weight so that small deviations from the correction are all that need be considered.

Let us carry through an example by defining a reference metal of density ρ_R.

Let

M_R^T = the true mass of an object made of reference material R. (This object need not actually exist.)

ρ_R = the density of the reference material

ρ_0 = 1.2 mg/cm^3 = the reference density of air. (This is very nearly the density of air at 20 °C, 50% humidity, and 760 mm Hg pressure.)

t_0 = 20 °C = reference temperature.

---

[2] If one actually were to weigh an object under vacuum conditions to determine its mass, one would have to be certain that the object weighed was sufficiently stable--that is, no gasses normally adsorbed on the surface at atmospheric pressure are vacuumed off, the weight itself does not "out-gas," etc. The equality between vacuum mass and true mass assumes that the object being weighed is stable.

We now define the "apparent" mass of an object as follows: The "apparent" mass M_X^A of an object X is equal to the "true" mass M_R^T of just enough reference material to produce a balance reading equal to that produced by X if the measurements are done at temperature t_0 in air of density ρ_0 (see fig. 2).

Stated another way, we first define specific, unique weighing conditions:

i. the air density is 1.2 mg/cm^3
ii. the temperature is 20 °C (it is necessary to specify temperature because the volume of a weight depends slightly on temperature.)

Next we imagine a mass comparison between an object X and an assembly of known objects R. The definition specifies that the R objects shall all have a density equal to ρ_R. The mass of the R objects is then adjusted until

$$\Delta\theta = \theta_R - \theta_X = 0 \quad .$$

When these conditions are all satisfied, M_R^T is by definition equal to the apparent mass of X. We never really have to carry out the experiment because from eq. (9) we have

$$M_X^T - M_R^T = k\Delta\theta + \rho_0 (V_X - V_R)$$

where V_R is given by $V_R = M_R^T/\rho_R$. We can rewrite this relation using density notation:

$$M_X^T - \rho_0 V_X - M_R^T + \rho_0 V_R = k\Delta\theta$$

so that

$$M_X^T (1 - \rho_0/\rho_X) - M_R^T (1 - \rho_0/\rho_R) = k\Delta\theta$$

and with

$$\Delta\theta = 0$$

$$M_R^T = \frac{M_X^T (1-\rho_0/\rho_X)}{(1-\rho_0/\rho_R)} \equiv M_X^A \quad . \tag{11}$$

Note that the denominator is simply a constant whose value is the same for any weight, X. Also, for ρ_X greater than $1 g/cm^3$ (the density of water), M_X^A will be within 0.2 percent of M_X^T. Furthermore, eq. (11) states that the apparent mass can be obtained by multiplying the true mass by a buoyancy factor.

The usefulness of apparent mass and hence eq. (11), as well as the subtle distinctions between true and apparent mass are shown in the next five subsections. We begin with an example.

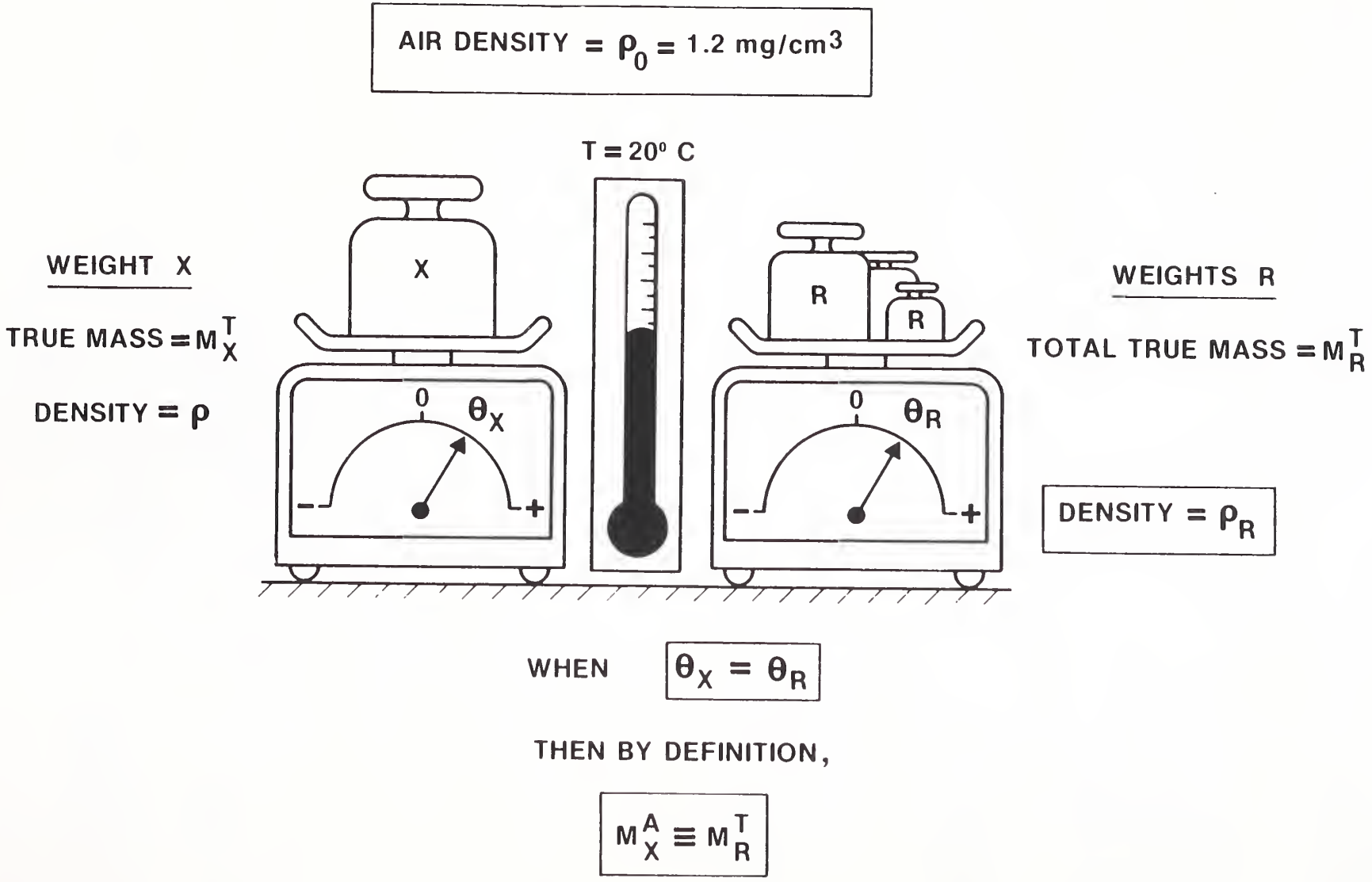


Figure 2. A schematic representation of the definition of apparent mass in the ρ_R basis. The boxed information is specified in the definition.

2.1.1. Example

Using the ideal balance described earlier, we wish to measure the mass of an unknown weight, X, in terms of standard weights, S, of known apparent masses M_S^A's. During the actual measurement, done in air of temperature t and air density ρ_A, it was found that a summation of standards produced the same reading on the balance as the unknown mass.

Under what conditions can we set

$$M_X^A \approx M_S^A \quad ?$$

Solution:

Starting with the balance equation for two masses (eq. (10)).

$$M_1 - M_2 = k\Delta\theta + \rho_A (V_1 - V_2)$$

we have

$$M_X^T - M_S^T = k(\theta_X - \theta_S) + \rho_A(V_X - V_S)$$
$$= \rho_A(V_X - V_S)$$

since the measurement yielded $\theta_X = \theta_S$.

Finally, we can write this as

$$M_X^T - M_S^T = \rho_A \Delta V \quad . \tag{E1}$$

Similarly, we have for the apparent masses, using eq. (11):

$$M_X^A - M_S^A = \frac{M_X^T (1-\rho_0/\rho_X)}{(1-\rho_0/\rho_R)} - \frac{M_S^T (1-\rho_0/\rho_S)}{(1-\rho_0/\rho_R)}$$

$$= \frac{(M_X^T - M_S^T) - \frac{M_X^T}{\rho_X}\rho_0 + \frac{M_S^T}{\rho_S}\rho_0}{(1-\rho_0/\rho_R)} \quad .$$

By setting

$$\frac{M_X^T}{\rho_X} = V_{X0} \quad , \quad \frac{M_S^T}{\rho_S} = V_{S0} \qquad \text{at } 20\ ^\circ C$$

we simplify to

$$M_X^A - M_S^A = \frac{(M_X^T - M_S^T) - \rho_0 (V_{X0} - V_{S0})}{(1-\rho_0/\rho_R)} \quad . \tag{E2}$$

Substituting for $M_X^T - M_S^T$ from (E1), we have

$$M_X^A - M_S^A = \frac{\rho_A(V_X - V_S) - \rho_0 (V_{X0} - V_{S0})}{(1-\rho_0/\rho_R)} \quad . \tag{E3}$$

Volumes have a slight temperature dependence but if the temperature of the measurement is close to 20 °C (generally true in the laboratory), then

$$V_X \approx V_{X0}$$

and

$$V_S \approx V_{S0}$$

so that (E3) reduces to

$$M_X^A - M_S^A = \frac{(\rho_A - \rho_0)(V_X - V_S)}{(1-\rho_0/\rho_R)} = \frac{\Delta\rho\Delta V}{(1-\rho_0/\rho_R)} \quad . \tag{E4}$$

The term $1/(1-\rho_0/\rho_R)$ can be approximated by 1 for the following reason: Via a Taylor expansion

$$\frac{1}{1-\rho_0/\rho_R} = 1 + \frac{\rho_0}{\rho_R} + \left(\frac{\rho_0}{\rho_R}\right)^2 + \left(\frac{\rho_0}{\rho_R}\right)^3 + \ldots$$

Now since ρ_0 is the density of air and ρ_R is the density of a reference metal, we have

$$\frac{\rho_0}{\rho_R} < 10^{-3}$$

so that

$$\frac{\Delta\rho\Delta V}{(1-\rho_0/\rho_R)} = \Delta\rho\Delta V + \Delta\rho\Delta V\ (\rho_0/\rho_R) + \ldots$$

and since $\Delta\rho\Delta V$ is already a small number, a correction of 10^{-3} (or 0.1 percent) is insignificant. Hence only the first term has to be retained. This argument is important in that it is used time and again in estimating buoyancy corrections.

With these approximations, we finally arrive at the equation for apparent mass differences

$$M_X^A - M_S^A = \Delta\rho\Delta V \quad . \tag{E5}$$

We now would like to find the conditions for which

$$M_X^A \approx M_S^A \quad .$$

This implies that $\Delta\rho\Delta V \approx 0$.

When one encounters such a relation, one must always decide its precise meaning. In this particular case the meaning is: $\Delta\rho\Delta V$ is smaller than the acceptable uncertainty of M_X^A.

For example, suppose M_X^A needs to be known to one part per million. Then

$$\Delta\rho\Delta V \approx 0$$

really means that

$$\left| \frac{\Delta\rho\Delta V}{M_X^A} \right| < 10^{-6} \quad . \tag{E6}$$

But we already know that M_X^A does not differ from M_X^T by more than a fraction of a percent (see comments after eq. (11)). We can therefore approximate

$$\rho_X = \frac{M_X^T}{V_X}$$

by

$$\rho_X \simeq \frac{M_X^A}{V_X} \quad .$$

Now (E6) becomes

$$\left| \frac{\Delta\rho}{\rho_X} \right| \left| \frac{\Delta V}{V_X} \right| < 10^{-6} \quad . \tag{E7}$$

The volume ratio $\Delta V/V_X$ can be written as

$$\frac{\Delta V}{V_X} = \frac{V_X - V_S}{V_X} \quad .$$

In the case of different alloys of stainless steel (SS), V_X will be close to V_S, so that

$$\left| \frac{\Delta V}{V_X} \right| < 10^{-1} \quad .$$

We now have

$$\left|\frac{\Delta\rho}{\rho_X}\right| \times 10^{-1} < 10^{-6} \quad . \tag{E7'}$$

The density ratio can be written as

$$\frac{\Delta\rho}{\rho_X} = \frac{\rho_A - \rho_0}{\rho_X}$$

where

$$\rho_X \simeq 8.0 \text{ g/cm}^3 \text{ for SS}$$

and

$$\rho_0 = 1.2 \times 10^{-3} \text{ g/cm}^3$$

so that

$$\frac{\Delta\rho}{\rho_X} = \frac{(\rho_A - 1.2\times10^{-3})}{8.0} \quad .$$

The inequality (E7') now reduces to

$$\left|\frac{\rho_A - 1.2\times10^{-3}}{8.0}\right| \times 10^{-1} < 10^{-6}$$

which is satisfied as long as

$$1.28 \times 10^{-3} \text{ g/cm}^3 \geq \rho_A \geq 1.12 \times 10^{-3} \text{ g/cm}^3.$$

This indicates that the air density during the measurement can be off by as much as 0.08 from 1.20 mg/cm^3 and still satisfy condition (E6).

In contrast, the buoyancy correction for true masses is much more critical. From eq. (E1), we want to satisfy the condition that

$$\left|\frac{\rho_A \Delta V}{M_X^T}\right| < 10^{-6}$$

or

$$\left|\frac{\rho_A}{\rho_X}\right| \left|\frac{\Delta V}{V_X}\right| < 10^{-6} \quad .$$

Using the same material as before, we have

$$\frac{\rho_A}{8.0} \times 10^{-1} < 10^{-6}$$

so that

$$\rho_A < 0.08 \times 10^{-3} \text{ g/cm}^3 \quad .$$

Note that in order to keep the desired accuracy, the air density during the measurement can only be ignored if it is very small, i.e. vacuum.

This has been a long example in which several generally useful techniques have been introduced. It is, therefore, worthwhile to look back and summarize what this example also teaches about apparent mass. The most important inference we can draw is that, for true mass comparisons, we almost always need to account for air buoyancy even though we might be able to use an assumed value for the air density (recall that air density is usually constant in a given laboratory to about ±3 percent). For apparent mass comparisons, the air buoyancy correction is much less important and can, therefore, sometimes be ignored. Apparent mass works the way it does because the formalism already has an assumed value for air density (i.e. 1.2×10^{-3} g/cm^3) built into it. In measurement situations where this assumed value is not accurate enough, the buoyancy term in eq. (E5) becomes important. This eventuality occurs at precisely the same point when ρ_A in eq. (E1) can no longer be sufficiently well approximated by 1.2×10^{-3} g/cm^3.

2.1.2. Reference Materials

So far we have not yet defined a reference material. At present, two different apparent mass bases are utilized by the NBS. One of them, the older one, is called "normal brass" and was the logical choice when most laboratory weights were made of brass.

Normal brass is defined by

$$\rho_{B0} = 8.4 \text{ g/cm}^3 \text{ at } 0 \text{ °C}$$

$$\alpha_B = 5.4 \times 10^{-5}/\text{°C} = \text{coefficient of cubical expansion}$$

$$\rho_0 = 1.2 \text{ mg/cm}^3 = \text{air density at } 20 \text{ °C}.$$

These parameters together with the expression for volumetric expansion

$$V_B(t) = V_{B0}[1+\alpha_B(t-0 \text{ °C})] \tag{12}$$

and the corresponding density correction

$$\rho_B(t) = \rho_{B0}/[1+\alpha_B(t-0 \text{ °C})] \tag{13}$$

completely determine the parameters at 20 °C. Note from above that ρ_{B0} is given at 0 °C and therefore has to be determined for 20 °C via eq. (13) to yield

$$\rho_B(20) = \frac{8.4}{[1+5.4\text{x}10^{-5}\text{x}20]} = 8.390938 \text{ g/cm}^3 \quad . \tag{13'}$$

The second apparent mass basis is referred to an arbitrary material with the following density:

$$\rho_0' = 8.0 \text{ g/cm}^3 \text{ at } 20 \text{ °C} \quad .$$

Once more,

$$\rho_0 = 1.2 \text{ mg/cm}^3 \text{ at } 20 \text{ °C}.$$

Note that this basis does not require any expansion coefficient since all parameters are defined at 20 °C. It is therefore comparable to the basis defined for normal brass via eq. (13').

It has now become common practice to report masses and their uncertainty in the true mass basis and then quote corrections to nominal values for "normal" brass and for $\rho_0' = 8.0$ g/cm^3 bases. Clearly, the latter apparent mass basis is beginning to be preferred by many laboratories since most weights used in a set are made of stainless steel which has a density very close to 8.0 at 20 °C. Thus the correction term in eq. (11) is small, making the apparent mass close to the true mass (see Appendix A.)

We are now in a position to write eq. (11) in the "normal brass" (ρ_B) and in the "8.0" (ρ_0') basis. Starting with eq. (11), we have

$$M_X^A = M_X^T \frac{(1-\rho_0/\rho_X)}{(1-\rho_0/\rho_R)}$$

and substituting for ρ_R with ρ_B from (13), we obtain

$$M_X^A = M_X^T \frac{(1-\rho_0/\rho_X)}{\left(1-\dfrac{\rho_0[1+\alpha_B(t-0\ °C)]}{\rho_{B0}}\right)} .$$

or

$$M_X^A = M_X^T \frac{(1-\rho_0/\rho_X)}{1-\dfrac{\rho_0}{\rho_{B0}}[1+20\alpha_B]} \qquad (14)$$

where the denominator is just a constant equal to 0.9998569886.

Similarly, we use

$$\rho_R = \rho_0'$$

for the 8.0 g/cm^3 basis to yield

$$M_X^A = M_X^T \frac{(1-\rho_0/\rho_X)}{(1-\rho_0/\rho_0')} \qquad (15)$$

where the denominator is once again a constant exactly equal to 0.99985000. It is interesting to note that apparent masses in the two bases we have examined, eqs. (14) and (15), have a constant ratio

$$\frac{(M^A)_{8.0}}{(M^A)_{BRASS}} = 1.00000699$$

That is, the apparent masses of the same weight in the two bases differ by only ~0.0007 percent.

Equation (15) indicates a very important point. If the object being weighed has a density ρ_X equal to 8.0 g/cm^3 (i.e. $\rho_X = \rho_0'$) at 20 °C then

$$M_X^A = M_X^T \quad .$$

For the brass basis, eq. (14), a similar argument has to be treated with more caution. In this case, if ρ_X is equal to the density defined by eq. (13'), then

$$M_X^A = M_X^T \quad .$$

2.1.3. Reporting Apparent Mass

Sometimes the apparent mass correction in a given basis is required. This is defined as

$$\text{Correction } (M_X^A) \equiv M_X^A - N_X \tag{16}$$

where N_X is the nominal value of the weight.

2.1.4. Comparing Two Apparent Masses

We now derive an equation similar to (10), appropriate for apparent mass differences. Using eq. (11) we have for mass 1

$$M_1^A = \frac{M_1^T - \rho_0 V_{01}}{1 - \rho_0/\rho_R}$$

where we have used $M_1^T/\rho_1 = V_{01}$.

Similarly we have for mass 2

$$M_2^A = \frac{M_2^T - \rho_0 V_{02}}{1 - \rho_0/\rho_R} \quad .$$

Subtracting the second from the first yields

$$M_1^A - M_2^A = \frac{M_1^T - M_2^T - \rho_0 (V_{01} - V_{02})}{1 - \rho_0/\rho_R}$$

From eq. (10), we have the true mass difference of the two weights

$$M_1^T - M_2^T = k\Delta\theta + \rho_A (V_1 - V_2)$$

so that, upon substitution, we obtain

$$M_1^A - M_2^A = \frac{k\Delta\theta + \rho_A (V_1 - V_2) - \rho_0 (V_{01} - V_{02})}{1 - \rho_0/\rho_R} \quad . \tag{17}$$

Now, employing eq. (12), we can use

$$V_1 = V_{01}\,[1+\alpha_1(t-t_0)]$$

and

$$V_2 = V_{02}\,[1+\alpha_2(t-t_0)]$$

where V_{01} and V_{02} are the respective volumes at $t_0 = 20$ °C.

Substituting for V_1 and V_2 into (17) yields

$$M_1^A - M_2^A = k\Delta\theta - \rho_0\,(V_{01}-V_{02}) + \rho_A\,[V_{01}+V_{01}\alpha_1(t-20) - V_{02}-V_{02}\alpha_2(t-20)]$$

where we have set

$$t_0 = 20\ °C$$

and have approximated $(1-\rho_0/\rho_R)^{-1}$ by 1 (see derivation of (E5)).

Simplifying further gives

$$M_1^A - M_2^A = k\Delta\theta + (\rho_A-\rho_0)(V_{01}-V_{02})+\rho_A[(t-20)(V_{01}\alpha_1-V_{02}\alpha_2)] \quad . \qquad (18)$$

Equation (18) is a fully corrected formula for calculating mass differences in the apparent mass basis, except for the one approximation mentioned above. At this point it is worthwhile to consider a specific example in order to get a feeling for the magnitude of the terms in eq. (18). Assume that two weights of nominal value 2g have the following properties:

$$V_{01} = 0.2564\ cm^3$$

$$V_{02} = 0.1884\ cm^3$$

also, $\alpha_1 \approx \alpha_2 = 4.5 \times 10^{-5}/°C;$

$$\rho_A = 1.17\ mg/cm^3$$

$$\rho_0 = 1.2\ mg/cm^3.$$

We then find that

$$(\rho_A - \rho_0)(V_{01} - V_{02}) = -0.03 \times 0.068 = -2.04 \times 10^{-3}\ mg$$

and

$$\rho_A(t - 20)(V_{01}\alpha_1 - V_{02}\alpha_2) = 1.17\ (t - 20) \times 3.06 \times 10^{-6}$$

$$= (3.58 \times 10^{-6})(t - 20) \quad .$$

Considering an extreme condition such that t is 10 °C, we have for the latter term

$$= -3.58 \times 10^{-5}\ mg \quad .$$

Combining this term with the first we have

$$-2.04 \times 10^{-3} - 0.0358 \times 10^{-3} = -2.08 \times 10^{-3}\ mg$$

so that the second term (under extreme conditions) contributes only 0.04 μg to the overall buoyancy correction. Usually the (t - 20) term is of the order of 0 °C to 4 °C so that the effect of the second term on the overall buoyancy correction of the above weights is even less. Hence in most cases the second term can be safely ignored and eq. (18) can be approximated by

$$M_1^A - M_2^A = k\Delta\theta + (\rho_A - \rho_0)(V_{01} - V_{02}) \quad . \tag{19a}$$

For true masses we start with eq. (10) and write

$$M_1^T - M_2^T = k\Delta\theta + \rho_A\,(V_1 - V_2) \quad .$$

Employing the same volume expansion formulae used for the apparent masses (after eq. (17)), we derive

$$M_1^T - M_2^T = k\Delta\theta + \rho_A\,(V_{01} - V_{02}) + \rho_A(t-t_0) \times (V_{01}\alpha_1 - V_{02}\alpha_2) \quad . \tag{10'}$$

Once again the second term is very small compared to the first (see arguments above) and the relation can usually be approximated by

$$M_1^T - M_2^T = k\Delta\theta + \rho_A\,(V_{01} - V_{02}) \quad . \tag{19b}$$

A quick comparison between eqs. (19a) and (19b) shows that the apparent mass difference has a correction term which contains the factor $(\rho_A - \rho_0)$ whereas for true masses the corresponding factor is ρ_A. This then shows that the former correction is much smaller. In both cases the unknown parameter is ρ_A.

2.1.5. Sources of Error in Buoyancy Corrections

It is important to determine the sources of error in buoyancy corrections in order to make the best measurements possible with the least amount of effort and expense. Let us look back to eqs. (19a) and (19b) and write the most important buoyancy terms:

True Mass	Apparent Mass
$\rho_A(V_{01} - V_{02})$	$(\rho_A - \rho_0)(V_{01} - V_{02})$

Since ρ_0 is defined exactly, uncertainties in the mass difference measurements arising from buoyancy corrections are due to errors in ρ_A, V_{01}, and V_{02}. Let us denote the uncertainty in a quantity by a preceding δ (i.e., the uncertainty in ρ_A is $\delta\rho_A$). Then, from eqs. (19a) and (19b) the sources of uncertainty in the buoyancy corrections are

True Mass	Apparent Mass
$\delta\rho_A(V_{01} - V_{02})$	$\delta\rho_A(V_{01} - V_{02})$
$\delta V_{01}\,\rho_A$	$\delta V_{01}(\rho_A - \rho_0)$
$\delta V_{02}\,\rho_A$	$\delta V_{02}(\rho_A - \rho_0)$

where we will assume that δV_{01}, δV_{02}, and $\delta\rho_A$ are uncorrelated.

In the calculation of true mass differences, uncertainty in the volume of the masses is much more important than it is for differences in apparent mass as can be seen by comparing similar terms:

$$(\delta V_{01} + \delta V_{02})\rho_A > (\delta V_{01} + \delta V_{02})(\rho_A - \rho_0) \quad .$$

Also we see that any uncertainty in V_{01} or V_{02} will lead to systematic uncertainty in mass measurements (even if $V_{01} = V_{02}$).

Example:

Two weights produce the same reading on an ideal balance in air of ρ_A = 1.17 mg/ and t = 20 °C. Both weights are said to be of density 7.89 g/cm^3. Find

1) their true mass difference
2) their apparent mass difference against ρ_0' = 8.0 g/cm^3
3) the errors in these numbers due to uncertainty in the volumes.

The weights are marked "1 kg".

Answers:

1) From eq. (19b)

$$M_1^T \left(1 - \frac{\rho_A}{7.89}\right) - M_2^T \left(1 - \frac{\rho_A}{7.89}\right) = k \cdot \Delta\theta = 0$$

because $\Delta\theta$ is given as 0.

Hence $M_1^T - M_2^T = 0$.

2) From eq. (19a)

$$M_1^A - M_2^A = (\rho_A - \rho_o)(V_{01} - V_{02}) + k\Delta\theta = 0$$

since $\Delta\theta = 0$ and $V_{01} = V_{02}$ (because $M_1^T = M_2^T$ and $\rho_{01}' = \rho_{02}'$).

3) the weights could have come from different lots of metal. Since no uncertainty is given for density, we must assume that

$$\rho_0' = 7.89 \pm 0.005 \text{ g/cm}^3 \quad ,$$

i.e. the density is uncertain in the first unreported decimal place.

As a worst case, let

$$\rho_{01}' = 7.895 \text{ and } \rho_{02}' = 7.885 \quad .$$

Then

$$\left|\delta(M_1^T - M_2^T)\right| \simeq \rho_A \left(\frac{M_2^T}{7.885} - \frac{M_1^T}{7.895}\right) \simeq \rho_A M_1^T \left(\frac{1}{7.885} - \frac{1}{7.895}\right)$$

$$\simeq 0.00117 \times 1000 \left(\frac{1}{7.885} - \frac{1}{7.895}\right) \simeq 0.270 \text{ mg}$$

$$\left|\delta(M_1^A - M_2^A)\right| \simeq 0.00003 \times 1000 \left(\frac{1}{7.885} - \frac{1}{7.895}\right) \simeq 0.005 \text{ mg}$$

Note that the nominal mass value of each weight is accurate enough to use in estimating the uncertainties in question.

CAUTION: In calibration certificates, it is normal practice not to include in the assignment of systematic errors any contribution from volume uncertainties: The volumes, usually calculated from densities provided by the weight manufacturers, are assumed to be without error. For the most critical work, volumes of individual weights must actually be measured. (Since 1982, NBS has been routinely measuring the volume of single-piece 1-kilogram weights which require the best possible mass calibration.)

We have yet to examine the effects on mass measurements of uncertainty in ρ_A. For mass differences, we have shown above that the errors are approximately $\delta\rho_A(V_{01} - V_{02})$. Therefore, unless $V_{01} = V_{02}$, we mst be concerned with uncertainty in ρ_A. The calculation of ρ_A along with an analysis of the attendant uncertainties are presented in the next section.

2.2. Air Density

We have seen from the derivations in Section 2.1 that the correction to the mass determination requires the precise knowledge of the air density (ρ_A) if the weighing is done in an air environment. (We will assume from now on that we are always working in air.) The buoyancy correction in eq. (19b) requires ρ_A for true masses, whereas for apparent masses we require $(\rho_A - \rho_0)$ in eq. (19a). It should be remembered that ρ_0 is defined as $\rho_0 = 1.2$ mg/cm^3 which is the approximate value for 20 °C with 1 atmosphere of pressure with 50% relative humidity.

The usual approach to determining ρ_A is to calculate the density of air based on an "equation-of-state." To use this equation, one must supply ambient values for temperature, barometric pressure, and relative humidity. For extremely exacting work, the concentration of carbon dioxide in the laboratory ambient is also required. Appendix B contains a derivation of the equation-of-state based on the work of Jones [2]. This formulation is close to those which are found in handbooks or older references but has three unique virtues:

1. It is based on the most recent reference data available.
2. It has been derived in such a way that uncertainties associated with the equation itself are known and stated.
3. Virtually the same equation has been endorsed for international use by the Consultative Committee for Mass of the International Committee of Weights and Measures (C.I.P.M.).[3]

The interested reader is urged to consult Appendix B and especially Ref. [2] for the complete derivation. We state here that[4]

$$\rho_A \left(\frac{mg}{cm^3}\right) = \frac{0.0034848}{(t+273.15)} \cdot (P - 0.0037960 \cdot U \cdot e_s) \qquad (20a)$$

where t = temperature in °C
P = barometric pressure in pascals (133.322 Pa = 1 mm Hg)
U = relative humidity in %
e_s = saturation vapor pressure.
$e_s = 1.7526 \times 10^{11} \exp(-5315.56/(273.15 + t))$ Pa is an excellent approximation.

If barometric pressure is measured in mm Hg, then

$$\rho_A \left(\frac{mg}{cm^3}\right) = \frac{0.46460}{(t+273.15)} (P - 0.0037960\ U \cdot e_s) \qquad (20b)$$

where $e_s = 1.3146 \times 10^9 \exp(-5315.56/(273.15 + t))$ mm Hg is an excellent approximation.

2.3. Sensitivity Arguments

The major correction normally required for mass determination is due to air buoyancy. As is evident from eqs. (19a) and (19b) the only parameter that has to be determined very carefully, aside from the volumes of the weights being compared, is the air density. Equation (20) clearly indicates that three parameters affect the air density, namely humidity, barometric pressure, and temperature. We will see in Section 2.3.5 that an overall uncertainty in ρ_A of ~ 0.0030 mg/cm^3, i.e. ~ 0.25 percent of ρ_A near sea level, is usually acceptable. In Appendix C we see that the total uncertainty in ρ_A is the square root of the sum of the squares of the individual uncertainties arising from measurements of humidity, temperature and pressure, (U, t and P). Thus one way of assuring that the total uncertainty of ρ_A is within acceptable limits is to make sure that the individual uncertainties from U, t and P are each less than $1/\sqrt{3}$ of the total acceptable uncertainty. That is, we set a goal that the contribution to uncertainty in ρ_A from U, P, or t should be less than 0.0017 mg/cm^3. We will see in Section 2.4 that this goal can be met or exceeded with a modest investment in equipment. We now show the relative importance of measurements of humidity,

[3]The CCM recommends that this equation referred to as "Equation for the determination of the density of moist air (1981): in anticipation of improved reference data. There are some stylistic differences in the CCM version of [2] but the calculated values are completely consistent at meaningful levels of precision.

[4]Equations (20a) and (20b) are somewhat simplified (see Appendix B). Weighing errors due to the simplifications are usually very small and are estimated in Table 5, p. 427 of ref. [2].

temperature, and pressure to the uncertainty in air density; we also indicate now accurately these must be measured to achieve our goal of 0.0017 mg/cm^3 for each one.

2.3.1. Humidity

It is easiest to discuss the effects of humidity (U) via examples. For practical purposes we can consider as standard a condition when t = 20 °C, P = 760.00 mm Hg, and U = 50 percent.

Example 1

Consider t = 20 °C, P = 760.00 mm Hg, U = 50 percent (i.e. standard condition). Using eq. (20b), we find

$$\rho_A = 1.2045 - 0.0053 = 1.1992 \text{ mg/cm}^3 \quad .$$

We note that the humidity correction, 0.0053 mg/cm^3, corresponds to a correction of only 0.44 percent in ρ_A.

Example 2

Same conditions as in Example 1 except t = 30 °C. We calculate

$$\rho_A = 1.1648 - 0.0093 = 1.1555 \text{ mg/cm}^3$$

and the humidity corrections is ∿0.8 percent of ρ_A.

Example 3

Same as Example 1 except U = 60 percent.

$$\rho_A = 1.2045 - 0.006330 = 1.1982 \text{ mg/cm}^3$$

so that the humidity correction is 0.5 percent.

Summarizing the results on humidity we can make the following observations:

i. The magnitude of the humidity correction to the air density at standard conditions is ∿0.4 percent.

ii. By increasing the temperature from 20 °C to 30 °C, the humidity correction amounts to 0.8 percent of the air density. Thus even under these high temperatures, very unlikely in a temperature-controlled laboratory, the humidity correction term is still quite small.

iii. A change of 10 percent in the relative humidity (i.e. $U = U_0 + 10$) changes the air density by only 0.08 percent (i.e. $\frac{1.1992-1.1982}{1.1992}$).

We can conclude, therefore, that if we want to know the air density to within 0.0017 mg/cm^3 then the relative humidity should be known to ±16 percent. (If a safety factor of 4 is required then the relative humidity should be known at the ±4 percent level).

2.3.2. Temperature

The effects of temperature on the humidity term were already demonstrated in Example 2. In particular we note that a 10 °C change, yielded a humidity correction of 0.8 percent.

The most sensitive effect of temperature on calculated air density, however, occurs in the first term

$$\rho_A = \frac{0.46460}{(t+273.15)} P$$

From Example 1, we note that at standard conditions the first term yields

$$\bar{\rho} = 1.2045 \text{ mg/cm}^3$$

Example 4

P = 760.0 mm Hg, t = 21 °C

In this case the first term of eq. (20b) gives

$$\rho_A = 1.2004 \text{ mg/cm}^3$$

and ρ_A has changed from $\bar{\rho}$ by 0.0041 mg/cm^3 or 0.34 percent.

To achieve an uncertainty of 0.0017 mg/cm^3, we must measure the temperature accurately to $\sim\pm$0.4 °C (which reduces to ±0.1 °C for a safety factor of 4.)

2.3.3. Pressure

The arguments for the pressure uncertainty follow those of the temperature. Once again we only have to consider the first term of the density equation

$$\rho_A = \frac{0.46460}{(t+273.15)} P$$

For standard conditions of t = 20 °C, and P = 760.0 mm Hg

$$\bar{\rho} = 1.2045 \text{ mg/cm}^3$$

Example 5

P = 764.00 mm Hg
t = 20 °C

In this case ρ_A = 1.2108 and has changed from $\bar{\rho}$ by 0.53 percent.

Thus in order to know ρ_A to 0.0017 mg/cm^3 we will need to have accurate barometric measurements to ±1.1 mm Hg (which reduces to ±0.3 mm Hg for a safety factor of 4).

2.3.4. Overall Uncertainty of Air Density

We have seen in eq. (20) that the air density is a function of the temperature, pressure, and humidity

$$\rho_A = \rho_A(t,P,U) \quad .$$

To evaluate the overall uncertainty of ρ_A, we assume uncorrelated errors and develop a final value in Appendix C:

$$\frac{\delta\rho_A}{\rho_A} \sim 0.25 \text{ percent} \quad .$$

Note in particular that this value was obtained using the errors established in Sections 2.3.1, 2.3.2, and 2.3.3 of this section (with no safety factor).

2.3.5. Effects of Small Changes in Air Density on the Mass Value

Having established the sensitivities of U, t, P on ρ_A, the air density, we can now proceed and see what effect a small change of ρ_A has on the mass determination. The arguments can best be carried through with the following example:

Assume that we are comparing the masses of two 1 kg weights.

Weight #1 has density of 8.4 g/cm^3
Weight #2 has density of 7.5 g/cm^3
Air density ρ_A = 1.20 mg/cm^3 .

If the air density is known to an accuracy of 0.25 percent, what uncertainty does this cause in the mass difference calculations?

Answer: We found in Section 2.1.5 that the uncertainty for true as well as apparent mass was given as

$$\text{uncertainty} = \pm\delta\rho_A\,(V_{01}-V_{02})$$

$$= \pm\delta\rho_A N\left(\frac{1}{\rho_1} - \frac{1}{\rho_2}\right)$$

where we have approximated the volumes by

$$V_1 \sim N/\rho_1;\ V_2 \sim N/\rho_2$$

with N being the nominal mass of each weight.

Substituting numerical values, yields

$$\text{Uncertainty} = \pm 0.0025 \times \rho_A \times 1000\left(\frac{1}{8.4} - \frac{1}{7.5}\right)$$

$$= \pm 0.0025 \times 1.2 \times 1000\left(\frac{1}{8.4} - \frac{1}{7.5}\right)$$

$$= \pm 0.043 \text{ mg.}$$

This is about one-half the calibration uncertainty limit provided by the NBS for high-quality 1-kg weights. The span 7.5 to 8.4 g/cm^3 includes the densities of 1-kg weights used in routine fine work.

It is usually true that the most stringent requirements placed on buoyancy measurements are in the precise comparison of kilograms. An uncertainty in ρ_A of 0.25 percent is usually sufficient even in this case.

CAUTION: Some weights are manufactured for purposes other than general mass metrology. For instance, a scientist may have a 1-kg aluminum weight which needs the best calibration possible. In such special situations, the requirements and limitations placed on buoyancy work must be examined on a case-by-case basis.

2.4. Measuring Equipment for Air Density

The previous section yielded information on the accuracies required for the three parameters needed for the air density. In this section we discuss briefly the types of equipment required to meet our accuracy goals.

2.4.1. Humidity

This parameter is usually measured with a psychrometer or hygrometer. Several commercial instruments can provide accuracies of ±2% relative humidity over a wide range such as 10-80%. The calibration of such instruments can be verified at fixed points by the user by means of an in-laboratory calibration station. Since an accuracy of ±16% (or ±4% using a "safety factor of 4") is required for mass metrology, these instruments are more than adequate.

2.4.2. Temperature

An accuracy of ±0.4 °C (or ±0.1 °C for a "safety factor of 4") is required. Simple, mercury-filled, glass thermometers can provide reading accuracies up to ±0.1 °C. More elaborate units can yield ±0.01 °C.

The user should be very careful in taking temperature readings. Generally, the physical presence of the metrologist tends to warm up the balance with respect to the rest of the room. It is therefore essential that, no matter what form of thermometer is used, its sensor be placed as close as possible to the balance pan. Also, every attempt should be made to ensure that objects being weighed are in thermal equilibrium with the balance. The latter requirement is especially important when comparing weights of large surface area or of different geometries.

2.4.3. Barometric Pressure

The required accuracy is ±1.1 mm Hg (or ±0.3 mm Hg for a "safety factor of 4"). One can obtain mercury manometers or aneroid barometers with reading accuracies of ±0.05 mm Hg. Since this is sufficient for usual mass calibrations, one does not have to push for higher accuracy instruments.

Present technology can provide defining instruments with overall uncertainty of ±0.01 mm Hg. Since such devices are the most accurate on the market, they are generally used in calibration of other manometers or aneroid barometers which consequently would have greater uncertainty.

Many laboratories possess either Fortin-type or aneroid barometers. These require calibration against some defining instrument--such as a mercury manometer or a piston gage; but such defining instruments are costly and may not be available in the laboratory. On the other hand, aneroid and Fortin-type barometers are best calibrated in place.

A solution to this problem has recently been demonstrated [9]. Two weights of well-known and nearly-equal mass but having very different volumes are compared in the laboratory on a sensitive balance. The measured difference in balance readings between these two weights, $k\Delta\theta$, determines the air density in the balance enclosure via eq. (9). The air density is also calculated from eq. (20) using a calibrated thermometer, a calibrated hygrometer, and the uncalibrated barometer. The difference in the two measurements of air density serves to calibrate the barometer. Calibration uncertainties of less than 1 mm Hg (three standard deviations plus known systematic errors) have been demonstrated.

In summary, we note that all the requirements for air density measurements can be met with presently available instrumentation. If, however, the accuracy of ±0.25 percent for ρ_A has to be improved, then the limiting instrument will be the pressure gage. (It does not make sense to improve temperature and humidity resolution by an order of magnitude, if the pressure cannot be read with higher precision.)

3. WEIGHING METHODS

Until this point, we have assumed that all mass comparisons have been carried out on the idealized balance defined in Section 2. In this section we develop the means of evaluating the constant, k of eq. (8). We also extend our analysis to actual balances currently in use. Just as one need not be a mechanical engineer to drive a car, so one need not have a detailed knowledge of balance design in order to carry out successful mass measurements. Nevertheless, a general knowledge of balance design is often useful. The review paper by Schoonover [10] provides an excellent introduction to this subject.

The idealized balance of Section 2 had no imprecision. Unfortunately, this is not the case for real balances. By saying a balance has imprecision, we mean two distinct contributions to uncertainty:

1) Resolution. A balance observation can be no more certain than one's ability to resolve the least-significant digit of the read-out.
2) Reproducibility. A balance observation can be no more certain than the tendency of the balance to produce identical read-outs under identical conditions. The lack of this property is called "scatter."

For balances used in the best metrology, scatter is by far the more important of the two. What we loosely term the scatter of the balance is not entirely intrinsic to the balance. That is, it may depend on the type of table on which the balance is placed, the type of air-conditioning in the laboratory, the skill of the balance operator, etc. We need a measure of the scatter which includes all these effects. Such a measure is provided by the "process standard deviation" [7,8].

The standard deviation for a particular weighing process is easily estimated by repeating the process a number of times. (Ten repetitions are usually sufficient to give a respectable estimate.) Two weights, representative of the weights normally measured on a particular balance, are intercompared n times by one of the methods described below. The n values obtained are $a_1, a_2, a_3, \ldots, a_{n-1}, a_n$. The average of the n values is

$$\bar{a} = \frac{\sum_{i=1}^{n} a_i}{n}$$

and the standard deviation (σ) of the balance is estimated to be

$$s = \left(\frac{\sum_{i=1}^{n} (a_i - \bar{a})^2}{n-1} \right)^{1/2}, \tag{21}$$

with $s \to \sigma$ as $n \to \infty$. (The estimate s approaches the actual standard deviation σ as n is made larger.)

The standard deviation should be estimated in this way at regular intervals in order to spot trends or shifts in the evolution of s with time. A slow rise in s could signal, for example, a slow deterioration of the balance. An abrupt shift could signal a problem with the balance, a change in the air-conditioning system of the laboratory, or some other problem which requires attention.

Before actually discussing various types of weighing procedures, we have to define some terms required in the descriptions. The following symbols will be utilized.

M = any unknown weight with true mass M^T whose value has to be determined

C = a counterpoise (or counterweight) with true mass C^T.

This weight is usually required on beam balances. It can be of any stable shape, form and composition. However, it has to be adjusted so that the scale (balance) does not deflect past its maximum allowable value when loaded with a weight. This mass appears in equations which are combined in such a way that C^T drops out of the final result.

S = a standard weight with true mass S^T.

This is a known mass against which the unknown is measured. It is always assumed that in any process S^T is known and is traceable to the NBS. In the following description S can be of a single unit or can be made up of

$$S = \sum_{i=1}^{n} S_i$$ such that the sum S^T roughly equals M^T or C^T.

Δ = a sensitivity weight with true mass Δ^T

This mass is usually of very small magnitude. In principle it is only used to measure the deflection of the scale per unit mass. This mass is very critical. It must be well calibrated and traceable to the NBS, although such calibration is a relatively simple matter.

Of the various weighing procedures, we will mention Direct Weighing, Direct-Reading Weighing, and discuss Substitution and Transposition Methods.

3.1. Direct Weighing

This type of weighing is seldom used in the mass calibration program. It requires the use of an equal-arm balance. Essentially, the reading steps are as follows:

1. Release brake and take null reading with both pans empty.
2. Load one pan with M and balance with S masses on the opposing pan until the null position is once again reached.

We have as final result, if buoyancy corrections are insignificant,

$$M^T = \sum_{i=1}^{n} S_i{}^T$$

3.2. Direct-Reading Weighing

This weighing procedure requires a one-pan balance that reads directly in S.I. mass units or some other commercially recognized unit (e.g. pounds or carats). The procedural steps are as follows:

1. Null the balance without any load on the pan.
2. Place M on pan and record the mass directly.

This method is obviously quite simple but usually is not accurate enough for metrology. It is appropriate for scales most commonly used in stores. Nevertheless, the method is quite similar in principle to that described in Section 3.4.5., below.

3.3. Substitution Weighing

This weighing procedure is the usual one employed for both single pan or equal-arm balances. The basic idea is the comparison of an unknown mass with a standard mass and a sensitivity mass.

3.3.1. Single Pan Balance

Known as "direct-reading analytical balances," they contain built-in weights which are usually manipulated by external knobs or dials. These we will refer to as "dial weights." The remaining balance reading is indicated by the rest point of a moving optical screen or by an automatic electronic display. This part of the balance output is called the "screen reading." Thus each balance indication is the summation of the dial weights and the screen reading. Calibration of the dial weights is unnecessary for the measurements described below. The majority of balance operations outside metrology (e.g., chemistry, metallurgy, etc.) do rely on the accuracy of the built-in dial weights, however. For calibration of these weights the reader is referred to Ref. [11] and Appendix A.

3.3.1.1. Single Substitution

To find the mass difference between two weights of nominally equal value, we will first describe the method of "single substitution". For the moment, let us assume that the weights are matched closely enough so that, when placed on the balance, they both require the same setting of dial weights. Therefore they differ only in their screen readings. If we look into the balance (fig. 3), we see that the pan is suspended from a beam which pivots about a very good bearing (the fulcrum, usually a knife on a flat). On the end of the beam opposite to the pan, is a fixed counterpoise C. (There will also be either a servo-motor on this end or a dashpot to damp the beam oscillations. These details are unimportant to our derivation.) Also suspended from the pan-end of the beam are the built-in dial weights. Thus when an object is placed on the pan, dial weights are removed so that the force of the total assembly on the pan side (object being weighed plus remaining dial weights) balances the counterweight as closely as possible. The remaining imbalance causes the beam to rock slightly out of its horizontal position. The angle of tilt is proportional to the remaining force imbalance. In a mechanical balance, this angle is directly proportional to the screen readings. In a servo-controlled balance, a motor is used to drive the beam back to horizontal. The screen readings are then proportional to the average force generated by the motor.

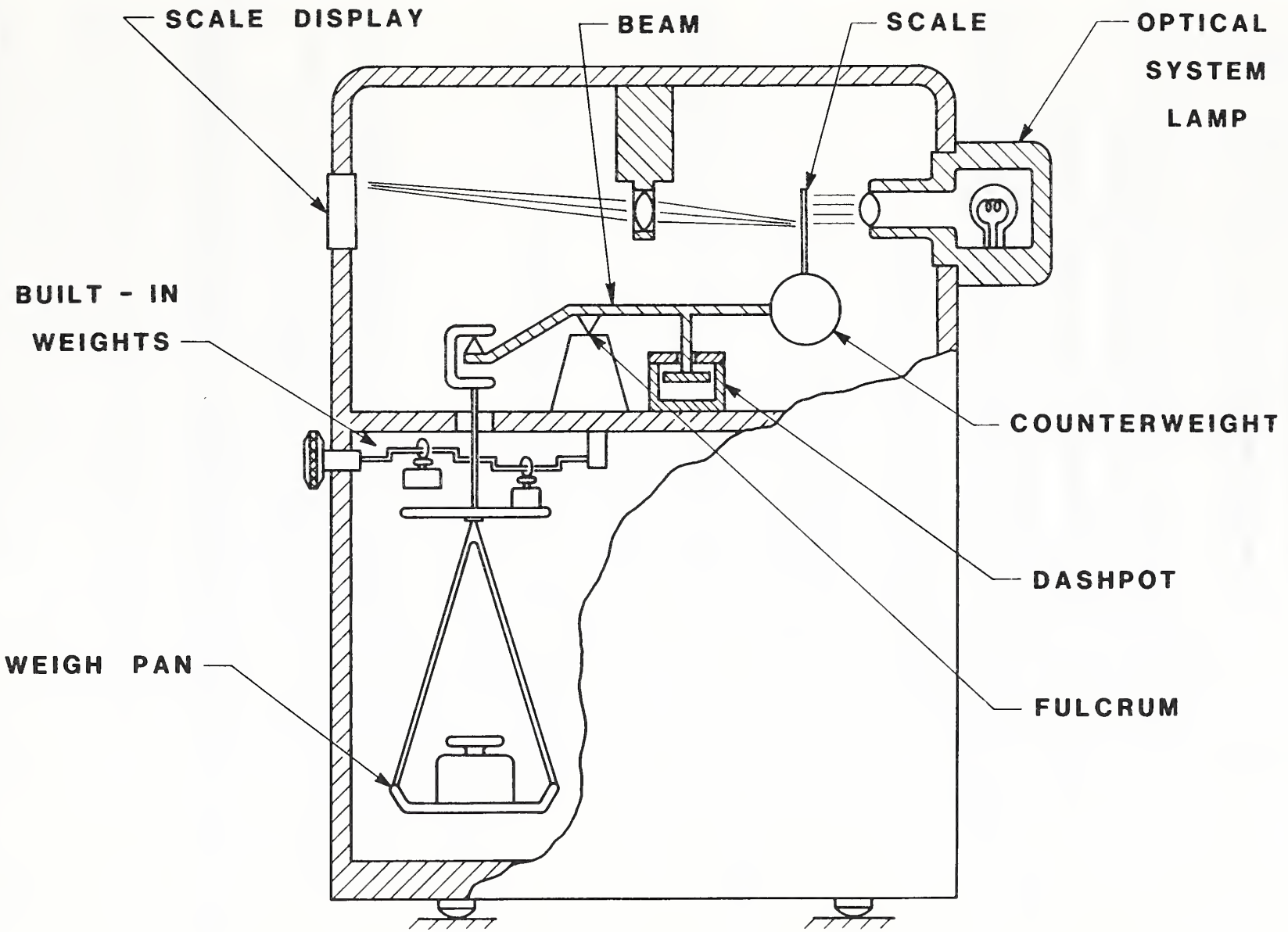


Figure 3. One type of single-pan balance. One dial for removing built-in weights is shown.

In the balances just described, we can define the following quantities:

ΣD_i^T = the true mass of the dial weights removed

ΣV_i = the volume of the dial weights removed

θ_1 = the screen reading

θ_0 = the screen reading when the pan is empty and the dial weights are replaced (this was always zero for the ideal balance of Section 2)

k = the proportionality constant between screen reading and mass.

(Recall the previous definitions for M, S, Δ.) We also recall from eq. (8) that for two weights, M_1 and M_2 with volumes V_1 and V_2, the weighing equation can be written as (using true masses)

$$\bar{g}\,(M_1^T - \rho_m V_1) = \bar{g}\,(M_2^T - \rho_m V_2) + \bar{g}\,k(\Delta\theta) \quad .$$

With

$$\rho_m = \rho_A = \text{air density};$$

$$(\Delta\theta) = \theta_1 - \theta_2;$$

and setting

$$M_2^T = \Sigma D_i^T$$

$$V_2 = \Sigma V_i$$

we arrive at[5]

$$\bar{g}(M_1^T - \rho_A V_1) = (\Sigma D_i^T - \rho_A\, \Sigma V_i)\bar{g} + \bar{g}k(\theta_1 - \theta_0)$$

where $\theta_0 \equiv \theta_2$ in this notation.

The following observations can be made:

- The counterpoise weight does not enter into the final result.
- The gravitational acceleration, $\bar{g}$, can be cancelled from both sides of the above balance equation.
- The above equation is identical to that for the true mass difference between weight #1 and the removed dial weights (see Appendix A).

In single substitution, we compare a weight M with a standard weight S. Either M or S may actually be a summation of weight pieces. A small calibration weight, sensitivity weight Δ, is also required.

---

[5] A full derivation would involve a balance equation for each weighing operation (i.e. the pan unloaded and the pan loaded). These equations would each involve the counterpoise C. Combining these relations would then give the desired result.

The steps in single substitution are:

1) Load M and record screen reading θ_1
2) Remove M
3) Load S and record screen reading θ_2
4) Add Δ to S and record screen reading θ_3

In general we adjust S and choose Δ so that

- The same dial weights are used in steps 1, 3, and 4.
- $|\theta_3 - \theta_2|$ is at least four times as large as $|\theta_2 - \theta_1|$
- $(\theta_3 - \theta_2) \sim 1/5$ the total range of the screen.

From the above steps 1, 3 and 4, we have the following relations:

$$k\theta_1 \doteq (M^T - \rho_A V_M) - (\Sigma D_i^T - \rho_A \Sigma V_i) + k\theta_0 \tag{22}$$

$$k\theta_2 \doteq (S^T - \rho_A V_S) - (\Sigma D_i^T - \rho_A \Sigma V_i) + k\theta_0 \tag{23}$$

$$k\theta_3 \doteq (S^T - \rho_A V_S) + (\Delta^T - \rho_A V_\Delta) - (\Sigma D_i^T - \rho_A \Sigma V_i) + k\theta_0 \ . \tag{24}$$

The symbol $\doteq$ indicates that we are now assuming that some or all of the measured values have a non-negligible uncertainty.

Subtracting (23) from (22) yields

$$M^T - S^T = \rho_A(V_M - V_S) + k(\theta_1 - \theta_2) \ . \tag{25}$$

Subtracting (24) from (23) yields

$$k = \frac{\Delta^T - \rho_A V_\Delta}{(\theta_3 - \theta_2)} \ . \tag{26}$$

Finally, substituting for k from (26) into (25) yields

$$M^T - S^T = \rho_A(V_M - V_S) + \frac{\Delta^T - \rho_A V_\Delta}{(\theta_3 - \theta_2)}(\theta_1 - \theta_2) \ . \tag{27}$$

Note that this equation is identical to (10), the crucial difference being that we have found the value for the proportionality constant k.

One may wonder why k must be evaluated for every measurement if it is a constant. The answer is that for most sensitive balances, k is truly constant only for short periods of time and only over reduced regions of the range of the screen.

In the derivation above we choose Δ so that $\theta_3 - \theta_2$ is about 20 percent of the screen range (fig. 4). In this reduced region, k will be sufficiently constant for metrology if one is using well-made balances.

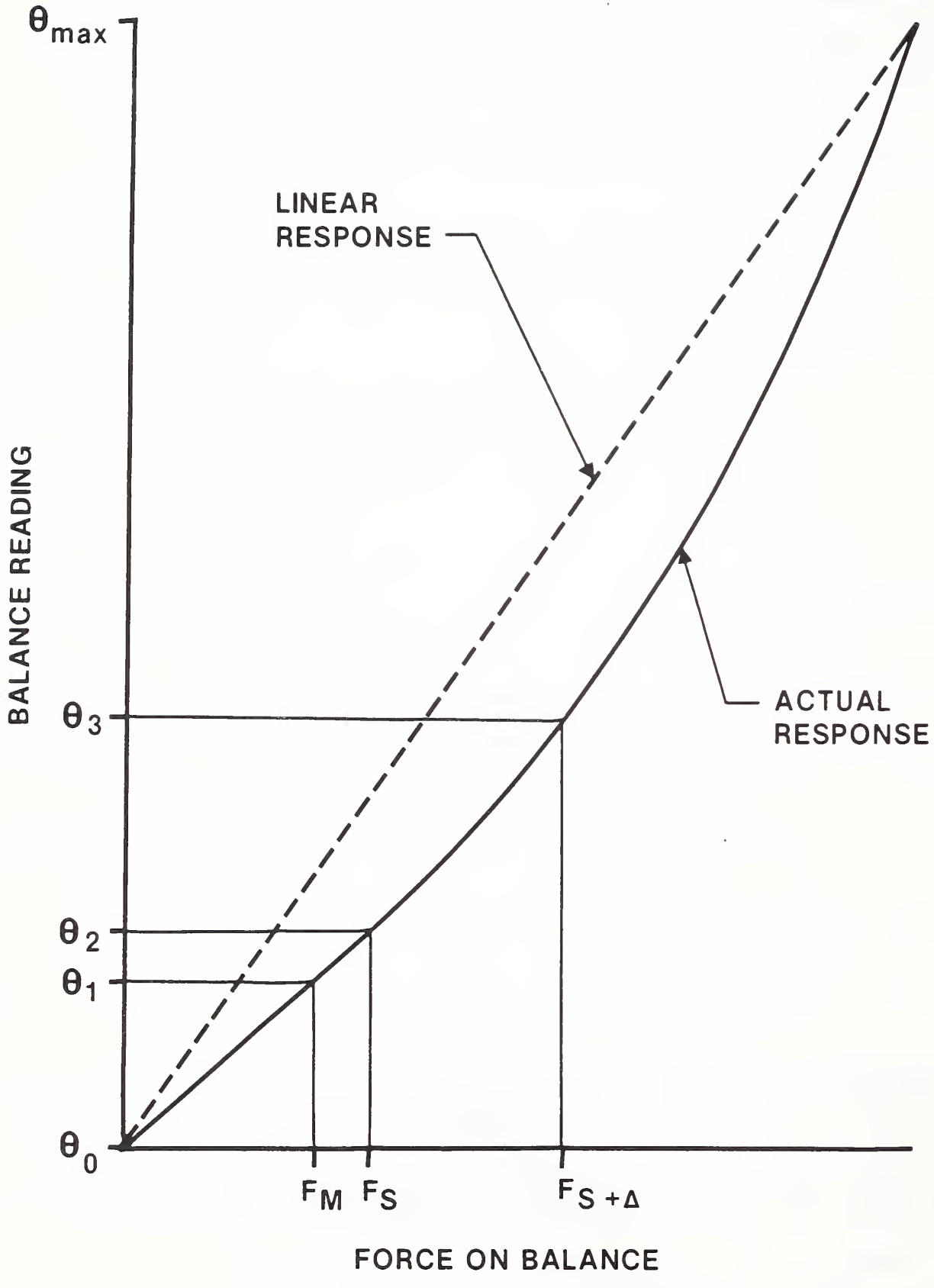


Figure 4. Balance indications as a function of the force on the pan. The response shown is that of the screen when the dial weights are not changed.

3.3.1.2 Double-Substitution

In the derivation of single-pan substitution given above, we have treated θ_0 as a constant. This is the same as assuming that the balance indication does not drift with time. But this is rarely the case. Usually, however, the zero-drift in the balance will be approximately linear over a short period of time. The method known as single-pan double-substituion has the following two advantages over single-substitution:

- linear drifts in the balance zero are eliminated if there is an equal time between successive observations;
- a second estimate of the mass difference between M and S is obtained.

Both benefits are obtained by a single additional measurement:

5) Remove S from pan, leave Δ where it is

6) Place M on the pan and record secreen reading θ_4. The last step gives the additional balance relation

$$k\theta_4 \doteq (M^T-\rho_A V_M) + (\Delta^T-\rho_A V_\Delta) - (\Sigma D_i^T-\rho_A \Sigma V_i) + k\theta_0 \quad . \tag{28}$$

Substracting (24) from (28) yields

$$M^T - S^T = \rho_A(V_M-V_S) + k(\theta_4-\theta_3) \quad .$$

Substituting for k from eq (26) then gives

$$M^T - S^T = \rho_A(V_M-V_S) + \frac{(\Delta^T-\rho_A V_\Delta)}{(\theta_3-\theta_2)} \quad (\theta_4-\theta_3) \quad . \tag{29}$$

Finally, by adding (29) and (27) we obtain

$$M^T - S^T = \rho_A(V_M-V_s) + \frac{(\Delta^T-\rho_A V_\Delta)}{2(\theta_3-\theta_2)} \quad (\theta_1-\theta_2-\theta_3+\theta_4) \quad . \tag{30}$$

One might wonder how eq. (30) would have to be changed if θ_0 were drifting linearly with time instead of being constant. The answer is: the term $(\theta_1-\theta_2-\theta_3+\theta_4)$ would still be strictly correct; the term $(\theta_3-\theta_2)$ would not be strictly correct. However, provided that the drift among readings is small compared to $(\theta_3-\theta_2)$, the error from this source will be negligible. This will in actuality always be the case. If it were not, the balance would be drifting so badly that no stable readings could be obtained. If instead of $(\theta_3-\theta_2)$ one substitutes the term $(3\theta_3-3\theta_2+\theta_1-\theta_4)/2$, linear drift in the sensitivity calculation will be removed [12]. This more complicated formulation is generally used when data are being analyzed by digital-computer software.

CAUTION: If θ_1 or θ_2 is very close to zero, balance drift may cause readings to become negative or to go off scale. To avoid this problem, a small, uncalibrated "tare weight" can be added to the pan during all measurements. Mathematically, this is equivalent to changing θ_0 by a constant amount and so has no effect on the final results.

3.3.2. Two-Pan Equal-Arm Balance

This is the oldest type of precision balance. Its basic design is represented by the "scales of justice" or by the zodiac sign for Libra.

3.3.2.1. Single Substitution

This method is very similar to single-substitution on a single-pan balance. The read-out provided by an equal-arm balance corresponds to just the screen-reading of the single-pan balance. An uncalibrated counterpoise which has nearly the same mass as M and S is placed on one of the pans throughout the measurement sequence. The size of the counterpoise is selected to ensure that the balance indication will be on scale with M on the second pan. The weights M, S and a sensitivity weight Δ are placed on the second pan in nearly the same sequence used in the one-pan case.

The required steps are as follows:

1) Place C on first pan, M on second pan and record indicator reading θ_1.
2) Remove M
3) Place S on second pan and record indicator reading θ_2. (Recall that S is a summation of calibrated weights which is chosen to be near the scale indication of M.)
4) Place the sensitivity weight Δ on that pan which will cause the indicator to move toward the center of the scale. Record indicator reading θ_3.

It is convenient to choose as the second pan (i.e. the one on which M and S are placed) the one for which increasing increments of load give increasingly positive indicator readings.

Going through detailed arguments similar to those for single-pan, single substitution, we arrive at

$$M^T - S^T = \rho_A(V_M - V_S) + \frac{(\Delta^T - \rho_A V_\Delta)}{|\theta_3 - \theta_2|}(\theta_1 - \theta_2) \qquad (31)$$

The absolute value $|\theta_3 - \theta_2|$ comes about because the Δ weight might be placed on either of the two pans. Note that if the arms of the balance are not equal, the sensitivity weight must be placed on the second pan.

3.3.2.2. Double-Substitution

As in the case of the single-pan balance, one additional reading is taken for double substitution.

5) Remove S from the balance; leave Δ where it is,
6) Place M on the second pan and record the indicator reading θ_4.

By obvious extension of the preceding arguments, we arrive at the equation for double-substitution on an equal-arm balance:

$$M^T - S^T = \rho_A(V_M - V_S) + \frac{(\Delta^T - \rho_A V_\Delta)}{|\theta_3 - \theta_2|} \frac{(\theta_1 - \theta_2 - \theta_3 + \theta_4)}{2} \quad . \tag{32}$$

Again, to be strictly correct in the presence of linear drfit, the term $\theta_3 - \theta_2$ can be replaced by $(3\theta_3 - 3\theta_2 + \theta_1 - \theta_4)/2$. Double substitution in this case has the same advantages as weighing on a single-pan balance.

In summary, we have derived the following equations for the four types of substitution weighing we have discussed:

Balance Type	Substitution Type	Equation
Single-Pan	Single	(27)
Single-Pan	Double	(30)
Two-Pan	Single	(31)
Two-Pan	Double	(32)

3.4. Transposition Weighing

Two-pan, equal-arm balances can also be used to do transposition weighing. This method does not require a separate counterpoise and also doubles the balance sensitivity.

3.4.1. Single Transposition

The following steps are required:

1. Place M on one of the pans, e.g., the left pan and S on the right pan. S should be adjusted so that the indicator reads on scale. Record the indication θ_1.
2. Remove M and S.
3. Replace M and S on the balance but in transposed position (that is, M on the right pan and S on the left pan). Record the reading θ_2.
4. Add Δ to that pan which will cause the indicator to deflect towards the center of the reading scale. Record the indication θ_3.

The three indications give the relations:

$$M^T - S^T \doteq \rho_A(V_M - V_S) + k(\theta_1 - \theta_0) \tag{33}$$

$$S^T - M^T \doteq \rho_A(V_S - V_M) + k(\theta_2 - \theta_0) \tag{34}$$

and

$$S^T - M^T \doteq \rho_A(V_S - V_M) \pm (\Delta^T + \rho_A V_\Delta) + k(\theta_3 - \theta_0) \quad . \tag{35}$$

Here again, θ_0 is the balance indication when both pans are empty. The ± sign is used to indicate that, in the last equation, the sign of $(\Delta^T - \rho_A V_\Delta)$ depends on which pan held the sensitivity weight.

Subtracting (34) from (35) yields

$$k = \frac{\Delta^T - \rho_A V_\Delta}{|\theta_3 - \theta_2|} \quad . \tag{36}$$

where, once again, the absolute value was used to take account of whichever pan the sensitivity weight is placed on.

By subtracting (34) from (33), we have

$$M^T - S^T = \rho_A(V_M - V_S) + \frac{k}{2}(\theta_1 - \theta_2) \quad . \tag{37}$$

Finally, we can substitute from (36) into (37) to obtain

$$M^T - S^T = \rho_A(V_M - V_S) + \frac{(\Delta^T - \rho_A V_\Delta)}{|\theta_3 - \theta_2|} \frac{(\theta_1 - \theta_2)}{2} \quad . \tag{38}$$

CAUTION: In transposition weighing, attention has to be paid to the sense of the balance scale. We have assumed above that the balance indication becomes more positive as the mass on the left pan is increased. For scales of opposite sense, the plus sign of eq.(38) should be changed to minus. Also, if the arms of the balance are insufficiently "equal," transposition weighing will result in error.

3.4.2. Double Transposition

The procedure for this method follows that outlined above but with the following additions:

5) Remove M and S, leaving Δ where it is.
6) Place M on the left pan and S on the right pan. Record indication θ_4.

These steps provide us with the additional equation

$$M^T - S^T \doteq \rho_A(V_M - V_S) \pm (\Delta^T - \rho_A V_\Delta) + k(\theta_4 - \theta_0) \tag{39}$$

where the ± sign is needed once again since the location of the Δ weight is ambiguous.

By subtracting (35) from (39) we have

$$M^T - S^T = \rho_A(V_M - V_S) + \frac{k}{2}(\theta_4 - \theta_3)$$

and substituting for k from (36) yields

$$M^T - S^T = \rho_A(V_M - V_S) + \frac{1}{2} \frac{(\Delta^T - \rho_A V_\Delta)}{|\theta_3 - \theta_2|} (\theta_4 - \theta_3) \quad . \tag{40}$$

Adding (40) to (38) finally results in

$$M^T - S^T = \rho_A(V_M - V_S) + \frac{\left(\Delta^T - \rho_A V_\Delta\right)}{4\left|\theta_3 - \theta_2\right|}(\theta_1 - \theta_2 - \theta_3 + \theta_4) \quad . \tag{41}$$

Again the reader is cautioned that: the plus sign in front of the second term applies only to balances whose indications become more positive as the mass in the left pan is increased; in the presence of linear drift, the term $\theta_3 - \theta_2$ can be replaced by $(3\theta_3 - 3\theta_2 + \theta_1 \theta_4)/2$.

3.5. Weighing on Electronic Balances

The most recent development in mass technology is the appearance of the completely electronic balance. This can be thought of as a single-pan balance whose screen span has become so large that dial weights are no longer necessary. These balances are rugged, easy to use, and most are directly interfaceable to computers or data loggers. Unfortunately at present they rarely attain the precision achieved by the best mechanical balances. For metrology purposes, these balances are used just as the single-pan balances discussed above and the same equations apply. If and only if experience verifies that the constant k is indeed constant, can weighing procedures be simplified so that the sensitivity weight need not be measured for every substitution or double substitution.

The results we have just derived can be generalized. First, although only true mass has been dealt with, the reader should by now be able to generate the corresponding equations for apparent mass. Next, since S is assumed to be an assembly of calibrated weights, all the measurements described in this section provide an estimate for M^T, although they do not constitute an actual calibration.

Recall that we had (eq. (10))

$$M^T = S^T + \rho_A\,(V_M - V_S) + k\Delta\theta \tag{10'}$$

or equivalently

$$M^T = \frac{S^T - \rho_A V_S + k\Delta\theta}{1 - \rho_A/\rho} \quad . \tag{10"}$$

Usually, the manufacturer of a weight supplies the buyer with the weight's density so that formulation (10") is more appropriate. Equation (10') is often used in computer analysis of data, however. In this case, one estimates V_M as $V_M \quad N/\rho$, where N is the nominal mass value of M. Equation (10') then gives a first order approximation for M^T. This first estimate can be used to find a better value for V_M and the process repeated until convergence is reached.

If S and M are both unknown and not sufficiently close in mass, known weights can be added to either S or M. These added weights are then included in the weighing equations.

Example

Sometimes M itself will be a summation of weights. Let us say, to take a specific example, that M is the sum of three weights whose nominal values are 50 mg, 30 mg, and 20 mg. Furthermore, the weight manufacturer specifies that the 50-mg and 30-mg weights are made of tantalum (ρ = 16.6 g/cm^3; α = 20 x 10^{-6}/°C) but the 20-mg weight is made of aluminum (ρ = 2.7 g/cm^3; α = 69 x 10^{-6}/°C). What is the effective density of the summation and what is its effective coefficient of expansion?

Answer:

As a first approximation, assume the mass of each weight is equal to its nominal value. Therefore, the total volume is approximated by

$$\frac{.050}{16.6} + \frac{0.030}{16.6} + \frac{0.020}{2.7} = 0.012227 \text{ cm}^3$$

The effective density is then approximated by

$$\rho \equiv M/V \approx \frac{.050 + .030 + .020}{0.012227} = 8.18 \text{ g/cm}^3 \quad .$$

The effective coefficient of expansion is a weighted average--the coefficients of each weight are added in proportion to the volume of the weight to the total volume:

$$\alpha = \frac{20 \times 10^{-6} \times \frac{0.050}{16.6} + 20 \times 10^{-6} \times \frac{0.030}{16.6} + 69 \times 10^{-6} \times \frac{0.020}{2.7}}{\frac{.050}{16.6} + \frac{0.030}{16.6} + \frac{0.020}{2.7}}$$

$$= 50 \times 10^{-6}/°C$$

After mass values for the weights comprising M have been found, the metrologist should verify that the approximations used to estimate ρ and α were adequate. Only very rarely will this not be the case, requiring an iteration: the mass values obtained for the weights comprising M are used to find new effective values of ρ and α; these latter two values are then used to calculate improved mass values.

4. PROGRAM TYPES

We will discuss two distinct programs in mass metrology: surveillance and calibration. As the name implies, surveillance attempts to monitor a calibrated assembly or set of weights. Surveillance looks for signs that one or more members of the set may have changed since the last calibration. Calibration, of course, attempts to assign the best possible value of mass to a weight by relying on a chain of traceability to universally recognized standards. Assignment of a realistic uncertainty to a calibration value is equal in importance to the calibration value itself. Several additional definitions will be useful.

A standard mass S^T was already defined in Section 3. We also have:

SC: Check standard weight with true mass $(SC)^T$.

This mass is basically required to monitor in-house procedures and accuracies. Usually (SC) is a calibrated standard S that has been added to a set during calibration. By measuring (SC) (treated as an unknown M) versus S and comparing the results with the known value one can then determine if the measuring process is either in or out of control. This type of self check should be done periodically and the values for (SC) should be monitored continuously via trend charts (see below).

Set: A group of weights in a specific order. To be of maximum utility, sets usually cover several decades of nominal mass; e.g., 1 mg to 100 g. Further smaller groups (called subsets) of the complete set can be selected to produce any nominal value within the span of the set in increments of the smallest set members. What this means is that, taking the example of a 100 g to 1 mg set, any of the nominal values 0.001 g, 0.002 g, 0.003 g, ..., 99.998 g, 99.999 g, 100 g can be obtained by selecting appropriate weights from the set. Usually a set will employ the smallest number of weights possible to cover each decade. A set with a sequence 5, 3, 2, 1 could for instance range from 100 g to 1 mg with the sequence repeated for each decade (e.g. 500 mg, 300 mg, 200 mg, 100 mg).

Design: This specifies explicitly how and in what order a set of weights is to be compared (measured). It states in detail what weights of the set have to be used at a particular step.

ST: Transfer standard weight with true mass $(ST)^T$.

As mentioned under "Set" above, in many cases a number of weights, e.g. from 1 kg to 10 g must be calibrated. These are usually divided into decade groupings; that is, weights from 1 kg to 100 g are calibrated at the same time using one design (A standard kilogram is also included in the design). Next the weights from 100 g to 10 g are calibrated--possibly using a different balance. In this procedure, it is common for the 100 g weight which was calibrated as a result of the first series of measurements to become the standard for the second series. The 100 g weight thus serves as a "transfer standard" because it is used to transfer our knowledge of a standard kilogram to weights of smaller denomination.

4.1. Surveillance

Two distinct types are outlined below as Type I and Type II. (For more complete detailed discussions of many different designs, the reader is referred to ref. [4]).

The basic idea of surveillance testing is to ensure the self-consistency of the weight set. For example, the 20 g, 30 g, and 50 g weights can be checked against the 100 g weight to see if the difference is within expected limits. It is also advantageous (though not essential) to compare one weight of the set (usually the largest) against an independent standard, S. The latter operation establishes whether the entire weight set has undergone a change--even though there may still be self-consistency within the set. The basic motivations for surveillance testing are:

- to verify the values of newly calibrated weights
- to establish the stability of a new weight set
- to determine if an accident (such as being dropped on the floor) has damaged the weights involved.

Buoyancy corrections may not be needed in surveillance testing. One should check the magnitude of such corrections compared to the surveillance limits (see below) to see whether it is worthwhile to make the buoyancy corrections.

4.1.1. Type I

The object of this surveillance method is to perform an intercomparison of all weights in a set using a minimum number of steps. It is preferable to have one standard weight as a member of the set. We will denote it as S for this discussion. The standard S should always be as large as the largest member of the set; or as large as is convenient.

For the comparison measurement one always starts out with the largest weight and compares it with a summation of weights next in magnitude such that the sum is equivalent to the largest. Next, a weight from the first summation is compared with a lower summation, and the process is continued until all the weights have been used. If a standard is included in the set, then the S has to be compared first with the largest mass. An example will help to visualize the procedure.

Example:

The set contains a standard weight S whose nominal value is 100 g and weights of 100, 50, 30, 20, 10, 5, 3, 2, 1, 0.5, 0.3, 0.2, 0.1, 0.05, 0.03, 0.02, 0.01, 0.005, 0.003, 0.002 and 0.001 g; called M_1 to M_{21}, respectively. Such a set is known as a set with mass ratios of 5, 3, 2, 1 from 100 g to 1 mg. For the measurement sequence, we start out with the standard S versus M_1 = 100 g and work down to 1 mg such that all masses are included via a minimum number of steps. The designation M_i refers to the nominal value of the ith weight whereas M_i^T refers to the true mass value of the ith weight.

1st Meas: $M_1^T - S^T = \delta_1$, where M_1 = 100 g, S = 100 g

2nd Meas: $M_1^T - M_2^{T'} = \delta_2$, where $M_2' = (M_2 + M_3 + M_4)$

$= (50 + 30 + 20)$ g

$= 100$ g

3rd Meas: $M_4^T - M_3^{T'} = \delta_3$, where $M_3' = (M_5 + M_6 + M_7 + M_8)$

$= (10 + 5 + 3 + 2)$ g

$= 20$ g

4th Meas: $M_8^T - M_4^{T'} = \delta_4$, where $M_4' = (M_9 + M_{10} + M_{11} + M_{12})$

$= (1 + 0.5 + 0.3 + 0.2)$ g

$= 2$ g

5th Meas: $M_{12}^T - M_5^{T'} = \delta_5$, where $M_5' = (M_{13} + M_{14} + M_{15} + M_{16})$

$= (0.1 + 0.05 + 0.03 + 0.02)$ g

$= 0.2$ g

6th Meas: $M_{16}^T - M_6^{T'} = \delta_6$, where $M_6' = (M_{17} + M_{18} + M_{19} + M_{20})$

$= (0.01 + 0.005 + 0.003 + 0.002)$ g

$= 0.02$ g

7th Meas: $M_{19}^T - M_7^{T'} = \delta_7$, where $M_7' = (M_{20} + M_{21})$

$= (0.002 + 0.001)$ g

$= 0.003$ g

Note that the last measurement is amended because only M_{21} is left to measure.

With a very simple software routine, the differences (δ's) can then be compared against the known (accepted) values. These new differences should then be plotted and compared chronologically with previous tests. Together with predetermined uncertainty limits one can then monitor the particular weight set. Usually one devotes one chart for each δ. For example, fig. 5 shows a surveillance chart for δ_4. The horizontal line is the value of δ_4 inferred from the most recent calibration report of the set. The points represent values of δ_4 which were derived from surveillance testing. As long as the points remain within the upper and lower horizontal lines, (known as the surveillance limits), we have no evidence that any of the weights involved in the measurement of δ_4 has changed from its reported calibration. If an obvious trend is apparent, however, a recalibration can be done before the surveillance limit is exceeded. Two questions remain to be answered: (1) How are the surveillance limits determined?; and (2) If a point lies outside the surveillance limits, how does one determine which of the weights (that is 2 g, 1 g, 500 mg, 300 mg, or 200 mg in the case of fig. 4) has changed? The answers to these questions will be deferred until after the description of Type II Surveillance.

4.1.2. Type II

For Type II surveillance, a more sophisticated approach to the mass comparisons is taken. Again, the procedure is best described by an example.

Suppose we consider the following three weight groupings from the set used in the example for Type I.

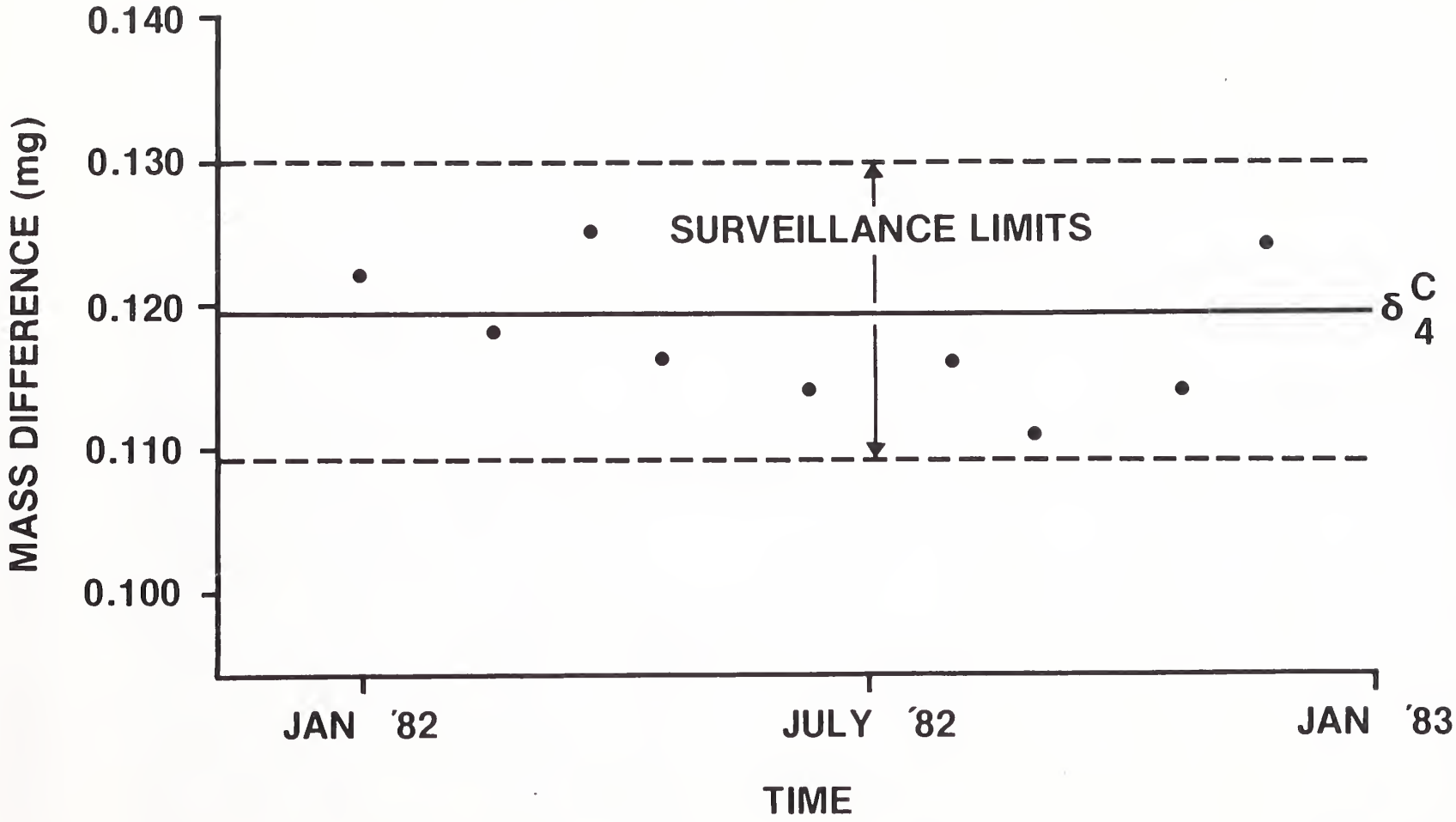


Figure 5. Results of nine type I surveillance checks of δ_4 over a period of one year. None of the checks falls outside the surveillance limits. δ_4^C is the value of δ_4 calculated from the most recent calibration report.

Weight Grouping	Designation
3 g	M_7
2 g + 1 g	$M_3'' = M_8 + M_9$
2 g + 500 mg + 300 mg + 200 mg	$M_3''' = M_8 + M_{10} + M_{11} + M_{12}$

For Type II surveillance, the following three mass differences are measured.

1. $M_7^T - M_3^{T''} = \delta_1$
2. $M_7^T - M_3^{T'''} = \delta_2$
3. $M_3^{T''} - M_3^{T'''} = \delta_3$

(Note that δ_3 cannot be measured by transposition weighing because M_8 is common to M_3'' and M_3''' .)

The three weighing operations shown above form a simple "design". The concept of weighing designs is crucial to calibration operations and will be discussed at length in Section 4.2.2. Here we may simply note that we are working with a "3-1's design" (i.e., 3 different weight groupings with 1 nominal mass).

If the weighing process had no scatter (standard deviation equal to zero), then $\delta_3 = \delta_2 - \delta_1$. In that case we would derive no additional information from the δ_3 measurement and its inclusion would, in fact, be a waste of time. Of course, the process does have some scatter and the δ_3 which we measure has only a slight probability of being equal to the measured value $\delta_2 - \delta_1$. Therefore, far from being useless, the measurement of the trio δ_1, δ_2, δ_3 can be used to good advantage in two ways:

First, a statistical technique known as "least squares" fitting [3,8] of the data provides better estimates of the three mass differences than can be gotten with a single measurement.

1. $M_7^T - M_3^{T''} = 1/3(2\delta_1 + \delta_2 - \delta_3) = \delta_1'$
2. $M_7^T - M_3^{T'''} = 1/3(\delta_1 + 2\delta_2 + \delta_3) = \delta_2'$
3. $M_3^{T''} - M_3^{T'''} = 1/3(-\delta_1 + \delta_2 + 2\delta_3) = \delta_3'$.

The least-squares estimates δ_1', δ_2', and δ_3' can be checked against the calibration report as in Type I surveillance (see refs. [3] and [8]).

Second, the differences between the least-squares estimates and the observations, $|\delta_1'-\delta_1|$, $|\delta_2'-\delta_2|$ and $|\delta_3'-\delta_3|$, are related to scatter in the measurement process.[6] If one of these differences is markedly higher (say three times higher)

---

[6] It is characteristic of the least squares solution to a 3-1's design that $|\delta_1'-\delta_1| = |\delta_2'-\delta_2| = |\delta_3'-\delta_3|$. This feature cannot be generalized to other designs.

than the measured standard deviation of the balance, one should redo the measurements of δ_1, δ_2, and δ_3. If trouble persists, it is a good indication that the precision of the balance (eq. (21)) has deteriorated.

As another example we will use a set similar to the one discussed under Type I, i.e. mass ratios 5, 3, 2, 1 from 100 g to 1 mg with S = 100 g.

The first 3 - 1's weighing series consists of

1a. $S^T - M_1^T = \delta_1$, where $M_1 = 100$ g

1b. $S^T - M_2^{T'} = \delta_2$, where $M_2' = (M_2 + M_3 + M_4)$

$= (50 + 30 + 20)$g

$= 100$ g

1c. $M_1^T - M_2^{T'} = \delta_3$

The next 3 - 1's weighing series is

2a. $M_3^T - M_3^{T'} = \delta_4$, where $M_3 = 30$ g

$M_3' = (M_4 + M_5)$

$= (20 + 10)$g

$= 30$ g

2b. $M_3^T - M_4^{T'} = \delta_5$, where $M_4' = (M_4 + M_6 + M_7 + M_8)$

$= (20 + 5 + 3 + 2)$g

$= 30$ g

2c. $M_3^{T'} - M_4^{T'} = \delta_6$

This is followed by the third 3 - 1's series as

3a. $M_7^T - M_5^{T'} = \delta_7$, where $M_7 = 3$ g

$M_5' = (M_8 + M_9)$

$= (2 + 1)$ g

$= 3$ g

3b. $M_7^T - M_6^{T'} = \delta_8$, where $M_6' = (M_8 + M_{10} + M_{11} + M_{12})$

$= (2 + 0.5 + 0.3 + 0.2)$ g

$= 3$ g

3c. $M_5^{T'} - M_6^{T'} = \delta_9$

For the fourth sequence we have:

4a. $M_{11}^{T} - M_{7}^{T'} = \delta_{10}$, where $M_{11} = 0.3$ g

$$M_{7}' = (M_{12} + M_{13})$$
$$= (0.2 + 0.1) \text{ g}$$
$$= 0.3 \text{ g}$$

4b. $M_{11}^{T} - M_{8}^{T'} = \delta_{11}$, where $M_{8}' = (M_{12} + M_{14} + M_{15} + M_{16})$

$$= (0.2 + 0.05 + 0.03 + 0.02) \text{ g}$$
$$= 0.3 \text{ g}$$

4c. $M_{7}^{T'} - M_{8}^{T'} = \delta_{12}$.

For the fifth series we measure:

5a. $M_{15}^{T} - M_{9}^{T'} = \delta_{13}$, where $M_{15} = 0.03$ g

$$M_{9}' = (M_{16} + M_{17})$$
$$= (0.02 + 0.01) \text{ g}$$
$$= 0.03 \text{ g}$$

5b. $M_{15}^{T} - M_{10}^{T'} = \delta_{14}$, where $M_{10}' = (M_{16} + M_{18} + M_{19} + M_{20})$

$$= (0.02 + 0.005 + 0.003 + 0.002) \text{ g}$$
$$= 0.03 \text{ g}$$

5c. $M_{9}^{T'} - M_{10}^{T'} = \delta_{15}$.

Finally, we arrive at the last series with

6a. $M_{19}^{T} - M_{11}^{T'} = \delta_{16}$ where $M_{19} = 0.003$ g

$$M_{11}' = (M_{20} + M_{21})$$
$$= (0.002 + 0.001) \text{ g} = 0.003 \text{ g}$$

6b. $M_{19}^{T} - M_{12}^{T'} = \delta_{17}$ where $M_{12}' = (M_{20} + M_{22}^{*})$

$$= (0.002 + 0.001^{*}) \text{ g} = 0.003 \text{ g}$$

6c. $M_{11}^{T'} - M_{12}^{T'} = \delta_{18}$.

---

*Note that in the last series a known (standard) 0.001-g weight was added in order to complete the design.

Once again the least squares fitting technique provides better estimates such that eqs. (1a), (1b), and (1c) yield

$$1a') \quad 1/3(2\delta_1 + \delta_2 - \delta_3) = \delta_1'$$

$$2b') \quad 1/3(\delta_1 + 2\delta_2 + \delta_3) = \delta_2'$$

$$3c') \quad 1/3(-\delta_1 + \delta_2 + 2\delta_3) = \delta_3'.$$

If for any reason a suitable standard, S, is unavailable, then any uncalibrated weight or summation of weights of the correct nominal value can be used in eqs. (1a) and (1b). However, in this case only δ_3' can be checked against the reported (known) calibration and only internal consistency of the entire set can be determined.

We also have

$$2a') \quad 1/3(2\delta_4 + \delta_5 - \delta_6) = \delta_4'$$

$$2b') \quad 1/3(\delta_4 + 2\delta_5 + \delta_6) = \delta_5'$$

$$2c') \quad 1/3(-\delta_4 + \delta_5 + 2\delta_6) = \delta_6'$$

and

$$3a') \quad 1/3(2\delta_7 + \delta_8 - \delta_9) = \delta_7'$$

$$3b') \quad 1/3(\delta_7 + 2\delta_8 + \delta_9) = \delta_8'$$

$$3c') \quad 1/3(-\delta_7 + \delta_8 + 2\delta_9) = \delta_9'$$

as well as

$$4a') \quad 1/3(2\delta_{10} + \delta_{11} - \delta_{12}) = \delta_{10}'$$

$$4b') \quad 1/3(\delta_{10} + 2\delta_{11} + \delta_{12}) = \delta_{11}'$$

$$4c') \quad 1/3(-\delta_{10} + \delta_{11} + 2\delta_{12}) = \delta_{12}'.$$

Finally

$$5a') \quad 1/3(2\delta_{13} + \delta_{14} - \delta_{15}) = \delta_{13}'$$

$$5b') \quad 1/3(\delta_{13} + 2\delta_{14} + \delta_{15}) = \delta_{14}'$$

$$5c') \quad 1/3(-\delta_{13} + \delta_{14} + 2\delta_{15}) = \delta_{15}'$$

and

$$6a') \quad 1/3(2\delta_{16} + \delta_{17} - \delta_{18}) = \delta_{16}'$$

$$6b') \quad 1/3(\delta_{16} + 2\delta_{17} + \delta_{18}) = \delta_{17}'$$

$$6c') \quad 1/3(-\delta_{16} + \delta_{17} + 2\delta_{18}) = \delta_{18}'.$$

The differences obtained for the first two results of each 3-1's series (i.e. a and b) should be compared to accepted values derived from the most recent calibration report.

The above examples of surveillance measurements for Type I and Type II present a simplified picture of the equations. All δ's must contain any necessary buoyancy corrections for true mass differences. Different δ's are obtained if apparent masses are used. From the simplified eq. (19a), one finds

$$M^A - S^A = k\Delta\theta + (\rho_A - \rho_0)\,(V_{OM} - V_{OS}) = \delta$$

where we assumed that the mass of the standard is given in apparent mass units, S^A.

Deviations of the measurements from values inferred from the most recent calibration report are evident from the previously mentioned surveillance charts.

4.1.3. Surveillance Limits

Upon carrying out a surveillance test of either Type I or II it will become evident that, after any necessary buoyancy corrections have been made, the measured values δ_i (Type I) or the predicted values δ_i' (Type II) do not exactly agree with those values, δ_i^c, calculated from the calibration certificate accompanying the set. To judge how serious the disagreement is, one must also calculate surveillance limits. The surveillance limits associated with each δ_i or δ_i' define the limits of credibility that the difference $\delta_i^c - \delta_i$ or $\delta_i^c - \delta_i'$ could be due to a combination of calibration uncertainties of the weight set and the random error of the measurement which is associated with the balance used in the surveillance measurements.

Let us designate to random and systematic error limits by SL;

$$(SL) = U + E \tag{42}$$

where

U = systematic error as determined from the calibration report
E = limit to random error = 3σ

and

σ = standard deviation of the measurement (refer to Appendix D for information on estimating this number).

Then any value δ_i or δ_i' should fall between $\delta_i^c \pm SL$, as shown in fig. 5. The value of 3σ gives a 99.7 percent confidence level for the random error (i.e. if the measurements were repeated a great many times there is a 99.7 percent chance the average would be within ± E of the result of a single surveillance test).

For surveillance, U is the root-sum-square of the individually reported calibration uncertainties such that

$$U = (\Sigma U_i^2)^{1/2} \tag{43}$$

Combining the uncertainties in this way is strictly valid only if they are uncorrelated. It is the nature of calibration designs, however, that the uncertainties of weights within the set generally are correlated. Thus one should view eq. (43) as an approximation which is adequate for surveillance limits. The reader is referred to the discussion of calibration uncertainties below for a fuller explanation.

The standard deviation used is known from many previous measurements (eq.(21)). It is a measure of the random errors in the balance being used.

At this stage it is worthwhile to quote a couple of examples and step through the process. From Type I surveillance, we have

$S = 10$ g

$M_1 = 100$ g

$M_2' = 100$ g $= 50$ g $+ 30$ g $+ 20$ g $= M_2 + M_3 + M_4$.

Assume the following previously reported calibration:

	Mass Value	Uncertainty
S^T	$= 100.0010196$ g	$U_S = 0.015$ mg
M_1^T	$= 100.0009407$ g	$U_{M_1} = 0.020$ mg
M_2^T	$= 50.0004628$ g	$U_{M_2} = 0.011$ mg
M_3^T	$= 30.0002926$ g	$U_{M_3} = 0.012$ mg
M_4^T	$= 20.0001578$ g	$U_{M_4} = 0.010$ mg .

Also assume that $\sigma = 0.026$ mg.

Since the first measurement is

$$M_1^T - S^T = \delta_1;\ \delta_1^c = -0.0000789 \text{ g} = -0.079 \text{ mg}$$

and we have

$$U = \sqrt{U_S^2 + U_{M_1}^2}$$

$$= \sqrt{(0.015)^2 + (0.020)^2} = 0.025 \text{ mg}$$

and

$$(SL) = U + 3\sigma$$

$$= 0.025 + 3 \times 0.026$$

$$= 0.103 \text{ mg.}$$

Hence $(M_1^T - S^T)$ values should fall in the range of (-0.079 ± 0.103) mg.

For the second measurement we have

$$M_1^T - M_2^{T'} = \delta_2;\ \delta_2^c = 0.028 \text{ mg}$$

so that

$$U_{M_2'} = \sqrt{U_{M_2}^2 + U_{M_3}^2 + U_{M_4}^2}$$

$$= 0.019 \text{ mg}$$

$$U = \sqrt{U_{M_2'}^2 + U_{M_1}^2}$$

$$= \sqrt{(0.019)^2 + (0.020)^2}$$

$$= 0.028 \text{ mg}$$

and $(SL) = 0.028 + 3 \times 0.026$

$= 0.106$ mg

and $M_1^T - M_2^{T'}$ should fall in the range

0.028 ± 0.106 mg.

Next, we consider the example from Type II. The first series of the design consists of S, M_1, and M_2' with $M_2' = M_2 + M_3 + M_4$. Let us assume the same calibration report as above, with the scale standard deviation at $\sigma = 0.026$ mg. In this case we have three measurements

$$S^T - M_1^T = \delta_1$$
$$S^T - M_2^{T'} = \delta_2$$
$$M_1^T - M_2^{T'} = \delta_3$$

Because we have used a 3 - 1's design, the standard deviation of our result is not simply the standard deviation of a single measurement, σ, but is instead $\sqrt{2/3}\,\sigma$. A proof of this conclusion is beyond the scope of this text but is explained fully in ref. [3].

There are several surveillance limits that can be calculated. For the three differences, we follow the calculations in Type I (but we have to remember that we are in Type II so that the standard deviation is $\sqrt{2/3}\,\sigma$).

Limits for $S^T - M_1^T$

$$U = \sqrt{U_S^2 + U_{M_1}^2} = \sqrt{(0.015)^2 + (0.020)^2} = 0.025 \text{ mg}$$

We then have

$$(SL) = U + 3 \times (\sqrt{2/3}\,\sigma) = 0.025 + 3\sqrt{2/3} \times 0.026$$
$$= 0.089 \text{ mg}$$

so that

$$(S^T - M_1^T) = (0.079 \pm 0.089) \text{ mg}.$$

Limits for $S^T - M_2^{T'}$

$$U = \sqrt{U_S^2 + U_{M_2}'^2}$$

where $U_{M_2}' = \sqrt{U_{M_2}^2 + U_{M_3}^2 + U_{M_4}^2} = \sqrt{(0.011)^2 + (0.012)^2 + (0.010)^2}$

$= 0.019$ mg

so that

$$U = \sqrt{(0.015)^2 + (0.019)^2} = 0.024 \text{ mg} .$$

Hence

$$(SL) = U + 3\sqrt{2/3}\,\sigma = 0.024 + 3\sqrt{2/3}\ 0.026 = 0.088 \text{ mg}$$

and $(S^T - M_2^{T'}) = (0.106 \pm 0.088)$ mg.

4.1.4. Identifying Weights Which Have Changed

If a measurement falls outside the surveillance limits, it becomes necessary to determine which individual weights are responsible. Simple deductive reasoning is all that is required, although a few extra weighing combinations may also be needed if Type I surveillance was used.

As an example, suppose that in the Type I measurements shown earlier (4.1.1) the value for δ_4 was outside but the value for δ_3 was inside the surveillance limits. All we know at this point is that there is some change in the subset $M_4' = 1\text{ g} + 0.5\text{ g} + 0.3\text{ g} + 0.2\text{ g}$. Three additional weighings are now required to pinpoint the cause of the discrepancy. Recall that $M_9 = 1$ g, $M_{10} = 0.5$ g, $M_{11} = 0.3$ g, $M_{12} = 0.2$ g, $M_{13} = 0.1$ g.

We then measure

4') $M_9^T - M_4^{T}{}'' = \delta_4''$, where $M_4'' = M_{10} + M_{11} + M_{12}$

4") $M_{10}^T - M_5^{T}{}'' = \delta_5''$, where $M_5'' = M_{11} + M_{12}$

4"') $M_{11}^T - M_6^{T}{}'' = \delta_6''$, where $M_6'' = M_{12} + M_{13}$

After the surveillance limits for δ_4'', δ_5'', δ_6'' are calculated one investigates the results:

1.) If δ_4'' lies outside the surveillance limits but δ_5'' and δ_6'' do not, then it is probable that M_9 has changed since its last calibration.

2.) If δ_4'' and δ_5'' lie outside the surveillance limits by opposite amounts and δ_6'' is inside, then it is probable that the M_{10} (0.5 g) weight has changed.

3.) If δ_4'' and δ_5'' lie outside the limits by about the same amount and δ_6'' is outside also by the same amount but in the opposite direction, then it is probable that the M_{11} (0.3 g) weight has changed.

4.) If δ_4'', δ_5'', and δ_6'' all lie outside the limits by about the same amount in the same direction, then it is probable the M_{12} (0.2 g) has changed.

If none of the above conditions are met, then it is probable that more than one weight has changed. The reader should then consult ref. [4] for a more thorough analysis of surveillance methods.

4.2. Calibration

The process of "calibration" assigns mass values to weights by comparing the unknown weights to recognized standards. An uncertainty limit--3 times the standard deviation of the measurement process, plus estimated uncertainties systematic to the measurement process--accompanies each calibrated value. To an even greater degree than in Type II surveillance, redundant information is gathered in order to determine whether the measurement scatter is acceptable. A powerful self-consistency check of the calibration process is also included.

Most of the following information can be found in much greater detail in refs. [1] and [3]. It is repeated here mainly for explaining the overall approach of the program.

Weight sets come in various denominations. The following groupings are frequently calibrated:

1. Nominally Equal Groups

These sets consist of weights, all of which have the same nominal value. The number of weights can go from 3 to as high as 13 or more. (Sets of up to 50 members have been utilized.) For large sets, subsets are usually analyzed. Otherwise the number of weighing operations would become impractically large.

2. Groups for 2, 2, . . . 1, 1 . . sets.

Several combinations are in use such as 2, 1, 1, 1, or 2, 2, 1, 1, 1, or 2, 1, 1, 1, or 2, 2, 1, 1, 1, etc. Here 2 and 1 imply the same decade of mass such as 2 kg and 1 kg.

3. Binary and Miscellaneous Groups

These sets are usually comprised of pound units. Many combinations are available, the use of which is slowly diminishing. A few examples provided in ref. [3] are 4, 3, 2, 1, 1, or 10, 5, 2, 2, 1, 1, or 6, 5, 4, 3, 2, 1, etc. Note that a combination of the lower masses always sums to the value of a higher mass unit.

4. The 5, 3, 2, 1 and 5, 2, 2, 1 Groups

These are the sets most widely used. In general a set of weights in this group can span many decades with each decade comprising, for example, the 5, 3, 2, 1 sequence. In many cases standards or "check standards" (see below) are added to the sets. Examples of such groups could be 5, 5, 3, 2, 1, 1, or 5, 3, 2, 1, 1, etc.

4.2.1. Trend Elimination in Direct Reading Balances

If a single-pan, direct reading balance has truly constant sensitivity and is subject to only a slight linear drift, the following simplification is possible:

Suppose there are four nominally equal weights, designated A, B, C, and D to be compared in a particular calibration scheme, the following eight "direct" weighings are done:

(1) Place A on balance and read θ_1
(2) Remove A
(3) Place B on balance and read θ_2
(4) Remove B
(5) Place C on balance and read θ_3
(6) Remove C
(7) Place D on balance and read θ_4
(8) Remove D
(9) Replace D on balance and read θ_5
(10) Remove D
(11) Place C on balance and read θ_6
(12) Remove C
(13) Place B on balance and read θ_7

(14) Remove B
(15) Place A on balance and read θ_8.

We have taken the weights in the following order: A,B,C,D,D,C,B,A. We then make the following calculations:

$$\theta_A = 1/2(\theta_1 + \theta_8)$$

$$\theta_B = 1/2(\theta_2 + \theta_7)$$

$$\theta_C = 1/2(\theta_3 + \theta_6)$$

$$\theta_D = 1/2(\theta_4 + \theta_5)$$

Then the estimated difference in mass between any two weights (say C and D) is

$$C^T - D^T = \rho_A(V_C - V_D) + k(\theta_C - \theta_D)$$

where we assume k is well-known and constant. The symmetry of the weighing sequence removes problems caused by any linear drifts in the balance or weighing conditions.

The above example is an instance of "trend elimination." (Double substitution is another instance.) When weighings are subject to large air buoyancy corrections, drifts in temperature or barometric pressure may lead to errors if those quantities are only read once during all the measurements required by the design. Some calibration schemes, however, have the property of "trend elimination" even for this problem [3].

4.2.2. Designs

Sets or subsets of weights are calibrated together by means of a weighing design. A design simply prescribes what weighings are to be made. Each weighing is used to estimate a mass difference between two nominally equal weights or groups of weights in the set.

A typical design for a group, say 5, 3, 2, 1, 1, 1, could be as follows:

	Mass					
Observation	M_5	M_3	M_2	M_1	$M_{\Sigma 1}$	SC
δ_1	+	-	-	+	-	
δ_2	+	-	-		+	-
δ_3	+	-	-	-		+
δ_4	+	-	-			
δ_5	+		-	-	-	-
δ_6		+	-	+	-	-
δ_7		+	-	-	+	-
δ_8		+	-	-	-	+
δ_9			+	-	-	
δ_{10}			+	-		-
δ_{11}			+		-	-

Here +/- signs indicate the weight starts out on the left/right pan, for transposition weighing, or that the weight is the first/second in substitution weighing. We notice that in this group there are six masses, which must be determined (K = 6), 11 measurements (N = 11), and suppose 1 restraint (e.g. the mass of the 5+3+2 summation is known from previous measurements, L = 1). We can then calculate the number of degrees of freedom (D.F.) for this design as

$$D.F. = N - K + L = 11 - 6 + 1 = \underline{6}.$$

From the 11 observations and the given value of the restraint one can then use the least squares method to solve all equations to obtain best values of all differences and therefore all masses, standard deviations, and variances. Such fitting techniques are well described in literature available from the NBS [3]. (See Appendix D.)

In general the more degrees of freedom provided by a particular design, the lower the estimates of standard deviation. Recall the previous section on surveillance testing. Each measurement in Type I surveillance had no degrees of freedom. That is, we needed to determine a quantity δ_i and we had only one measurement of that quantity (K = 1, N = 1, L = 0; D.F. = 0). In Type II surveillance, however, we used a simple 3-1 design: Three quantities were determined, for example δ_1', δ_2', δ_3' (N = 3), three measurements were made (δ_1, δ_2, δ_3, so that K = 3), and there was one restraint ($\delta_3' = \delta_2' - \delta_1'$). Thus D.F. = 3 - 3 + 1 = 1. The appearance of a degree of freedom allowed a crude statistical check (e.g., whether $\left|\delta_1 - \delta_1'\right| <$ balance standard deviation) and also led to a slightly reduced standard deviation ($\sqrt{2/3}\,\sigma$ instead of σ).

The design illustrated in this section is much more sophisticated than the 3-1 but is motivated by the same twin desires for good statistical analysis and small standard deviation of the calibrated masses. The design in the above example is shown in figure 6, where it is designated C.2. This figure is reproduced from ref. [3]. Note that two different restraints are considered: restraint A is that the mass of the summation of the 5, 3, and 2 weights is known;[7] restraint B is that the mass of a single "1" weight is the known standard.[8] We will discuss the results based on restraint A in detail. That is, we will assume that the true mass of $M_5+M_3+M_2$ is known. Let us designate this mass as R^T.

In the 532111 design shown above, we have labeled one of the weights as $\Sigma 1$. This indicates that we could calibrate the 5+3+2 summation of the next lower decade of the same weight set which could then serve as a transfer standard for a subsequent design. By "SC" we designate a weight which is external to the set being calibrated and which will serve as check standard.

---

[7]This would be useful, for instance, if the 5+3+2 summation had been calibrated in a previous design. In this case, the 5+3+2 summation would be a transfer standard. This restraint is used in working down in mass from 1 kg.

[8]This would be useful in working up in mass from 1 kg.

The least-squares solutions for the five unknown weight are [3]:

$$\hat{M}_5^T = (1/920)\ \{100(\delta_1+\delta_2+\delta_3+\delta_4)+60\delta_5 - 20(\delta_6+\delta_7+\delta_8+\delta_9+\delta_{10}+\delta_{11})+460R^T\}$$

$$\hat{M}_3^T = (1/920)\{-68(\delta_1+\delta_2+\delta_3+\delta_4)-4\delta_5+124(\delta_6+\delta_7+\delta_8)-60(\delta_9+\delta_{10}+\delta_{11})+276R^T\}$$

$$\hat{M}_2^T = (1/920)\{-32(\delta_1+\delta_2+\delta_3+\delta_4)-56\delta_5-104(\delta_6+\delta_7+\delta_8)+80(\delta_9+\delta_{10}+\delta_{11})+184R^T\}$$

$$\hat{M}_1^T = (1/920)\{119\delta_1+4\delta_2-111\delta_3+4\delta_4-108\delta_5+128\delta_6-102(\delta_7+\delta_8)-125(\delta_9+\delta_{10})-10\delta_{11}+92R^T\}$$

$$\hat{M}_{\Sigma 1}^T = (1/920)\{-111\delta_1+119\delta_2+4(\delta_3+\delta_4)-108\delta_5-102\delta_6+128\delta_7-102\delta_8-125\delta_9-10\delta_{10}-125\delta_{11}+92R^T\}$$

In addition, the least-squares solution provides a value for the check standard:

$$(\hat{SC})^T = (1/920)\{4\delta_1-111\delta_2+119\delta_3+4\delta_4-108\delta_5-102(\delta_6+\delta_7)+128\delta_8-10\delta_9-125(\delta_{10}+\delta_{11})+92R^T\}$$

The metrologist must determine the values δ_1, δ_2, ... δ_M and the value of R^T by experiment. The least-squares solution shows how to combine the observations with the value of the restraint in order to arrive at mass values for the weights used in the design.

In addition to mass values, the least-squares solution provides fitted values for the observations, which when compared with the measured values give:

$$\delta_1-\delta_1' = \delta_1-\hat{M}_5^T+\hat{M}_3^T+\hat{M}_2^T-\hat{M}_1^T+\hat{M}_{\Sigma 1}^T$$

$$\delta_2-\delta_2' = \delta_2-\hat{M}_5^T+\hat{M}_3^T+\hat{M}_2^T-\hat{M}_{\Sigma 1}^T+(\hat{SC})^T$$

$$\delta_3-\delta_3' = \delta_3-\hat{M}_5^T+\hat{M}_3^T+\hat{M}_2^T+\hat{M}_1^T-(\hat{SC})^T$$

$$\delta_4-\delta_4' = \delta_4-\hat{M}_5^T+\hat{M}_3^T+\hat{M}_2^T$$

$$\delta_5-\delta_5' = \delta_5-\hat{M}_5^T+\hat{M}_2^T+\hat{M}_1^T+\hat{M}_{\Sigma 1}^T+(\hat{SC})^T$$

$$\delta_6-\delta_6' = \delta_6-\hat{M}_3^T+\hat{M}_2^T-\hat{M}_1^T+\hat{M}_{\Sigma 1}^T+(\hat{SC})^T$$

$$\delta_7-\delta_7' = \delta_7-\hat{M}_3^T+\hat{M}_2^T+\hat{M}_1^T-\hat{M}_{\Sigma 1}^T+(\hat{SC})^T$$

$$\delta_8-\delta_8' = \delta_8-\hat{M}_3^T+\hat{M}_2^T+\hat{M}_1^T+\hat{M}_{\Sigma 1}^T-(\hat{SC})^T$$

$$\delta_9-\delta_9' = \delta_9-\hat{M}_2^T+\hat{M}_1^T+\hat{M}_{\Sigma 1}^T$$

$$\delta_{10}-\delta_{10}' = \delta_{10}-\hat{M}_2^T+\hat{M}_1^T+(\hat{SC})^T$$

$$\delta_{11}-\delta_{11}' = \delta_{11}-\hat{M}_2^T+\hat{M}_{\Sigma 1}^T+(\hat{SC})^T$$

These deviations are useful for two reasons. First, the estimated standard deviation of the least-squares fit is equal to

$$\frac{1}{N-K+1}\left(\sum_{i=1}^{N}(\delta_i-\delta_i')^2\right)^{1/2}.$$

Second, a glance at each of the eleven values of $\delta_i-\delta_i'$ can often pinpoint the source of a blunder in entering the raw data into a computer--that is, if a blunder has been made, the value of $\delta_i-\delta_i'$ affected by the mistake will often appear much larger than all the other values.

Finally, the least-squares solution also provides estimates of the standard deviation for each computed mass and each combination of mass. Below the design in fig. 6 is a table called "Factors for Computing Standard Deviations." Suppose the standard deviation of the measurement process is σ. Then if design C.2 is used to calibrate the mass of the "5" weight, for example, the table tells us that the standard deviation assigned to this mass, σ_5, will be:

$0.2331 \times \sigma$ if restraint A were used
$1.7846 \times \sigma$ if restraint B were used.

Also from the table we see that the standard deviation of the sum of the 5 and 3 weights, σ_8, is given by:

$0.2638 \times \sigma$ if restraint A were used
$2.8284 \times \sigma$ if restraint B were used.

It is typical of least squares results that, for example,

$$\sigma_{m+n} \neq (\sigma_m^2 + \sigma_n^2)^{1/2}.$$

In this case,

$$\sigma_8 < (\sigma_5^2 + \sigma_3^2)^{1/2} \text{ for restraint A}$$

$$\sigma_8 > (\sigma_5^2 + \sigma_3^2)^{1/2} \text{ for restraint B.}$$

Thus the metrologist must choose both design and restraint carefully to minimize the resulting standard deviations.

4.2.3. Statistical Checks

Computer programs, such as those developed by NBS, are routinely utilized in mass calibration laboratories. The user supplies all the measured data for the set, all environmental conditions, and other necessary data. The program then provides:

1. a detailed listing of data provided
2. the least squares fit, i.e. the desired mass values
3. the "F ratio" and the "t value".

Two crucial assumptions underlie the calibration of unknown weights by least-squares fitting of design data:

DESIGN C.1 5.3.2.1.1 K= 5
N= 8
(HAYFORD) D.F.= 4

OBSERVATIONS	5	3	2	1	1
Y(1)	+	-	-	+	-
Y(2)	+	-	-	-	+
Y(3)	+	-	-		
Y(4)	+	-		-	-
Y(5)		+	-	-	
Y(6)		+	-		-
Y(7)			+	-	-
Y(8)				+	-
RESTRAINT A	+	+	+		
RESTRAINT B					+

FACTORS FOR COMPUTING ST DEVS

WT	RESTRAINTS A	B	5	3	2	1	1
1	.4175	.0000					+
1	.4175	.5345				+	
2	.3546	.9258			+		
3	.3338	1.3093		+			
4	.5237	1.6903		+		+	
5	.2673	2.0702	+				
6	.5237	2.4495	+			+	
7	.3338	2.9277	+		+		
8	.3546	3.3381	+	+			
9	.5632	3.7033	+	+		+	
10	.0000	4.1748	+	+	+		

DESIGN C.3 5.3.2.2.1.1 K= 6
N=11
D.F.= 6

OBSERVATIONS	5	3	2	2	1	1
Y(1)	+	-	+	-	-	-
Y(2)	+	-	-	+	-	-
Y(3)	+	-	-	-	+	+
Y(4)	+	-	-		+	-
Y(5)	+	-		-	+	-
Y(6)	+	-	-		-	+
Y(7)	+	-		-	-	+
Y(8)	+	-			-	-
Y(9)	+		-	-	-	
Y(10)	+		-	-		-
Y(11)			+	-		
RESTRAINT A	+	+	+			
RESTRAINT B						+

FACTORS FOR COMPUTING ST DEVS

WT	RESTRAINTS A	B	5	3	2	2	1	1
1	.3592	.0000						+
1	.3592	.4714					+	
2	.4259	.7553				+		
2	.3625	.7553			+			
3	.3835	1.2536		+				
3	.6116	1.1123				+	+	
4	.6130	1.4254			+	+		
5	.2266	1.7412	+					
6	.4790	2.0893	+				+	
7	.3835	2.4199	+		+			
8	.3625	2.9493	+	+				
9	.4711	3.2708	+	+			+	
10	.0000	3.5922	+	+	+			

DESIGN C.2 5.3.2.1.1.1 K= 6
N=11
D.F.= 6

OBSERVATIONS	5	3	2	1	1	1
Y(1)	+	-	-	+	-	
Y(2)	+	-	-		+	-
Y(3)	+	-	-	-		+
Y(4)	+	-	-			
Y(5)	+		-	-	-	-
Y(6)		+	-	+	-	-
Y(7)		+	-	-	+	-
Y(8)		+	-	-	-	+
Y(9)			+	-	-	
Y(10)			+	-		-
Y(11)			+		-	-
RESTRAINT A	+	+	+			
RESTRAINT B						+

FACTORS FOR COMPUTING ST DEVS

WT	RESTRAINTS A	B	5	3	2	1	1	1
1	.3551	.0000						+
1	.3551	.5000					+	
1	.3551	.5000				+		
2	.2638	.7802			+			
3	.2985	1.0885		+				
4	.4778	1.4781		+		+		
5	.2331	1.7846	+					
6	.4299	2.1644	+			+		
7	.2985	2.5216	+		+			
8	.2638	2.8284	+	+				
9	.4616	3.2016	+	+		+		
10	.0000	3.5509	+	+	+			

DESIGN C.4 5.3.2.2.1.1 K= 6
N=12
D.F.= 7

OBSERVATIONS	5	3	2	2	1	1
Y(1)	+	-	-			
Y(2)	+	-		-		
Y(3)	+	-			-	-
Y(4)	+		-	-	-	
Y(5)	+		-	-		-
Y(6)		+	-		-	
Y(7)		+	-			-
Y(8)		+		-	-	
Y(9)		+		-		-
Y(10)			+	-		
Y(11)			+		-	-
Y(12)				+	-	-
RESTRAINT A	+	+	+			
RESTRAINT B						+

FACTORS FOR COMPUTING ST DEVS

WT	RESTRAINTS A	B	5	3	2	2	1	1
1	.4796	.0000						+
1	.4796	.8165					+	
2	.4055	1.0541				+		
2	.3127	1.0541			+			
3	.2848	1.4530		+				
3	.6172	1.7638				+	+	
4	.4372	2.0276			+	+		
5	.2887	2.3805	+					
6	.5888	3.1091	+				+	
7	.5608	3.3830	+			+		
8	.3127	3.8006	+	+				
9	.6119	4.5216	+	+			+	
10	.0000	4.7958	+	+	+			

Figure 6. Least squares information for the design described in the text. (Reproduced from ref. [3]).

1. The scatter in the data just taken is typical of the scatter found in previous measurements.

2. The mass of the standard weight used in the calibration design has not changed from its accepted value.

The F ratio and t value provide important tests of these two assumptions.

The F ratio essentially monitors the precision of the measuring process. A detailed discussion is presented in Appendix D. We note that

$$F = s^2/\sigma^2$$

where σ^2 is the variance of the particular balance being utilized (based on a large collection of previous measurements). Generally, we expect F to be close to 1, but we must not be surprised if a particular value of F is somewhat larger than 1.

In particular, a simple check for F can be (and has been) established for most mass metrology laboratories. By comparing F with a fixed ratio F_t, which could be defined at the 99 percent confidence level one can then easily monitor whether the measured ratio F is greater than F_t

i.e. $$F > F_t.$$

If this check holds true, then the measurement process is considered "out of control" and further studies have to be conducted. The quantity F_t depends only on the degrees of freedom (k_1) in s^2 and on the probability level at which we wish to conduct the test. The calculations are outlined in Appendix D.

The second test is called the "t test" which monitors the accuracy of the measurements. For this test, an extra mass, the check standard, is included in the particular design. A simple check is then performed to see whether the mass value assigned to the check standard by the calibration agrees with the accepted value. The t test thus monitors the systematic errors of the measuring process.

In particular one calculates

$$t_c = \left|(\hat{SC})^T - \text{accepted } (SC)^T\right|/\sigma_c$$

where

$(\hat{SC})^T$ = observed mass of the check standard as found by least-squares fitting;

accepted $(SC)^T$ = accepted mass of the check standard;

and

$$\sigma_c^2 = \ell\sigma^2 + \sigma_T^2 + (N_{SC}/N_R)^2\sigma_R^2 \quad .$$

Here

σ_T^2 = accepted variance of the measurement process between runs (see below)

ℓ = multiplier which is determined by the least squares process.

σ = "process" standard deviation of the balance

N_{sc} = the nominal value of the check standard

N_R = the nominal value of the standard or transfer standard used as the restraint in the least squares solution

σ_R = the standard deviation of the value of the restraint. σ_R is taken as zero unless the restraint is a transfer standard, whose value was determined as part of the calibration of the complete weight set [1].

For simplicity, we will discuss a case for which $\sigma_R = 0$. In general one checks to see that

$$t_c < 3.$$

The probability is less than one in 100 that this inequality will be violated by chance. Thus $t_c = 3$ is taken as the control limit. Any measurement for which $t_c > 3$ is considered "out of control" and is repeated.

CAUTION: The t test is one of the best statistical measures of systematic errors available. Nevertheless, by their very nature, systematic errors are difficult to detect. If, for instance, buoyancy corrections were not important in assignment of mass to the check standard but were important in assigning mass values to the weights being calibrated, the t test will not detect systematic errors in the buoyancy correction. Also, the t test cannot detect an identical change in the standard and check standard. We have taken a very simple example of a t test. The most general case is shown in ref. [1].

The variance of the measurement process between runs, σ_T^2, is an important concept [1,8]. Recall that the process standard deviation, σ, is a measure of scatter over the period of time it takes to make a mass measurement: i.e., a few minutes to a few hours. There may, however, be sources of random uncertainty which fluctuate more slowly (i.e., over days or months); but still rapidly compared with intervals between recalibration of a given weight. How can we estimate the process standard deviation for this longer period of time? The easiest way is to monitor the values of mass assigned to the check standard each time it is used in a run. After many runs, over a span of many months, we can estimate the variance, σ_c^2, of the observed values of the check standard about their mean value.

If the process standard deviation has no "between-run" component, then

$$\sigma_c^2 = \ell\sigma^2 \quad (\sigma_T^2 \approx 0)$$

where σ^2 is the process variance which was found by pooling variances estimated from the least squares fits to many individual runs. That is, σ is the within-run process variance based on combining calculations of s^2 from many runs.

It will generally be true, however, that

$$\sigma_c^2 > \ell\sigma^2$$

indicating that there is an additional source of scatter from run to run. We define this between-run variance for the measurement process as

$$\sigma_T^2 = \sigma_c^2 - \ell\sigma^2 \quad .$$

If σ_T^2 is not negligible, the source of between-run scatter should, of course, be sought and eliminated if possible.[9] Failing this, the process standard deviation must be expanded so that the over-all uncertainty assigned in the calibration report includes the between-run component.

The cautionary statement given above for the t test also applies to the use of the check standard in estimating between-run scatter. That is, if a weight being calibrated differs markedly in its construction from the weight(s) used in the check standard, a σ_T component may go undetected if one relies only on check-standard behavior.

Suppose a series of design data is "in control." True mass values have been assigned to the unknown weights and uncertainty limits have been given. What do these uncertainties mean? In a practical sense, the meaning is that the uncertainty bands assigned to each mass calibrated by the methods of ref. [1] should almost always overlap the uncertainty bands which would have been obtained had the masses been calibrated at NBS.

Note that this specification, while satisfying the expectation of the overwhelming majority of users, is somewhat less stringent than asserting that: the true mass as defined in eq. [1] and reported in S.I. units is almost surely within the uncertainty band provided by the calibration.

---

[9] σ_T has been found to be negligible for the NBS calibration process.

5. CONCLUSION

This document has been less of a complete treatise on mass metrology than a guide for the mass metrologist. We have indicated appropriate references where the various topics touched on are treated in greater detail.

What we have tried to offer is a coherent development of the fundamental concepts of mass metrology as a means of explaining the relevance to the metrologist of the references found at the end of this publication.

In brief, starting with the basic law of classical mechanics, we derived:

- Relations for the true mass differences between nominally equal weights (eqs. (10') and (19b))
- Relations for apparent mass differences between nominally equal weights (eqs (18) and (19a))

We stressed the importance of measuring or calculating air density in order to make proper buoyancy corrections.

We then provided:

- An explanation of how the basic mass relations apply to measurements on several types of commonly used balances.
- A demonstration of how these relations are used in two types of surveillance testing.
- A brief sketch of how these relations are used in mass calibrations.

We concluded with a discussion of the statistical checks crucial to a calibration.

In a document such as this, there is a danger of losing the basic outlines of a mass measurement program in the many details essential to carrying out a successful program. Every program must have a goal--for example, the calibration of weights in a certain range of masses to a desired uncertainty.

One must then acquire balances equal to the task and house them in a room which will not degrade their performance. Control charts should be used to establish the longterm reliability of the balances. Calibrated sets of weights to be used as working standards should be acquired. Surveillance testing can establish the stability of these weights to within the surveillance limits.

Anticipated levels of buoyancy correction should be estimated. Auxillary equipment such as barometers, hygrometers,and thermometers should be acquired, if necessary, to achieve adequate capability in determining the density of air. These instruments must be calibrated periodically.

When calibrating weight sets, weighing designs should be used. These are selected on the basis of providing the necessary accuracy with the least number of weighing operations. Great care must be exercised in assigning a total uncertainty to calibration results.

The techniques outlined in this document form the core of such a mass program.

We would like to express our appreciation to Dr. J.D. Simmons of NBS for his support. One of us (K.J.) wishes to thank R. Weber and Dr. D. Cauffman, both of Lockheed, for their encouragement, help, and overall support for this endeavor. Also, one of us (R.D.) extends thanks to his colleagues at NBS for their cooperation and technical guidance. We especially wish to thank Mrs. Betty King for typing the manuscript.

6. REFERENCES

[1] Varner, R.N.; Raybold, R.C. National Bureau of Standards mass calibration computer software. Nat. Bur. Stand. (U.S.) Tech. Note 1127; 1980 July. 164 p. NTIS PB80203441

[2] Jones, Frank E. The air density equation and the transfer of the mass unit. J. Res. Nat. Bur. Stand. (U.S.) 83(5): 419-428; 1978 September-October.

[3] Cameron, J. M.; Croarkin, M. C.; Raybold, R. C. Designs for the calibration of standards of mass. Nat. Bur. Stand. (U.S.) Tech. Note 952; 1977 June. 64 p. NTIS PB268499

[4] Almer, H. E., Keller, Jerry, ed. Surveillance test procedures. Nat. Bur. Stand. (U.S.) NBSIR 76-999; 1977 May. 77 p. NTIS PB268130

[5] Pontius, Paul E. Mass and mass values. Nat. Bur. Stand. (U.S.) Monogr. 133; 1974 January. 39 p. NTIS COM7450309

[6] Almer, H. E. Method of calibrating weights for piston gages. Nat. Bur. Stand. (U.S.) Tech. Note. 577; 1971 May. 54 p. NTIS COM7150264

[7] Precision measurement and calibration statistical concepts and procedures, Special Publication 300, Volume 1, Pontius, P.E., Cameron, J.M.; pp. 2-20. GPO Stock No. 003-003-00072-8

[8] Pontius, P.E. Measurement philosophy of the pilot program for mass calibration. Nat. Bur. Stand. (U.S.) Tech. Note 288; 1966 May. 39 p. (Reprinted January, 1968).

[9] Schoonover, Randall M. A simple gravimetric method to determine barometer corrections. J. Res. Nat. Bur. Stand. (U.S.)85 (5) 341-345, 1980-September-October.

[10] Schoonover, Randall M. A look at the electronic analytical balance. Anal. Chem. 52: 973A-980A; 1982 July.

[11] Schoonover, Randall M.; Jones, Frank E. Air buoyancy correction in high-accuracy weighing on analytical balances. Anal. Chem. 53: 900-902; 1981 May.

[12] Cameron, Joseph M.; Hailes, Geraldine E. Designs for the calibration of small groups of standards in the presence of drift. Nat. Bur. Stand. (U.S.) Tech. Note 844; 1974 August. 32 p. NTIS COM7450762

[13] Natrella, Mary Gibbons. Experimental statistics. Nat. Bur. Stand. (U.S.) Handb. 91; 1963 August (reprinted 1966 October with corrections). GPO Stock No. 0003-003-00135-0

[14] ASTM E617-81. Standard specifications for laboratory weights and precision mass standards. 1983 Annual Book of ASTM Standards Vol. 14.02: 455-474.

[15] Measurement Assurance Programs Part II: Development and Implementation, Nat. Bur. Stand. Special Publication 676II, Croarkin, C; 1984, 124 p.

References with NTIS numbers are available from:

The National Technical Information Service
Springfield, VA 22161

References with GPO stock numbers are available from:

The Superintendent of Documents
U.S. Government Printing Office
Washington, DC 20402

APPENDIX A

APPARENT MASS OF BUILT-IN BALANCE WEIGHTS

Balance manufacturers almost always adjust the built-in dial weights of single-pan balances so that their apparent masses equal the value of the dial position. No matter of what density material the manufacturer has actually made his weights, the user generally assumes that the true masses equal the dial values and the densities of the built-in weights equal ρ_R, the basis density.

Let us see why this scheme works and what its limitations are. These considerations can be demonstrated by referring to a single dial-weight (of density ρ_D) that exactly balances a weight X, which is on the pan.

We begin, since we are discussing a balance of the type shown in fig. 3 with eq. (22):

$$k\theta_1 \doteq (M_X^T - \rho_A V_X) - (\Sigma D_i^T - \rho_A \Sigma V_i) + k\theta_0,$$

where we have renamed M by M_X and V_M by V_X. We will assume that θ_0 was adjusted to zero before X was placed on the balance. We have chosen a special case: after X was placed on the balance, the screen reading θ_1 returned to zero with the removal of a single dial weight. Let us refer to this dial weight as "D" and say that it has a true mass D^T and a volume V at the balance temperature. In this special case, with $\theta_0 = \theta_1 = 0$, eq. (22) becomes

$$0 = (M_X^T - \rho_A V_x) - (D^T - \rho_A V) \; . \tag{A1}$$

Since $\rho_D = D^T/V$, we can rewrite (A1) as

$$M_X^T - \rho_A V_X = D^T(1 - \rho_A/\rho_D)$$

or,

$$M_X^T = D^T(1 - \rho_A/\rho_D) + \rho_A V_X \equiv M_1 \; . \tag{A2}$$

Now we ask the question, "What if one assumes the mass of D is equal to N_D, its nominal value, and the density of D is equal to ρ_R, the basis density of the apparent mass scale to which the balance weights have been adjusted?" This assumption is not correct because

$$\rho_R \neq \rho_D$$

and

$$D^T \neq N_D.$$

Nevertheless, the incorrect assumption would tell us that

$$M_X^T = N_D(1-\rho_A/\rho_R) + \rho_A V_X \equiv M_2 \quad . \tag{A3}$$

We can next ask ourselves "Under what conditions does $M_1 = M_2$?"

Recalling that M_D^A is made equal to N_D, the nominal dial value, we subtract eq. (A3) from (A2).

$$M_1-M_2 = D^T\ (1-\rho_A/\rho_D) - M_D^A(1-\rho_A/\rho_R) \quad . \tag{A4}$$

Also, by definition, (see eq.(11) in the main text),

$$M_D^A = D^T\ (1-\rho_0/\rho_{D0})/(1-\rho_0/\rho_{R0}) \quad . \tag{A5}$$

where the subscript zeros refer to the standard conditions specified in the definition of apparent mass. Substituting (A5) into (A4),

$$M_2-M_1 = D^T[(1-\rho_0/\rho_{D0})(1-\rho_A/\rho_R)/(1-\rho_0/\rho_{R0}) - (1-\rho_A/\rho_0)] \quad .$$

The first approximation we will make is that $\rho_{R0} = \rho_R$ and $\rho_{D0} = \rho_D$. That is, the temperature is sufficiently close to 20 °C that the expansion in volume of the balance weights is negligible. Thus

$$M_2-M_1 = D^T[(1-\rho_0/\rho_D)\ (1-\rho_A/\rho_R)/(1-\rho_0/\rho_R) - (1-\rho_A/\rho_D)] \quad .$$

Next, we use the relation developed in the main text:

$$(1-\rho_0/\rho_R)^{-1} = 1 + \rho_0/\rho_R + (\rho_0/\rho_R)^2 + (\rho_0/\rho_R)^3 + \ldots$$

so that

$$M_2-M_1 = D^T[(1-\rho_0/\rho_D)(1-\rho_A/\rho_R)\ (1+\rho_0/\rho_R+(\rho_0/\rho_R)^2 + (\rho_0/\rho_R)^3 + \ldots) - (1-\rho_A/\rho_D)] \quad .$$

Finally, we multiply out but throw away all terms of order $(\rho_0/\rho_R)^2$, $(\rho_A\rho_0/\rho_R^2)$, $\rho_0^2/\rho_D\rho_R$, and smaller. Since $\rho_0 \sim \rho_A \leq 1.2$ mg/cm^3 and $\rho_D \sim \rho_R > 7.5$ g/cm^3 (the balance manufacturer has seen to this), then neglecting these terms leads to errors of 3×10^{-6} percent or less.

Thus, we now have

$$M_2-M_1 = D^T\ (\rho_0/\rho_R - \rho_0/\rho_D - \rho_A/\rho_R + \rho_A/\rho_D)$$

or, more simply

$$M_2-M_1 = D^T\ (\rho_0-\rho_A)\ (1/\rho_R-1/\rho_D) \quad . \tag{A6}$$

The balance manufacturer has selected the metal for his weights from stock having a density ρ_D sufficiently close to ρ_R so that, near sea level, M_1-M_2 will not exceed the balance tolerance.

In recent years, balance manufacturers have been making weights of stainless steel. Equation (A6) shows why it is, therefore, desirable for these weights to be adjusted on the 8.0 basis.

The chief virtue of the apparent mass scheme is that, even though different balance manufacturers may make their weights of different alloys, a user can be completely unaware of these subtleties and still derive reasonably accurate true mass results on any balance so long as the correct basis density is used.

Example:

A balance has 20 g of built-in dial weights. The density of these weights is actually 7.8 g/cm^3. The weights have been adjusted so that the dial readings correspond to apparent mass on the brass basis. What is the value of M_2-M_1 at maximum load if ρ_A = 1.16 mg/cm^3? What if the laboratory is at high elevation so that ρ_A = 1.00 g/cm^3?

Answer:

a) $M_2-M_1 = 20\ g\ (0.00120 - 0.00116)\ [\frac{1}{8.4} - \frac{1}{7.8}] = -7\ \mu g.$

b) $M_2-M_1 = 20\ g\ (0.00120 - 0.00100)\ [\frac{1}{8.4} - \frac{1}{7.8}] = -37\ \mu g.$

By comparison, the calibration uncertainty for high-quality 20 g-weights as measured by NBS is about 10 µg; the tolerance for 20-g, Class 1, metric weights is 74 µg [14].

APPENDIX B

OUTLINE OF THE DERIVATION OF ρ_A, THE DENSITY OF AIR

In order to find ρ_A, an equation of state involving temperature, humidity, and barometric pressure has been developed. We will follow the definitive derivation of F. E. Jones [2] and briefly describe his arguments to arrive at a simplified relation sufficiently accurate for our applications. The reader is urged to consult [2] for full details.

In order to derive the density for a mixture of two gasses, let us start with the ideal gas law.

$$pV = nRT$$

and

$$n = \frac{m}{M}$$

where

n = number of moles
m = mass of gas
M = molecular weight of gas
R = universal gas constant
T = temperature in kelvins = (273.15 + t) for t in degrees celsius.

From the above, we can rewrite

$$pV = \frac{m}{M} RT \text{ so that } m = \frac{pVM}{RT}$$

and

$$\frac{M}{RT} = \frac{m}{pV} = \frac{\rho_A}{p} \text{ since } \rho_A = \frac{m}{V} .$$

Specifically, consider air consisting of dry air and water vapor. For the dry air, we have

$$m_D = \frac{p_D V M_D}{RT}$$

and for the water vapor, we have

$$m_w = \frac{p_w V M_w}{RT} .$$

Since the total density is given by

$$\rho_A = \frac{m_D + m_w}{V}$$

we obtain

$$\rho_A = \left(\frac{p_D VM_D}{RT} + \frac{p_W VM_W}{RT} \right) / V$$

$$= \frac{1}{RT} (p_D M_D + p_W M_W) \quad .$$

Using Dalton's Law for partial pressures

$$P = p_1 + p_2$$

we can substitute for

$$p_D = P - p_W$$

so that

$$\rho_A = \frac{1}{RT} (PM_D - p_W M_D + p_W M_W)$$

$$= \frac{M_D}{RT} \left(P + p_W \left(\frac{M_W}{M_D} - 1 \right) \right)$$

and using

$$\varepsilon = \frac{M_W}{M_D}$$

$$\rho_A = \frac{M_D}{RT} (P + p_W(\varepsilon - 1)) \quad .$$

At this point it is important to consider for a moment the correction necessary for a gas which is not ideal. To do so we rewrite the last relation as

$$P = \frac{\rho_A RT}{M_D} \frac{1}{\left[1 + \frac{p_W}{P} (\varepsilon - 1) \right]} \quad .$$

For an ideal gas the ratio

$$\frac{P}{\rho_A \frac{RT}{M_D} \left[\frac{1}{1 + \frac{p_W}{P} (\varepsilon - 1)} \right]} = Z$$

has to be equal to 1. If we want to consider non-ideal gas corrections, then we have to incorporate Z into our formalism. Hence,

$$P = \frac{\rho_A RTZ}{M_D} \frac{1}{\left[1 + \frac{p_W}{p}(\varepsilon - 1)\right]}$$

where Z is called the "compressibility factor."

Rewriting this relation for the mixture density, we now have

$$\rho_A = \frac{M_D}{RTZ}\left[(P + (\varepsilon - 1)p_W\right] .$$

Let us now turn to the dependence of this relation on the relative humidity. We define

$$U = \frac{p_W}{e'_S} \times 100$$

where

U = relative humidity in percent

p_W = effective vapor pressure of water in moist air

e'_S = effective saturation vapor pressure of water in moist air.

Furthermore, we know that

$$e'_s > e_s$$

where

e_s = the saturation vapor pressure of pure-phase water.

The ratio of the two pressures is called the "enhancement factor" for saturated water vapor and is given by

$$f = \frac{e'_s}{e_s}$$

so that

$$e'_s = e_s \cdot f$$

and since

$$p_W = \frac{U \cdot e'_s}{100}$$

we have

$$p_W = \frac{U \cdot e_s \cdot f}{100} \quad .$$

Substitution into the mixture density relation yields

$$\rho_A = \frac{M_D}{RTZ} \left[P + (\varepsilon - 1) \frac{U \cdot e_s \cdot f}{100} \right]$$

where U must be given in %.

This relation constitutes the final formula. The parameters P, T, and U must be measured by the user.

Substituting the best available values for R anf M_W; and choosing reasonable values for M_D, Z, and f, which are approximated as constant parameters, we have

$$R = 8.31441 \frac{\text{joules}}{\text{mole-K}} = 8,314.4 \frac{\text{joules}}{\text{kmole-K}}$$

$$M_D = 28.964 \text{ g/mole}$$

$$Z = 0.9996$$
$$f = 1.0042$$
$$M_W = 18.0152 \text{ g/mole.}$$

Since

$$\varepsilon = \frac{M_W}{M_D} = \frac{18.0152}{28.964} \quad ,$$

we obtain

$$\rho_A = \frac{3.4848}{T} \left(P - 0.0037960 \, Ue_s \right) \times 10^{-3}$$

$$\rho_A \left(\frac{mg}{cm^3} \right) = \frac{0.0034848}{(t+273.15)} \left(P - 0.0037960 \, Ue_s \right) \quad .$$

where

t is in °C
P is in pascals (133.3224 Pa = 1 mm Hg) .

Converting to mm Hg pressure, we have

$$\rho_A \left(\frac{mg}{cm^3} \right) = \frac{0.46460}{(t+273.15)} \left(P - 0.0037960 \, Ue_s \right) \quad .$$

The parameter e_s has been determined by fitting measured data between 288.15 K and 301.15 K. The relationship developed is

$$e_s = (1.3146 \times 10^9)\, e^{\frac{-5315.56}{(t+273.15)}} \text{ mm Hg.}$$

It should also be noted that the value for the enhancement factor used above, i.e. f = 1.0042 can be approximated more accurately by

$$f = 1.00070 + 4.150 \times 10^{-6} \times P + 5.4 \times 10^{-7} t^2 \quad .$$

In the above final equation, the relative humidity is given in percent

$$U = \% \text{ (e.g. 51.2, 43.7 etc.)}$$

and the pressure P in mm Hg.

M_D (and, therefore, ε) depends on the mixture of gases, other than water vapor, which makes up ambient air. The chief variability in this mixture comes from CO_2 and O_2 levels. These are assumed perfectly correlated--that is, CO_2 levels can only increase locally at the expense of O_2 and vice versa--as in processes of combustion, respiration and photosynthesis.

Fortunately, the variation has little effect on M_D. In eqs. (20a) and (20b) (main text) we have assumed the ambient level of CO_2 typically found in the NBS mass laboratories (0.00042 mol of CO_2/mol of air). A 100 percent increase in this level would raise ρ_A by less than 0.02 percent.

Z is slightly dependent on barometric pressure, temperature and relative humidity [2]. Between 19 °C and 26 °C; 525 mm Hg and 825 mm Hg; and 0 and 100% R.H., Z varies by less than 0.03 percent.

APPENDIX C

PROPAGATION OF ERROR THROUGH THE AIR DENSITY EQUATION

Using eq. (20b) (developed in Appendix B), we have

$$\rho_A = \frac{0.46460}{(t+273.15)} (P-0.0037960 \cdot U \cdot e_s) \quad .$$

For ease of writing, we will set

$$\rho = \rho_A$$
$$a = 0.46460$$
$$b = 273.15$$
$$c = 0.0037960$$
$$e = e_s$$

such that

$$\rho = \frac{a}{(t+b)}(P - c \cdot U \cdot e) \quad . \tag{C1}$$

We want to determine the uncertainty for ρ using

$$d\rho = \frac{\partial \rho}{\partial t} dt + \frac{\partial \rho}{\partial P} dP + \frac{\partial \rho}{\partial U} dU \quad . \tag{C2}$$

$\frac{\partial \rho}{\partial t}$:

Using (C1), we find

$$\frac{\partial \rho}{\partial t} = a(P - c \cdot U \cdot e) \frac{\partial}{\partial t} \left(\frac{1}{t+b}\right)$$

$$= \left(\frac{-1}{t+b}\right) \rho \tag{C3}$$

$\frac{\partial \rho}{\partial P}$:

Using (C1), we find

$$\frac{\partial \rho}{\partial P} = \frac{a}{t+b}$$

and since

$$\rho \approx \frac{aP}{(t+b)}$$

where we ignored the second term in (C1) since it is only a secondary source of error, we have

$$\frac{\partial \rho}{\partial P} \approx \frac{\rho}{P} \quad . \tag{C4}$$

$\frac{\partial \rho}{\partial U}$:

From (C1), we obtain

$$\rho = \frac{aP}{t+b} - \frac{aceU}{t+b}$$

so that

$$\frac{\partial \rho}{\partial U} = - \frac{ace}{(t+b)}$$

and again using $\rho \quad \frac{aP}{t+b}$

we have

$$\frac{\partial \rho}{\partial U} \approx \frac{-ce\rho}{P} \quad . \tag{C5}$$

Next, we have to estimate the combined uncertainties due to t, P, and U. In Section 2.3.2 (main text) we asserted that the temperature should be known to ±0.4 °C so that

$$dt = \pm 0.4 \text{ °C} \quad .$$

Similarly,

$$dP = \pm 1.1 \text{ mm Hg}$$

and

$$dU = \pm 16\% \quad .$$

All these values assume that we want to know the air density to 0.0017 mg/cm^3.

Substitution of all terms into (C2) yields

$$d\rho = \frac{-\rho}{t+b} (\pm 0.4) + \frac{\rho}{P} (\pm 1.1) - \frac{ce}{P} \rho \, (\pm 16) \quad .$$

In the worst case, these uncertainties would add linearly. However, we assume the errors are uncorrelated so that a better estimate of the total uncertainty is given by

$$\delta\rho = \sqrt{\left(\frac{\partial \rho}{\partial t} dt\right)^2 + \left(\frac{\partial \rho}{\partial P} dP\right)^2 + \left(\frac{\partial \rho}{\partial U} dU\right)^2}$$

$$= \sqrt{\frac{\rho^2 \cdot (0.4)^2}{(t+b)^2} + \frac{\rho^2 (1.1)^2}{p^2} + \frac{c^2 e^2 \rho^2 (16)^2}{p^2}}$$

$$= \rho \sqrt{\frac{0.16}{(t+b)^2} + \frac{1.21}{p^2} + \frac{c^2 e^2 256}{p^2}} \quad .$$

At standard conditions we have

$$t = 20\ ^\circ C$$
$$P = 760\ mm\ Hg \quad .$$

We now have

$$\delta\rho = \rho \sqrt{\frac{0.16}{(20+b)^2} + \frac{1.21}{(760)^2} + \frac{c^2 e^2 256}{(760)^2}} \quad .$$

Substituting $\rho = \rho_A$ and replacing b and c with their numerical values, we have

$$\frac{\delta\rho_A}{\rho_A} = \sqrt{\frac{0.16}{(20+273.15)^2} + \frac{1.21}{(760)^2} + \frac{(0.0037960)^2 \cdot 256 \cdot e_s^2}{(760)^2}} \quad .$$

Also e_s at 20 °C is 17.54 mm Hg, so that finally

$$\frac{\delta\rho_A}{\rho_A} \simeq 0.0024$$

or

$$\frac{\delta\rho_A}{\rho_A} \simeq 0.24\% \quad .$$

APPENDIX D

SOME STATISTICAL CONCEPTS

We start out with an infinitely large population of a random variable; to be practical, let us say a very large number of similar balance observations. We assume that this population is normally distributed with a mean of μ and a variance of σ^2.

Let us take a random sample of size n from the population. Then the mean of the sample is defined as

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \qquad (D1)$$

with an estimated variance for a single observation of

$$s^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1} \qquad (D2)$$

As n grows infinitely large $\bar{X} \to \mu$ and $s^2 \to \sigma^2$. (This, in fact, defines what we mean by μ and σ).

The treatment of least squares data is more complicated than analysis of repeated measurements of the same quantity. In this case, average values of several different masses are calculated at the same time from data obtained during one weighing design. If N measurements were necessary to complete the design, the metrologist has in hand the results of these N mass comparisons: δ_1, δ_2, ... δ_N. The least squares analysis provides a set of "best" estimates: δ_1', δ_2',...δ_N', as well as a set of "best" estimates: M_1^T, M_2^T,...M_K^T, of the K unknown masses calibrated by means of the design. These "best" estimates are linear combinations of the measured values δ_1, δ_2,...δ_N with the mass of the known standard used as the restraint. The linear combinations are uniquely determined by the choice of design and restraint. Least squares solutions to the most useful weighing designs are tabulated in ref. [3].

The standard deviation of the least squares fit to the design data is estimated from the formula[1]

$$s^2 = \frac{\sum_{i=1}^{N} (\delta_i - \delta_i')^2}{N-K+1} \qquad (D3)$$

where the denominator is given by the degrees of freedom (= number of observations - number of unknowns + number of restraints). If the same measurement design were repeated m times, a better estimate for the standard deviation of the process could be obtained:

---

[1] Least squares is so named because the least squares solution: δ_1', δ_2',...δ_N', minimizes the value of s^2 in (D3).

$$s^2 = \frac{s_1^2 + s_2^2 + \ldots + s_m^2}{m} \qquad \text{(D4)}$$

where s_i^2 is the estimated variance of the i^{th} run. As $m \to \infty$, $s^2 \to \sigma^2$. This defines what we mean by σ^2 in the case of least squares. Note: the number of degrees of freedom (D.F.) in s^2 as defined by (D4) is $m(N-K+1)$. Also, this σ is the same σ discussed in the text, the "process standard deviation."

Besides σ, we would also like to know the standard deviation of each individual mass, M_j, computed from the least squares analysis. The answer involves matrix algebra and depends both on the design and restraint which were used, as well as on σ. All of the commonly used cases are tabulated in ref. [3]. In general, the standard deviation of a measured mass $\hat{M}_j$ is some value $\ell_j\sigma$, where ℓ_j is a number that depends only on the particular design and restraint. This number can be found in ref. [3]; σ is defined above.

When two or more weights--j and p for example--are used in combination, their combined standard deviation is not simply $(\ell_j^2 + \ell_p^2)^{1/2}\,\sigma$ but is often somewhat greater. This is because data taken from a design are usually correlated. Both the recipe for handling weight summations and tabulated values for many special cases can be found in ref. [3].

From the above discussion, it should be evident that care must be taken in choosing the design and restraints for any set of weights to be calibrated. Some choices will minimize individual ℓ_i values but weight summations may have large uncertainties.

Assume that an estimate of σ^2 has been computed from m designs by application of eq. (D4). Future designs will produce individual values s_j^2 which will also be estimates of σ^2 as long as the measurement process remains stable. An F statistic

$$F = \frac{s_j^2}{s^2} \qquad \text{(D5)}$$

can be computed for each new s_j^2. Its purpose is to test the agreement between s_j^2 and s^2 --or more precisely to test whether or not s_j^2 comes from the same distribution of measurements that produced s^2.

Given a large number of F statistics, each of which is based on the same design, a histogram of these F statistics will fall very nearly on the universal curve known as the F distribution. The theoretical calculation of this curve depends only on the degrees of freedom in each s_j^2, i.e. N-K-1, and on the degrees of freedom in s^2, i.e. $m(N-K-1)$. The distribution is scaled so that the area under the curve is equal to 1, making it possible, for example, to find the point F_t such that 99 percent of all F values will be less than F_t. As a consequence, the next time that we carry out the weighing design, we expect that there is a 99 percent chance that the F statistic computed from that particular design will be <u>less than</u> the percent point F_t. If, in fact, it turns out that the F statistic is greater than the percent point F_t, the

precision for that design is poorer than is expected, and the design should be repeated.

The percent point F_t is often represented in tables as $F(\nu_1, \nu_2, 1-\alpha)$ because it depends on:

ν_1 = the degrees of freedom in s_j^2 which is N-K+1

ν_2 = the degrees of freedom in s^2 which is m(N-K+1)

α = the significance level such as $\alpha = 0.01$ (i.e. 99 percent)

For mass calibrations at NBS, the significance level is chosen to be $\alpha = 0.01$, and because the measurement process has been tracked for a long time resulting in an estimate of σ^2 that has a very large number of degrees of freedom, the value that is used for the test is

$$F_t = F(\nu_1, \infty, 0.99)$$

In this special case, F_t can be well-approximated by*:

$$F_t = \left(1 - \frac{2}{9(N-K+1)} + 2.32635 \sqrt{\frac{2}{9(N-K+1)}}\right)^3$$

for

$$\nu_1 \geq 2$$

and

$$F_t = 6.64$$

for

$$\nu_1 = 1 \quad .$$

Tables and detailed discussion on the F value and t ratio can be found in ref. [13].

*A derivation of this result is well beyond the scope of this work. The interested reader may wish to consult: Paulson, Edward. An approximate normalization of the analysis of variance distribution. Annals of Mathematical Statistics. 13: 223-235; 1942.

NBS-114A (REV. 2-80)

U.S. DEPT. OF COMM. **BIBLIOGRAPHIC DATA SHEET** *(See instructions)*	1. PUBLICATION OR REPORT NO. NBS SP 700-1	2. Performing Organ. Report No.	3. Publication Date November 1984

4. TITLE AND SUBTITLE

Industrial Measurement Series

A Primer for Mass Metrology

5. AUTHOR(S)

K. B. Jaeger and R. S. Davis

6. PERFORMING ORGANIZATION *(If joint or other than NBS, see instructions)*

NATIONAL BUREAU OF STANDARDS
DEPARTMENT OF COMMERCE
GAITHERSBURG, MD 20899

7. Contract/Grant No.

8. Type of Report & Period Covered

Final

9. SPONSORING ORGANIZATION NAME AND COMPLETE ADDRESS *(Street, City, State, ZIP)*

Same as item 6.

10. SUPPLEMENTARY NOTES

Library of Congress Catalog Card Number: 84-601090.

☐ Document describes a computer program; SF-185, FIPS Software Summary, is attached.

11. ABSTRACT *(A 200-word or less factual summary of most significant information. If document includes a significant bibliography or literature survey, mention it here)*

This paper attempts to fill the need for a coherent guide to the many publications which document the NBS program in mass metrology. The topics we emphasize are generally those which experience has shown to present the greatest difficulties for metrologists new to the field of mass measurements. Thus we have included many worked examples and have retained steps often omitted in more scholarly treatments of the same subjects. A full bibliography is included so that the reader may also consult the primary sources of this work.

12. KEY WORDS *(Six to twelve entries; alphabetical order; capitalize only proper names; and separate key words by semicolons)*

Air buoyancy; apparent mass; calibration; mass; metrology; surveillance testing.

13. AVAILABILITY

☒ Unlimited

☐ For Official Distribution. Do Not Release to NTIS

☒ Order From Superintendent of Documents, U.S. Government Printing Office, Washington, D.C. 20402.

☐ Order From National Technical Information Service (NTIS), Springfield, VA. 22161

14. NO. OF PRINTED PAGES

85

15. Price

NBS Technical Publications

Periodicals

Journal of Research—The Journal of Research of the National Bureau of Standards reports NBS research and development in those disciplines of the physical and engineering sciences in which the Bureau is active. These include physics, chemistry, engineering, mathematics, and computer sciences. Papers cover a broad range of subjects, with major emphasis on measurement methodology and the basic technology underlying standardization. Also included from time to time are survey articles on topics closely related to the Bureau's technical and scientific programs. As a special service to subscribers each issue contains complete citations to all recent Bureau publications in both NBS and non-NBS media. Issued six times a year.

Nonperiodicals

Monographs—Major contributions to the technical literature on various subjects related to the Bureau's scientific and technical activities.

Handbooks—Recommended codes of engineering and industrial practice (including safety codes) developed in cooperation with interested industries, professional organizations, and regulatory bodies.

Special Publications—Include proceedings of conferences sponsored by NBS, NBS annual reports, and other special publications appropriate to this grouping such as wall charts, pocket cards, and bibliographies.

Applied Mathematics Series—Mathematical tables, manuals, and studies of special interest to physicists, engineers, chemists, biologists, mathematicians, computer programmers, and others engaged in scientific and technical work.

National Standard Reference Data Series—Provides quantitative data on the physical and chemical properties of materials, compiled from the world's literature and critically evaluated. Developed under a worldwide program coordinated by NBS under the authority of the National Standard Data Act (Public Law 90-396). NOTE: The Journal of Physical and Chemical Reference Data (JPCRD) is published quarterly for NBS by the American Chemical Society (ACS) and the American Institute of Physics (AIP). Subscriptions, reprints, and supplements are available from ACS, 1155 Sixteenth St., NW, Washington, DC 20056.

Building Science Series—Disseminates technical information developed at the Bureau on building materials, components, systems, and whole structures. The series presents research results, test methods, and performance criteria related to the structural and environmental functions and the durability and safety characteristics of building elements and systems.

Technical Notes—Studies or reports which are complete in themselves but restrictive in their treatment of a subject. Analogous to monographs but not so comprehensive in scope or definitive in treatment of the subject area. Often serve as a vehicle for final reports of work performed at NBS under the sponsorship of other government agencies.

Voluntary Product Standards—Developed under procedures published by the Department of Commerce in Part 10, Title 15, of the Code of Federal Regulations. The standards establish nationally recognized requirements for products, and provide all concerned interests with a basis for common understanding of the characteristics of the products. NBS administers this program as a supplement to the activities of the private sector standardizing organizations.

Consumer Information Series—Practical information, based on NBS research and experience, covering areas of interest to the consumer. Easily understandable language and illustrations provide useful background knowledge for shopping in today's technological marketplace.

Order the **above** *NBS publications from: Superintendent of Documents, Government Printing Office, Washington, DC 20402.*

Order the **following** *NBS publications—FIPS and NBSIR's—from the National Technical Information Service, Springfield, VA 22161.*

Federal Information Processing Standards Publications (FIPS PUB)—Publications in this series collectively constitute the Federal Information Processing Standards Register. The Register serves as the official source of information in the Federal Government regarding standards issued by NBS pursuant to the Federal Property and Administrative Services Act of 1949 as amended, Public Law 89-306 (79 Stat. 1127), and as implemented by Executive Order 11717 (38 FR 12315, dated May 11, 1973) and Part 6 of Title 15 CFR (Code of Federal Regulations).

NBS Interagency Reports (NBSIR)—A special series of interim or final reports on work performed by NBS for outside sponsors (both government and non-government). In general, initial distribution is handled by the sponsor; public distribution is by the National Technical Information Service, Springfield, VA 22161, in paper copy or microfiche form.

U.S. Department of Commerce
National Bureau of Standards
Gaithersburg, MD 20899

Journal of Research of the National Bureau of Standards

Note on the Choice of a Sensitivity Weight In Precision Weighing

Volume 92 Number 3 May–June 1987

R. S. Davis

National Bureau of Standards
Gaithersburg, MD 20899

Good weighing practice usually dictates that, when using double-substitution weighing to determine the mass difference between two weights, the nominal value of the sensitivity weight used to calibrate the optical scale of the mass comparator be at least four times greater than the difference of the two weights being compared. However, there are times when other considerations must override this rule. We examine the theoretical basis for the rule and the penalty for violating it. Finally, we propose a modi-fied weighing scheme which imposes a much less stringent rule for the size of the sensitivity weight. The new scheme requires an additional balance reading, but does not increase the overall measurement time significantly.

Key words: mass metrology; precision weighing; sensitivity weight; substitution weighing; transposition weighing; weighing.

Accepted: November 28, 1986

1. Introduction

Many precision mass comparisons, especially in the realm of metrology, still rely on mechanical balances. These balances may be either one-pan or two-pan. In both cases, however, weighing is done by double substitution between the unknown and an external standard. The procedure in use in most metrology laboratories is shown in table 1.

Table 1. Four observation scheme.

Operation	Load on Balance	Balance Indication
1	Y	I_1
2	X	I_2
3	$X+d$	I_3
4	$Y+d$	I_4

where Y represents the standard, X the unknown, and d the sensitivity weight. We are assuming that for two-pan balances double substitution has been used rather than double transposition. The arguments that follow apply with modification to the latter technique.

The difference in mass between Y and X, ΔM, (ignoring buoyancy corrections) is sometimes computed as [1][1]:

$$\Delta M = \frac{I_1-I_2-I_3+I_4}{2(I_3-I_2)}\, m_d = \frac{\Delta I}{\Delta I_d}\, m_d\,. \qquad (1)$$

We may think of eq (1) as the product of the difference between Y and X in scale units, $(I_1-I_2-I_3+I_4)/2$, multiplied by the balance sensitivity, $m_d/(I_3-I_2)$. The sensitivity is the proportionality factor which converts differences in scale indication to units of mass. Here m_d is the known mass of d.

About the Author: R. S. Davis is a physicist in the Length and Mass Division of NBS' Center for Basic Standards.

[1]Figures in brackets indicate literature references.

Most balance indications drift with time. Often the time dependence of the drift can be assumed to be linear. Based on this assumption, one usually tries to make the time intervals between the four weighing operations equal. If this is done, the estimate of ΔI, the difference between Y and X in scale units, will be unbiased by the drift. This will not be true of I_3-I_2 however. The latter quantity estimates ΔI_d, the value of the sensitivity weight in scale units.

In order to remove the bias, a modified equation is used:

$$\Delta M = \frac{I_1-I_2-I_3+I_4}{I_1-3I_2+3I_3-I_4} m_d \,. \tag{2}$$

This is the equation found in the NBS MASSCODE [2] and has been advocated for general use if the added computational complexity can be handled by computer [3].

2. Variance of ΔM

There is a general rule [1,3] which states that the metrologist should take care that

$$\frac{\Delta I}{\Delta I_d} \leqslant 0.25 \,. \tag{3}$$

If the rule is violated, the NBS MASSCODE prints a warning message along with the final calculation [2]. Since the author has not found a rigorous theoretical basis for the rule in the literature, one will now be given.

Each reading of scale indication is subject to random error. Let us assume this error can be characterized by a variance σ_I^2 which is the same for all measurements. Then the variance of ΔM as computed by eq (2) using first order propagation of error techniques is

$$\mathrm{var}(\Delta M) = S^2\sigma_I^2 \left[1+5\left(\frac{\Delta I}{\Delta I_d}\right)^2\right] \tag{4}$$

where $S=2m_d/(I_1-3I_2+3I_3-I_4)$ is the nominal value of the balance sensitivity (the quantity m_d is treated as a constant in this computation since its variance is usually much smaller than σ_I^2). Therefore, the rule represented by eq (3) implies that the variance in a single measurement of ΔM should not be allowed to increase by more than a factor of 1.31 above its minimum value. The choice of 1.31 is, of course, somewhat arbitrary. Reasonable people might all agree that a factor of 2, for instance, would be intolerably large, while a factor of 1.1 would be impractically small. We choose 1.31 because it is the *de facto* choice of the NBS MASSCODE. The important point is that we now have a rational criterion by which to compare various weighing procedures with respect to their demands on the value of the sensitivity weight.

The absence of a term linear in $\Delta I/\Delta I_d$ in eq (4) shows that the estimate of ΔI is uncorrelated with the estimate of ΔI_d. It is also evident that the variance of ΔM increases monotonically as the ratio $\Delta I/\Delta I_d$ becomes larger. In particular, if $\Delta I/\Delta I_d$ is of the order of 0.5 then the variance of ΔM increases to 2.25 times its minimum value. This is unacceptably large in many cases. A value of 0.5 for $\Delta I/\Delta I_d$ was the unavoidable case, however, for a series of important measurements made several years ago on our best kilogram comparator [4]. In order to cope with such a large value of $\Delta I/\Delta I_d$ it was necessary to use a modified weighing scheme.

3. The Five Observation Scheme

The weighing scheme used is identical to that of table 1 except for the addition of a fifth operation which is a repeat of the first.[2] The scheme is shown in table 2.

Table 2. Five observation scheme.

Operation	Load on Balance	Balance Indication
1	Y	I_1
2	X	I_2
3	$X+d$	I_3
4	$Y+d$	I_4
5	Y	I_5

The apparent difference in mass between Y and X is then estimated as follows:

$$\Delta M = \frac{I_1-I_2-I_3+I_4}{-I_2+I_3+I_4-I_5} m_d \,. \tag{5}$$

Equation (5) is also unbiased for a linear drift between measurements (though eq (5) is not the least squares solution for a linear drift model). The real virtue of eq (5) is that it is also an unbiased solution for a model which assumes only that the drift between operations 1 and 2 equals the drift between operations 3 and 4; and that the drift be-

[2] To the author's knowledge, the first reported use of this weighing scheme was in a 1967 paper by Bowman, Schoonover, and Jones [8]. These authors used a five-observation scheme to compare an external object with the built-in weights of a single-pan, mechanical balance.

tween operations 2 and 3 equals the drift between operations 4 and 5 [5]. The first drift occurs between operations which exchange the test weights on the balance pans. The second drift occurs when the sensitivity weight is added or removed. This model frees the operator from having to wait equal times between all measurements. Since the addition or removal of the sensitivity weight is a faster operation than the exchange of test weights, it is usually possible to accomplish the scheme of table 2 (where one need not wait equal intervals between operations) in about the same time as it takes to carry out the scheme of table 1 (where one must take measurements at equally spaced intervals).

When one computes the variance of ΔM based on eq (5) one discovers a remarkable result:

$$\mathrm{var}(\Delta M) = S^2\sigma_I^2 \left[1-\frac{1}{2}\frac{\Delta I}{\Delta I_d}+\left(\frac{\Delta I}{\Delta I_d}\right)^2\right] \quad (6)$$

where $S=2m_d/(-I_2+I_3+I_4-I_5)$.

The appearance of a term linear in $\Delta I/\Delta I_d$ indicates that, unlike eq (2), the estimate of ΔI in eq (5) is not independent of the estimate of ΔI_d. The result of a negative term in eq (6) is that the variance of ΔM is insensitive to the ratio $\Delta I/\Delta I_d$ for ratios between 0 and 0.5. Within this range, the variance of ΔM is actually below what it would be if the ratio $\Delta I/\Delta I_d$ were zero (table 3). The minimum

Table 3. Comparison of variances with respect to $\Delta I/\Delta I_d$ for results derived from eqs (1, 2, and 5).

	$\mathrm{var}(\Delta M) = \mathrm{k}S^2\sigma_I^2$ k		
$\Delta I/\Delta I_d$	eq (1)	eq (2)	eq (5)
0	1.00	1.00	1.00
1/4	1.12	1.31	0.94
1/3	1.22	1.54	0.94
1/2	1.50	2.25	1.00
1	3.00	6.00	1.50

value for the variance of ΔM occurs for the ratio $\Delta I/\Delta I_d=0.25$, although this minimum is only 6 percent below the variance for a ratio of zero. Finally, if we want to ensure the variance of ΔM not exceed $(1.31)\cdot S^2\sigma_I^2$, we would make the rule that

$$\frac{\Delta I}{\Delta_d} < 0.86$$

This should be compared with eq (3).

4. Averaging

One of the ways to lessen the dependence of results obtained from table 1 on the ratio $\Delta I/\Delta I_d$ is by averaging. For N double substitutions at the same nominal load, one can average the N estimates of sensitivity and use the average value in the calculations of the various ΔM's. The NBS MASSCODE takes this approach and amends the rule for the ratio of $\Delta I/\Delta I_d$ to:

$$\frac{1}{N^{1/2}}\frac{\Delta I}{\Delta I_d} < 0.25 \quad (7)$$

to cover cases where $N>1$.

The amended algorithm leads to the following variance:

$$\mathrm{var}(\Delta M) = S^2\sigma_I^2 \left[1+\frac{5}{N}\left(\frac{\Delta I}{\Delta I_d}\right)^2\right]. \quad (8)$$

There are two possible objections to this approach. First, although the quadratic term in eq (8) is a factor of $1/N$ smaller than the same term in eq (4), it has been converted from a "within" to a "between-time" component [6]. Second, and more serious, the sensitivity of precision mechanical balances may be a function of time. This is certainly the case for NBS-2, the kilogram comparator which was designed and built at NBS and is now in use at the International Bureau of Weights and Measures (BIPM) [7]. In such cases, use of an average value for the sensitivity is unjustified.

5. Conclusion

The usual admonition that the ratio $\Delta I/\Delta I_d$ not exceed 0.25 ensures that the variance of a double substitution does not grow by more than 31 percent above its minimum value. We have examined a five-operation weighing scheme and have shown that use of this scheme relaxes the rule to the ratio $\Delta I/\Delta I_d$ not exceeding 0.86. We have also argued that the five-operation scheme can usually be performed in the same amount of time as the more usual four-operation scheme.

As a final comment, we emphasize that this analysis applies to un-servoed mechanical balances. For balances under servo control, the linear range of the scale is usually so large that it is never a problem to meet the conventional ratio rule. In addi-

tion, the sensitivities of servo-controlled balances are usually very stable over the course of a series of measurements.

References

[1] Almer, H. E., Method of Calibrating Weights for Piston Gages. NBS Tech. Note 577 (1971 May).

[2] Varner, R. N., and R. C. Raybold, National Bureau of Standards Mass Calibration Software. NBS Tech. Note 1127 (1980 July).

[3] Jaeger, K. B., and R. S. Davis, A Primer for Mass Metrology. NBS Spec. Publ. 700-1 (1984 November).

[4] Davis, R. S., Recalibration of the U.S. national prototype kilogram. J. Res. Natl. Bur. Stand. **90** (4) (1985 July-August) pp. 263-283.

[5] Carré, P., and R. S. Davis, Note on weighings carried out on the NBS-2 balance. J. Res. Natl. Bur. Stand. **90** (5) (1985 September-October) pp. 331-339.

[6] Pontius, P. E., Measurement Philosophy of the Pilot Program for Mass Calibration. NBS Tech. Note 288 (1966 May).

[7] Almer, H. E., National Bureau of Standards kilogram balance NBS no. 2. J. Res. Natl. Bur. Stand. **76C** (1,2) (1972 January-June) pp. 1-10.

[8] Bowman, H. A. and R. M. Schoonover (with appendix by M. W. Jones), Procedure for high precision density determinations by hydrostatic weighing. J. Res. Natl. Bureau Stand. **71C** (3) (1967 July-August) pp. 179-198.

CALIBRATION DESIGN AND STATISTICAL ANALYSIS

UNITED STATES DEPARTMENT OF COMMERCE
Alexander B. Trowbridge, *Secretary*
NATIONAL BUREAU OF STANDARDS • A. V. Astin, *Director*

Realistic Uncertainties and the Mass Measurement Process

An Illustrated Review

P. E. Pontius and J. M. Cameron

Institute for Basic Standards
National Bureau of Standards
Washington, D.C. 20234

National Bureau of Standards Monograph 103

Issued August 15, 1967

For sale by the Superintendent of Documents, U.S. Government Printing Office
Washington, D.C. 20402 - Price 20 cents

Library of Congress Catalog Card Number: 67–60056

Realistic Uncertainties and the Mass Measurement Process

An Illustrated Review

Paul E. Pontius and Joseph M. Cameron

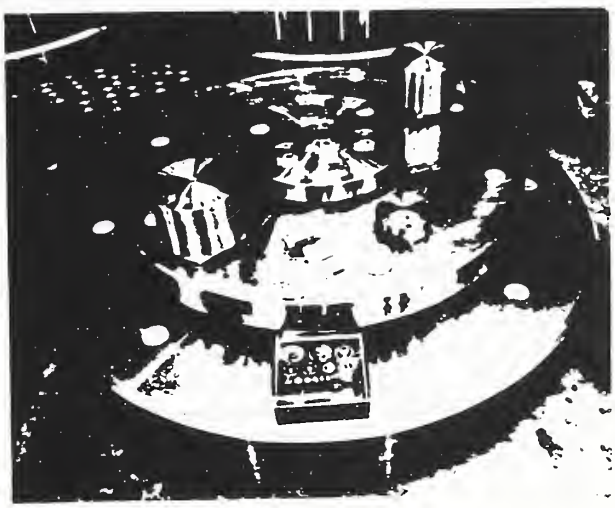
This paper gives a review of the concepts and operations involved in measuring the mass of an object. The importance of viewing measurement as a production process is emphasized and methods of evaluating process parameters are presented. The use of one of the laboratory's standards as an additional unknown in routine calibration provides an accuracy check and, as time goes on, the basis for precision and accuracy statements.

Key Words: Measurement, measurement process, uncertainty, mass measurement, precision, accuracy, statistical control.

Introduction

This paper is a condensed version of a lecture on "Error of Measurement" presented by Paul E. Pontius and Joseph M. Cameron at the Seminar on Mass Measurement, held at the National Bureau of Standards, Washington, D. C., November 30, December 1 and 2, 1964, and is essentially as presented by Paul E. Pontius at the 20th Annual ISA Conference held at Los Angeles, California, October 4–7, 1965.

It is a review of the mass measurement process from the initial basic concept to the statement of a measured mass value, examining in more or less detail certain important elements which are apt to be misunderstood, or perhaps misused. The importance of viewing measurement as a production process is emphasized and methods of evaluating process parameters are presented. The use of one of the laboratory's standards as an additional unknown in routine calibration provides an accuracy check and, as time goes on, the basis for precision and accuracy statements.



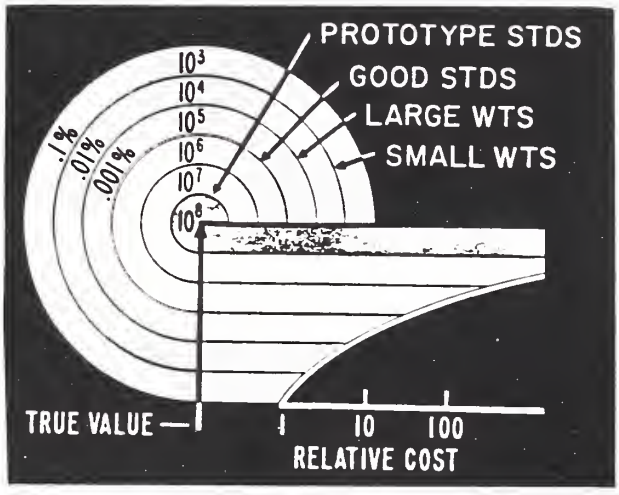
Mass Measurement Requirements

One role of the Bureau is to provide an extension of the mass measurement unit into the facilities of those who must use mass values to do other useful work. . . . These large weights, for example, are for use by another part of the Bureau to calibrate force measuring devices.

The calibration service provides values for single, selected groups, and ordered sets of standards, the values being with reference to the national

standard of mass. These values, together with a value for their uncertainty, allow each user to determine, in combination with his measurement process, the uncertainty of his measurements.

The three photographs above started with a group of standards whose cumulative total mass was in excess of one million pounds, and ends with a micropound standard, a range in excess of ten to the twelfth power (10^{12}).



The accuracy requirements for a measurement are set partly by experience, partly by discussions with others, and partly by analysis. For a particular purpose, the accuracy requirement must be established with care, as it provides a point of departure for the entire measurement process. Frequently we tend to lose perspective in regard to what we are measuring, or what the measurements mean, particularly if we concentrate on routine procedures or are remote to the actual measurement.

The aiming point for our measurement is to establish the mass, or true value, of a particular object for it is, in concept at least, unique and invariant. If, for example, accuracy within .01 percent is sufficient for our purpose, the target center is the area within the next to the last circle. Our measurements may group on either side of dead center, or may be randomly scattered across the center of the target, but as long as the spread is essentially within the target circle, the process is satisfactory for its intended use. Troubles arise when realistic requirements are divided by large arbitrary constants as specifications pass through various groups of people in a complex organization. Measurements accurate to better than .01 percent require attention to many details under more or less ideal conditions, and may not be obtainable under adverse conditions, consequently the entire measurement effort may be lost if the end use involves measurement processes of questionable precision. In the case of calibration, for example, in order to utilize the accuracy inherent in a good calibration, the user must work just as hard in his measurement process as the calibration facility did to determine the value of the standard originally.

The importance of incorporating the properties of the measurement process in setting up requirements or specifications is illustrated by the problem of adjustment tolerances for different classes of weights.

	TYPICAL PROCESS PARAMETERS			CLASS ADJ TOL.	
NOMINAL VALUE	UNCERTAINTY (SYS. ERROR) OF STD. VALUE	S.D. OF SINGLE MEAS.	SINGLE MEAS PROCESS UNCERTAINTY*	CLASS M (mg)	CLASS S (mg)
10 g	.0087 mg	.0074 mg	.031 mg	.050	.074
5 g	.0050	.004	.017	.034	.054
1 g	.0047	.004	.017	.034	.054
500 mg	.0024	.0007	.005	.010	.025
100 mg	.0009	.0007	.003	.010	.025
10 mg	.0008	.0007	.003	.010	.014

*3 S.D. + SYS. ERROR

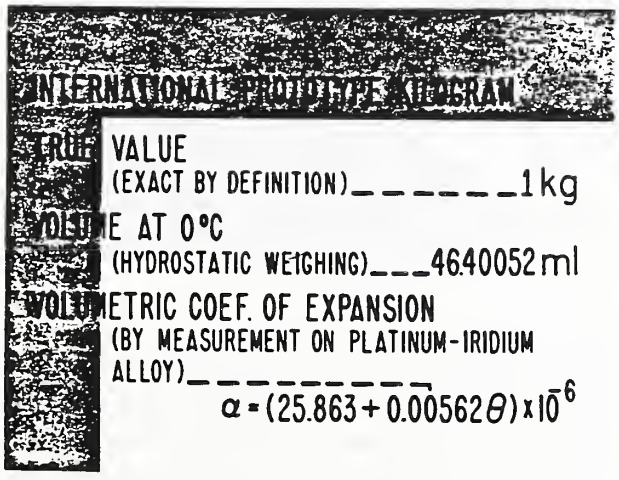
The Class M and Class S adjustment tolerance limits for selected weights are shown in the two right hand columns. The uncertainty associated with the stated value for standards of the same nominal value is shown in the 2d column and the precision for a single measurement is shown in the 3d column. If one tries to establish the compliance with Class M adjustment tolerances by a single weighing against a known standard, the uncertainty of the process would be as shown in the 4th column. This uncertainty, compared with the quantity we are trying to detect, is such that in

the first 4 cases the measurement uncertainty is a large fraction of the tolerance so that only those items well inside of tolerance have a good chance of being passed. A measurement procedure more sophisticated than a single comparison with a known standard may be desirable.

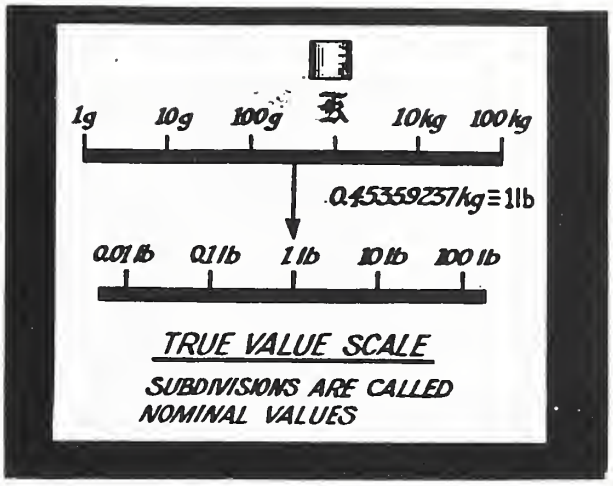
	TYPICAL PROCESS PARAMETERS			CLASS ADJ. TOL.	
NOMINAL VALUE	UNCERTAINTY OF CLASS M (WITHIN TOL.)	S.D. OF SINGLE MEAS.	SINGLE MEAS. PROCESS UNCERTAINTY	CLASS S (mg)	CLASS S-1 (mg)
10 g	.050	.0074	.072	[illegible]	.18
5g	.034	.004	.048	[illegible]	.18
1g	.034	.004	.048	[illegible]	.10
500mg	.010	.0007	.012	.025	.08
100mg	.010	.0007	.012	.025	.05
10mg	.010	.0007	.012	[illegible]	.03

We would be in greater difficulties if we were to try to establish compliance with Class S adjustment tolerances in the same manner with reference to Class M standards, which are known only to be within the Class M tolerance limits. In 4 of the 6 examples, the process uncertainty is of the same order of magnitude as the quantity we are trying to check. These examples illustrate the necessity for a careful evaluation before venturing a commitment on the performance of a particular measurement process.

The Unit of Mass



By practically universal agreement, the mass of the International Prototype Kilogram is the basic unit for mass measurement. It is a particular object, defined to have an exact invariant mass of one kilogram, that is to say, the true value is one kilogram. The volume and the coefficient of volumetric expansion are necessary to determine the best estimate of the true value of other objects compared with this standard.



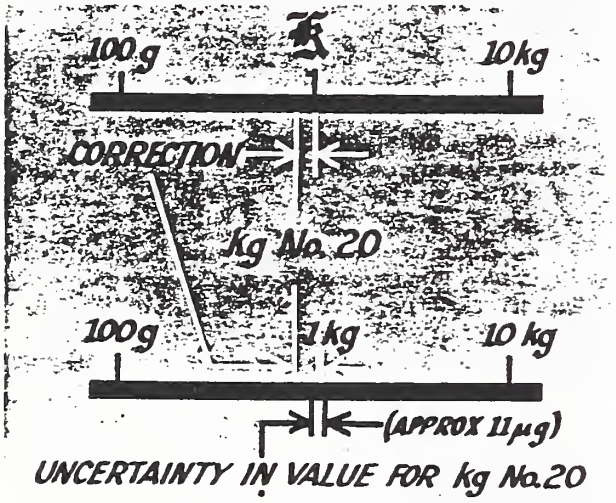
With the unit defined, we can logically construct a true value scale which has the property that some point on the scale will correspond to the mass of any chosen object. We call the major subdivisions of this scale nominal values. Other customary units, such as the pound, are not ambiguous if they have an exact definition relative to the basic unit. An intermediate point on the scale can be described either relative to the whole scale, as for example, 9.995 grams, or relative to the closest nominal value, in which case the point would be described as 10 grams minus 5 milligrams. The minus 5 milligrams may be called a correction or error, depending on one's viewpoint. The use of a nominal value and a correction is often convenient in computations, however, the word "correction", or "error", overly emphasizes the importance of the nominal value. Interpretation of tolerance limits on the value of the standard as the error automatically disregards the primary benefits of a good calibration. Only an ideal measurement method or process can produce true values of multiples and subdivisions of the basic unit which will exactly coincide with nominal values on the true value scale. It should be emphasized that, from a measurement standpoint, adjustment to nearly coincide with a nominal value is necessary only to assure an "on scale" condition when intercomparing equal nominal summations.

In our previous example, we elected to interpret the adjustment tolerance limits associated with our Class M set as the uncertainty of the value. While this may be appropriate with respect to the nominal value, such an interpretation raised serious doubts as to our ability to test the Class S weight set. If we had used the actual value and its uncertainty as a basis for our tests, the doubt essentially disappears. With minor modification at the 10 g level, the uncertainty of the values established for the Class S weights by our single measurement is clearly suitable for the task at hand. It must be emphasized that our apparent increase in measure-

NOMINAL VALUE	TYPICAL PROCESS PARAMETERS: UNCERTAINTY (SYS. ERROR) OF STD. VALUE	S.D. OF SINGLE MEAS.	SINGLE MEAS. PROCESS UNCERTAINTY *	CLASS ADJ. TOL: CLASS S (mg)	CLASS S-1 (mg)
10 g	.0087 mg	.0074 mg	.031 mg	.074	.18
5 g	.0050	.004	.017	.054	.18
1 g	.0047	.004	.017	.054	.10
500 mg	.0024	.0007	.005	.025	.08
100 mg	.0009	.0007	.003	.025	.05
10 mg	.0008	.0007	.003	.014	.03

*3 S.D. + SYS. ERROR

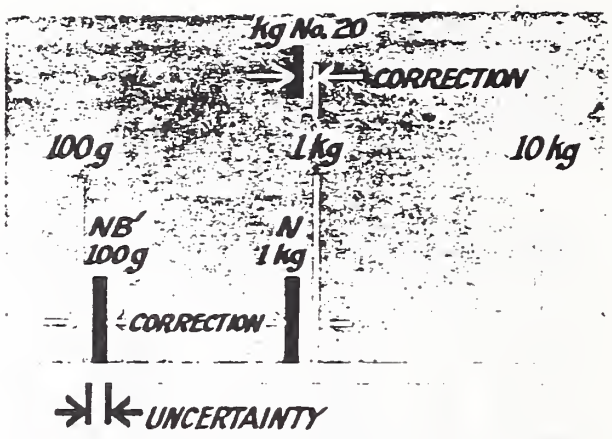
ment capability did not require any change in our process hardware. It has been achieved, for the most part, by a change in philosophy.



UNCERTAINTY IN VALUE FOR kg No. 20

Our access to the true value scale as established by the international standard is through prototype kilogram number 20. The estimated true value of number 20 is 1 kilogram minus 19 micrograms, based on several measurements. We can construct an accessible true value scale by setting off from the value of kg 20 an amount equal to the correction. Practically, the stated value is assumed to be exact, the uncertainty of the value introducing only a slight systematic error in our reconstructed scale.

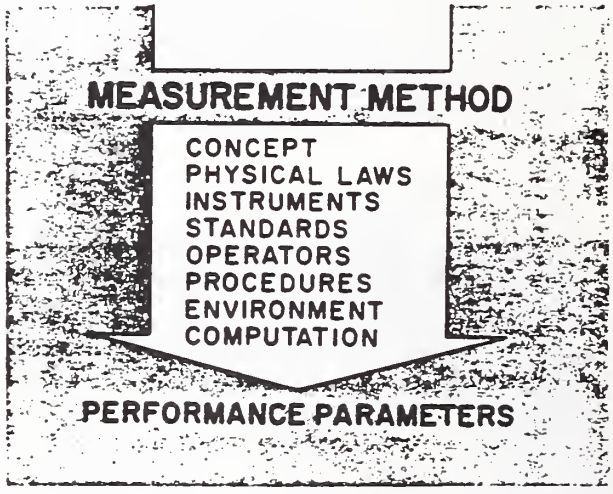
By comparing other objects with kilogram 20, either singly or in combination, we can assign values relative to our accessible scale. A sufficient number of well calibrated standards which can be intercompared, and which may occasionally be compared with our prototype standard, serve to maintain our scale with perhaps a greater precision than was available in the starting measurements. All mass values on NBS Reports of Calibration are with reference to a minimum number of selected mass standards. For example, practically all sets



TRUE VALUE OR IDEAL SCALE

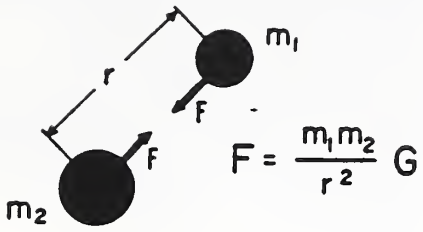
of metric weights are calibrated with reference to a pair of 1 kg or a pair of 200 g or a pair of 100 g weights. The national reference standards group does *not* include weights of all denominations.

Measurement Method

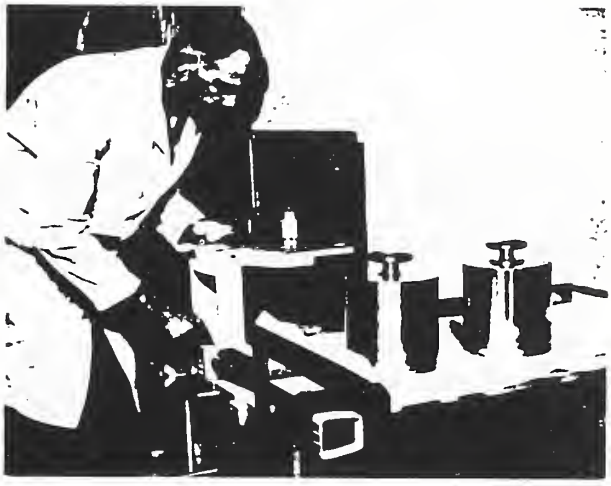


A practical measurement method is easy to visualize in the form of a broad outline of the elements of the method such as, the concept of the quantity to be measured, pertinent physical laws, various instruments, standards, the operators, procedures to be used, the environment in which the measurements are to be made, the computations which are to be made, and a means of establishing some parameters of performance. As we briefly review some of these elements, we will find that every mass measurement facility has many things in common.

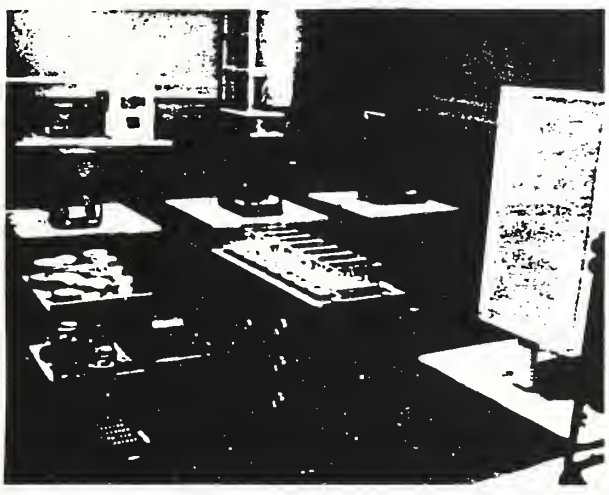
Mass is an inertial property of an object, which, within the framework in which our measurements apply, is considered to be proportional to the amount of material. Mass is generally thought of as being measured through some application of



F – ATTRACTIVE FORCE
m_1, m_2 – MASS OF BODIES
r – DISTANCE BETWEEN C.G.'s
G – UNIVERSAL CONSTANT



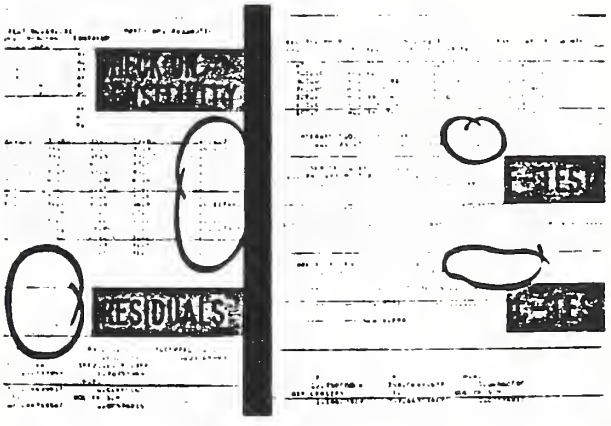
Newton's law of gravitational attraction, however, it is perhaps more precise to say that measurements are made by comparing the forces attracting suspended bodies toward the earth—that is the net vertical forces including the effects of G, air buoyancy, rotation of the earth, etc.

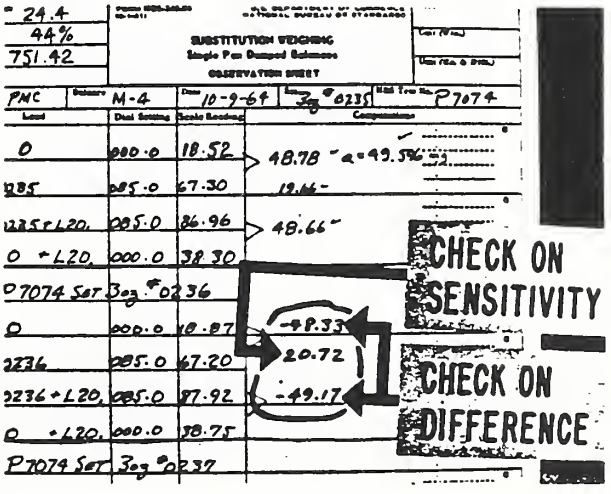


The environment in which the measurements are made does not vary substantially between calibration facilities. Weighing rooms are almost universally clean, with restricted access, and relatively free of vibration. With the possible exception of freedom from vibration, these desirable features are easily obtained.

People operate the equipment, following prescribed procedures. Operator skill increases with practice, and in time, operators in a given group approach a uniform level of skill.

Each comparison, or weighing, consists of a sequence of operations, more or less formalized. Detailed procedures and weighing designs, ranging from simple to complex, are available for a wide variety of requirements. Modern computation equipment ranging from desk calculator to electronic computer are now widely available so that laborious long hand computations are no longer necessary.





While perhaps not generally considered so, analysis is a part of the measurement method. Whether done by machine . . .
. . . or by hand, the analysis verifies that such parameters continue to be applicable.

A PARTICULAR MEASUREMENT METHOD

INSTRUMENT . . . AI
STANDARDS . . . $200_1, 200_2, 100_1$
PROCEDURES . . . CLEAN & WEIGH USING 52-I SERIES
OPERATOR . . . P. CRONE
ENVIRONMENT . . . ROOM 1, SOUTH
COMPUTATION . . . COMPUTER PROGRAM
ANALYSIS . . . F-TEST, t-TEST

A particular measurement method is like a specification for a particular measurement. The specific instrument, the standards to be used, the specific operations to be performed and the planned sequence in which they are to be carried out, the operator, the location, and the method of computation and analysis, collectively define a particular measurement method. Until the measurement has actually been made and analyzed, the performance is only "on paper" and therefore ideal.

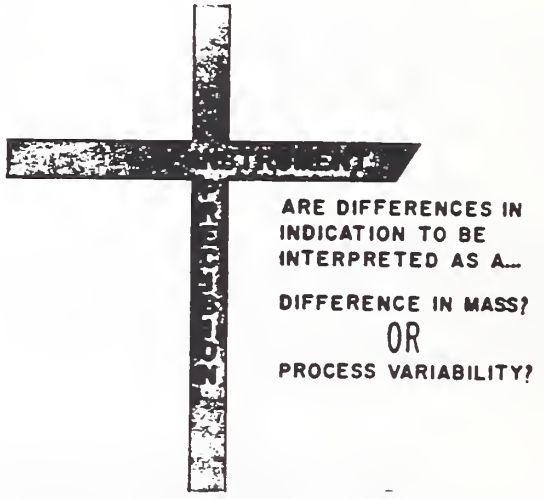
A MEASUREMENT PROCESS

PRODUCES:

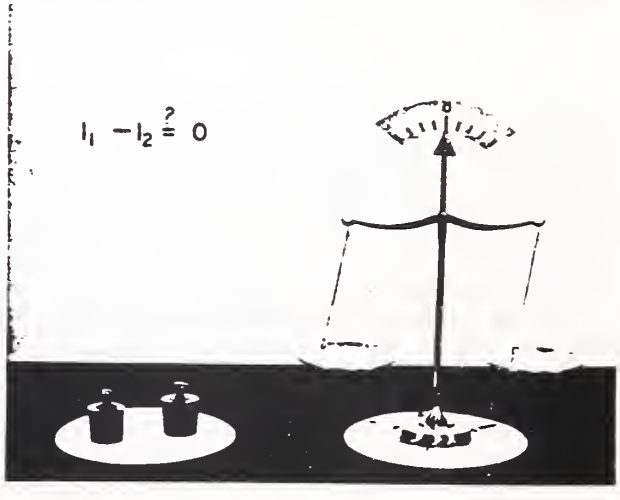
1. A USEFUL MEASURED VALUE
2. AN ESTIMATE OF UNCERTAINTY FOR THAT VALUE

A measurement process involves the actual physical operation of the specified equipment following the procedures as closely as possible. It is subject to the many variations that can and do occur during the operation. The end result is an estimated best value, which, in order to be useful, must be accompanied by the uncertainty with reference to known performance parameters.

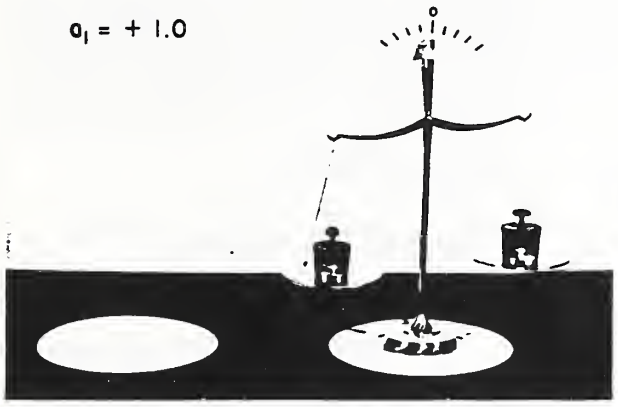
Changes in any one or in a group of elements of the method constitutes, in effect, a different particular method and a different process which will in turn produce a different result and a different uncertainty. Small changes can make the difference between a useful value or a wasted effort.



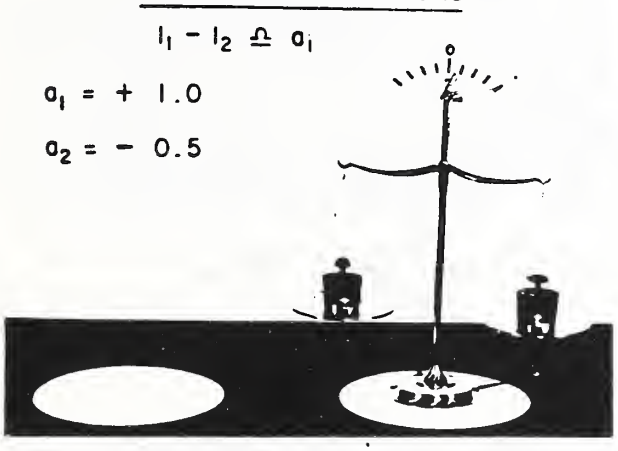
Because we must establish the mass of the object in question by measuring the mass difference between it and some known standard, the comparator is a vital element in the process. The inherent characteristic of the comparator is precision—not accuracy. The fundamental question is whether the indicated difference is really a mass difference, or an indication of some other variability. While we may be able to identify large sources of variability, in the limit, we cannot differentiate between instrument precision, variability from extraneous sources, or variability of the standard.



We start by determining the indicated difference between two objects that are nearly alike.



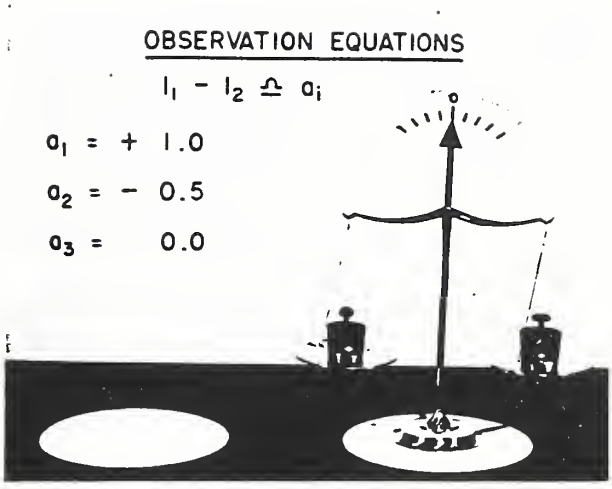
From our first comparison, it appears that the round knob weight on the left is clearly heavier than the flat knob weight by one scale division. If we stop here, we would simply state the value of one object in terms of another, however, we have no way of knowing the uncertainty to associate with this value.

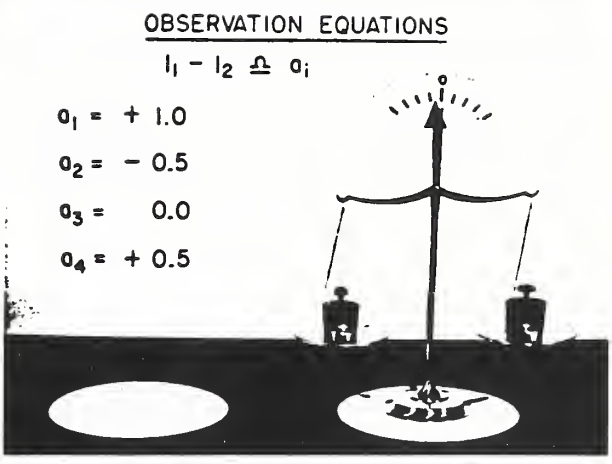


If we repeat the comparison at some other time, we are quite likely to obtain a different result. This raises a serious question—which of the two results is correct?

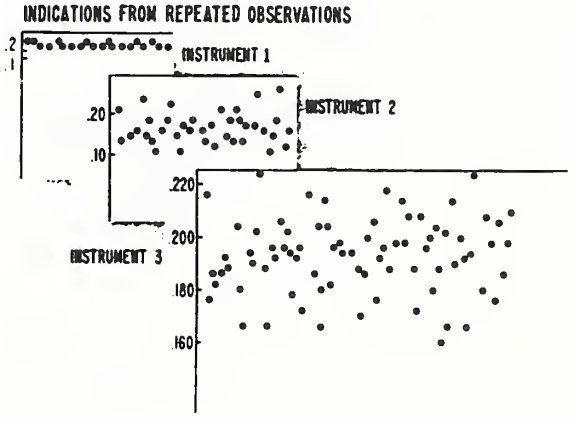
We repeat the comparison again . . .

. . . and again. Now there are four different values, none of which alone can be considered the best measure of the difference, but considered as a group they can tell us something about the



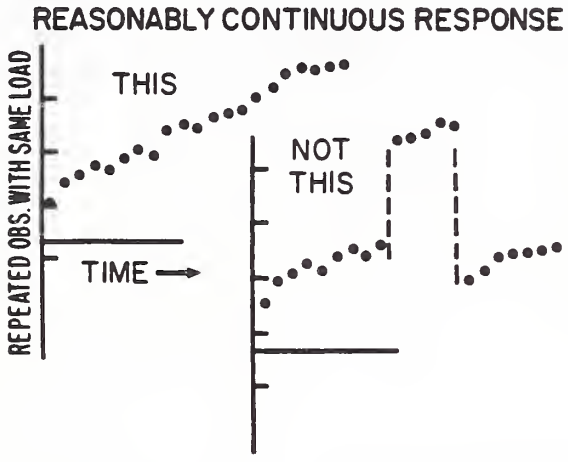


instrument. Continuing to record the indicated difference between two similar objects, and preferably making the comparisons in the environment in which the instrument is to be used, a plot is made against time of the differences which may look like this.



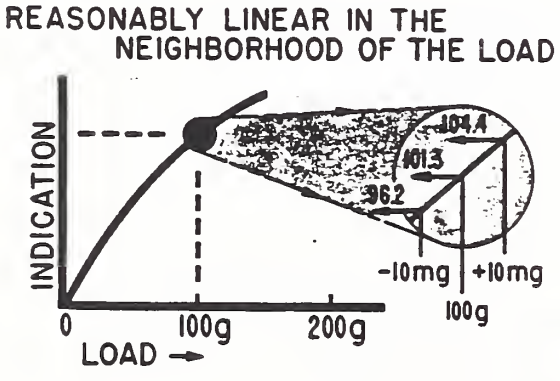
257-257 O - 67 - 2

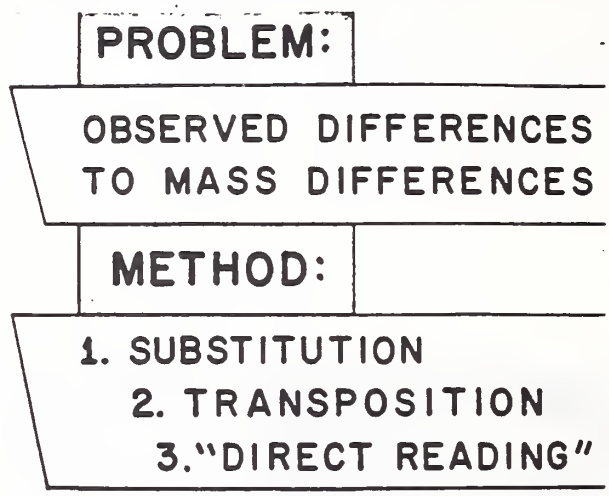
The first plot indicates a severe rounding off, which may be from several causes. Such a response clearly lacks the appearance of randomness. The second plot at least appears to be random. The third plot, while perhaps appearing to be random, obviously lacks the precision of the second plot. The range of the differences as plotted gives us an idea of the smallest mass difference that can be detected with assurance, and is obviously related to the requirements our measurements must meet. Repeated independent measurements of the same mass difference are essential to the evaluation of the instrument.



The operator, or manufacturer, must search for cause and effect until repeated indications for the same load, or differences are reasonably consistent. Effects which are periodic in nature, but with a period significantly longer than the period of the instrument, can be minimized in the design of the weighing method.

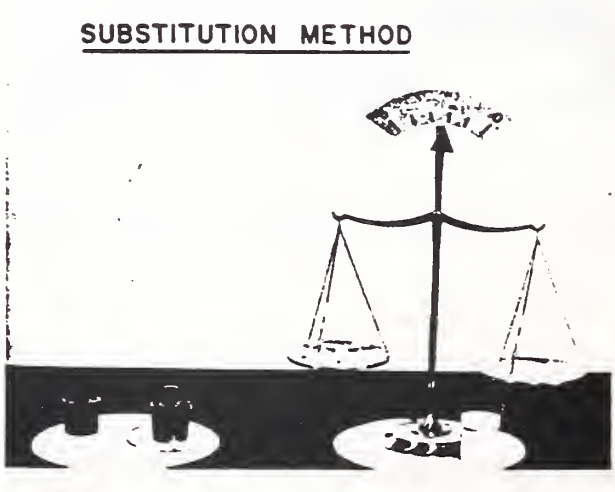
One additional requirement, generally beyond the control of the operator, is that of linearity. An instrument, used as a comparator rather than a direct reading device, requires linearity only in the neighborhood of the actual load.





The problem of establishing the correspondence between observed differences and mass differences is a part of the weighing method. The first two methods, substitution and transposition, are comparative methods. That is to say, the method requires observations relative to a suitable standard along with the unknown. With these methods, the measurement equipment need be continuous only over the time interval required for making a group of observations and linear only over the range of the difference between the standard and the unknown. Most direct reading equipment is in a sense a substitute standard, that is, at some point in time it is calibrated with reference to a standard, and from that point until recalibration, it is generally assumed to have a long term constancy approaching that of the standard. Most mass measurement equipment can be used either way. The smallest uncertainties invariably will be associated with the comparative mode of operation.

Weighing Method



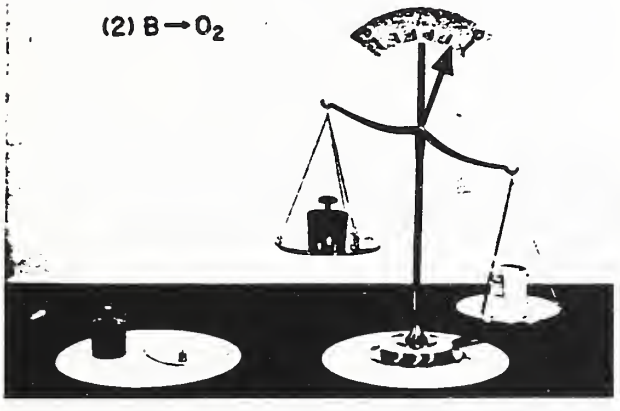
To illustrate the principle, the double substitution method is performed as follows: We start with a simulated equal arm balance, a tare weight—the white cylinder near the base of the balance, a sensitivity weight of known value immediately in front of the dark weight near the center, and two nearly equal brass weights, one with a flat knob in the center and one with a round knob on the left. The scale indication is in arbitrary numbers and the tare weight is necessary to establish an "on scale" condition.

(1) $A \rightarrow O_1$

The first observation is that produced with the round knob weight on the pan.

(1) $A \rightarrow O_1$

(2) $B \rightarrow O_2$



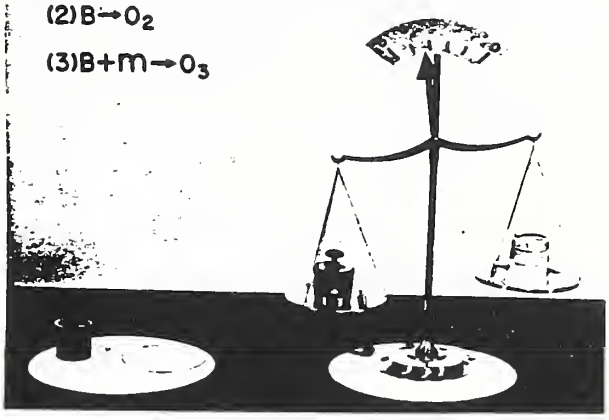
The second observation is that produced with the flat knob weight, which might be a standard, replacing, or substituted for, the round knob weight.

The third observation is that produced by repeating the previous step and adding the sensitivity weight to the pan load.

(1) $A \rightarrow O_1$

(2) $B \rightarrow O_2$

(3) $B+m \rightarrow O_3$

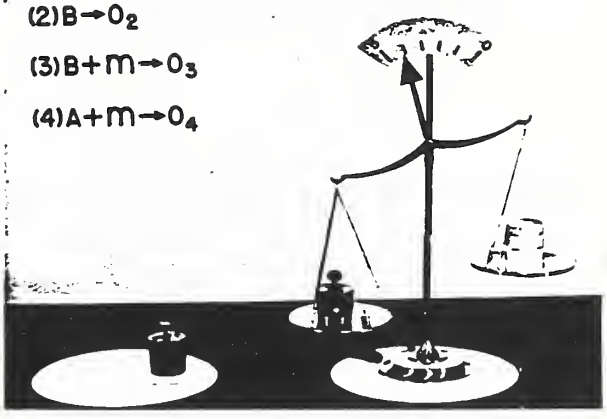


(1) $A \rightarrow O_1$

(2) $B \rightarrow O_2$

(3) $B+m \rightarrow O_3$

(4) $A+m \rightarrow O_4$



The fourth observation is a repetition of the first step including the sensitivity weight.

$$A-B \triangleq K\left\{\frac{(1)-(2)+(4)-(3)}{2}\right\}$$

$$\triangleq K\left(\frac{O_1-O_2+O_4-O_3}{2}\right)$$

$$B+m-B \triangleq K\{(3)-(2)\}$$

$$m \triangleq K(O_3-O_2)$$

$$A-B \triangleq \frac{m}{(O_3-O_2)}\left(\frac{O_1-O_2+O_4-O_3}{2}\right)$$

Using the requirement for continuity, a relation can be established for A minus B from the average of the two sets of differences as shown. Using the linearity requirement, the constant of proportionality K, or the mass value of the indicating scale division can be determined from the second and third observation. Finally, the difference A minus B is expressed as a function of the observations, in ratio form and the value of the sensitivity weight.

$$A-B \triangleq \frac{m}{(O_3-O_2)}\left[\frac{O_1-O_2+O_4-O_3}{2}\right]$$

$$A-B \triangleq \frac{m}{(O_3-O_2)}\left[\frac{O_1-O_2}{2}\right]$$

$$A-B \triangleq \frac{m}{(O_3-O_2)}\left[\frac{O_1-O_2+O_4-O_3}{4}\right]$$

All usual methods result in very similar relations expressing the difference between two objects being compared. In all cases, A minus B is expressed as a ratio between sets of observations multiplied by the value of the sensitivity weight. Obviously requirements for knowledge of the value of M are minimized when the size of the ratio involving the observation is small. The constant of proportionality, K, is really the ratio in front of the bracket terms which we call the value of the division. The strange equal sign is used to indicate that the relations shown are observational equations and not mathematical identities.

With the measurement method agreed upon, let us now discuss its performance—we put it into production and see how it works out as a measurement process.

Measurement as a Process

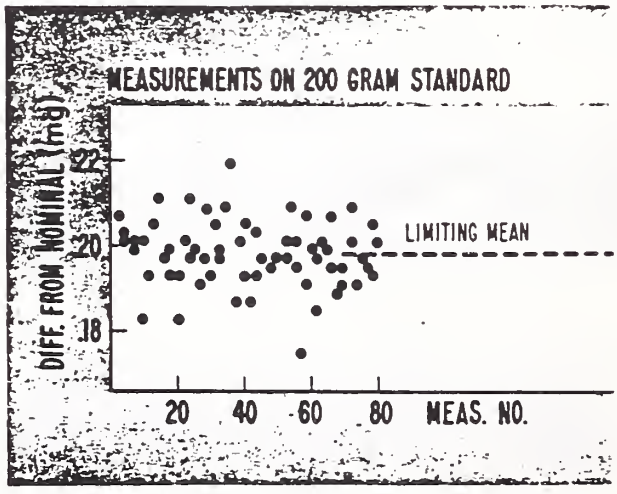
MEASUREMENT PROCESS

OUTPUT	MEASUREMENT
PROCESS AVG.	LIMITING MEAN
VARIABILITY	PRECISION
BIAS	SYSTEMATIC ERROR
PROCESS LIMITS	UNCERTAINTY OR ACCURACY

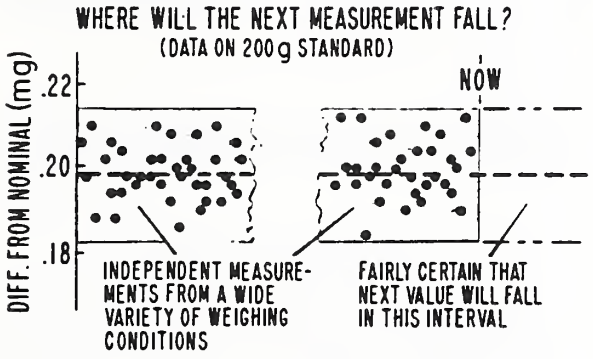
A measurement process is essentially a production process, the "product" being numbers, that is, the measurements. A characteristic of a measurement process is that repeated measurements of the same thing result in a series of non-identical numbers. To specify a measurement process involves ascertaining the limiting mean of the process; its variability due to random imperfections in the behavior of the system, that is, its precision; possible extent of systematic errors from known sources, or bias; and overall limits to the uncertainty of independent measurements.

PARTICULAR MEASUREMENT METHOD

INSTRUMENT	AI
STANDARDS	$200_1, 200_2, 100_1$
PROCEDURES	CLEAN & WEIGH USING 52-I SERIES
OPERATOR	P. CRONE
ENVIRONMENT	ROOM 1, SOUTH
COMPUTATION	COMPUTER PROGRAM
ANALYSIS	F-TEST, t-TEST

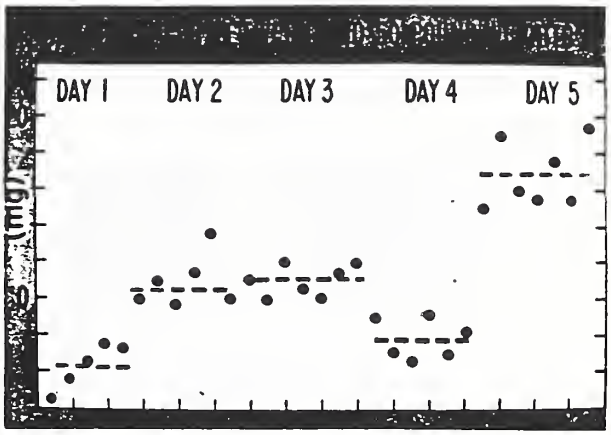


The chart shows measurements on a 200 g weight, plotted in the order in which they were taken. Despite the presence of one or two stragglers, the measurements tend to cluster around the central line—the process average or limiting mean. Our confidence that the process has settled down to a single limiting mean is strengthened as the length of the record is increased. We may have satisfied ourselves regarding the mean but what about the next measurement?

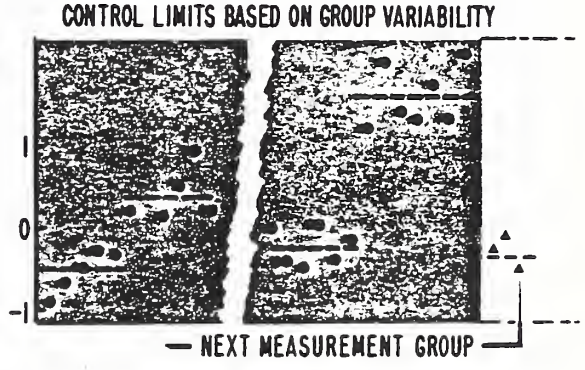


It seems clear that we cannot give an exact answer but will have to content ourselves with a statement that allows for the scatter of the results. Our goal is to make a statement with respect to a new measurement that is independent of all those that have gone before. As indicated in the chart, if we had a sufficiently long record of measurements we could set limits within which we were fairly certain that the next measurement would lie. Such a statement should be based on a collection of independent determinations, each one similar in character to the new observation, that is to say, so that each observation of the collection and also the new observation can be considered as random drawings from the same probability distribution. These conditions will be satisfied if the collection of points is independent, that is free of patterns, trends and so forth; and provided it is from a sufficiently broad set of environmental and operating conditions to allow all the random effects to which the process is subject, to have a chance to exert their influence on the variability. Suitable collections of data can be obtained by incorporating an appropriate measurement into daily routine weighing procedures, for example, a daily measurement of the difference between two laboratory weights, or in the regular calibration of the same weight.

If the measurements tend to cluster when taken close together in time, like the results shown on the chart, some systematic effect is present and certainly the results are not independent. This may

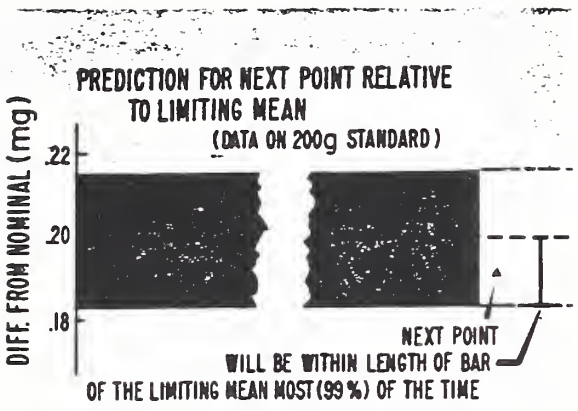


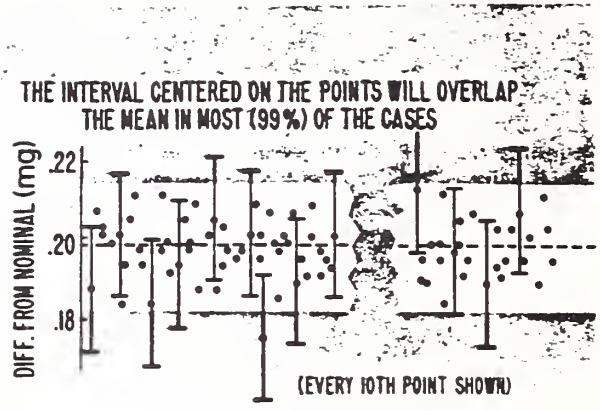
be due to some as yet undetermined cause, and the group means may have the appearance of randomness of the previous chart.

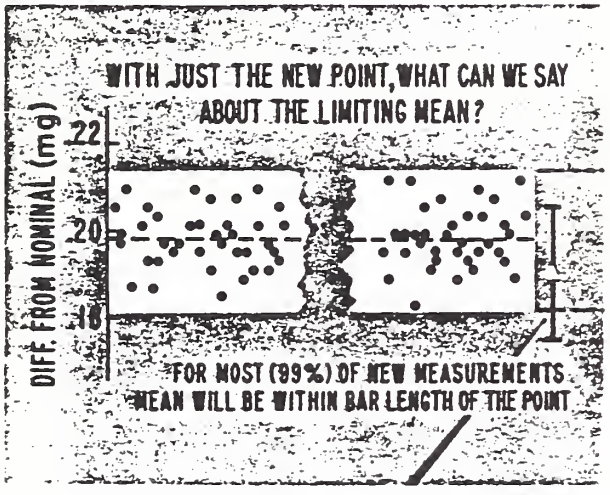


The group means may tend to a limit and the process may have all the properties of a good measurement system, once the allowance is made for the grouping. It is important that grouping be properly handled in determining the precision of the process. By modifying the process or changing the schedule of measurements to give the effect of independent measurements, we can arrive at a situation like the values on the 200 g standard. The shaded band is meant to suggest a limit, not an artistic slide.

From a study of a sequence of such independent measurements, we can use control chart techniques to set up limits within which the next value should lie. In the case where we have an extremely long sequence, a bar, as illustrated in the chart, can be marked off on either side of the mean so that some suitable fraction, say 99 percent, of the observations are within the interval represented by its length.





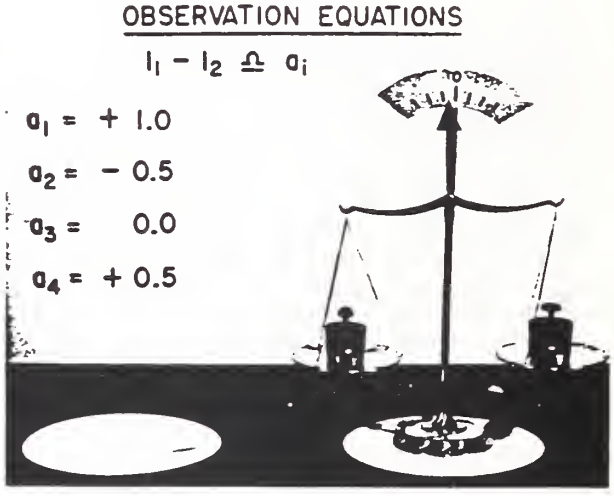


We can reverse the process and say that the probability is 99 percent, that the true value, or limiting mean, will not be more than the width of the bar from any observation chosen at random. This will be true of the next observation as well, provided it is an independent measurement from the same process. The probability statement attaches to the sequence of such statements. For each individual new observation the statement is either true or false but in the long run 99 percent of such statements will be true.

Assuming that the limits on the chart are based on large numbers of observations, we would find that very nearly the intended percentage of all such bars, centered on the observed values, would in fact overlap the mean. Only in those cases, such as the points in the area outside of the control limits, will the bar fail to overlap the mean. This is expected in only 1 percent of the cases. More frequent occurrence is a clear indication of either loss of control or that the limits were not properly set. Once we are satisfied that the process has a limiting mean value and is stable enough to permit prediction we turn our attention to evaluating its precision.

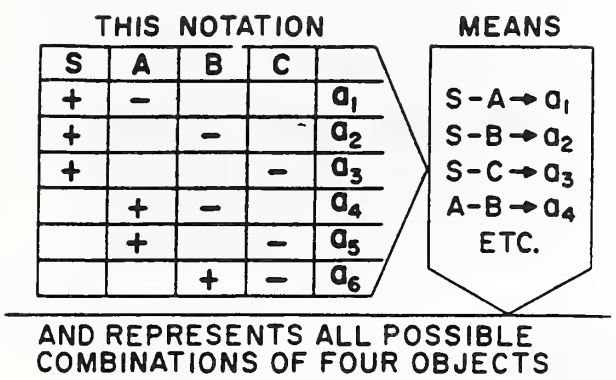
Process Precision

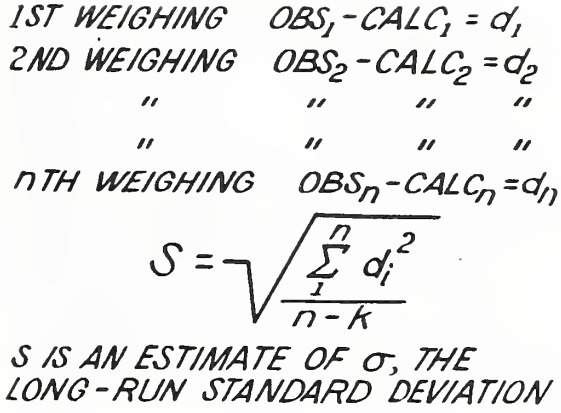
Let us now take a look at the situation in weighing to see what is involved in the study of the precision of the process.



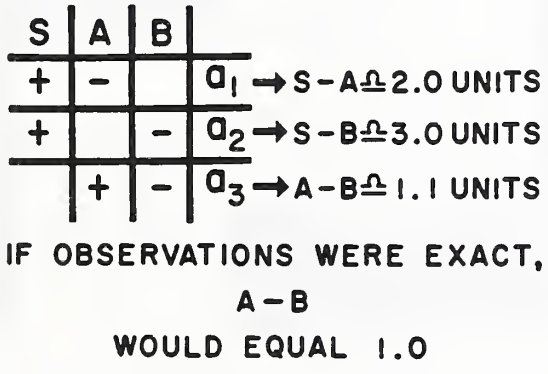
A characteristic of a measurement process is that it produces non-identical results. In our previous charts we had measurements of a 200 g weight, here are shown four measurements of the difference in mass. Through the redundancy—here 3 extra measurements—we get our grip on precision. In weight calibration we do not rely on repeated measurements of the same quantity but achieve the same result in another way.

When we intercompare four objects, for example, four 1-kg standards, we could use six observations. Weight S is compared with A for a_1, S with B for a_2 and so on. If S were a standard and the rest



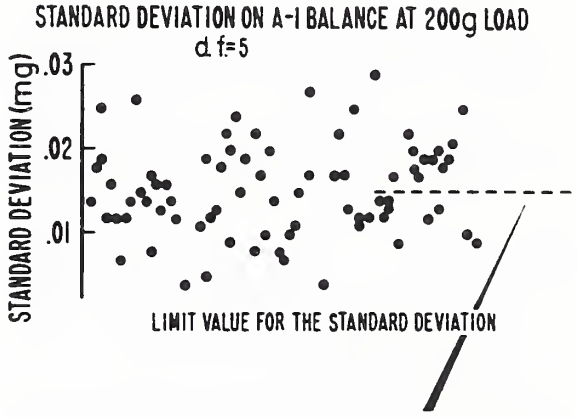


unknowns, we again have 3 more measurements than we need and these serve to tell us of the precision of the process.

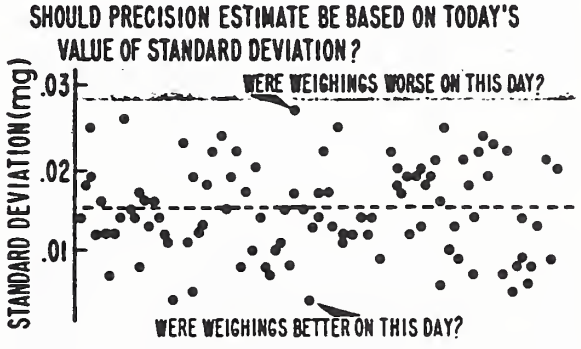


A simple example, using only three of the observations of the previous series, with S as the standard, A as the unknown, and B as the check standard, might give rise to the values shown. If everything were perfect, all equations representing the weighings would be satisfied exactly. Their lack of agreement would give a measure of the variability.

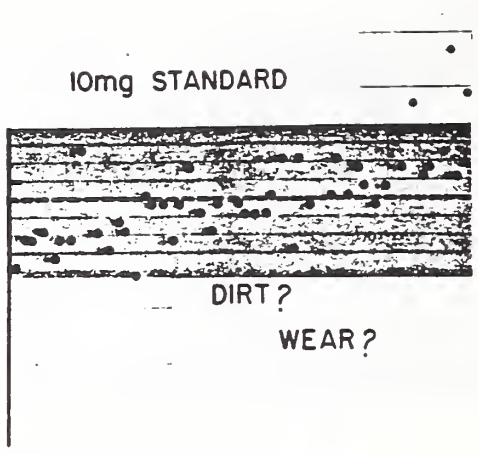
In general, for such weighing, there will be a discrepancy between the observed value and the best value calculated from the data, "best" meaning in most cases the value obtained in the method of least squares. If all is going well, none of these deviations will be too large, and also certain combinations of them, such as the sum of the squares, will also be well behaved. For statistical analysis the standard deviation, S, is used as the measure for describing variability. The quantity, S, is a function of the observational errors and will change with each set of data just as the values for the unknown weights do. (The quantity, k, is the number of unknowns in the system.)



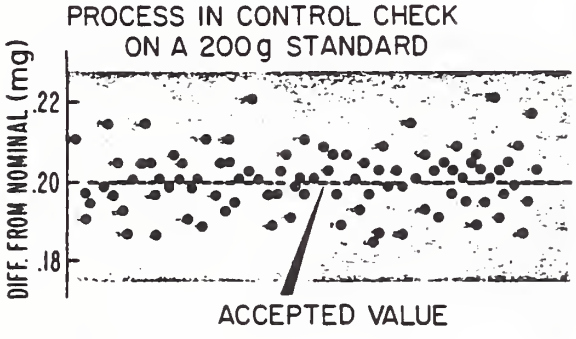
If the process is in a state of control these values of s will scatter about some value which is the true or long run standard deviation of the process.

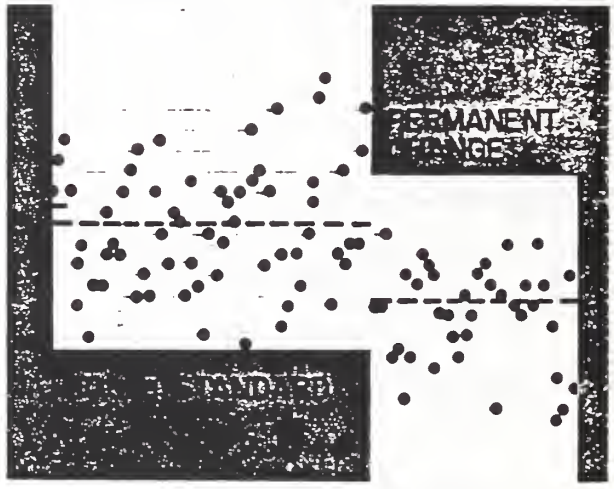


The argument that the uncertainty should be based on the internal agreement of today's values on the grounds that each day is unique or that weighing conditions are better on one day than on another may well be true. However, it will be expensive to make enough measurements on a given day to be sure that the variability has indeed changed from its long run average or to provide a reliable enough value to represent today's results. If the process did not change, using today's value would be analogous to keeping the last value of a sequence rather than using the mean represented by the dotted line. It is a sign that weighing conditions are not being reproduced, i.e., that the process is not in control if the standard deviation does not stay within predicted limits. Let us now look again at the check standard.



Process Mean



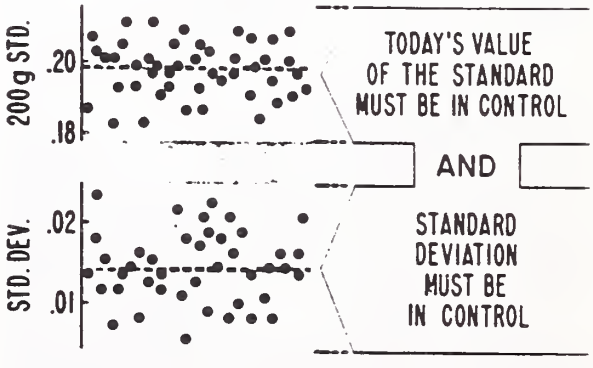


Each value obtained for the check standard serves not only a check on the process mean, but also can be used for evaluating the process variability. The same check standard, perhaps one of a group reserved for this purpose, is used consecutively in a given procedure until many independent values are obtained.

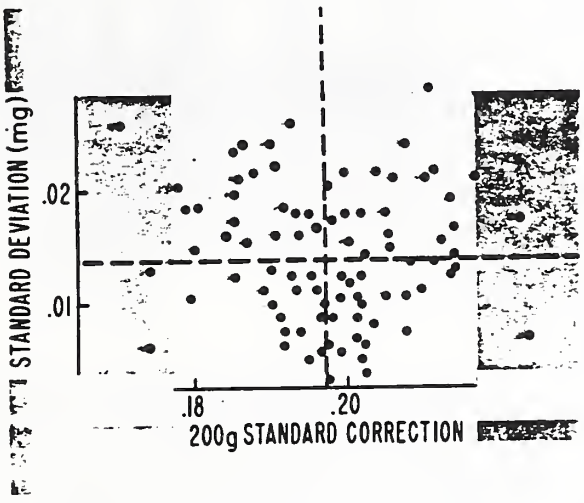
The importance of randomness cannot be overemphasized. As the collection of independent measurements on the check standard grows, it must be continually re-evaluated with reference to predicting the band within which the next point will lie. Slow drifts or sharp discontinuities are cause for concern until corrected, or satisfactorily explained.

If values return to normal after cleaning, one can rest easy, knowing the process is behaving properly. Indication of permanent changes are sometimes harder to explain, and even the most careful laboratories must occasionally repeat measurements because of troubles with foreign material adhering to or falling off the standard. If the new mean value persists over a sufficient number of measurements, it is proper to assume the standard has changed for some reason.

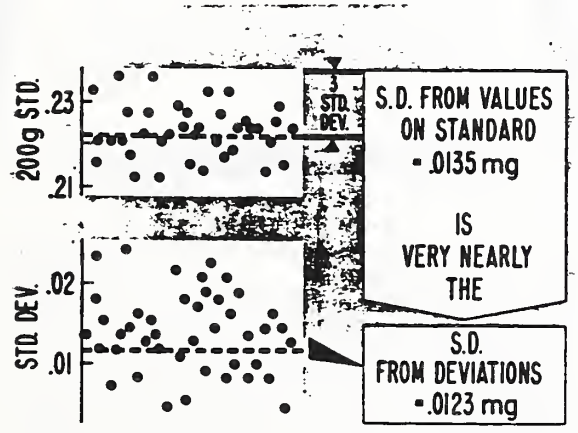
Process Control



A check on *just* the value of the standard or *just* the precision is not enough. It turns out that the value for the precision and the value for the check standard are generally independent, that is, when s is small the deviation of the value determined for the check standard from the accepted value is equally often big and small. For control we need both conditions.



For a given set of observations the precision must be proper as shown on vertical scale and we must have a check on a known weight to establish that the limiting mean has not changed as shown on horizontal scale. Until these conditions are fulfilled, we cannot be sure exactly what it is that we are measuring. These are necessary conditions, and in perhaps most cases, also sufficient conditions to proclaim that the measurement process is in a state of control, as indicated by points within the central rectangle.



Because the check on the standard is spread over a considerable time interval, the variability will include the proper diversity of environmental and other factors and the sequence will, in the absence of seasonal or other systematic trouble, approximate a sequence of independent values. If the weighing conditions are reproducible, then the daily standard deviation, s, and the variability as computed from the values of the check standard will be in agreement, i.e., the long run average of the variability as estimated from the control chart on the standard deviation should approach the corresponding value from the control chart based on the variability of the values of the check standard. Frequently, one is not in as good a shape as that indicated on the slide. When the measurements are spread out in time or space, an additional component of variation enters so that the lower chart gives an overly optimistic view of the process. A realistic estimate of process variability has to be based on that from the upper chart which reflects the total variation to which the measurements are subject. One would still use the within occasion variability for checking on control of the process, of course.

DETERMINING THE MASS OF AN OBJECT AND THE ASSOCIATED UNCERTAINTY IS A CALIBRATION.

ROUTINELY, THE CALIBRATION MUST BE LIMITED TO A FEW MEASUREMENTS.

If in calibration we could measure the difference between the standard and the unknown again and again we could make an uncertainty statement similar to those just discussed for the case of measurements of a fixed difference, but in fact, we cannot routinely make enough measurements of this type to permit reliable estimates of the uncertainties.

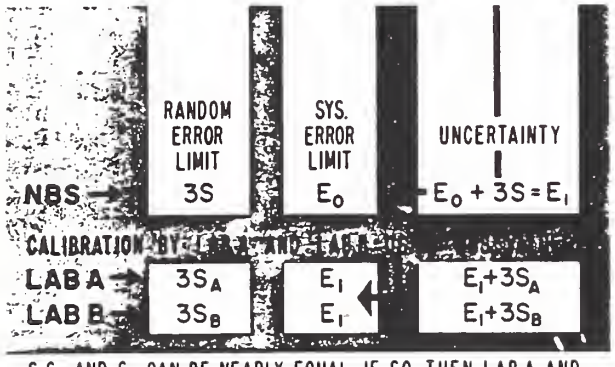
Process Parameters and Uncertainty of Calibration

If we could be sure that our measurements of the difference between the unknown and the standard came from a process in a state of statistical control, that is to say a stable process with a known variability, then we could transfer the properties of the process to the individual measurement and be correct a stated percentage of the time.

THE MEASUREMENT PROCESS REMAINS, AND IS, IN A SENSE, A CAPITAL INVESTMENT.

THE MEASUREMENTS, LIKE PRODUCTS, PASS ON TO OTHER DESTINATIONS.

All who weigh, or make other measurements, should concentrate on the properties of the measurement process—the degree to which the process re-creates the same value for its standards and exhibits the same level of variability. These are the properties that remain. The weights that are calibrated pass on to other destinations.



S, S_A AND S_B CAN BE NEARLY EQUAL. IF SO, THEN LAB A AND LAB B CAN CALIBRATE THIER OWN SET FROM SELECTED STANDARD WEIGHTS

At every stage in the extension of a measurement unit from an accepted standard to the ultimate user, there are three items of interest—a standard item, or items, with announced values and associated uncertainty, an assembly of equipment and procedures necessary for making the necessary comparisons, and the items which must be measured to accomplish some useful task. The uncertainty of the values established for the user are of paramount importance. This uncertainty has two components—one associated with the value of the starting standard and one reflecting the contribution of the local measurement process. The total uncertainty at any particular place becomes the systematic error for those who must use the service provided.

NATIONAL BUREAU OF STANDARDS — INSTITUTE FOR BASIC STANDARDS
REPORT OF CALIBRATION TEST NO. 188180 — U.S. DEPARTMENT OF COMMERCE

FOR LOCKHEED-CALIFORNIA COMPANY
BURBANK, CALIFORNIA

ITEM SET OF MASS STANDARDS 10KG TO 1G DESIGNATED SET C
MFR. SARTORIUS-WERKE AG SERIAL NO. 781163
STATED DENSITY 10KG TO 1G 7.89 G PER CM3 AT 20C

THE ABOVE ITEMS HAVE THE MASS VALUES SHOWN WITH REFERENCE TO THE NATIONAL STANDARD OF MASS. SEE ATTACHED SUPPLEMENT FOR LIMITATIONS IN USE OF APPARENT MASS VALUE AND UNCERTAINTY FROM ... IN DENSITY.

ITEM	NOMINAL	APP MASS CORR (MG)	TRUE MASS CORR (MG)	UNCERTAINTY (MG)	VOL AT 20 (CM)3
1KG	1000.00	0.045700	9.128909	0.110227	26.74388883
500G	500.00	0.278974	4.819581	0.069271	82.37196884
300G	300.00	-0.965581	2.458779	0.063098	38.02512489
200G	200.00	-0.018969	1.797272	0.052346	25.34877014
100G	100.00	-0.093402	0.814718	0.064880	12.87437446
50G	50.00	-0.013024	0.441037	0.032909	6.33719146
30G	30.00	-0.000559	0.271878	0.020735	3.80231580
20G	20.00	0.075999	0.257824	0.014434	2.53488587
10G	10.00	-0.118775	-0.027964	0.010710	1.26742357
5G	5.00	0.002680	0.048086	0.005702	0.63371985
3G	3.00	0.015487	0.042730	0.004076	0.38025355
2G	2.00	0.043916	0.062078	0.003082	0.25349289
1G	1.00	[illegible]	[illegible]	[illegible]	0.12673927

THE UNCERTAINTY FIGURE IS AN EXPRESSION OF THE OVERALL UNCERTAINTY USING THREE STANDARD DEVIATIONS AS A LIMIT TO THE EFFECT OF RANDOM ERRORS OF MEASUREMENT, THE MAGNITUDE OF SYSTEMATIC ERRORS FROM KNOWN SOURCES BEING NEGLIGIBLE.

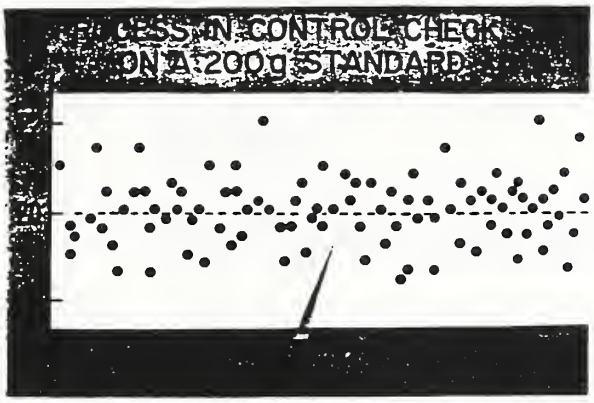
FOR THE DIRECTOR,
PAUL E. PONTIUS, CHIEF

TEST COMPLETED AUGUST 17, 1966 — MASS AND VOLUME SECTION
WASHINGTON, D.C. 20234

Any report of calibration or report of test must state a realistic uncertainty based on actual process performance. All of the pertinent data must be included so that the local processes can minimize the introduction of additional systematic errors. The random component of the uncertainty is a function of the measurement effort in the local process, reflecting the actual performance of that particular measurement process.

REPEATED CALIBRATION OF SAME OBJECT CLEARLY DEMONSTRATES THE PERFORMANCE OF A PARTICULAR MEASUREMENT PROCESS

There is no substitute for the evidence provided by the repeated calibration of the same object, over an extended time period, in demonstrating what the measurement process can do. These measurements should be independent repetitions, made under all the diversity of condition by which the method is affected so as to represent the set of conditions to which we wish our prediction to apply. The internally based precision estimate is applicable only to a narrower range of conditions, and it is only when the measurement conditions are highly reproducible that the two estimates of precision become equal.



The routine calibration of one of the laboratory's weights, used as check standard, tells us what the process can do—it is not just a simulation of the calibration process—it is the real thing—without the need for any assumptions. It provides the basis for the precision statement or gives us a check on any internally based statement. We can say to our clients: "If we calibrate your weight a large number of times the results would look like those on the chart. We did it only once so that your value is like one of these points. Which one, we cannot say but we are fairly certain that it is within the indicated uncertainty."

U.S. GOVERNMENT PRINTING OFFICE : 1967 OL—257-257

NBSIR 76-999

SURVEILLANCE TEST PROCEDURES

H. W. Almer

Edited by: Jerry Keller

Institute for Basic Standards
National Bureau of Standards
Washington, D.C. 20234

February 1976

Final

Issued May 1977

U.S. DEPARTMENT OF COMMERCE, Juanita M. Kreps, *Secretary*

Dr. Betsy Ancker-Johnson, *Assistant Secretary for Science and Technology*

NATIONAL BUREAU OF STANDARDS, Ernest Ambler, *Acting Director*

SURVEILLANCE TEST PROCEDURES

H. E. Almer

Abstract

Surveillance tests are designed to monitor the values of mass standards between calibrations. Two types are described; both consist of comparisons of the weights of an ordered set of mass standards with each other. The differences found are compared with those computed from the reported mass values. Surveillance limits based on the precision of both the calibration and the surveillance test processes are computed. These limits are estimates of the departure of the measured differences from the expected, or predicted, differences as computed from the reported values. A larger change is considered significant. Additional measurements to identify individual weights which have changed are required when a given comparison indicates that the mass of one or more of the weights involved has changed. Buoyancy corrections are used to correct for the difference in the buoyant effect on weights of differing densities. Records document the surveillance test results, and control charts help detect trends. Judgments concerning recalibration can be made based on the constancy of the weights relative to the use requirements.

Key words: Apparent mass; buoyancy; buoyancy correction; change; comparison; difference; mass; records; set; surveillance limits; surveillance test; test interval; true mass; value; weighing design; weights.

CONTENTS

1. INTRODUCTION

Surveillance test procedures are designed to monitor the values of mass standards between calibrations. This is important because the problem of the continuing validity of the values contained in the report of calibration is always present, and especially so for those who look to others for calibration service. Surveillance test procedures, if properly implemented, so provide a means of detecting gross changes as soon as possible with a minimum expenditure of time and effort.

Two types of surveillance tests are described. The first type, designated Type I, uses a minimum number of measurements that involve all of the weights in the set. The second type, designated Type II, requires a larger number of measurements which are grouped so that they are a series of 3-1's weighing designs.[1] This method has some redundancy.

Included in the surveillance test procedures are methods of identifying any weights whose mass values may have changed since they were calibrated, and methods of correcting for the buoyant effect of the atmosphere.

2. MEASUREMENT PROCEDURES

A surveillance test consists of a series of comparisons of the weights of an ordered set of mass standards with each other, according to an appropriate weighing design, and comparing the differences in mass value found by these comparisons with those computed from the values contained in the report of calibration [1]*. Ideally a suitable known weight, other than one of the weights in the set being tested, is used as the standard on which the values found by the surveillance test are based. This also establishes whether or not the whole set has changed proportionally. For sets where the largest weight is one kilogram or less, the nominal value of the weight used as a standard should be that of the largest weight in the set. For example, a set whose largest weight is 100g is being tested. For this set, a 100g weight whose mass value is known would be suitable for use as a standard. For sets having weights greater than one kilogram, a suitable one kilogram weight may be used as the standard. For sets in the avoirdupois system having weights greater than one pound, a suitable one pound weight may be used as the standard. Generally the uncertainty of the mass value of a one kilogram, or a one pound standard, is less than the uncertainty of the value of a larger standard.

---

[1] A title given to the three intercomparisons of three objects A, B, and C, namely the measurements of the differences A-B, A-C, and B-C.

* Figures in brackets refer to similarly numbered references at the end of this paper.

If a weight of the suggested denomination is not available, a suitable known weight of a different denomination, if available, may be used to establish whether or not the whole set has changed. The nominal value of this weight should be equal to that of one of the larger weights in the set being tested, say not less than 20g for a set beginning at 100g, or less than 100g for a set beginning at 1kg. Where the weight used as the standard has the same nominal value as the largest weight in the test set, up to one kilogram, the comparison between the standard and the largest weight of the set is a part of the surveillance test weighing design. Where the nominal value of the weight used to establish whether the whole set has changed is not the same as the largest weight of the test set, the comparison between it and the corresponding weight of the test set is a side measurement and not a part of the surveillance test weighing design.

Where a suitable known weight, other than the weights in the set being tested, is not available, the usual procedure is to base the values found by the surveillance test on the largest weight of the set under test, up to one kilogram. Weights larger than one kilogram may be based on the largest weight of the set. The weighings may be made by either the substitution or the transposition method of weighing [2].

In general, the capacities of the balances selected for surveillance tests should be the smallest available that will accommodate the maximum load to be placed on it. For example, when testing a set of weights ranging from 100g to 1mg, a balance having a capacity of from 100g to 200g would be used for loads from 100g to 20g, and a balance of 20g capacity for loads under 20g. If a balance of say 1g and 2g capacity were available, it would be used for the fractional weights.

2.1 Type I Surveillance Test

In a type I surveillance test, the first measurement is the comparison between the largest weight of the set and a summation of the next smaller weights, from the set, the sum of whose nominal values is equal to that of the largest weight. The next comparison would be between a selected weight from the summation, that is, the summation used in the first comparison, and another summation whose nominal value is equal to that of the selected weight.

This procedure of selecting a weight from each summation and comparing it with a summation of the next smaller weights is repeated until all of the weights of the set have been involved in a comparison. Any given comparison should involve the fewest weights that will permit all of the weights of the set to be included in the chain of comparisons.

If a suitable weight having the same nominal value as the largest weight of the set is available for use as a standard, then the first comparison would be between this weight and the largest weight of the set.

If, for example, a set of weights ranging from 100g to 1mg is to be tested using the Type I surveillance test procedures where another 100g weight is to be used as a standard, the ratios of the weights to each other are 5, 3, 2, 1. The first comparison would be:

$$100g - S100g = a_1$$

The second comparison would be:

$$100g - \Sigma 100g = a_2$$

where $\Sigma 100g = 50g + 30g + 20g$

The third comparison would be:

$$20g - \Sigma 20g = a_3$$

where $\Sigma 20g = 10g + 5g + 3g + 2g$

This procedure is continued until all of the weights have been compared.

In this example the last comparison would be:

$$3mg - \Sigma 3mg = a_n$$

where $\Sigma 3mg = 2mg + 1mg.$

The observed differences in mass values ($a_1, a_2, \ldots a_n$) found by these comparisons are compared with the accepted differences, as computed from the reported values, to determine the degree of agreement between the observed and the accepted differences. If the agreement is within the limits for surveillance (see section 3) any indicated changes may be regarded as being insignificant, and the continuing validity of the reported values may be assumed. If the agreement between the observed and the accepted differences is not within the surveillance limits, the indicated changes should be regarded as significant, and the weights exhibiting a significant change should be recalibrated. When the result of a comparison indicates that one or more of the weights has changed significantly, additional measurements are made to identify the weight, or weights, that have changed.

2.2 Type II Surveillance Test

In a Type II surveillance test, the measurements of the first 3-1's weighing design series are between the largest weight of the set, another weight of the same nominal value, and a summation of the next smaller weights from the set also having the same nominal value as the largest weight of the set. The comparisons of the next 3-1's weighing would be between a selected weight from the summation, used in the first 3-1's series, and two other summations, of the next smaller weights, whose nominal values are the same as that of the selected weight. This procedure of selecting a weight from a summation and comparing it with other summations of the next smaller weights according to the 3-1's weighing design is repeated until all of the weights of the set have been involved in the comparisons.

For example, a set ranging from 100g to 1mg is to be tested using the Type II surveillance test procedures, where another 100g weight[1] is to be used as a standard. The ratios of the weights to each other are 5, 3, 2, and 1. The first series according to the 3-1's weighing design would be:

$$S100g - 100g = a_1$$

$$S100g - \Sigma 100g = a_2$$

$$100g - \Sigma 100g = a_3$$

where S100g is the standard

100g is the 100g of the set being tested

$\Sigma 100g = 50g + 30g + 20g$

---

[1] If a suitable known 100g weight is not available for use as a standard, the first series according to the 3-1's weighing design would be:

$$100g - 100g' = a_1$$

$$100g - \Sigma 100g - a_2$$

$$100g' - \Sigma 100g = a_3$$

where

100g' is any 100g weight, or a summation whose nominal value is 100g, used to fill the series $\Sigma 100g = 50g + 30g + 20g$. The other series remain as indicated.

The second series would be:

$$30g - \Sigma 30g_1 = a_1$$
$$30g - \Sigma 30g_2 = a_2$$
$$\Sigma 30g_1 - \Sigma 30g_2 = a_3$$

where $\Sigma 30g_1 = 20g + 10g$

$\Sigma 30g_2 = 20g + 5g + 3g + 2g$

This procedure is continued for each decade until all of the weights in the set have been compared. Unless the set contains an extra 1mg weight or another 1mg weight whose mass value is known is available, the 3-1's weighing design cannot be used for the last decade. Where the set has only one 1mg weight and another is not available to fill the series, the comparisons for the last decade are:

$$5mg - 3mg - 2mg = a_1$$
$$3mg - 2mg - 1mg = a_2$$

These two comparisons are treated as the comparisons in Type I surveillance test. Where the set has two 1mg weights, or another 1mg weight whose value is known, is available, the last series is:

$$3mg - \Sigma 3mg_1 = a_1$$
$$3mg - \Sigma 3mg_2 = a_2$$
$$\Sigma 3mg_1 - \Sigma 3mg_2 = a_3$$

where $\Sigma 3mg_1 = 2mg + 1mg_1$

$\Sigma 3mg_2 = 2mg + 1mg_2$

The $1mg_2$ may be either the second 1mg weight of the set or another 1mg weight whose mass value is known.

If a weight other than one of the same denomination as the largest weight in the set is used to establish whether or not all the weights of the set have changed proportionately, then some other known weight must be compared to a weight of the set or (e.g. in this case) a known 30g is compared with the 30g of the set.

$$30g - S30g = a$$

If this difference agrees with the expected difference as computed from the reported values of the two weights, within the surveillance limit, (see section 3), and the observed differences of the other comparisons are in agreement with the predicted differences, it may be assumed that the set as a whole has not changed significantly.

Because in most of the series used in a Type II surveillance test one of the weights is part of both summations used in a given series, the weighings are made by the substitution method of weighing. For example, in the series involving the 30g weight, $\Sigma 30g_1$, and $\Sigma 30g_2$, the 20g weight is part of both summations.

3. SURVEILLANCE LIMITS

3.1 Uncertainties of Each of the Summations from the Calibration Process Known [1], [3]

Ideally the surveillance limits are calculated from the standard deviations of the calibration process and surveillance test process as follows:

$$sl = U_c + 3\sigma_d \qquad (1)$$

where U_c = uncertainty of calibration process

σ_d = standard deviation of one weighing of the surveillance test process

sl = surveillance limit

3.2 Uncertainties for Individuals but not Summations from the Calibration Process Known

Sometimes only the uncertainties associated with the mass values of the weights, as reported in the Calibration Report, are available for estimating the uncertainties of the summations. In this situation, an approximate estimate of the uncertainties is found by taking the square root of the sum of the squares of the uncertainties of the values of the weights in a given comparison [4].

Suppose that the comparison is between a selected weight, W_1, and a summation consisting of three weights, W_2, W_3 and W_4, whose nominal value is equal to that of the selected weight, W_1. The uncertainty of each value is U_1, U_2, U_3 and U_4 respectively.

An approximate estimate of the uncertainty, U_c, for these weights is:

$$U_c = \sqrt{U_1^2 + U_2^2 + U_3^2 + U_4^2} \qquad (2)$$

where U_c is the uncertainty of the calibration mass measurement process and

U_i is the uncertainty for the individual weights as reported on the Report of Calibration.

With this procedure, the expression for the surveillance limit is:

$$sl = U_c + 3\sigma_d \qquad (3)$$

where U_c is the uncertainty as defined above, and

sl and σ_d have the same meaning as in equation (1).

This process is equally applicable for any number of weights.

For most designs, this procedure gives a somewhat smaller uncertainty than the uncertainties from the calibration process.

3.2.1 Numerical Example

Assume that the following weights and their associated uncertainties are involved in the comparison 100g - Σ100g.

Weight	Uncertainty
100g	0.015
50g	0.011
30g	0.012
20g	0.010

$$U_c = \sqrt{0.015^2 + 0.011^2 + 0.012^2 + 0.010^2}$$

$$= \sqrt{0.00025 + 0.000121 + 0.000144 + 0.0001}$$

$$= \sqrt{0.00059}$$

$$= 0.024 \text{ mg}$$

This is an approximate estimate of the uncertainty of the calibration process for this comparison.

Now let us assume that the standard deviation of one weighing of the surveillance test process is 0.015 mg.

Then the surveillance limit, sl, is:

$$sl = 0.024 + 3(0.015)$$

$$sl = 0.024 + 0.045$$

$$= \underline{0.069 \text{ mg}}$$

4. Identifying the Weights Which Have Changed

If, in any comparison, the observed difference differs from the predicted value of the difference by more than the surveillance limits for that comparison, the weight, or weights, that have changed must be identified so that they can be recalibrated. The identity of the weights that have changed may be established by additional measurements. In general, these additional measurements are comparisons between the weights making up the summation that was compared with the selected weight.

Suppose, for example, that the observed difference of

$$20g - \Sigma 20g = a$$

where $\Sigma 20g = 10g + 5g + 3g + 2g$

differs from the predicted value of the difference by more than the surveillance limits. Assume, also, that the observed differences in the comparison in which the 20g weight was a part of the summation, $100g - \Sigma 100g$, and the comparison in which the 2g weight was the selected weight, $2g - \Sigma 2g$, are in good agreement with their predicted, or accepted, differences as computed from the reported values. This indicates that neither the 20g weight nor the 2g weight have changed significantly. The following measurements are made and their results analyzed to identify the weight, or weights, whose masses have changed:

$$10g - (5g + 3g + 2g) = a'$$

$$5g - (3g + 2g) = a''$$

$$3g - (2g + 1g) = a'''$$

4.1 Analysis of Measurement Results

If a' differs from the predicted value by more than the surveillance limits and a'' and a''' agree with the predicted value within the surveillance limits, it is probable that the 10g weight has changed. If both a' and a'' differ from the corresponding predicted values by more than the surveillance limits by about the same amount, numerically, but with opposite signs, and a''' agrees with the predicted value within the surveillance limit, it is probable that the 5g weight has changed. If a' and a'' differ from the corresponding predicted values by markedly difference amounts which are greater than the corresponding surveillance limits, and a''' agrees with the corresponding predicted value within the surveillance limit, it is probable that both the 10g and the 5g weights have changed.

If a', a'', and a''' all differ from the corresponding predicted values by more than the surveillance limits, but by about the same amount, it is probable that the 3g weight is the one that has changed. If a' and a'' differ from the corresponding predicted values by about the same amount, but a''' differs from the corresponding predicted value by a markedly different amount, it is probable that both the 5g weight and the 3g weights have changed.

If the results of all three measurements differ from the corresponding predicted values by more than the corresponding surveillance limits, by markedly different amounts, it is probable that all three weights have changed and may require recalibration.

If all three (a', a'', and a''') of the observed differences are in good agreement with the predicted differences, it is still possible that the weights involved in either of the comparisons

$$100g - \Sigma 100g = a_1 \quad \text{or} \quad 2g - \Sigma 2g = a_3$$

experienced compensating changes in mass, even though the agreement between the observed differences and the predicted differences were within the surveillance limits. However, this is an unlikely situation. But, if it does occur, the weights that have changed may be identified in the manner described for the comparison between the 20g and Σ20g weights, as may the weights involved in any measurements where the observed difference does not agree with the predicted difference within the surveillance limits.

In any event, if it is determined that several weights of a given set require recalibration (more than, say, three or four weights in a 100g to 1mg set, or more in a larger set) the entire set should be recalibrated.

4.2 Numerical Example

The following numerical example, using difference measurement 20g - Σ20g, discussed above, illustrates the procedure.

The observed value of the difference:

$$20g - \Sigma 20g = +0.084mg$$

The predicted value is +0.052mg. The surveillance limit is +0.028mg. The difference between the observed value and the predicted value is:

$$+0.084mg - 0.052mg = 0.032mg$$

This difference exceeds the surveillance limits and indicates that the mass of one or more of the weights involved has changed. Three weighings were made to determine which weight, or weights, have changed. The results of these measurements are:

	Observation	Observed Value of Difference	Predicted Value of Difference	Surveillance Limit
a'	10g - (5g + 3g + 2g) =	-0.025 mg	-0.057 mg	0.024 mg
a''	5g - (3g + 2g) =	-0.065 mg	-0.032 mg	0.018 mg
a'''	3g - (2g + 1g) =	+0.031 mg	+0.034 mg	0.015 mg

Examining these results, we find that the agreement between the observed value and the predicted value for a''' is well within the surveillance limit, thus virtually ruling out any change in the masses of the 3g and 2g weights. But, the observed values for a' and a'' do not agree with the predicted values within the surveillance limits. Further, the observed values for both a' and a'' differ from the predicted values by about the same amount, but with opposite signs.

For a' -0.025 - (-0.057) = +0.032 mg

For a'' -0.065 - (-0.032) = -0.033 mg

Had it been only for a' that the observed value did not agree with the predicted value, within the surveillance limit, it would be logical to conclude that the mass of the 10g weight had changed. But, for both a' and a'', the observed values of the differences do not agree with the predicted values by about the same amount, numerically, though with opposite signs. Therefore, the conclusion is that the mass of the 5g weight has changed because it is involved in both a' and a'', while the 10g weight is involved only in a'. Further, the 5g weight is in opposed positions in the two equations.

5. BUOYANCY CORRECTIONS

Buoyancy corrections are used to account for the difference in the buoyant effect of the air on weights of differing densities [5]. In some instances it will be necessary to apply buoyancy corrections to the measured differences between weights in surveillance tests because the buoyant effect on the weights may mask real changes in their masses, or apparent changes in mass may be indicated when there is no change. This is true whether the computations of the results are made on the true mass or the apparent mass basis. In general, the buoyancy corrections computed on the true mass basis are numerically greater than buoyancy corrections computed on the apparent mass basis when weights having widely different densities are involved in a given comparison.

It is always good practice to compute, at least roughly, the magnitude of the correction to establish the order of magnitude with reference to the uncertainty of the surveillance test measurement [1]. If the correction is not significant, it can be ignored.

5.1 Buoyancy Corrections Computed on True Mass Basis

When the results of the surveillance test weighings are computed on the true mass basis, the expected differences being computed from the reported mass (true mass) values, the true mass buoyancy correction term, $\rho\Delta V$, for the measured difference may be derived from the weighing equation for the difference between two weights.

$$(M_C - \rho V_C)g - (M_D - \rho V_D)g = ag \quad \text{weighing equation (1)}$$

where: M_C and M_D = the masses of weights C and D, respectively

V_C and V_D = the volumes of C and D, respectively, from the Report of Calibration

ρ = air density when weighing was made

a = the indicated difference in mass units

g = acceleration of gravity

The derivation of the buoyancy correction term, $\rho\Delta V$, for the true mass difference between the two masses C and D is:

$(M_C - \rho V_C)g - (M_D - \rho V_D)g = ag$	weighing equation	(1)
$M_C - \rho V_C - M_D + \rho V_D = a$	dividing by g	(2)
$M_C - M_D = a + \rho(V_C - V_D)$	transposing and collecting terms	(3)
$M_C - M_D = a + \rho\Delta V$	substituting ΔV for $(V_C - V_D)$	(4)

It is better to use the form of the buoyancy correction term, $\rho(V_C - V_D)$, in equation (3) above when computing the buoyancy correction because its sign is more readily apparent. The following example illustrates this.

The measured difference, a, between 2g and Σ2g is 0.0388mg.

Weight	Volume		
2 g		0.2564 cm^3	from Report of Calibration
1 g	0.12820 cm^3		" " " "
500 mg	0.03012 cm^3		" " " "
300 mg	0.01807 cm^3		" " " "
200 mg	0.01205 cm^3		" " " "
Σ2g		0.1884 cm^3	" " " "

ρ = 1.17 mg/cm^3

The true mass difference:

$$2g - \Sigma 2g = +0.0388 + 1.17(0.2564 - 0.1884)$$

$$= +0.0388 + 0.0796 = +0.1184 \text{ mg}$$

If volumes are not listed on the Report of Calibration, they may be computed from:

$$\text{Volume} = \frac{\text{Mass}}{\text{Density}}$$

5.2 Buoyancy Corrections Computed on Apparent Mass Basis

When the results of the weighings are computed on the apparent mass[1] basis [5], the expected differences being computed from the reported apparent mass values, the apparent mass buoyancy correction term, $\Delta\rho\Delta V$, for the measured differences may be derived from the expression for finding the apparent mass when the true mass and the volume are known.

$$AM_W = M_W - \rho_n(V_W - V_R) \tag{5}$$

where AM_W = apparent mass value of weight "W" versus the reference material (R)

M_W = mass (true mass) of weight "W"

ρ_n = density of normal air

V_W = volume of weight "W" at 20 °C

V_R = volume of equivalent mass of the reference material (R) at 20 °C

The derivation of the buoyancy correction term, $\Delta\rho\Delta V$, for the apparent mass difference between the weights C and D is:

$$AM_C = M_C - \rho_n(V_C - V_b) \tag{6}$$

$$AM_D = M_D - \rho_n(V_D - V_b) \tag{7}$$

---

[1] In the United States, the apparent mass is usually expressed as apparent mass versus normal brass in normal air. Normal brass is defined as brass having a density of 8.4 g/cm^3 at 0 °C and a coefficient of cubical expansion 0.000054 per degree C. Normal air is defined as air having a density of 1.2 mg/cm^3 at 20 °C.

$$AM_C - AM_D = M_C - \rho_n(V_C - V_b) - M_D + \rho_n(V_D - V_b) \quad \text{subtracting} \quad (8)$$

$$AM_C - AM_D = M_C - M_D - \rho_n V_C + \rho_n V_b + \rho_n V_D - \rho_n V_b$$

$$= a + \rho(V_C - V_D) - \rho_n(V_C - V_D)$$

substituting $a + \rho(V_C - V_D)$ for $(M_C - M_D)$ (see equation (3)) (9)

$$AM_C - AM_D = a + (\rho - \rho_n)(V_C - V_D) \quad \text{combining terms} \quad (10)$$

$$= a + \Delta\rho\Delta V \quad \text{substituting } \Delta\rho\Delta V \text{ for } (\rho - \rho_n)(V_C - V_D) \quad (11)$$

where AM_C and AM_D = the apparent mass of weights C and D

M_C and M_D = the masses of weights C and D

V_C and V_D = the volumes of C and D, respectively, from the Report of Calibration

V_b = the volume of equivalent mass of normal brass, the reference material

ρ = the air density when the weighing was made

ρ_n = the density of normal air at 20 °C.

It is better to use the form in equation (10) above when computing the buoyancy correction term because its sign is more readily apparent.

The following example illustrates this:

The measured difference, a, between 2g and Σ2g is 0.0388 mg.

Weight	Volume	
2 g		0.2564 cm^3 from Report of Calibration
1 g	0.12820 cm^3	" " " "
500 mg	0.03012 cm^3	" " " "
300 mg	0.01807 cm^3	" " " "
200 mg	0.01205 cm^3	" " " "
Σ2 g		0.1884 cm^3 " " " "

ρ = 1.17 mg/cm^3

ρ_n = 1.20 mg/cm^3

The apparent mass difference

$$2g - \Sigma 2g = + 0.0388 + (1.17 - 1.20)(0.2564 - 0.1884)$$

$$= + 0.0388 + (-0.03)(0.0680)$$

$$= + 0.0388 - 0.0020$$

$$= + 0.0368 \text{ mg}$$

5.3 Application of Buoyancy Correction

The buoyancy correction terms derived above are correct when the mass difference and the volume difference of the weights are taken in the same direction. That is, if the difference between the masses of weights C and D is taken as $M_C - M_D$ then their volume difference must be taken as $V_C - V_D$ or the buoyancy correction will have the wrong sign.

If, when assigning a mass value to one of the two weights being compared with each other, the other weight being used as the standard, a buoyancy correction is used, it is essential that the correct sign be used for the buoyancy correction term.

5.3.1 Buoyancy Correction Application for True Mass

Consider the relationship

$$C - D = a + \rho(V_C - V_D) \qquad (1)$$

If D is the standard then

$$C = a + \rho(V_C - V_D^{\prime}) + D \qquad (2)$$

substituting for a, ρ, V_C, and V_D and D their values, we get the true mass value of C, provided the true mass value of D was used.

If C, the first weight in the difference, C - D, is the standard (this is the situation in many weighing designs) then,

$$-D = a + \rho(V_C - V_D) - C$$

and

$$D = -a - \rho(V_C - V_D) + C \qquad (3)$$

Substituting for a, ρ, V_C, V_D and C their values, we get the true mass value of D, provided the true mass value of C was used.

Note that the sign of the buoyancy correction term in (3) above is minus. This application is illustrated on the computation sheet for the 3-1's weighing design.

5.3.2 Buoyancy Correction Application for Apparent Mass

Consider the relationship

$$C - D = a + (\rho - \rho_n)(V_C - V_D) \tag{4}$$

If D is the standard, then

$$C = a + (\rho - \rho_n)(V_C - V_D) + D \tag{5}$$

Substituting for a, ρ, ρ_n, V_C, V_D, and D their values, we get the apparent mass value of C, provided the apparent mass value of D was used.

If C, the first weight in the difference, C - D, is the standard (this is the situation in many weighing designs) then,

$$-D = a + (\rho - \rho_n)(V_C - V_D) - C$$

and

$$D = -a - (\rho - \rho_n)(V_C - V_D) + C \tag{6}$$

Substituting for a, ρ, ρ_n, V_C, V_D and C their values, we get the apparent mass value of D, provided the apparent mass value of C was used.

Note that the sign of the buoyancy correction term in (6) above is minus. This application is illustrated on the computation sheets for the 3-1's weighing design as used in the example for the Type II surveillance test (see appendix 2).

6. RECORDS

Records are an essential part of any measurement program. In a surveillance test program, adequate records are necessary to document the continuing validity of the reported mass values and to realize the full value of the program. Such records may be simple, or elaborate, as long as they contain the information needed to document the claimed validity of the mass values. A notebook or card file should be maintained containing a description of the test system. This should include a statement of the procedures, a list of standards (if any) and weighing instruments, test intervals, and a tabulation of the accumulated results of tests. The records should also include the identity of the weights, the expected, or predicted, values of the differences measured as computed from the reported values, and the surveillance limits. The calibration report should be an integral part of the records. In addition, where the Type II Surveillance Test is used, the estimate of the standard deviation should be compared for each 3-1's series, compared with the long term estimate of the standard deviation and recorded. This information, combined with the original data sheets, forms an adequate record. A large operation may require a more elaborate record keeping system.

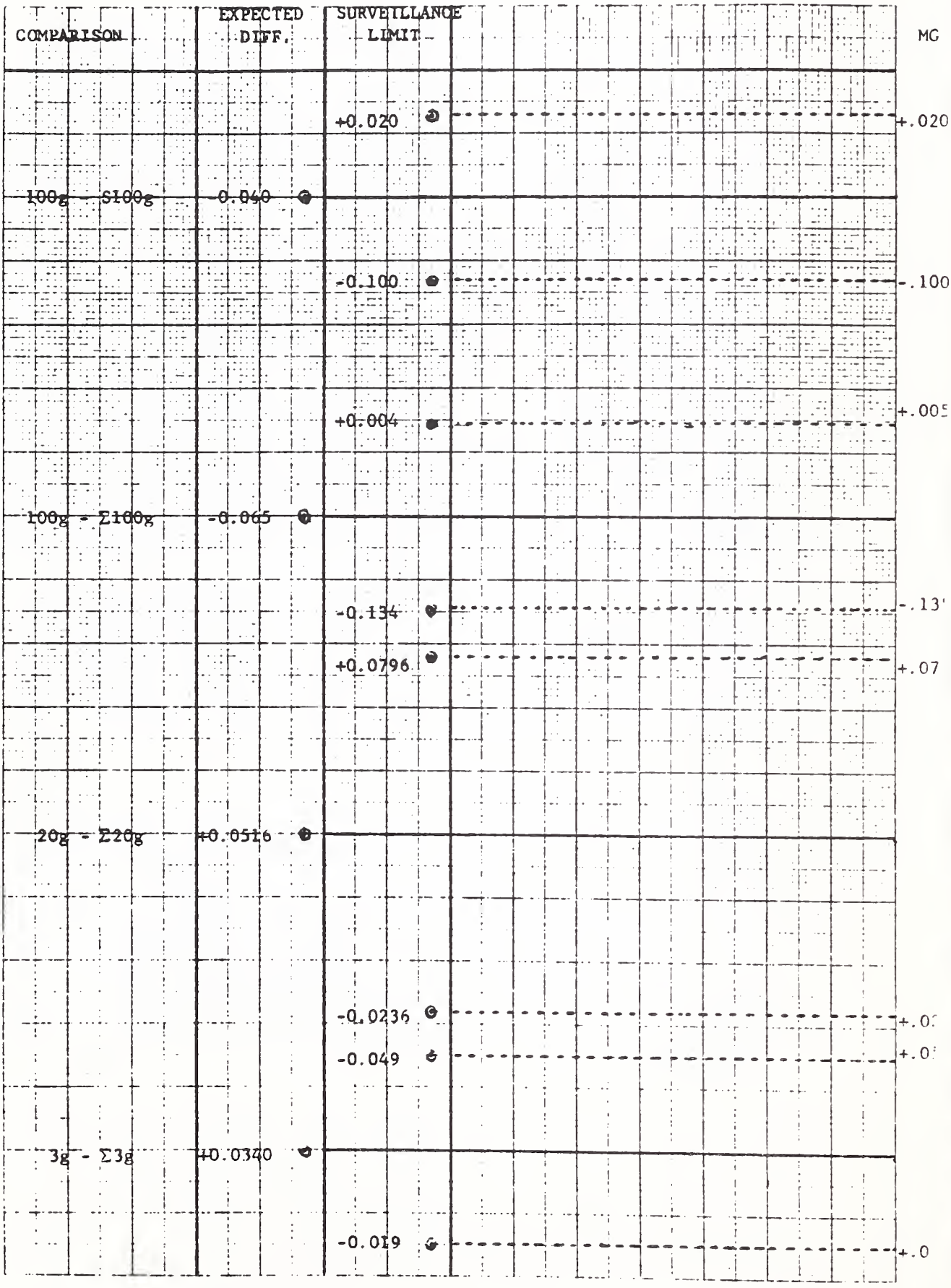
Control charts [3] similar to the one illustrated on page 18 are a useful addition to the surveillance test records. Control charts show more readily than tabulations whether a trend in the values of the differences being measured is developing. Such trends, when detected, can signal the need for recalibration before the values of the mass standards become invalid.

7. SURVEILLANCE TEST INTERVAL

The purpose of surveillance test procedures is to assure continuing validity of the values contained in the calibration report and to prevent, or at least minimize, the possibility of using the weights as standards when their reported values are no longer valid. But, when and how frequently should the surveillance test procedures be used in order to achieve this goal? Because of the many variables affecting the stability of the weights, such as the type of weights, the use to which the weights are put, the care they receive, etc., a categorical answer covering all situations cannot be given.

SURVEILLANCE TEST
CONTROL CHART

SET RANGE: 100g - 1g
CALIBRATION TEST NO. NBS 200390



The following suggestions, where they are applicable, may serve as general guide lines for the use of surveillance tests and the interval between surveillance tests.

1. Immediately upon the receipt of a newly calibrated set of weights, comparisons should be made to verify the values reported.

2. If this is a set for which no history exists, the comparisons should be repeated monthly, or bimonthly until the degree of stability of the weights has been demonstrated.

3. Where sufficient information about a set of weights has been developed to predict their performance with some degree of certainty, this information may be used in determining the interval between surveillance tests.

4. If there has been an accident with the weights, such as dropping them on the floor, at least the weights involved in the accident should be given a surveillance test before being used as standards to be sure that their reported values are still valid.

5. If a facility performs a large number of calibrations, its procedures should provide "built-in" checks on standards and if the standards checked on are part of the set in question, the information developed from these "built-in" checks can be used to determine when a surveillance test is needed.

6. Where the number of calibrations performed is small, the standards may be given a surveillance test just prior to using the standards in the calibration of other weights.

REFERENCES

[1] Pontius, P. E., Measurement Philosophy of the Pilot Program for Mass Calibration, Nat. Bur. Stand. (U.S.), Tech. Note 288, 39 pages (1968).

[2] Almer, H. E., Method of Calibrating Weights for Piston Gages, Nat. Bur. Stand. (U.S.), Tech. Note 577, 49 pages (1971).

[3] Pontius, P. E., Cameron, J. M., Realistic Uncertainties and The Mass Measurement Process, Nat. Bur. Stand. (U.S.), Monogr. 103, 17 pages (1967).

[4] Youden, W. J., Statistical Methods for Chemists, John Wiley & Sons, New York (1541).

[5] Nat. Bur. Stand. (U.S.), Handbook 77, Precision Measurement and Calibration, Volume III, Optics, Metrology and Radiation, Circular 3, pp. 671/53 to 683/65. This handbook is available for reference in most Government Depository Libraries throughout the United States.

APPENDIX I. WEIGHING DESIGNS FOR SURVEILLANCE TESTS

The weighing design used in a given surveillance test depends on the range of the set and ratio of the weights in the set to each other. Some suggested weighing designs for weight sets having 5, 3,2,1; 5,2,2,1 and 5,2,1,1,Σ1 ratios are shown for various ranges. Other designs may be developed by using the principles outlines in section 2 for situations where the suggested weighing designs do not apply. The surveillance test weighing designs are shown with metric units of mass. But with a given design, customary units of mass can be substituted for the metric units, provided the ratios of the weights to each other are the same in both systems.

Weighing Designs for Type I Surveillance Test

Design 1

Set - Range: 1kg to 1mg Ratio 5, 3, 2, 1

1kg - S1kg = a_1*

S1kg = Standard 1kg

1kg - Σ1kg = a_2

Σ1kg = 500g + 300g + 200g

200g - Σ200g = a_3

Σ200g = 100g + 50g + 30g + 20g

20g - Σ20g = a_4

Σ20g = 10g + 5g + 3g + 2g

2g - Σ2g = a_5

Σ2g = 1g + 500mg + 300mg + 200mg

200mg - Σ200mg = a_6

Σ200mg = 100mg + 50mg + 30mg + 20mg

20mg - Σ20mg = a_7

Σ20mg = 10mg + 5mg + 3mg + 2mg

3mg - Σ3mg = a_8

Σ3mg = 2mg + 1mg

* If a known 1kg weight suitable for use as a standard is not available, this "a" is omitted and 1kg - Σ1kg becomes the first "a", 200g - Σ200g = a_2, 20g - Σ20g = a_3, etc.

Design 2

Set - Range: 100g to 1mg Ratio: 5, 3, 2, 1

$$100g - S100g = a_1^*$$
$$S100g = \text{Standard 100g Weight}$$

$$100g - \Sigma 100g = a_2$$
$$\Sigma 100g = 50g + 30g + 20g$$

$$20g - \Sigma 20g = a_3$$
$$\Sigma 20g = 10g + 5g + 3g + 2g$$

$$2g - \Sigma 2g = a_4$$
$$\Sigma 2g = 1g + 500mg + 300mg + 200mg$$

$$200mg - \Sigma 200mg = a_5$$
$$\Sigma 200mg = 100mg + 50mg + 30mg + 20mg$$

$$20mg - \Sigma 20mg = a_6$$
$$\Sigma 20mg = 10mg + 5mg + 3mg + 2mg$$

$$3mg - \Sigma 3mg = a_7$$
$$3mg = 2mg + 1mg$$

For sets in which the smallest weight is 1g, the last "a" would be:

$$3g - \Sigma 3g = a$$
$$\Sigma 3g = 2g + 1g$$

* If a known 100g weight suitable for use as a standard is not available, this "a" is omitted and 100g - Σ100g becomes the first "a", 20g - Σ20g = a_2, etc.

Design 3

Set - Range: 1kg to 1mg Ratio: 5, 2, 2, 1

$$1kg - S1kg = a_1^*$$
$$S1kg = \text{Standard 1kg Weight}$$

$$1kg - \Sigma 1kg = a_2$$
$$\Sigma 1kg = 500g + 200g_1 + 200g_2 + 100g$$

$$100g - \Sigma 100g = a_3$$
$$\Sigma 100g = 50g + 20g_1 + 20g_2 + 10g$$

$$10g - \Sigma 10g = a_4$$
$$\Sigma 10g = 5g + 2g_1 + 2g_2 + 1g$$

$$1g - \Sigma 1g = a_5$$
$$\Sigma 1g = 500mg + 200mg_1 + 200mg_2 + 100mg$$

$$100mg - \Sigma 100mg = a_6$$
$$\Sigma 100mg = 50mg + 20mg_1 + 20mg_2 + 10mg$$

$$10mg - \Sigma 10mg = a_7$$
$$\Sigma 10mg = 5mg + 2mg_1 + 2mg_2 + 1mg$$

* If a known 1kg weight suitable for use as a standard is not available, this "a" is omitted and 1kg - Σ1kg becomes the first "a", and 100g - Σ100g = a_2, 10g - Σ10g = a_3, etc.

Design 4

Set - Range: 100g - 1mg Ratio: 5, 2, 2, 1

$$100g - S100g = a_1^*$$
$$S100g = \text{Standard 100g Weight}$$

$$100g - \Sigma 100g = a_2$$
$$\Sigma 100g = 50g + 20g_1 + 20g_2 + 10g$$

$$10g - \Sigma 10g = a_3$$
$$\Sigma 10g = 5g + 2g_1 + 2g_2 + 1g$$

$$1g - \Sigma 1g = a_4$$
$$\Sigma 1g = 500mg + 200mg_1 + 200mg_2 + 100mg$$

$$100mg - \Sigma 100mg = a_5$$
$$\Sigma 100mg = 50mg + 20mg_1 + 20mg_2 + 10mg$$

$$10mg - \Sigma 10mg = a_6$$
$$\Sigma 10mg = 5mg + 2mg_1 + 2mg_2 + 1mg$$

For the set in which the smallest weight is 1g, the last "a" would be:

$$5g - \Sigma 5g = a$$
$$\Sigma 5g = 2g_1 + 2g_2 + 1g$$

* If a known 100g weight suitable for use as a standard is not available, this "a" is omitted and $100g - \Sigma 100g$ becomes the first "a", and $10g - \Sigma 10g = a_2$, etc.

Design 5

Set - Range: 30kg - 1mg Ratio: 5, 3, 2, 1

$30kg - \Sigma 30kg = a_1$

$\Sigma 30kg = 20kg + 10kg$

$10kg - \Sigma 10kg = a_2$

$\Sigma 10kg = 5kg + 3kg + 2kg$

$2kg - \Sigma 2kg = a_3$

$\Sigma 2kg = 1kg_1 + 1kg_2$

$1kg_1 - \Sigma 1kg = a_4$

$\Sigma 1kg = 500g + 300g + 200g$

$200g - \Sigma 200g = a_5$

$\Sigma 200g = 100g + 50g + 30g + 20g$

$20g - \Sigma 20g = a_6$

$\Sigma 20g = 10g + 5g + 3g + 2g$

$2g - \Sigma 2g = a_7$

$\Sigma 2g = 1g + 500mg + 300mg + 200mg$

$200mg - \Sigma 200mg = a_8$

$\Sigma 200mg = 100mg + 50mg + 30mg + 20mg$

$20mg - \Sigma 20mg = a_9$

$\Sigma 20mg = 10mg + 5mg + 3mg + 2mg$

$2mg - \Sigma 2mg = a_{10}$

$\Sigma 2mg = 1mg_1 - 1mg_2$

Weighing Designs for Type I Surveillance Tests

Design 6

Set - Range: 100g - 1mg Ratio: 5, 2, 1, 1, Σ1

$100g - S100g = a_1$*

$S100g$ = Standard 100g weight

$100g - \Sigma 100g = a_2$

$\Sigma 100g = 50g + 20g + 10g_1 + 10g_2 + \Sigma 10g$

$10g_1 - \Sigma 10g = a_3$

$\Sigma 10g = 5g + 2g + 1g_1 + 1g_2 + \Sigma 1g$

$1g_1 - \Sigma 1g = a_4$

$\Sigma 1g = 500mg + 200mg + 100mg_1 + 100mg_2 + \Sigma 100mg$

$100mg_1 - \Sigma 100mg = a_5$

$\Sigma 100mg = 50mg + 20mg + 10mg_1 + 10mg_2 + \Sigma 10mg$

$10mg_1 - \Sigma 10mg = a_6$

$\Sigma 10mg = 5mg + 2mg + 1mg_1 + 1mg_2 + 1mg_3$

When comparing the unit of weight of a given decade of weights with summation of smaller weights from a weight set in which the ratio of the weights to each other 5, 2, 1, 1, 1, it is necessary to include all of the set's weights smaller than the unit weight to which summation is being compared. For example: in the comparison 100g - Σ100g, the Σ100g includes all of the weights in the set smaller than 100g; and in the comparison 10g - Σ10g the Σ10g includes all of the weights smaller than 10g and so on.

* If a known 100g weight suitable for use as a standard is not available, this "a" is omitted and 100g - Σ100g becomes the first "a" and 10g - Σ10g = a_2, etc.

Weighing Designs for Type II Surveillance Tests

Design 7

Set - Range: 1kg to 1mg Ratio: 5, 3, 2, 1

Series 1

$$S1kg - 1kg = a_1^*$$
$$S1kg - \Sigma 1kg = a_2$$
$$1kg - \Sigma 1kg = a_3$$

S1kg = Standard 1kg

$\Sigma 1kg = 500g + 300g + 200g$

Series 2

$$300g - \Sigma 300g_1 = a_1$$
$$300g - \Sigma 300g_2 = a_2$$
$$\Sigma 300g_1 - \Sigma 300g_2 = a_3$$

$\Sigma 300g_1 = 200g + 100g$

$\Sigma 300g_2 = 200g + 50g + 30g + 20g$

Series 3

$$30g - \Sigma 30g_1 = a_1$$
$$30g - \Sigma 30g_2 = a_2$$
$$\Sigma 30g_1 - \Sigma 30g_2 = a_3$$

$\Sigma 30g_1 = 20g + 10g$

$\Sigma 30g_2 = 20g + 5g + 3g + 2g$

Series 4

$$3g - \Sigma 3g_1 = a_1$$
$$3g - \Sigma 3g_2 = a_2$$
$$\Sigma 3g_1 - \Sigma 3g_2 = a_3$$

$\Sigma 3g_1 = 2g + 1g$

$\Sigma 3g_2 = 2g + 500mg + 300mg + 200mg$

Series 5

$$300mg - \Sigma 300mg_1 = a_1$$
$$300mg - \Sigma 300mg_2 = a_2$$
$$\Sigma 300mg_1 - \Sigma 300mg_2 = a_3$$

$$\Sigma 300mg_1 = 200mg + 100mg$$
$$\Sigma 300mg_2 = 200mg + 50mg + 30mg + 20mg$$

Series 6

$$30mg - \Sigma 30mg_1 = a_1$$
$$30mg - \Sigma 30mg_2 = a_2$$
$$\Sigma 30mg_1 - \Sigma 30mg_2 = a_3$$

$$\Sigma 30mg_1 = 20mg + 10mg$$
$$\Sigma 30mg_2 = 20mg + 5mg + 3mg + 2mg$$

Series 7

$$3mg - \Sigma 3mg_1 = a_1$$
$$3mg - \Sigma 3mg_2 = a_2$$
$$\Sigma 3mg_1 - \Sigma 3mg_2 = a_3$$

$$\Sigma 3mg_1 = 2mg + 1mg_1$$
$$\Sigma 3mg_2 = 2mg + 1mg_2**$$

* If a known 1kg weight, suitable for use as a standard is not available, any 1kg or Σ1kg may be used to fill the series. Then the 1kg of the set is used as the standard and the first series of measurements is:

$$1kg - 1kg' = a_1$$
$$1kg - \Sigma 1kg = a_2$$
$$1kg' - \Sigma 1kg = a_3$$

where 1kg' is either the 1kg weight or the Σ1kg used to complete the series.

** The $1mg_2$ is an extra 1mg weight used to fill the last series.

Design 8

Series - Range: 100g to 1mg Ratio: 5, 3, 2, 1

Series 1

$S100g - 100g = a_1$*

$S100g - \Sigma 100g = a_2$

$100g - \Sigma 100g = a_3$

$\Sigma 100g = 50g + 30g + 20g$

Series 2

$30g - \Sigma 30g_1 = a_1$

$30g - \Sigma 30g_2 = a_2$

$\Sigma 30g_1 - \Sigma 30g_2 = a_3$

$\Sigma 30g_1 = 20g + 10g$

$\Sigma 30g_2 = 20g + 5g + 3g + 2g$

Series 3

$3g - \Sigma 3g_1 = a_1$

$3g - \Sigma 3g_2 = a_2$

$\Sigma 3g_1 - \Sigma 3g_2 = a_3$

$\Sigma 3g_1 = 2g + 1g$

$\Sigma 3g_2 = 2g + 500mg + 300mg + 200mg$

Series 4

$300mg - \Sigma 300mg_1 = a_1$

$300mg - \Sigma 300mg_2 = a_2$

$\Sigma 300mg_1 - \Sigma 300mg_2 = a_3$

$\Sigma 300mg_1 = 200mg + 100mg$

$\Sigma 300mg_2 = 200mg + 50mg + 30mg + 20mg$

Series 5

$$30mg - \Sigma 30mg_1 = a_1$$
$$30mg - \Sigma 30mg_2 = a_2$$
$$\Sigma 30mg_1 - \Sigma 30mg_2 = a_3$$

$$\Sigma 30mg_1 = 20mg + 10mg$$
$$\Sigma 30mg_2 = 20mg + 5mg + 3mg + 2mg$$

Series 6

$$3mg - \Sigma 3mg_1 = a_1$$
$$3mg - \Sigma 3mg_2 = a_2$$
$$\Sigma 3mg_1 - \Sigma 3mg_2 = a_3$$

$$\Sigma 3mg_1 = 2mg + 1mg_1$$
$$\Sigma 3mg_2 = 2mg + 1mg_2**$$

* If a known 100g weight suitable for use as a standard is not available, any 100g weight or Σ100g weight may be used to fill the first series. Then the 100g of the set is used as the standard and the first series of measurement is:

$$100g - 100g' = a_1$$
$$100g - \Sigma 100g = a_2$$
$$100g' - \Sigma 100g = a_3$$

where 100g' is either the 100g or the Σ100g used to complete the series.

** The $1mg_2$ is an extra 1mg weight used to fill the last series.

Design 9

Set - Range: 1kg to 1mg Ratio: 5, 2, 2, 1

Series 1

$$S1kg - 1kg = a_1{}^*$$
$$S1kg - \Sigma 1kg = a_2$$
$$1kg - \Sigma 1kg = a_3$$

$$\Sigma 1kg = 500g + 200g_1 + 200g_2 + 100g$$

Series 2

$$200g_1 - 200g_2 = a_1$$
$$200g_1 - \Sigma 200g = a_2$$
$$200g_2 - \Sigma 200g = a_3$$

$$\Sigma 200g = 100g + 50g + 20g_1 + 20g_2 + 10g$$

Series 3

$$20g_1 - 20g_2 = a_1$$
$$20g_1 - \Sigma 20g = a_2$$
$$20g_2 - \Sigma 20g = a_3$$

$$\Sigma 20g = 10g + 5g + 2g_1 + 2g_2 + 1g$$

Series 4

$$2g_1 - 2g_2 = a_1$$
$$2g_1 - \Sigma 2g = a_2$$
$$2g_2 - \Sigma 2g = a_3$$

$$\Sigma 2g = 1g + 500mg + 200mg_1 + 200mg_2 + 100mg$$

Series 5

$$200mg_1 - 200mg_2 = a_1$$
$$200mg_1 - \Sigma 200mg = a_2$$
$$200mg_2 - \Sigma 200mg = a_3$$

$$\Sigma 200mg = 100mg + 50mg + 20mg_1 + 20mg_2 + 10mg$$

Series 6

$$20mg_1 - 20mg_2 = a_1$$
$$20mg_1 - \Sigma 20mg = a_2$$
$$20mg_2 - \Sigma 20mg = a_3$$

$$\Sigma 20mg = 10mg + 5mg_1 + 2mg_1 + 2mg_2 + 1mg$$

Series 7

$$2mg_1 - 2mg_2 = a_1$$
$$2mg_1 - \Sigma 2mg = a_2$$
$$2mg_2 - \Sigma 2mg = a_3$$
$$\Sigma 2mg = 1mg_1 + 1mg_2 **$$

* If a known 1kg weight suitable for use as a standard is not available, any 1kg weight or Σ1kg weight may be used to fill the series. Then the 1kg of the set is used as the standard and the first series of measurements is:

$$1kg - 1kg' = a_1$$
$$1kg - \Sigma 1kg = a_2$$
$$1kg' - \Sigma 1kg = a_3$$

where 1kg' is either the 1kg weight or the Σ1kg used to complete the series.

** The $1mg_2$ is an extra 1mg weight used to fill the last series.

Design 10

Set - Range: 30kg to 1mg Ratio: 5, 3, 2, 1

Series 1

$$30kg - \Sigma 30kg_1 = a_1$$
$$30kg - \Sigma 30kg_2 = a_2$$
$$\Sigma 30kg_1 - \Sigma 30kg_2 = a_3$$

$$\Sigma 30kg_1 = 20kg + 10kg$$
$$\Sigma 30kg_2 = 20kg + 5kg + 3kg + 2kg$$

Series 2

$$3kg - \Sigma 3kg_1 = a_1$$
$$3kg - \Sigma 3kg_2 = a_2$$
$$\Sigma 3kg_1 - \Sigma 3kg_2 = a_3$$

$$\Sigma 3kg_1 = 2kg + 1kg.$$
$$\Sigma 3kg_2 = 2kg + 1kg..$$

Series 3

$$1kg. - 1kg.. = a_1$$
$$1kg. - \Sigma 1kg = a_2$$
$$1kg.. - \Sigma 1kg = a_3$$

$$\Sigma 1kg = 500g + 300g + 200g$$

Series 4

$$300g - \Sigma 300g_1 = a_1$$
$$300g - \Sigma 300g_2 = a_2$$
$$\Sigma 300g_1 - \Sigma 300g_2 = a_3$$

$$\Sigma 300g_1 = 200g + 100g$$
$$\Sigma 300g_2 = 200g + 50g + 30g + 20g$$

Series 5

$$30g - \Sigma 30g_1 = a_1$$
$$30g - \Sigma 30g_2 = a_2$$
$$\Sigma 30g_1 - \Sigma 30g_2 = a_3$$

$$\Sigma 30g_1 = 20g + 10g$$
$$\Sigma 30g_2 = 20g + 5g + 3g + 2g$$

Series 6

$$3g - \Sigma 3g_1 = a_1$$
$$3g - \Sigma 3g_2 = a_2$$
$$\Sigma 3g_1 - \Sigma 3g_2 = a_3$$

$$\Sigma 3g_1 = 2g + 1g$$
$$\Sigma 3g_2 = 2g + 500mg + 300mg + 200mg$$

Series 7

$$300mg - \Sigma 300mg_1 = a_1$$
$$300mg - \Sigma 300mg_2 = a_2$$
$$\Sigma 300mg_1 - \Sigma 300mg_2 = a_3$$

$$\Sigma 300mg_1 = 200mg + 100mg$$
$$\Sigma 300mg_2 = 200mg + 50mg + 30mg + 20mg$$

Series 8

$$30mg - \Sigma 30mg_1 = a_1$$
$$30mg - \Sigma 30mg_2 = a_2$$
$$\Sigma 30mg_1 - \Sigma 30mg_2 = a_3$$

$$\Sigma 30mg_1 = 20mg + 10mg$$
$$\Sigma 30mg_2 = 20mg + 5mg + 3mg + 2mg$$

Series 9

$$3mg - \Sigma 3mg_1 = a_1$$
$$3mg - \Sigma 3mg_2 = a_2$$
$$\Sigma 3mg_1 - \Sigma 3mg_2 = a_3$$

$$\Sigma 3mg_1 = 2mg + 1mg_1$$
$$\Sigma 3mg_2 = 2mg + 1mg_2$$

Appendix 2. Surveillance Test Examples

Type I Surveillance Test Example

Example of a Type I surveillance test for a set of metric mass standards according to design 2, appendix 1. This set was calibrated by the National Bureau of Standards. The National Bureau of Standards Report of Calibration Test No. 200390 is reproduced on pages 45–48. The standards used (other than the set) are listed below together with their apparent mass corrections and uncertainties. The balances used and their standard deviation for one double substitution weighing are also listed. The double substitution weighing method is used. The standard deviation of the calibration mass measurement process is not known, so an estimate is computed from the reported uncertainties, as described in section 3.2, and used in computing the surveillance limits.

Standards	Apparent Mass Corr. (mg)	Volume (cm^3)	Uncertainty (mg)
S100g	- 0.019	12.822	0.015
h 10mg	+ 0.0450	0.0037	0.0006
h 5mg	+ 0.0045	0.0018	0.0005

Balance Laboratory Designation	Standard Deviation (mg)	Capacity (g)
H - 200	0.015	200g
M - 10	0.003	20g

Computation of Surveillance Limits (sl)

In the following, "U_s" will denote the uncertainty of the standard and "S.D." the standard deviation of the process:

For 100g - S100g

$U_s = 0.015mg$

S.D. = 0.015mg

sl = 0.015 +3(0.015) = <u>0.060mg</u>

In the following, "U_c" will denote the uncertainty (see Equation (2) Page 7, and "S.D." the standard deviation of the process:

For 100g - Σ100g

$$U_c = \sqrt{.015^2 + .011^2 + .012^2 + .010^2}$$

$$= \sqrt{.000225 + .000121 + .000144 + .0001}$$

$$= \sqrt{.00059}$$

$U_c = 0.024mg$

S.D. = 0.015mg

sl = 0.024 + 3(0.015) = <u>0.069mg</u>

For 20g - Σ20g

$$U_c = \sqrt{.010^2 + .013^2 + .007^2 + .004^2 + .003^2}$$

$$= \sqrt{.0001 + .000169 + .000049 + .000016 + .000009}$$

$$= \sqrt{.000343}$$

$U_c = 0.019mg$

S.D. = 0.003mg

sl = 0.019 + 3(0.003) = <u>0.028mg</u>

For $2g - \Sigma 2g$

$$U_c = \sqrt{.0032^2 + .0030^2 + .0016^2 + .0011^2 + .0008^2}$$

$$= \sqrt{.00001 + .000009 + .00000256 + .0000012 + .00000064}$$

$$= \sqrt{.00002341}$$

$$U_c = 0.0048$$

$$S.D. = 0.003mg$$

$$sl = 0.0048 + 3(0.003) = \underline{0.014mg}$$

For $200mg - \Sigma 200mg$

$$U_c = \sqrt{.0008^2 + .0008^2 + .0005^2 + .0005^2 + .0005^2}$$

$$= \sqrt{.00000064 + .00000064 + .00000025 + .00000025 + .00000025}$$

$$= \sqrt{.00000203}$$

$$U_c = 0.0014mg$$

$$S.D. = 0.003mg$$

$$sl = 0.0014 + 3(0.003) = \underline{0.010mg}$$

For $20mg - \Sigma 20mg$

$$U_c = \sqrt{.00042^2 + .00059^2 + .00049^2 + .00052^2 + .00045^2}$$

$$= \sqrt{.000000176 + .000000348 + .000000240 + .000000270 + .00000020}$$

$$= \sqrt{.000001234}$$

$$U_c = 0.0011mg$$

$$S.D. = 0.003mg$$

$$sl = 0.0011 + 3(0.003) = \underline{0.010mg}$$

For 3mg - Σ3mg

$$U_c = \sqrt{.00052^2 + .00045^2 + .00059^2}$$

$$= \sqrt{.000000270 + .000000202 + .000000349}$$

$$= \sqrt{.00000082}$$

$$U_c = 0.00091\text{mg}$$

$$S.D. = 0.003\text{mg}$$

$$sl = 0.00091 + 3(0.003) = \underline{0.0099\text{mg}}$$

For the weighings made on the smaller balance, the uncertainty of the values of the weights is small compared to the standard deviation of that balance. Therefore, for all practical purposes, three times the standard deviation of the balance may be taken as the surveillance limit for these weighings.

Computation of Buoyancy Corrections

The buoyancy corrections, $\Delta\rho\Delta V$, are computed according to the procedure set forth in section 5.2, using the formula:

$$\text{buoyancy correction} = (\rho - \rho_n)(V_C - V_D)$$

Consider the weighing 100g - S100g = a

where $\rho = 1.17 mg/cm^3$ air density at time of weighing

$\rho_n = 1.20 mg/cm^3$ density of normal air

$V_C = 12.821 cm^3$ volume of 100g weight of set under test, from Report of Calibration

$V_D = 12.822 cm^3$ volume of S100g standard

$$\text{buoyancy correction} = (1.17 - 1.20)(12.821 - 12.822)$$
$$= (-0.03)(-0.001)$$
$$= +0.00003 mg$$

This amount is insignificant compared to the surveillance limit of 0.060mg and may be ignored. Similarly, the differences in the volumes in the weighings 100g - Σ100g and 20g - Σ20g are small enough so that the buoyancy corrections are negligible. But, in the weighings 2g - Σ2g and 200mg - Σ200mg weights of differing densities are involved. Consequently, the volumes of the individual weight and the summation of weights are different.

The volumes are:

For the weighing 2g - Σ2g:

Weights		Volumes
2g		$0.2564 cm^3$
1g	$0.1282 cm^3$	
500mg	$0.0301 cm^3$	
300mg	$0.0181 cm^3$	
200mg	$0.0120 cm^3$	
Σ2g		$0.1884 cm^3$

The actual air density, ρ, is the same as for 100g - Σ100g.

$$\text{buoyancy correction} = (1.17 - 1.20)(0.2564 - 0.1884)$$
$$= (-0.03)(+.0680)$$
$$= -0.0020 mg$$

This buoyancy correction, while relatively small compared to the surveillance limit, is not insignificant and must be applied to the measured difference between 2g and Σ2g.

For the weighing 200mg - Σ200mg:

Weights		Volumes
200mg		$0.01205 cm^3$
	100mg	$0.00602 cm^3$
	50mg	$0.00301 cm^3$
	30mg	$0.01111 cm^3$
	20mg	$0.00741 cm^3$
Σ200mg		$0.02755 cm^3$

$$\text{buoyancy correction} = (1.17 - 1.20)(0.01205 - 0.02755)$$
$$= (-0.03)(+0.01550)$$
$$= -0.00046 mg$$

This buoyancy correction is small compared to the surveillance limits for this comparison and in most cases can be ignored.

For the weighing 20mg - Σ20mg:

Weights		Volumes
20mg		$0.00741 cm^3$
	10mg	$0.00371 cm^3$
	5mg	$0.00185 cm^3$
	3mg	$0.00111 cm^3$
	2mg	$0.00074 cm^3$
Σ20mg		$0.00741 cm^3$

The volumes of the two masses are equal, therefore the buoyancy correction is zero.

For the weighing 3mg - Σ3mg:

Weights		Volumes
3mg		$0.00111 cm^3$
	2mg	$0.00074 cm^3$
	1mg	$0.00037 cm^3$
Σ3mg		$0.00111 cm^3$

The volumes of the two masses are equal, therefore the buoyancy correction is zero.

Computation of Predicted or Expected Differences

The expected differences are computed from the reported values as follows: (see report on page 47).

For the weighing 100g - S100g:

Weights	Values	
	100g	S100g
100g	-0.058mg	
S100g		-0.019mg
Sums	-0.058mg	-0.019mg
Expected Diff. =	-0.039mg	

For the weighing 100g - Σ100g:

Weights	Values	
	100g	Σ100g
100g	-0.0589mg	
50g		-0.0133mg
30g		-0.0134mg
20g		+0.0330mg
Sums	-0.0589mg	+0.0063mg
Expected Diff. =	-0.065 mg	

For the weighing 20g - Σ20g:

Weights	Values	
	20g	Σ20g
20g	+0.0330mg	
10g		-0.0378mg
5g		-0.0065mg
3g		+0.0191mg
2g		+0.0066mg
Sums	+0.0330mg	-0.0186mg
Expected Diff. =	+0.0516mg	

For the weighing 2g - Σ2g:

Weights	Values	
	2g	Σ2g
2g	+0.0066mg	
1g		-0.0216mg
500mg		-0.0005mg
300mg		-0.0041mg
200mg		-0.0049mg
Sums	+0.0066mg	-0.0311mg
Expected Diff. =	+0.0377mg	

For the weighing 200mg - Σ200mg:

Weights	Values	
	200mg	Σ200mg
200mg	-0.0049mg	
100mg		+0.0008mg
50mg		+0.0074mg
30mg		-0.0049mg
20mg		-0.0020mg
Sums	-0.0049mg	+0.0013mg
Expected Diff. =	-0.0062mg	

For the weighing 20mg - Σ20mg:

Weights	Values	
	20mg	Σ20mg
20mg	-0.0020mg	
10mg		+0.0028mg
5mg		+0.0065mg
3mg		+0.0030mg
2mg		-0.0142mg
Sums	-0.0020mg	-0.0019mg
Expected Diff. =	-0.0001mg	

For the weighing 3mg - Σ3mg:

Weights	Values	
	3mg	Σ3mg
3mg	+0.0030mg	
2mg		-0.0142mg
1mg		+0.0091mg
Sums	+0.0030mg	-0.0051mg
Expected Diff. =	+0.0081mg	

COMPANY X
NEW YORK, NEW YORK
SET OF MASS STANDARDS
TEST NUMBER

SUMMARY

FOR CONVENIENCE, THE RESULTS OF THIS WORK ARE SUMMARIZED IN TABLES I and II. THE VALUES ASSIGNED ARE WITH REFERENCE TO THE STANDARDS IDENTIFIED ON THE DATA SHEETS. THE UNCERTAINTY FIGURE IS AN EXPRESSION OF THE OVERALL UNCERTAINTY USING THREE STANDARD DEVIATIONS AS A LIMIT TO THE EFFECT OF THE RANDOM ERRORS OF THE MEASUREMENT ASSOCIATED WITH THE MEASUREMENT PROCESSES. THE MAGNITUDE OF SYSTEMATIC ERRORS FROM SOURCES OTHER THAN THE USE OF ACCEPTED VALUES FOR CERTAIN STARTING STANDARDS ARE CONSIDERED NEGLIGIBLE. IT SHOULD BE NOTED THAT THE MAGNITUDE OF THE UNCERTAINTY REFLECTS THE PERFORMANCE OF THE MEASUREMENT PROCESS USED TO ESTABLISH THESE VALUES. THE MASS UNIT, AS REALIZABLE IN ANOTHER MEASUREMENT PROCESS, WILL BE UNCERTAIN BY AN AMOUNT WHICH IS A COMBINATION OF THE UNCERTAINTY OF THIS PROCESS AND THE PROCESS IN WHICH THESE STANDARDS ARE USED.

THE ESTIMATED MASS VALUES LISTED IN TABLE I ARE BASED ON AN EXPLICIT TREATMENT OF DISPLACEMENT VOLUMES, E.G., "TRUE MASS", "MASS IN VACUO", MASS IN THE NEWTONIAN SENSE. THE DISPLACEMENT VOLUME ASSOCIATED WITH EACH VALUE IS LISTED AS WELL AS THE VOLUMETRIC COEFFICIENT OF EXPANSION. THESE VALUES SHOULD BE USED, TOGETHER WITH APPROPRIATE CORRECTION FOR THE BUOYANT EFFECTS OF THE ENVIRONMENT, TO ESTABLISH CONSISTENT MASS VALUES FOR OBJECTS WHICH DIFFER SIGNIFICANTLY IN DENSITY AND/OR FOR MEASUREMENTS WHICH MUST BE MADE TO DIFFERING ENVIRONMENTS. THE RELATION 1 LB AVDP = .45359237KG IS USED AS REQUIRED.

THE ESTIMATED MASS VALUES LISTED IN TABLE II ARE BASED ON AN IMPLICIT TREATMENT OF DISPLACEMENT VOLUMES, E.G., "APPARENT MASS", "APPARENT MASS VERSUS BRASS", "APPARENT MASS VERSUS DENSITY 8.0". THE VALUES ARE LISTED AS CORRECTIONS TO BE APPLIED TO THE LISTED NOMINAL VALUE (A POSITIVE CORRECTION INDICATES THAT THE MASS IS LARGER THAN THE STATED NOMINAL VALUE BY THE AMOUNT OF THE CORRECTION). THESE VALUES ARE COMPUTED FROM THE VALUES BASED ON AN EXPLICIT TREATMENT OF DISPLACEMENT VOLUMES USING THE FOLLOWING DEFINING RELATIONS AND ARE UNCERTAIN BY THE AMOUNT SHOWN IN TABLE I.

THE ADJUSTMENT OF WEIGHTS TO MINIMIZE THE DEVIATION FROM NOMINAL ON THE BASIS OF "NORMAL BRASS" (IN ACCORDANCE WITH COR. A BELOW) IS WIDESPREAD IN THIS COUNTRY AND IN MANY PARTS OF THE WORLD. VALUES STATED ON EITHER BASIS ARE INTERNALLY CONSISTENT AND DEFINITE. THERE IS, HOWEVER, A SYSTEMATIC DIFFERENCE BETWEEN THE VALUES ASSIGNED ON EACH BASIS, THE VALUE ON THE BASIS OF "DENSITY 8.0' BEING 7 MICROGRAMS/GRAM LARGER THAN THE VALUE ON THE BASIS OF NORMAL BRASS. THIS SYSTEMATIC DIFFERENCE IS CLEARLY DETECTABLE ON MANY DIRECT READING BALANCES.

CORRECTION A - "APPARENT MASS VERSUS BRASS" OR "WEIGHT IN AIR AGAINST BRASS" IS DETERMINED BY A HYPOTHETICAL WEIGHING OF THE WEIGHT AT 20 CELSIUS IN AIR HAVING A DENSITY OF 1.2 MG/CM3, WITH A (NORMAL BRASS) STANDARD HAVING A DENSITY OF 8.4 G/CM3 AT 0 CELSIUS WHOSE COEFFICIENT OF VOLUMETRIC EXPANSION IS 0.000 054 PER DEGREE CELSIUS, AND WHOSE VALUE IS BASED ON ITS TRUE MASS OR WEIGHT IN VACUO.

COMPANY X 1/30/70
NEW YORK, NEW YORK
SET OF MASS STANDARDS 100G TO 1MG
TEST NUMBER 232.09/200390

CORRECTION B - 'APPARENT MASS VERSUS DENSITY 8.0' IS DETERMINED BY A HYPOTHETICAL WEIGHING OF THE WEIGHT, IN AIR HAVING A DENSITY OF 1.2 MG/CM3, WITH A STANDARD HAVING A DENSITY OF 8.0 G/CM3 AT 20 CELSIUS, AND WHOSE VALUE IS BASED ON ITS TRUE MASS OR WEIGHT IN AIR.

COMPANY X
NEW YORK, NEW YORK
SET OF MASS STANDARDS
TEST NUMBER

TABLE I

ITEM	MASS (G)	UNCERTAINTY (G)	VOI AT 20 (CM3)	COEFF OF EXP
100G	100.00102471	.00001506	12.82064	.000045
50G	50.00052848	.00001128	6.41032	.000045
30G	30.00031168	.00001166	3.84619	.000045
20G	20.00024971	.00000996	2.56413	.000045
10G	10.00007056	.00001287	1.28206	.000045
5G	5.00004767	.00000667	.64103	.000045
3G	3.00005156	.00000446	.38462	.000045
2G	2.00002831	.00000325	.25641	.000045
1G	.99998924	.00000296	.12820	.000045
500MG	.49996408	.00000155	.03012	.000020
300MG	.29997469	.00000107	.01807	.000020
200MG	.19998090	.00000079	.01205	.000020
100MG	.09999375	.00000076	.00602	.000020
50MG	.05000386	.00000054	.00301	.000020
30MG	.03000416	.00000054	.01111	.000069
20MG	.02000406	.00000046	.00741	.000069
10MG	.01000583	.00000059	.00371	.000069
5MG	.00500803	.00000049	.00185	.000069
3MG	.00300395	.00000052	.00111	.000069
2MG	.00198636	.00000045	.00074	.000069
1MG	.00100943	.00000059	.00037	.000069

COMPANY X
NEW YORK, NEW YORK
SET OF MASS STANDARDS
TEST NUMBER

TABLE II

ITEM	COR. A (MG)	COR. B (M
100G	-.05893	.64003
50G	-.01333	.33615
30G	-.01342	.19627
20G	.03298	.17277
10G	-.03781	.03209
5G	-.00651	.02844
3G	.01905	.04002
2G	.00664	.02062
1G	-.02160	-.01461
500MG	-.00055	.00294
300MG	-.00409	-.00199
200MG	-.00495	-.00356
100MG	.00082	.00152
50MG	.00740	.00775
30MG	-.00488	-.00468
20MG	-.00198	-.00184
10MG	.00282	.00289
5MG	.00652	.00656
3MG	.00305	.00307
2MG	-.01424	-.01423
1MG	.00913	.00913

Temperature 23.5 °C | Humidity 40 % | Barometer 748.2 | ρ = 1.17

FORM NBS-345.06 (6-1-61) — U.S. DEPARTMENT OF COMMERCE NATIONAL BUREAU OF STANDARDS

SUBSTITUTION WEIGHING
Single Pan Damped Balances
OBSERVATION SHEET

Sheet 1 | Unit (Wts.) g | Unit (Cr. & Dill.) mg

Observer C.B.G. | Balance H-200 | Date 10-12-72 | Set | NBS Test No. 200390

Load	Dial Setting	Scale Reading	Computations
100 g	100 g	2.34	Q_1
S 100 g	"	2.38	-0.04 -0.05 $-0.04 \times \frac{10.045}{10.06} = -0.04$ mg
S 100 g + h 10 mg	"	12.44	
100 g + h 10 mg	"	12.39	Expected Diff. $= -0.039$ mg
100 g	"	2.36	Q_2
Σ 100 g	"	2.43	-0.07 -0.07 $-0.07 \times \frac{10.045}{10.05} = -0.07$ mg
" + h 10 mg	"	12.48	
100 g + h 10 mg	"	12.41	Expected Diff $= -0.065$ mg
20 g	20.0	2.52	Q_3
Σ 20 g	"	2.47	$+0.05$ $+0.06$ $+0.06 \times \frac{10.045}{10.05} = +0.06$
" + h 10 mg	"	12.51	
20 g + h 10 mg	"	12.57	Expected Diff. $= +0.516$
			Q
			Q
			Obs. D_s =
			M_s =

Temperature 23.6 °C	FORM NBS-345.06 (6-1-61)	U.S. DEPARTMENT OF COMMERCE NATIONAL BUREAU OF STANDARDS	Sheet 2
Humidity 40 o/o			Unit (Wts.) g and mg
Barometer 748.0	SUBSTITUTION WEIGHING Single Pan Damped Balances OBSERVATION SHEET		Unit (Cr. & Dill.) mg
ρ = 1.17			

Observer C.B.G	Balance M-10	Date 10-12-72	Set	NBS Test No. 200390

Load	Dial Setting	Scale Reading	Computations
2g	2.0g	0.342	Q_4 +0.040 +0.041 +0.040 × 5.0051/5.005 = +0.090 mg Buoyancy cr. −0.002 " Diff. +0.038 " Expected Diff. = +0.0375 mg
Σ2g	"	0.302	
Σ2g + h 5mg	"	5.307	
2g + h 5mg	"	5.348	
200mg	0.20	0.223	Q_5 −0.007 −0.006 −0.006 × 5.0054/5.006 = −0.006 mg Buoyancy cr. +0.0046 " Diff. −0.0014 " Expected Diff. = −0.0062 "
Σ200 mg	"	0.230	
Σ200 mg + h 5mg	"	5.236	
200 mg + h 5mg	"	5.230	
20 mg	0.02	0.228	Q_6 −0.001 0.000 0.000 × 5.0054/5.006 = 0.000 mg Expected Diff. = −0.0001 mg
Σ20 mg	"	0.229	
Σ20mg + h 5mg	"	5.235	
20mg + h 5mg	"	5.235	
3 mg	0.00	3.338	Q_7 +0.008 +0.010 +0.009 × 5.0054/5.005 = +0.009 mg Expected Diff. = +0.0081 mg
Σ3 mg	"	3.330	
Σ3 mg + h 5mg	"	8.335	
3 mg + h 5mg	"	8.345	
			Q
			Obs. D_s =
			M_s =

Type II Surveillance Test Example

This example of a Type II Surveillance Test is for a set of metric mass standards according to Design 7, appendix 1. This set was calibrated by the National Bureau of Standards and reported under National Bureau of Standards Report of Calibration, Test No. 200390. The report is reproduced on page 47. This is the same set which was used for the example of the Type I Surveillance Test. The only standards (other than the set under test) used in this example are the sensitivity weights listed below, with their apparent mass corrections and uncertainties. The balances used and their standard deviation for a double substitution weighing are also listed. The double substitution method of weighing is used. The standard deviation of the calibration mass measurement is not known, so an estimate is computed from the reported uncertainties of the weights being tested, as described in section 3.2, and used in computing the surveillance limits.

Sensitivity Weight	Apparent Mass Corr. (mg)	Uncertainty (mg)
h10mg	+ 0.0450	0.0006
h 5mg	+ 0.0045	0.0005

Balance (Laboratory Designation)	Standard Deviation (mg)	Capacity g
H-200	0.015	200g
M-10	0.003	20g

Computation of Surveillance Limit (sl)

In the following "U_c" will denote the uncertainty (see Equation (2) Page 7) and "S.D." the estimate of the standard deviation. σ_B denotes the standard deviation of the balance used.

For series 1: 100g, 100g', Σ100g

Weight	U_c(mg)
100g	0.015
100g'	UNKNOWN MASS USED TO FILL SERIES
50g	0.011
30g	0.012
20g	0.010

Standard deviation of balance H-200 = 0.015mg

$$U_c = \sqrt{.015^2 + .011^2 + .012^2 + .010^2}$$

$$= \sqrt{.000225 + .000121 + .000144 + .0001}$$

$$= \sqrt{.00059}$$

$$U_c = 0.024\text{mg}$$

$$\text{S.D.} = 3\sqrt{2/3}\sigma_B$$

$$= 3\sqrt{2/3}(.015)$$

$$= 3\sqrt{.66666}\ (.015)$$

$$= 3(.81649)(.015)$$

$$\text{S.D.} = 0.037\text{mg}$$

sl for Σ100g = 0.024 + 0.037 = <u>0.061mg</u>

Since 100g' is assumed to be an unknown weight, a surveillance limit for it cannot be computed.

For series 2: 30g, $\Sigma 30g_1$, $\Sigma 30g_2$

Weight	U_c (mg)
30g	0.012
20g	0.010
10g	0.013
5g	0.0067
3g	0.0045
2g	0.0032

Standard deviation of balance H-200 = 0.015mg

$$U_c = \sqrt{.012^2 + .010^2 + .013^2}$$

$$= \sqrt{.000144 + .0001 + .000169}$$

$$= \sqrt{.000413}$$

$$U_c = 0.020\text{mg}$$

S.D. = SAME AS IN SERIES 1 (0.037mg)

sl for $\Sigma 30g_1$ = 0.020 + 0.037 = <u>0.057mg</u>

$$U_c = \sqrt{.012^2 + .010^2 + .0067^2 + .0045^2 + .0032^2}$$

$$= \sqrt{.000144 + .0001 + .0000448 + .00002025 + .00001024}$$

$$= \sqrt{.000319}$$

$$U_c = 0.018\text{mg}$$

S.D. = SAME AS IN SERIES 1 (0.037mg)

sl for $\Sigma 30g_2$ = 0.018 + 0.037 = <u>0.055mg</u>

For series 3: $3g$, $\Sigma 3g_1$, $\Sigma 3g_2$

Weight	U_c (mg)
3 g	.0045
2 g	.0032
1 g	.0030
500mg	.0016
300mg	.0011
200mg	.0008

Standard deviation of balance M-10 = 0.003mg

$$U_c = \sqrt{.0045^2 + .0032^2 + .0030^2}$$

$$= \sqrt{.00002025 + .00001024 + .000009}$$

$$= \sqrt{.00003949}$$

$$U_c = 0.0063\text{mg}$$

$$S.D. = 3\sqrt{2/3}\sigma_B$$

$$= 3\sqrt{2/3}(.003)$$

$$= 3\sqrt{.66666}\ (.003)$$

$$= 3(.81649)(.003)$$

$$S.D. = 0.0073\text{mg}$$

sl for $\Sigma 3g_1$ = 0.0063 + 0.0073 = <u>0.014mg</u>

$$U_c = \sqrt{.0045^2 + .0032^2 + .0016^2 + .0011^2 + .0008^2}$$

$$= \sqrt{.00002025 + .00001024 + .00000256 + .00000121 + .00000064}$$

$$= \sqrt{.0000349}$$

$$U_c = 0.0059\text{mg}$$

S.D. = SAME AS FOR $\Sigma 3g_1$ (0.0073mg)

sl for $\Sigma 3g_2$ = 0.0059 + 0.0073 = <u>0.013mg</u>

For series 4: 300mg, $\Sigma 300mg_1$, $\Sigma 300mg_2$

Weight	U_c (mg)
300mg	0.0011
200mg	0.0008
100mg	0.0008
50mg	0.0005
30mg	0.0005
20mg	0.0005

Standard deviation of balance M-10 = 0.003mg

$$U_c = \sqrt{.0011^2 + .0008^2 + .0008^2}$$

$$= \sqrt{.00000121 + .00000064 + .00000064}$$

$$= \sqrt{.00000249}$$

$$U_c = 0.0016mg$$

S.D. = SAME AS IN SERIES 3 (0.0073mg)

sl for $\Sigma 300g_1$ = 0.0016 + 0.0073 = 0.0089mg

$$U_c = \sqrt{.0011^2 + .0008^2 + .0005^2 + .0005^2 + .0005^2}$$

$$= \sqrt{.00000121 + .00000064 + .00000025 + .00000025 + .00000025}$$

$$= \sqrt{.0000026}$$

$$U_c = 0.0016mg$$

S.D. = SAME AS FOR $\Sigma 300mg_1$ (0.0073mg)

sl for $\Sigma 300mg_2$ = 0.0016 + 0.0073 = 0.0089mg

For series 5: 30mg, $\Sigma 30mg_1$, $\Sigma 30mg_2$

Weight	U_c (mg)
30mg	0.00054
20mg	0.00046
10mg	0.00059
5mg	0.00049
3mg	0.00052
2mg	0.00045

Standard deviation for balance M-10 = 0.003mg

$$U_c = \sqrt{.00054^2 + .00046^2 + .00059^2}$$

$$= \sqrt{.000000291 + .0000002116 + .0000003481}$$

$$= \sqrt{.0000008513}$$

$$U_c = 0.00092mg$$

S.D. = SAME AS IN SERIES 4 (0.0073mg)

sl for $\Sigma 30mg_1$ = 0.00092 + 0.0073 = <u>0.0082mg</u>

$$U_c = \sqrt{.00054^2 + .00046^2 + .00049^2 + .00052^2 + .00045^2}$$

$$= \sqrt{.0000002916 + .0000002116 + .0000002401 + .0000002704 + .0000002025}$$

$$= \sqrt{.0000012162}$$

$$U_c = 0.0011mg$$

S.D. = SAME AS FOR $\Sigma 30mg_1$ (0.0073mg)

sl for $\Sigma 30mg_2$ = 0.0011 + 0.0073 = <u>0.0084mg</u>

For 5mg - Σ5mg:

Weight	U_c (mg)
5mg	0.00049
3mg	0.00052
2mg	0.00045

Standard Deviation of Balance M-10 = 0.003mg

$$U_c = \sqrt{.00049^2 + .00052^2 + .00045^2}$$

$$= \sqrt{.0000002401 + .0000002704 + .0000002025}$$

$$= \sqrt{.000000713}$$

$$U_c = 0.00084\text{mg}$$

$$sl = .00084 + 3(.003)$$
$$= .00084 + .009 = \underline{0.0098\text{mg}}$$

For 3mg - Σ3mg

Weight	U_c (mg)
3mg	0.00052
2mg	0.00045
1mg	0.00059

Standard Deviation of Balance M-10 = 0.003mg

$$U_c = \sqrt{.00052^2 + .00045^2 + .00059^2}$$

$$= \sqrt{.0000002704 + .0000002025 + .0000003481}$$

$$= \sqrt{.000000821}$$

$$U_c = 0.00091\text{mg}$$

$$sl = 0.00091 + 3(.003)$$
$$= 0.00091 + .009 = \underline{0.0099\text{mg}}$$

Buoyancy Corrections

The buoyancy corrections $\Delta\rho\Delta V$ are computed according to the procedure set forth in section 5.2 using the formula:

$$\text{Buoyancy Correction} = (\rho - \rho_n)(V_C - V_D)$$

In this example, only two of the comparisons, $3g - \Sigma 3g_2$ and $300mg - \Sigma 300mg$, are between weights having different densities. A buoyancy correction need be computed only for these two comparisons. All of the other comparisons are between weights having the same density, so their volume differences are virtually zero and the buoyancy corrections are also virtually zero.

The buoyancy correction for the comparison $3g - \Sigma 3g_2$, a_2 of series 3, is:

Weights		Volumes	
3g		$0.3846 cm^3$	from Report of Calibration
2g	$0.2564 cm^3$		"
500mg	$0.0301 cm^3$		"
300mg	$0.0181 cm^3$		"
200mg	$0.0120 cm^3$		"
$\Sigma 3g$		$0.3166 cm^3$	"

$\rho = 1.16 mg/cm^3$ Air density at time of weighing

$\rho_n = 1.20 mg/cm^3$ Normal air density

Buoyancy correction = $(1.16 - 1.20)(0.3846 - 0.3166) = -0.0027mg$

This is the figure entered on the 3-1's computation sheet under the $\Sigma 3g_2$ column on the $-\Delta\rho\Delta V$ line, Sheet 2, Series 3. Note that it is $-\Delta\rho\Delta V$ that is called for and the buoyancy correction is -0.0027mg, therefore, the buoyancy correction is entered as +0.0027mg. (See section 5.3.2).

The buoyancy correction for the comparison $300mg - \Sigma 300mg_2$, a_2 of series 4, is:

Weight			Volume	
300mg			$0.0181 cm^3$	From Report of Calibration
	200mg	$0.0120 cm^3$		"
	50mg	$0.0030 cm^3$		"
	30mg	$0.0111 cm^3$		"
	20mg	$0.0074 cm^3$		"
Σ300mg			$0.0335 cm^3$	"

ρ = $1.16 mg/cm^3$ Air density at time of weighing

ρ_n = $1.20 mg/cm^3$ Normal air density

Buoyancy Correction = $(1.16 - 1.20)(0.0181 - 0.0335) = +0.0006$mg

This is the figure entered on the 3-1's computation sheet under the $\Sigma 300mg_2$ column on the $-\Delta\rho\Delta V$ line, Sheet 2, Series 4. Note that it is $-\Delta\rho\Delta V$ that is called for and that the buoyancy correction is +0.0006mg, therefore the buoyancy correction is entered as - 0.0006 mg. (See section 5.3.2).

Computation of Predicted or Expected Values

The expected differences are computed from the reported values as follows: (See report on page 47).

Series 1: Σ100g

Weight		Apparent Mass Value
50g		-0.0133mg
30g		-0.0134mg
20g		+0.0330mg
Σ100g	Expected Value	+0.0063mg

Series 2: $\Sigma 30g_1$

20g		+0.0330mg
10g		-0.0378mg
$\Sigma 30g_1$	Expected Value	-0.0048mg

Series 2: $\Sigma 30g_2$

20g		+0.0330mg
5g		-0.0065mg
3g		+0.0191mg
2g		+0.0066mg
$\Sigma 30g_2$	Expected Value	+0.0522mg

Series 3: $\Sigma 3g_1$

2g		+0.0066mg
1g		-0.0216mg
$\Sigma 3g_1$	Expected Value	-0.0150mg

Series 3: $\Sigma 3g_2$

2g		+0.0066mg
500mg		-0.0005mg
300mg		-0.0041mg
200mg		-0.0049mg
$\Sigma 3g_2$	Expected Value	-0.0029mg

Series 4: $\Sigma 300mg_1$

Weight		Apparent Mass Value
200mg		-0.0049mg
100mg		+0.0008mg
$\Sigma 300mg_1$	Expected Value	-0.0041mg

Series 4: $\Sigma 300mg_2$

200mg		-0.0049mg
50mg		+0.0074mg
30mg		-0.0049mg
20mg		-0.0020mg
$\Sigma 300mg_2$	Expected Value	-0.0044mg

Series 5: $\Sigma 30mg_1$

20mg		-0.0020mg
10mg		+0.0028mg
$\Sigma 30mg_1$	Expected Value	+0.0008mg

Series 5: $\Sigma 30mg_2$

20mg		-0.0020mg
5mg		+0.0065mg
3mg		+0.0030mg
2mg		-0.0142mg
$\Sigma 30mg_2$	Expected Value	-0.0067mg

Series 6: For weighing 5mg - Σ5mg

Weight	Apparent Mass Value 5mg	Σ5mg
5mg	+0.0065mg	
3mg		+0.0030mg
2mg		-0.0142mg
Sums	+0.0065mg	-0.0112mg
Expected Difference		-0.0047mg

Series 6: For weighing 3mg - Σ3mg

	3mg	Σ3mg
3mg	+0.0030mg	
2mg		-0.0142mg
1mg		+0.0091mg
Sums	+0.0030mg	-0.0051mg
Expected Difference	-0.0021mg	

Temperature 24.1 °C	FORM NBS-345 .6 (6-1-61)	U.S. DEPARTMENT OF COMMERCE NATIONAL BUREAU OF STANDARDS	Sheet 1 Series 1
Humidity 40%			Unit (Wts.)
Barometer 747.2 mm		SUBSTITUTION WEIGHING Single Pan Damped Balances	Unit (Cl. & Dill.)
ρ = 1.16 mg/cm^3		OBSERVATION SHEET	

Observer S.V.O.	Balance H-200	Date 11-1-72	Set	NBS Test No. 200390

Load	Dial Setting	Scale Reading	Computations
100g	100.0g	2.38	a_1
100'g	"	2.01	$\frac{+0.37}{+0.38}$ / $+0.38 \times \frac{10.045}{10.05} = +0.38$ mg
" + h 10 mg	"	12.06	
100g + h 10 mg	"	12.44	
100g	"	2.39	a_2
Σ 100g	"	2.47	$\frac{-0.08}{-0.09}$ / $-0.085 \times \frac{10.045}{10.06} = -0.08$
" + h 10 mg	"	12.53	
100g + h 10 mg	"	12.44	
100'g	"	2.03	a_3
Σ100g	"	2.47	$\frac{-0.44}{-0.46}$ / $-0.45 \times \frac{10.045}{10.06} = -0.45$
" + h 10 mg	"	12.53	
100'g + h 10 mg	"	12.07	
			a
100'g = any 100g weight nominal is 100g.			a summation of weights whose used to fill the series
Σ 100g = 50g + 30g + 20g			
			Obs. D_s = ... M_s =

Sheet 2 Series 1

Date 11-1-72

Std. 1, l_1, l_2

	Std. 1	l_1	l_2	Observations
	+	-		a_1
	+		-	a_2
		+	-	a_3

K - A M Cor. Std. 1 = - 0.058 mg

	Std. 1 (100g)	l_1 (100g)	l_2 (Σ 100g)	Check
a_1 + 0.38	0	-2	-1	-3
a_2 - 0.08	0	-1	-2	-3
a_3 - 0.45	0	1	-1	0
K - 0.058	3	3	3	9
Sum =	-0.174	-1.304	+0.056	Σ down / Σ across
d =	3	3	3	
Sum/d	- 0.058	-0.435	+0.019	
-ΔρAV	———	———	———	
Est. True Mass Cor.	———	———	———	
App. Mass vs. Brass Cor.	- 0.058 mg	- 0.435 mg	+0.019 mg	
Accepted Cor. from Report	- 0.058 mg	*	+0.007 mg	

	Δ_1	Δ_2	Δ_3	Check
a_1 +0.38	1	-1	1	1
a_2 - 0.08	-1	1	-1	-1
a_3 -0.45	1	-1	1	1
Sum =	+0.01	-0.01	+0.01	Σ down / Σ across
d =	3	3	3	
Sum/d =	+0.0033	-0.0033	+0.0033	

$S = \sqrt{(\Delta_1)^2 + (\Delta_2)^2 + (\Delta_3)^2}$ = 0.0057 mg

* This is an unknown

Temperature: 24.1 °C
Humidity: 40%
Barometer: 747.3 mm
ρ = 1.16 mg/cm²

FORM NBS-345.06 (6-1-61) U.S. DEPARTMENT OF COMMERCE NATIONAL BUREAU OF STANDARDS

SUBSTITUTION WEIGHING
Single Pan Damped Balances
OBSERVATION SHEET

Sheet: 1 Series 2
Unit (Wts.):
Unit (Cr. & Dill.): 200390

Observer: S.V.O. | Balance: H-200 | Date: 11-1-72 | Set: | NBS Test No.: 200390

Load	Dial Setting	Scale Reading	Computations
30g	30.0g	2.30	$30g - \Sigma 30g_1 =$ -0.03 -0.01 $-0.02 \times \frac{10.045}{10.06} = -0.02$ mg $= a_1$
$\Sigma 30g_1$	"	2.33	
" + h 10 mg	"	12.39	
30g + h 10 mg	"	12.38	
30g	"	2.28	$30g - \Sigma 30g_2 =$ -0.06 -0.07 $-0.065 \times \frac{10.045}{10.06} = -0.06$ mg $= a_2$
$\Sigma 30g_2$	"	2.34	
" + h 10 mg	"	12.39	
30g + h 10 mg	"	12.32	
$\Sigma 30g_1$	"	2.30	$\Sigma 30g_1 - \Sigma 30g_2 =$ -0.07 -0.08 $-0.075 \times \frac{10.045}{10.06} = -0.07$ mg $= a_3$
$\Sigma 30g_2$	"	2.37	
" + h 10 mg	"	12.43	
$\Sigma 30g_1$ + h 10 mg	"	12.35	
			a
$\Sigma 30g_1 = 20g +$	10g		a
$\Sigma 30g_2 = 20g +$	5g + 3g + 2g		
			Obs. D_s = M_s =

Sheet 2 Series 2

Date 11-1-72

Std. 1, l_1, l_2

Std. 1	l_1	l_2	Observations
+	-		a_1
+		-	a_2
	+	-	a_3

K - A M Cor. Std. 1 = - 0.013 mg

	30g Std. 1	Σ 30g. l_1	Σ 30g. l_2	Check
a_1 -0.02	0	-2	-1	-3
a_2 -0.06	0	-1	-2	-3
a_3 -0.07	0	1	-1	0
K -0.013	3	3	3	9
Sum =	-0.039	-0.009	+0.171	Σ down / Σ across
d =	3	3	3	
	-0.013	-0.003	+0.057	
Sum/d				
$-\Delta\rho\Delta V$				
Est. True Mass Cor.				
App. Mass vs. Brass Cor.	-0.013 mg	-0.003 mg	+0.057 mg	
Accepted Cor. from Report	- 0.013 mg	-0.005 mg	+0.053 mg	

	Δ_1	Δ_2	Δ_3	Check
a_1 -0.02	1	-1	1	1
a_2 -0.06	-1	1	-1	-1
a_3 -0.07	1	-1	1	1
Sum =	-0.03	+0.03	-0.03	Σ down / Σ across
d =	3	3	3	
	-0.01	+0.01	-0.01	
Sum/d =				

$$S = \sqrt{(\Delta_1)^2 + (\Delta_2)^2 + (\Delta_3)^2} = \sqrt{3 \times 10^{-4}} = .0.17 \text{ mg}$$

Temperature 24.2 °C	FORM NBS-345 J6 (6-1-61)	U.S. DEPARTMENT OF COMMERCE NATIONAL BUREAU OF STANDARDS	Sheet 1 Series 3
Humidity 40 %			Unit (Wts.)
Barometer 747.2	SUBSTITUTION WEIGHING Single Pan Damped Balances		Unit (Cr. & Dills.)
ρ = 1.16 mg/cm^3	OBSERVATION SHEET		

Observer	Balance	Date 11-1-72	Set	NBS Test No. 200390

Load	Dial Setting	Scale Reading	Computations
3g	3.00g	0.340	$3g - \Sigma 3g_1 =$ a_1
$\Sigma 3g_1$	"	0.301	$+0.034$
" + h 5mg	"	5.306	$+0.037$
3g + h 5mg	"	5.343	$+0.038 \times \frac{5.0045}{5.005} = +0.038$ mg
3g	"	0.324	$3g - \Sigma 3g_2 =$ a_2
$\Sigma 3g_2$	"	0.305	$+0.019$
" + h 5mg	"	5.310	$+0.017$
3g + h 5mg	"	5.327	$+0.018 \times \frac{5.0045}{5.005} = +0.018$
$\Sigma 3g_1$	"	0.304	$\Sigma 3g_1 - \Sigma 3g_2 =$ a_3
$\Sigma 3g_2$	"	0.319	-0.015
" + h 5mg	"	5.324	-0.014
$\Sigma 3g_2$ + h 5mg	"	5.310	$-0.0145 \times \frac{5.0045}{5.005} = -0.014$
			a
$\Sigma 3g_1 = 2g + 1g$			a
$\Sigma 3g_2 = 2g + 500mg + 300mg + 200mg$			
			Obs. D_s =
			M_s =

Sheet 2 Series 3

Date 11-1-72

Std. 1, 1_1, 1_2

Std. 1	1_1	1_2	Observations
+	-		a_1
+		-	a_2
	+	-	a_3

K - A M Cor. Std. 1 = +0.0191 mg

	$3g$ Std. 1	$\Sigma 3g_1$ 1_1	$\Sigma 3g_2$ 1_2	Check
a_1 +0.038 mg	0	-2	-1	-3
a_2 +0.018	0	-1	-2	-3
a_3 -0.014	0	1	-1	0
K +0.0191	3	3	3	9
Sum =	+0.0573	-0.0507	-0.0027	Σ down Σ across
d =	3	3	3	
Sum/d	+0.0191	-0.0139	-0.0009	
-ΔρΔV	———	———	+0.0027	
Est. True Mass Cor.	———	———	———	
App. Mass vs. Brass Cor.	+0.0191 mg	-0.0139	+0.0018 mg	
Accepted Cor. from Report	+0.0191 mg	-0.0150	-0.0027 mg	

	Δ_1	Δ_2	Δ_3	Check
a_1 +0.038	1	-1	1	1
a_2 +0.018	-1	1	-1	-1
a_3 -0.014	1	-1	1	1
Sum =	+0.006	-0.006	+0.006	Σ down Σ across
d =	3	3	3	
Sum/d =	+0.002	-0.002	-0.002	

$$S = \sqrt{(\Delta_1)^2 + (\Delta_2)^2 + (\Delta_3)^2} = \sqrt{12 \times 10^{-6}} = 0.0035 \text{ mg}$$

Temperature 24.2 °C	FORM NBS-345.06 (6-1-61)	U.S. DEPARTMENT OF COMMERCE NATIONAL BUREAU OF STANDARDS	Sheet 1 Series 4
Humidity 40 %	**SUBSTITUTION WEIGHING**		Unit (Wts.)
Barometer 747.2 mm	**Single Pan Damped Balances**		Unit (Cr. & Dill.)
ρ = 1.16 mg/cm³	**OBSERVATION SHEET**		

Observer S. V. O.	Balance M-10	Date 11-1-72	Set	NBS Test No. 200390

Load	Dial Setting	Scale Reading	Computations
300 mg	0.30	0.291	$300\text{ mg} - \Sigma 300\text{ mg}_1 =$ a_1
Σ 300 mg$_1$	"	0.289	$+0.082$
" + h 5 mg	"	5.294	$+0.001$
300 mg + h 5 mg	"	5.295	$+0.0015 \times \frac{5.0045}{5.005} = +0.001$ mg
300 mg	"	0.297	$300\text{ mg} - \Sigma 300\text{ mg}_2 =$ a_2
Σ 300 mg$_2$	"	0.296	$+0.081$
" + h 5 mg	"	5.300	-0.001
300 mg + h 5 mg	"	5.299	$0.000 \times \frac{5.0045}{5.004} = 0.000$ mg
Σ 300 mg$_1$	"	0.295	$\Sigma 300\text{ mg}_1 - \Sigma 300\text{ mg}_2 =$ a_3
Σ 300 mg$_2$	"	0.294	$+0.081$
" + h 5 mg	"	5.299	$+0.003$
Σ 300 mg$_1$ + h 5 mg	"	5.232	$+0.002 \times \frac{5.0045}{5.005} = +0.002$ mg
	.	.	a
	.	.	
	.	.	
	.	.	
Σ 300 mg$_1$ = 200 mg + 100 mg			a
Σ 300 mg$_2$ = 200 mg + 50 mg + 30 mg + 20 mg			
	.	.	
	.	.	
	"	.	Obs. D_s = ... M_s = .

Sheet 2 Series 4

Date 11-2-72

Std. 1, 1_1, 1_2

	Std. 1	1_1	1_2	Observations
	+	-		a_1
	+		-	a_2
		+	-	a_3

K - A K Cor. Std. 1 = - 0.0040 mg

	300 mg Std. 1	Σ 300 mg, 1_1	Σ 300 mg, 1_2	Check
a_1 +0.001	0	-2	-1	-3
a_2 0.000	0	-1	-2	-3
a_3 +0.002	0	1	-1	0
K -0.0040	3	3	3	9
Sum =	-0.0120	-0.0120	-0.0150	Σ down, Σ across
d =	3	3	3	
Sum/d	-0.0040	-0.0040	-0.0050	
-ΔρΔV	-------	-------	-0.0006	
Est. True Mass Cor.	-------	-------	-------	
App. Mass vs. Brass Cor.	-0.0040 mg	-0.0040 mg	-0.0056 mg	
Accepted Cor. from Report	-0.0040 mg	-0.0041 mg	-0.0044 mg	

	Δ_1	Δ_2	Δ_3	Check
a_1 +0.001	1	-1	1	1
a_2 0.000	-1	1	-1	-1
a_3 +0.003	1	-1	1	1
Sum =	+0.003	-0.003	+0.003	Σ down, Σ across
d =	3	3	3	
Sum/d =	+0.001	-0.001	+0.001	

$$S = \sqrt{(\Delta_1)^2 + (\Delta_2)^2 + (\Delta_3)^2} = \sqrt{3 \times 10^{-6}} = 0.0017 \text{ mg}$$

Temperature 24.2 °C	FORM NBS-345 (6-1-61)	U.S. DEPARTMENT OF COMMERCE NATIONAL BUREAU OF STANDARDS	Sheet 1 Series 5
Humidity 40 %			Unit (Wts.)
Barometer 747.1 mm	SUBSTITUTION WEIGHING Single Pan Damped Balances OBSERVATION SHEET		Unit (Cr. & Dill.)
ρ = 1.16 mg/cm^3			

Observer S.V.O.	Balance M-10	Date 11-1-72	Set	NBS Test No. 200390

Load	Dial Setting	Scale Reading	Computations
30 mg	0.03g	0.285 mg	$30\text{ mg} - \Sigma 30\text{ mg}_1 =$ a_1 -0.806 -0.004 $-0.005 \times \frac{5.0045}{5.005} = -0.005$ mg
$\Sigma 30$ mg$_1$	0.03	0.291	
" + h 5 mg	"	5.296	
30 mg + h 5 mg	"	5.292	
30 mg	"	0.283	$30\text{ mg} - \Sigma 30\text{ mg}_2 =$ a_2 -0.804 -0.003 $-0.0035 \times \frac{5.0045}{5.006} = -0.003$ mg
$\Sigma 30$ mg$_2$	"	0.287	
" + h 5 mg	"	5.293	
30 mg + h 5 mg	"	5.290	
$\Sigma 30$ mg$_1$	"	0.291	$\Sigma 30\text{ mg}_1 - \Sigma 30\text{ mg}_2 =$ a_3 $+0.006$ $+0.005$ $+0.0055 \times \frac{5.0045}{5.005} = +0.005$
$\Sigma 30$ mg$_2$	"	0.285	
" + h 5 mg	"	5.290	
$\Sigma 30$ mg$_1$ + h 5 mg	"	5.295	
	.	.	a
	.	.	
	.	.	
	.	.	
	.	.	a
	.	.	
	.	.	
	.	.	
	.	.	
	.	.	Obs. D_s = ... M_s = .

Sheet 2 Series 5

Date 11-2-72

Std. 1, l_1, l_2

Std. 1	l_1	l_2	Observations
+	-		a_1
+		-	a_2
	+	-	a_3

K - A M Cor. Std. 1 = -0.0049 mg

	30 mg Std. 1	Σ30 mg, l_1	Σ 30 mg, l_2	Check
a_1 -0.005	0	-2	-1	-3
a_2 -0.003	0	-1	-2	-3
a_3 +0.005	0	1	-1	0
K -0.0049	3	3	3	9
Sum =	-0.0147	+0.0033	-0.0087	Σ down, Σ across
d =	3	3	3	
Sum/d	-0.0049	+0.0011	-0.0029	
-ΔρΔV	———	———	———	
Est. True Mass Cor.	———	———	———	
App. Mass vs. Brass Cor.	-0.0049 mg	+0.0011 mg	-0.0029 mg	
Accepted Cor. from Report	-0.0049 mg	+0.0008 mg	-0.0067 mg	

	Δ_1	Δ_2	Δ_3	Check
a_1 -0.005	1	-1	1	1
a_2 -0.003	-1	1	-1	-1
a_3 +0.005	1	-1	1	1
Sum =	+0.003	-0.003	+0.003	Σ down, Σ across
d =	3	3	3	
Sum/d =	+0.001	-0.001	+0.001	

$$S = \sqrt{(\Delta_1)^2 + (\Delta_2)^2 + (\Delta_3)^2} = \sqrt{3 \times 10^{-6}} = 0.0017 \text{ mg}$$

Temperature 24.3 °C | Humidity 40 % | Barometer 747.3 mm | ρ = 1.16 mg/cm³

FORM NBS-345.06 (6-1-61) U.S. DEPARTMENT OF COMMERCE NATIONAL BUREAU OF STANDARDS

SUBSTITUTION WEIGHING
Single Pan Damped Balances
OBSERVATION SHEET

Sheet 1 Series 6 | Unit (Wts.) | Unit (Cr. & Diff.) 200390

Observer S.V.O. | Balance M-10 | Date 11-1-72 | Set | NBS Test No.

Load	Dial Setting	Scale Reading	Computations
5 mg	0.00	0.257	a_1: 5 mg − Σ 5 mg = +0.018, +0.016 / +0.017 × 5.0045/5.005 = +0.017 mg
Σ 5 mg	"	0.239	
" + h 5 mg	"	5.244	
5 mg + h 5 mg	"	5.260	
3 mg	"	0.241	a_2: 3 mg − Σ 3 mg = +0.004, +0.008 / +0.0085 × 5.0045/5.006 = +0.008 mg
Σ 3 mg	"	0.232	
" + h 5 mg	"	5.238	
3 mg + h 5 mg	"	5.246	
	.	.	a
	.	.	
	.	.	
	.	.	
	.	.	a
	.	.	
	.	.	
	.	.	
	.	.	a
	.	.	
	.	.	
	.	.	
	.	.	
		.	Obs. D_s = ... M_s = .

NBS-114A (REV. 7-73)

U.S. DEPT. OF COMM. BIBLIOGRAPHIC DATA SHEET	1. PUBLICATION OR REPORT NO. NBSIR 76-999	2. Gov't Accession No.	3. Recipient's Accession No.

4. TITLE AND SUBTITLE

SURVEILLANCE TESTS PROCEDURES

5. Publication Date

February 1976

6. Performing Organization Code

7. AUTHOR(S)

H. E. Almer Edited by: Jerry Keller

8. Performing Organ. Report No.

NBSIR 76-999

9. PERFORMING ORGANIZATION NAME AND ADDRESS

NATIONAL BUREAU OF STANDARDS
DEPARTMENT OF COMMERCE
WASHINGTON, D.C. 20234

10. Project/Task/Work Unit No.

11. Contract/Grant No.

12. Sponsoring Organization Name and Complete Address (Street, City, State, ZIP)

Same as No. 9

13. Type of Report & Period Covered

Final

14. Sponsoring Agency Code

15. SUPPLEMENTARY NOTES

16. ABSTRACT *(A 200-word or less factual summary of most significant information. If document includes a significant bibliography or literature survey, mention it here.)*

Surveillance tests are designed to monitor the values of mass standards between calibrations. Two types are described; both consist of comparisons of the weights of an ordered set of mass standards with each other. The differences found are compared with those computed from the reported mass values. Surveillance limits based on the precision of both the calibration and the surveillance test processes are computed. These limits are estimates of the departure of the measured differences from the expected, or predicted, differences as computed from the reported values. A larger change is considered significant. Additional measurements to identify individual weights which have changed are required when a given comparison indicates that the mass of one or more of the weights involved has changed. Buoyancy corrections are used to correct for the difference in the buoyant effect on weights of differing densities. Records document the surveillance test results, and control charts help detect trends. Judgments concerning recalibration can be made based on the constancy of the weights and the use requirements.

17. KEY WORDS *(six to twelve entries; alphabetical order, capitalize only the first letter of the first key word unless a proper name; separated by semicolons)*

Apparent mass; buoyancy; buoyancy correction; change; comparison; difference; mass; records; set; surveillance limits; surveillance test; test interval; true mass; value; weighing design; weights.

18. AVAILABILITY

[X] Unlimited

[ ] For Official Distribution. Do Not Release to NTIS

[ ] Order From Sup. of Doc., U.S. Government Printing Office Washington, D.C. 20402, SD Cat. No. C13

[X] Order From National Technical Information Service (NTIS) Springfield, Virginia 22151

19. SECURITY CLASS (THIS REPORT)	21. NO. OF PAGES
UNCLASSIFIED	77
20. SECURITY CLASS (THIS PAGE)	**22. Price**
UNCLASSIFIED	$5.00

USCOMM-DC 29042-P74

Designs for the Calibration of Small Groups of Standards in the Presence of Drift

Joseph M. Cameron and
Geraldine E. Hailes

Institute for Basic Standards
National Bureau of Standards
Washington, D.C. 20234

U.S. DEPARTMENT OF COMMERCE, Frederick B. Dent, *Secretary*
NATIONAL BUREAU OF STANDARDS, Richard W. Roberts, *Director*

Issued August 1974

DESIGNS FOR THE CALIBRATION OF SMALL GROUPS OF STANDARDS IN THE PRESENCE OF DRIFT

by

Joseph M. Cameron and Geraldine E. Hailes

The process of calibrating a small number of "unknown" standards relative to one or two reference standards involved determining differences among the group of objects. Drift, due most often to temperature effects, or a "left-right" polarity effect can bias both the values assigned to the objects and the estimate of the effect of random errors. This note presents schedules of measurements of differences that eliminate the bias from these sources in the assigned value and variances and at the same time gives estimates of the magnitude of these extraneous components. The use of these designs in measurement process control is discussed and a computer program in BASIC is presented.

Key Words: Calibration; calibration design; experiment design; instrumental drift; measurement process; statistical analysis; trend elimination

1. Introduction

In very few processes can the effect of time be ignored. Instability in the object being measured, inability to maintain constant conditions or procedures, and variations in the detector or comparator all contribute to changes with time. A number of approaches have been suggested for reducing or eliminating the effects of these temporal effects on the validity of one's measurements. One way is to make measurements far enough apart in time (usually with some formal randomization procedure to guarantee statistical independence of the measurements) that the cumulative effects from the various sources appear in the random error component. At the other extreme, one can go to great lengths to eliminate these time dependent effects by achieving better environmental control, better instruments, better procedures, etc. If the measurements are to be transferred, as with instrument calibrations, then the first procedure leads to error bounds in which the random error limits include a between-time component whereas the latter procedure suppresses such a component. Neither of these represent the conditions of use adequately.

A compromise consists of arranging the experiment under its normal conditions so that it is as nearly as possible free of time dependent effects. The classic example of this is afforded by the calibration of thermometers in a bath with a gradually rising temperature using the schedule whose structure is as follows for a standard S, and 4 unknowns, T_1, T_2, T_3, T_4.

$$S\ T_1\ T_2\ T_3\ T_4\ S\ T_4\ T_3\ T_2\ T_1\ S$$

If the measurements are evenly spread in time, then the average of the bath temperature for all thermometers are the same (see [3] for a discussion of this practice). A similar procedure has been followed in weighing where in the substitution method one measures in scale units

$$A,\ B,\ B+S,\ A+S$$

to obtain the difference A-B and the deflection corresponding to the sensitivity weight, S.

The calibration of a small number of "unknown" objects relative to one or two reference standards involves determining differences among the group of objects. Instrumental drift, due most often to temperature effects, or a "left-right" polarity effect can bias both the values assigned to the objects and the estimate of the effect of random errors. This note presents schedules of measurements of differences that eliminate the bias from these sources and at the same time gives estimates of the magnitude of these extraneous components. The use of these designs in measurement process control is discussed and a computer program in BASIC is presented in this report.

2. Measurement as a Process

A single isolated measurement, like a single event in history, is difficult to interpret unless it can be regarded as a part of a continuing process. When the measurement is looked upon as the output of a process--a production process whose output is the measured values--then one can attribute to the single measurement the properties of the process from which it arose (for a discussion of this approach, see Eisenhart [2]). Just as with any production process, the operating characteristics are determined by building some redundancy into the system. Redundancy is needed to assure oneself that he has indeed measured the sought after quantity, uncontaminated by extraneous factors related to the operator, instrument, environment, or other items.

Among the characteristics of the process are those associated with the ability to repeat a measurement both in the short term and in the long term. Repetitions made within a few hours, such as with designs having more observations than unknowns, usually exhibit less variation than those made at long time intervals. This additional long-term component of variance can be measured from the agreement among repeated measurements on the same quantity. In addition to these properties related to variability, one needs to incorporate checks on the systematic errors which may possibly affect the process, and, if possible, measurements that provide information as to the adequacy of the assumptions in the underlying physical model.

In calibration it is often convenient to measure a check standard along with the calibration of one or more unknowns. One thus has a value for monitoring the process that is on an equal footing with the unknowns. By tracking its long-run performance, one can determine not only the presence of components of variance, but also by recording ancillary information on environmental and other factors one can develop information for assessing the adequacy of the assumed physical model and for setting bounds to the effect from known sources of possible systematic error. This "check standard" need not be the value of a single item but may take the form of a difference between two such items or some linear combination of several.

The effect of some sources of systematic error can be eliminated by "balancing out" the effect by repeating the measurement of a difference, $(x - y)$ in the reverse order, $(y - x)$. Time dependent effects can be balanced out by using the techniques of this report. For others, it is sometimes possible to alter the conditions to levels of a factor beyond that known to have been in effect at the time of the measurement and to use the changes produced in the output at these extremes as a bounds to the effect of the factor.

In all cases one has to continually monitor the process output just as one does with an industrial production process if he is to have assurance that the calibrations are correct.

3. Substitution Weighing

Consider first the simple situation of scale deflections produced on a balance by adding weights A and B and a sensitivity weight S. One could use either of the following sequences.

Sequence 1	Sequence 2
A	A
A+S	B
B	B+S
B+S	A+S

In high precision work one invariably finds a change in balance response with time so that the value for the difference (A-B) will obviously be contaminated by whatever time effects exist for Sequence 1. If Sequence 2 is used, it may be represented as follows.

Quantity	Effect of Drift	Scale Divisions
A	-3Δ	x_1
B	$-\Delta$	x_2
B+S	Δ	x_3
A+S	3Δ	x_4

and the quantity

$$\frac{1}{2}(x_1 - x_2 - x_3 + x_4)$$

can be seen to give an unbiased value for (A-B) because the drift effect (a 2Δ change in scale reading between each observation) cancels out. The least squares values for A-B, S, and Δ in scale divisions are

$$\widehat{(A-B)} = \frac{1}{2}(x_1 - x_2 - x_3 + x_4)$$

$$\hat{S} = \frac{1}{2}(x_1 - 3x_2 + 3x_3 - x_4)$$

$$\hat{\Delta} = \frac{1}{4}(-x_1 + x_2 - x_3 + x_4)$$

4. Thermometry

At NBS, the calibration of liquid in glass thermometers is usually carried out in a controlled bath which is continually heated so as to give a slight temperature increase with time. The temperature of the bath is measured by resistance thermometry at the start, middle, and end of the run with the test thermometers being run once each in the first interval and once again in reverse order in the second. The time sequence for the resistance measurements R_1, R_2, R_3 and the two series of test thermometer values denoted by $T_1\ T_2 \cdots T_k$ are as follows:

$$R_1\ T_1\ T_2 \cdots T_k\ R_2\ T'_k \cdots T'_2\ T'_1\ R_3$$

If equal time intervals are maintained between readings of the test thermometers, then one would expect an increase, ΔT in temperature with each interval except perhaps the middle one in which the resistance thermometer reading, R_2, is made. The analysis of this form of data is given in Appendix A.

5. Polarimeter Data

In determining the optical rotation of a quartz control plate used as reference standards in polarimeters, one measures the voltage response of a synchronous detector as the angle is varied. However, the response, y, of the system has a nearly linear drift with the angle so that one can represent this drift effect relative to the centroid of the data as being either $\cdots -3\Delta, -2\Delta, -\Delta, 0, \Delta, 2\Delta, 3\Delta, \cdots$ with Δ being the increment to the response added in each time interval. [For even n it is convenient to use $\cdots -3\Delta, -\Delta, \Delta, 3\Delta \cdots$ or 2Δ increment per time interval.]

In the polarimeter experiment the response is a linear function of angle so that the observation becomes

$$y_i = \alpha + \beta x_i + (i - \tfrac{n+1}{2})\Delta + \text{random error}$$

where the x_i are evenly spaced deviations from the nominal angle, e.g., x = 0", 10", 20", 30", $\cdots$ If the usual estimate of α and β are to remain unbiased and unchanged in precision, then one must have

$$\Sigma x_i\ (i - \tfrac{n+1}{2}) = 0$$

so that the estimates are orthogonal to the drift in the detector. The following orderings have this property:

	n = 4		n = 5	
Measurement Number	Quantity to be Measured	Setting for Polarimeter	Quantity to be Measured	Setting for Polarimeter
1	$\alpha + 2\beta$	20"	$\alpha + \beta$	10"
2	α	0"	$\alpha + 4\beta$	40"
3	$\alpha + 3\beta$	30"	$\alpha + 2\beta$	20"
4	$\alpha + \beta$	10"	α	0"
5	---	---	$\alpha + 3\beta$	30"

6. Calibration Designs

The term calibration design has been applied [1] to experiments where only differences between nominally equal objects or groups of objects can be measured. Perhaps the simplest such experiment consists in measuring the differences between the two objects of the $n(n-1)/2$ distinct pairing that can be formed from n objects. Ordinarily the order in which these measurements are made is of no consequence. However, when the response of the comparator is time dependent, attention to the order is important if one wishes to minimize the effect of these changes. When this effect can be adequately represented by a linear drift, it is possible to balance out the effect by proper ordering of the observations. As with the polarimeter data, this drift can be represented by the series $\cdots -3\Delta, -2\Delta, -\Delta, 0, \Delta, 2\Delta, 3\Delta, \cdots$ if $n(n-1)/2$ is odd or by $\cdots -5\Delta, -3\Delta, -\Delta, \Delta, 3\Delta, 5\Delta \ldots$ if $n(n-1)/2$ is even.

For $n = 4$, $n(n-1)/2 = 6$, and it turns out that it is not possible to balance out the drift effect with 6 measurements. However, with 8 measurements the balance can be achieved by the following order, denoting the four objects by A, B, C, D.

Observation	Observation is a Measurement of				
	A	B	C	D	Δ
y_1	+	-	0	0	-7
y_2	-	0	0	+	-5
y_3	0	0	+	-	-3
y_4	0	+	-	0	-1
y_5	0	+	0	-	1
y_6	-	0	0	+	3
y_7	+	0	-	0	5
y_8	0	-	+	0	7

The notation used here, the plus and minus signs, indicate the items entering into the difference measurement. Thus, y_2 is a measurement of the difference (D-A).

To see how the drift effect is balanced out, consider item C which occurs in the third, fourth, seventh, and eighth observations. In the third and eighth observations the item occurs positively and the corresponding drift effects are -3Δ and 7Δ respectively. For the fourth and seventh observations, item C occurs negatively while the corresponding

drift effects are $-\Delta$ and 5Δ. The overall effect can be represented by the sum of cross products of the columns for C and Δ, namely

$$[1](-3\Delta) + [-1](-\Delta) + [-1]\ (5\Delta) + [1]\ (7\Delta) = 0$$

using square brackets to denote the coefficient attached to the direction of the difference and parenthesis for the drift effect. For A, one has

$$[1](-7\Delta) + [-1](-5\Delta) + [-1](3\Delta) + [1](5\Delta) = 0$$

In general, if the cross products sum out to zero, then the drift effect is said to be completely "balanced out" or "orthogonal" to the items being measured.

7. Restraints

In calibration designs only differences between items are measured so that unless one or more of them are standards for which values are known, one cannot assign values for the remaining "unknown" items. Algebraically, one has a system of equations that is not of full rank and needs the value for one item or the sum of several items as the restraint to lead to a unique solution.

In the design of Section 6, for example, if one has a single standard and three unknowns, the standard can be assigned to any one of the letters. (The same would be true of three standards and one unknown.) If there are two standards and two unknowns, the choice of which pair of letters to assign for the standards is important in terms of minimizing the uncertainty in the unknown.

It turns out that the pairing of A with D or of B with C is slightly less efficient (see Appendix B) than the other pairings A with B or C with D. This results from the fact that the observation on the differences (A-D) and (B-C) are repeated and it is usually better (to achieve smaller variance for the test items) to measure differences between standards and unknowns than between pairs of standards.

8. Use of Calibration Design in Gage Block Calibration

The calibration design of Section 6 is used in gage block calibration at the National Bureau of Standards and the analysis and interpretation of the design for this application is representative of the principles involves in the use of the design in other applications.

At NBS two master sets of gage blocks are maintained for transferring length calibration to users gage blocks, these are designated A and B and their sum is designated by K. These are combined with two sets of unknowns, designated C and D. The difference (A-B) is used as the check standard.

If we denote the values determined for A B C and D by $\hat{A}$ $\hat{B}$ $\hat{C}$ $\hat{D}$ in accordance with the statisticians' practice of distinguishing the value from the experiment from the sought-after or long-run value, we may then write

$$\hat{A} = \frac{1}{24}(5y_1 - 2y_2 - y_3 - 2y_4 - 3y_5 - 2y_6 + 3y_7 + 2y_8) + \frac{K}{2}$$

$$\hat{B} = \frac{1}{24}(-5y_1 + 2y_2 + y_3 + 2y_4 + 3y_5 + 2y_6 - 3y_7 - 2y_8) + \frac{K}{2}$$

$$\hat{A-B} = \frac{1}{24}(10y_1 - 4y_2 - 2y_3 - 4y_4 - 6y_5 - 4y_6 + 6y_7 + 4y_8)$$

$$\hat{C} = \frac{1}{24}(-y_1 + 2y_2 + 5y_3 - 6y_4 - y_5 + 2y_6 - 7y_7 + 6y_8) + \frac{K}{2}$$

$$\hat{D} = \frac{1}{24}(y_1 + 6y_2 - 5y_3 - 2y_4 - 7y_5 + 6y_6 - y_7 + 2y_8) + \frac{K}{2}$$

where $\hat{A} + \hat{B}$ necessarily sum to K.

These values have the following standard deviations in terms of the long run precision as represented by the process standard deviation σ.

$$\text{s.d. } (\hat{A}) = \text{s.d. } (\hat{B}) = \sigma \sqrt{\frac{5}{48}}$$

$$\text{s.d. } (\hat{A-B}) = \sigma \sqrt{\frac{5}{12}}$$

$$\text{s.d. } (\hat{C}) = \text{s.d. } (\hat{D}) = \sigma \sqrt{\frac{13}{48}}$$

One also obtains values $\hat{\Delta}$ for Δ where

$$\hat{\Delta} = \frac{1}{168}(-7y_1 - 5y_2 - 3y_3 - y_4 + y_5 + 3y_6 + 5y_7 + 7y_8)$$

$$\text{s.d. } (\hat{\Delta}) = \sigma \sqrt{\frac{1}{168}}$$

Because this is an overdetermined system (more observations than unknowns) the deviation between observed and computed value is, in general, different from zero and reflects the random errors of measurement. These deviations, $d_1\ d_2 \cdots d_8$ are as follows:

$$d_1 = \frac{1}{168}(49y_1 - 7y_2 - 7y_3 + 21y_4 + 49y_5 + 49y_6 - 7y_7 + 21y_8)$$

$$d_2 = \frac{1}{168}(-7y_1 + 87y_2 + 13y_3 - 5y_4 + 33y_5 - 41y_6 + 53y_7 + 35y_8)$$

$$d_3 = \frac{1}{168}(-7y_1 + 13y_2 + 89y_3 + 25y_4 - 39y_5 + 37y_6 + 57y_7 - 7y_8)$$

$$d_4 = \frac{1}{168}(21y_1 - 5y_2 + 25y_3 + 111y_4 - 27y_5 + 3y_6 - 23y_7 + 63y_8)$$

$$d_5 = \frac{1}{168}(49y_1 + 33y_2 - 39y_3 - 27y_4 + 97y_5 + 25y_6 + 9y_7 + 21y_8)$$

$$d_6 = \frac{1}{168}(49y_1 - 41y_2 + 37y_3 + 3y_4 + 25y_5 + 103y_6 + 13y_7 - 21y_8)$$

$$d_7 = \frac{1}{168}(-7y_1 + 53y_2 + 57y_3 - 23y_4 + 9y_5 + 13y_6 + 73y_7 - 7y_8)$$

$$d_8 = \frac{1}{168}(21y_1 + 35y_2 - 7y_3 + 63y_4 + 21y_5 - 21y_6 - 7y_7 + 63y_8)$$

These deviations provide the information needed to obtain a value, s, which is the experiment's value for the process standard deviation, σ.

$$s = \sqrt{\frac{\Sigma(\text{dev})^2}{4}} \qquad \text{degrees of freedom} = 4$$

The number of degrees of freedom results from taking the number of observations less the number of unknowns then adding one (for the restraint) to give 8 - 5 + 1 = 4.

9. Example

Routine calibration of gage blocks is carried out with two NBS master blocks (designated S. and S..) and two test blocks (designated X and Y). The blocks are placed close together on a metal platen for a sufficiently long time to insure temperature equilibrium. A mechanical intercomparator is used to determine the difference between the blocks by first determining a reading for the block indicated by "+" then following with the block indicated by "-". The difference between these two readings is the observation, y (all values are in micro-inch). For a set of 0.101 in. blocks, the following data was obtained.

DATA FROM NBS CALIBRATION OF FOUR 0.101 INCH GAGE BLOCKS

i	Schedule	Difference Measured	First Reading	Second Reading	Difference y(i)	Deviation
1	+ - 0 0	S.-S..	52.0	52.5	-0.5	0.029
2	- 0 0 +	Y-S.	45.2	52.1	-6.9	-0.046
3	0 0 + -	X-Y	50.0	45.1	4.9	0.113
4	0 + - 0	S..-X	53.1	50.0	3.1	0.571
5	0 + 0 -	S..-Y	52.3	45.2	7.1	-0.238
6	- 0 0 +	Y-S.	45.1	52.0	-6.9	-0.079
7	+ 0 - 0	S.-X	52.0	50.1	1.9	-0.154
8	0 - + 0	X - S..	50.1	52.3	-2.2	0.304

S.+S.. = 6.4 used as restraint

S.-S.. = -0.133 used as check standard

σ = .32 accepted standard deviation

The values for the blocks and the drift effect, Δ, are

$$\hat{S}. = \frac{1}{24}[5(-0.5) - 2(-6.9) - (4.9) - 2(3.1) - 3(7.1) - 2(-6.9) + 3(1.9) + 2(-2.2)] + \frac{(6.4)}{2}$$

$$= \frac{1}{24}(-6.0) + 3.2 = 2.9500$$

$$\hat{S}.. = \frac{1}{24}[-5(-0.5) + 2(-6.9) + (4.9) + 2(3.1) + 3(7.1) + 2(-6.9) - 3(1.9) - 2(-2.2)] + \frac{(6.4)}{2}$$

$$= \frac{1}{24}(6.0) + 3.2 = 3.4500$$

$S_{\cdot}\hat{-}S_{\cdot\cdot} = -0.5$

$$\hat{X} = \frac{1}{24}[-(-0.5) + 2(-6.9) + 5(4.9) - 6(3.1) - (7.1) + 2(-6.9) - 7(1.9) + 6(-2.2)] + \frac{(6.4)}{2}$$

$$= \frac{1}{24}(-54.8) + 3.2 = 0.9167$$

$$\hat{Y} = \frac{1}{24}[(-0.5) + 6(-6.9) - 5(4.9) - 2(3.1) - 7(7.1) + 6(-6.9) - (1.9) + 2(-2.2)] + \frac{(6.4)}{2}$$

$$= \frac{1}{24}(-170.0) + 3.2 = -3.8833$$

$$\hat{\Delta} = \frac{1}{168}[-7(-0.5) - 5(-6.9) - 3(-4.9) - 1(3.1) + (7.1) + 3(-6.9) + 5(1.9) + 7(-2.27)$$

$$= \frac{1}{168}(.7) = 0.0042$$

The accepted standard deviation for the process is 0.32 μ-in so that one can compare the observed standard deviation, s,

$$s = \sqrt{\Sigma dev^2/4} = \sqrt{\frac{.5208}{4}} = 0.361$$

to the accepted value by computing

$$F = (\frac{s}{\sigma})^2 = \frac{0.1302}{0.1024} = 1.27$$

Had the ratio $(s/\sigma)^2$ exceeded 3.32 (the critical value for the 1% probability level of the F distribution), then the measurements would be regarded as being "out of control" and would be repeated. The other check on process performance is provided by the check standard for which the difference between $(S_{\cdot}\hat{-}S_{\cdot\cdot})$ and its accepted value should be less than 3 times the standard deviation of $(S_{\cdot}\hat{-}S_{\cdot\cdot})$. See Section 10 for a discussion of this test.

The drift term, $\hat{\Delta}$, has a standard deviation of $\sigma/\sqrt{168}$ or 0.025. The statistical significance of $\hat{\Delta}$ can be judged by forming the ratio $\frac{\hat{\Delta}}{\sigma/\sqrt{168}}$.

If this ratio exceeds 3, then $\hat{\Delta}$ would be regarded as significant. However, because the design has eliminated the effect of drift on the

values of the blocks, one would not be concerned about a "significant" $\hat{\Delta}$ unless it was greatly in excess of previously encountered values.

The deviations are computed as shown in Section 8, for example, for the deviation corresponding to y_8 is given by

$$(\text{dev})_8 = \frac{1}{168}[21(-0.5) + 35(-6.9) - 7(4.9) + 63(3.1) + 21(7.1) - 21(-6.9) -$$

$$7(1.9) + 63\ (-2.2)]$$

$$= \frac{1}{168}[51.1] = 0.304$$

10. Process Control

As mentioned in Section 2, continued monitoring of the measurement process is required to assure that predictions based on the accepted values for process parameters are still valid. For gage block calibration, the process is monitored for precision by comparison of the observed standard deviation to the accepted value, σ_w, by means of the F-test. In the case of the design of Section 6, the square of the ratio of the two standard deviations is compared to the critical value, $F(4, \infty, \alpha)$, which is the α probability point of the F distribution for degrees of freedom 4 and ∞. [For calibrations at NBS, α is chosen as .01 to give $F(4, \infty, .01) = 3.32$].

The check for systematic error is given by comparison of the observed value of the difference, S. - S.., between the two standards with its accepted value. The uncertainty of this difference is given by $\sigma_T = \sqrt{(5/12)\sigma_w^2 + 2\sigma_B^2}$ where σ_w is the "within run" standard deviation and σ_B is the component of variance arising from variations from run-to-run. The value of σ_T is obtained directly from the sequence of values of S. - S.. arising in regular calibrations. The check standard test is therefore,

$$t = \frac{\left|\text{observed } (S. - S..) - \text{accepted } (S. - S..)\right|}{\sigma_T} < 3$$

i.e., t is compared to the critical value 3.0 which would correspond to the .003 probability level for the normal distribution.

If both the "precision" (F-test) and "accuracy" (t-test) criteria are satisfied, the process is regarded as being "in control" and values for the unknowns, X and Y, and their associated uncertainties are regarded as valid. Failure on either criterion is an "out-of-control" signal and the measurements are repeated.

When the between run component, σ_B, is present, the standard deviation associated with the values for the unknowns are given by *

$$\sigma(S.) = \sigma(S..) = \sqrt{\frac{5}{48}\sigma_w^2 + \frac{1}{2}\sigma_B^2} = \frac{1}{2}\sigma_T$$

$$\sigma(X) = \sigma(Y) = \sqrt{\frac{13}{48}\sigma_w^2 + \frac{3}{2}\sigma_B^2} = \sqrt{\frac{3}{4}\sigma_T^2 - \frac{2}{48}\sigma_w^2}$$

The value for the drift serves as an indicator of possible trouble if it changes markedly from its usual range of values. However, because any linear drift is balanced out, a change in the value does not of itself vitiate the results.

*See M.C. Croarkin, "An Extended Error Model for Comparison Calibration" for an explanation.

If the uncertainty attached to the restraint value is not negligible, this will lead to a possible systematic error in all measurements based on this restraint. Therefore, as a bound to this error one should, for the design of section 6, add to the uncertainty from random error an allowance of one-half the uncertainty in the sum ($S_i + S_j$). This is shown in the computer example.

11. Computer Program

Appendix C lists a computer program in BASIC for carrying out the calculation for the gage block example. The program can be used with any design provided one has the arrays of coefficients for the determination of the values of the unknowns and the deviations corresponding to the two arrays given in Section 8 for the gage block example.

The program calls for input of:

a) Administrative data--designation of blocks, operator, date, etc.

b) Process parameters--standard deviations, value for check standard, etc.

c) Comparator readings

The computer programs provide in the output:

a) Deviations, s.d.

b) Values for unknowns, drift, and associated undertainties.

c) Statistical tests as to whether process can be regarded as "in control": on standard deviation and on value of check standard.

12. Other Designs for Elimination of Drift With Order of Measurement

The number of observations over which a linear drift could be expected to be valid varies with the type of measurement, but experience indicates that it is unusual if it is as large as 20. For all distinct pairings of n items $n(n-1)/2$ exceeds 20 for $n \geq 7$. The table below gives designs for n = 5, 6, 7 which are balanced for linear drift.

n = 5 10 Observations	n = 6 18 Observations	n = 7 21 Observations
+ - 0 0 0 1-2	+ - 0 0 0 0 1-2	+ - 0 0 0 0 0 1-2
0 + - 0 0 2-3	0 + - 0 0 0 2-3	0 + - 0 0 0 0 2-3
0 0 + - 0 3-4	0 0 + - 0 0 3-4	0 0 + - 0 0 0 3-4
0 0 0 + - 4-5	0 0 0 + 0 - 4-6	0 0 0 + - 0 0 4-5
- 0 0 0 + 5-1	0 0 0 0 - + 6-5	0 0 0 0 + - 0 5-6
	- 0 0 0 + 0 5-1	0 0 0 0 0 + - 6-7
- 0 0 + 0 4-1		- 0 0 0 0 0 + 7-1
0 + 0 - 0 2-4	0 0 + 0 0 - 3-6	
0 - 0 0 + 5-2	- 0 0 + 0 0 4-1	0 + 0 - 0 0 0 2-4
0 0 + 0 - 3-5	0 + 0 0 - 0 2-5	0 0 + 0 - 0 0 3-5
+ 0 - 0 0 1-3	0 0 - 0 0 + 6-3	0 0 0 + 0 - 0 4-6
	+ 0 0 - 0 0 1-4	0 0 0 0 + 0 - 5-7
	0 - 0 0 + 0 5-2	- 0 0 0 0 + 0 6-1
		0 - 0 0 0 0 + 7-2
	0 0 0 - + 0 5-4	+ 0 - 0 0 0 0 1-3
	0 0 - 0 0 + 6-3	
	0 - 0 + 0 0 4-2	0 0 - 0 0 0 + 7-3
	- 0 + 0 0 0 3-1	0 - 0 0 0 + 0 6-2
	0 + 0 0 - 0 2-5	- 0 0 0 + 0 0 5-1
	+ 0 0 0 0 - 1-6	0 0 0 + 0 0 - 4-7
		0 0 + 0 0 - 0 3-6
		0 + 0 0 - 0 0 2-5
		+ 0 0 - 0 0 0 1-4

An alternate form of displaying the design is shown for n = 5 and is used for the other designs.

Designs that involve a subset of all possible pairings are given below:

n = 6 12 Observ.	n = 7 14 Observ.	n = 8 16 Observ.	n = 9 18 Observ.
1-2	1-2	1-2	1-2
5-1	2-3	2-3	2-3
2-3	3-4	3-4	6-5
4-6	4-5	4-5	3-1
3-4	5-6	5-6	5-4
6-5	6-7	6-7	8-9
	7-1	7-8	4-7
2-4		8-1	9-6
4-5	3-1		7-8
6-2	5-3	4-1	
3-1	7-5	7-4	7-1
1-6	2-7	2-7	4-6
5-3	4-2	5-2	9-7
	6-4	8-5	1-4
	1-6	3-8	3-9
		6-3	5-8
		1-6	6-3
			2-5
			8-2

REFERENCES

1. Bose, R. C. & Cameron, J. M., The Bridge Tournament Problem and Calibration Designs for Comparing Pairs of Objects, NBS J. of Res. 69B (1965), 323-332.

2. Eisenhart, C., Realistic Evaluation of the Precision and Accuracy of Instrument Calibration Systems, NBS J. of Res. 67C (1963), 161-187.

3. Swindells, J. F., Calibration of Liquid-in-Glass Thermometers, NBS Monograph 90, GPO, 1965.

4. Zelen, M., Linear Estimation and Related Topics, Chapter 17 of Survey of Numerical Analysis edited by J. Todd, McGraw Hill, New York City (1962), 558-584.

APPENDIX A

Thermometer Calibration

Liquid in glass thermometers are calibrated at NBS in a controlled bath in which the temperature is increasing in a nearly linear fashion with time. The temperature of the bath is measured by platinum resistance thermometry at the beginning, middle, and end of a run with the test thermometers being read once in the first interval and again in reverse order in the second interval. The time sequence for the resistance measurements, R_1, R_2, R_3 and the two series of thermometer values denoted by T_i' and T_i are as follows:

$$R_1\ T_1'\ T_2'\ \cdots\ T_k'\ R_2\ T_k\ \cdots\ T_2\ T_1\ R_3$$

with uniform time intervals between the thermometer readings. Figure shows a schematic of the situation with the increment to the bath temperature being Δ for each time period except for the middle reading involving resistance thermometry where a step of α in temperature is assumed.

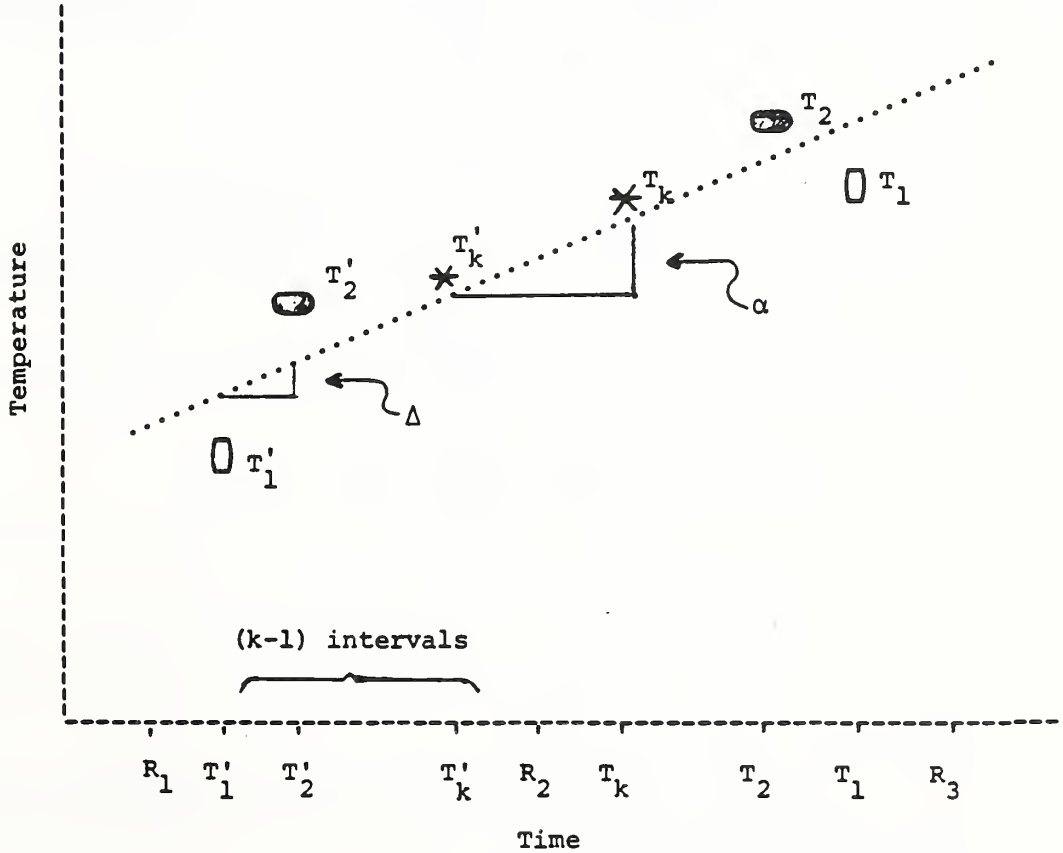


Figure Thermometer reading at fixed time intervals in a bath with linear drift.

The average $(T_i' + T_i)/2$ will be the indication of the i-th thermometer at the temperature implied by $(R_1 + R_2 + R_3)/3$. The differences, $d_i = T_i - T_i'$ will be a measure of $\alpha + 2(k - i)\Delta$ so that the observational equations may be written

$$E(d) = E\begin{bmatrix} T_1 - T_1' \\ T_2 - T_2' \\ \cdot \\ \cdot \\ T_{k-1} - T_{k-1}' \\ T_k - T_k' \end{bmatrix} = \begin{bmatrix} \alpha + 2(k-1)\Delta \\ \alpha + 2(k-2)\Delta \\ \cdot \\ \cdot \\ \alpha + 2\Delta \\ \alpha \end{bmatrix} = \begin{bmatrix} 1 & 2(k-1) \\ 1 & 2(k-2) \\ \cdot & \cdot \\ \cdot & \cdot \\ 1 & 2 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \alpha \\ \Delta \end{bmatrix} = X \begin{bmatrix} \alpha \\ \Delta \end{bmatrix}$$

where X stands for the indicated matrix, and where E( ) stands for the "expected value of," i.e., the limiting value if the effects of random error were eliminated.

The least squares estimates of α and Δ are given by the solution to the normal equations

$$(X'X)\begin{bmatrix} \alpha \\ \Delta \end{bmatrix} = X'd = \begin{bmatrix} \Sigma d \\ 2\Sigma d\ (k-i) \end{bmatrix}$$

where the inverse of the matrix of normal equations is

$$(X'X)^{-1} = \begin{bmatrix} k & k(k-1) \\ k(k-1) & 2k(k-1)(2k-1)/3 \end{bmatrix}^{-1} = \frac{1}{k(k^2-1)} \begin{bmatrix} 2(k-1)(2k-1) & -3(k-1) \\ -3(k-1) & 3 \end{bmatrix}$$

The estimates of α, Δ and σ^2, the variance of the observations are given by

$$\hat{\alpha} = \frac{2}{k(k+1)}\,[3\Sigma id - (k+1)\Sigma d]$$

$$\hat{\Delta} = \frac{3}{k(k^2-1)}\,[(k+1)\Sigma d - 2\Sigma id]$$

$$\hat{\sigma}^2 = \frac{1}{k-2}\,[\Sigma d^2 - \hat{\alpha}\Sigma d - 2\hat{\Delta}\Sigma(k-i)d] = \frac{1}{k-2}\,\Sigma(dev)^2$$

where $dev_i = d_i - \hat{\alpha} - 2(k-i)\hat{\Delta}$.

The standard deviation of the value for the test thermometer is $\sigma/\sqrt{2}$ and for α and Δ,

$$\text{s.d. } (\alpha) = \sigma\sqrt{2(2k-1)/k(k+1)}$$

$$\text{s.d. } (\Delta) = \sigma\sqrt{3/k(k^2-1)}$$

Control on the measurement process is maintained by two forms of redundancy--one to check on the process average and the other to check on process variability. The first of these is provided by incorporating an NBS standard thermometer among the k thermometers and requiring that its value be within random error of its accepted value. The variability check is given by comparing $\hat{\sigma}$ with the long run value established for the process. When these conditions are satisfied, then one can regard the process as being in a state of control.

A typical set of data for this type of calibration is given in the following table. For simplicity the resistance measurements have been suppressed and the temperature reported directly.

Calibration of Thermometers
Data From NBS Calibration of 22 August 1972
Provided by J. Wise, NBS, Thermometry Section

Thermometer	Observation	Averages	PRT - OBS = Correction at 40°
Reference (PRT)	39.9378		
T_1'	39.983	T_1 39.9870	-0.0436
T_2'	39.913	T_2 39.9150	0.0284
T_3'	39.966	T_3 39.9675	-0.0241
T_4' (check standard)	39.840	T_4 39.8410	0.1024*
Reference (PRT)	39.9422	PRT 39.9434	
T_4	39.842		
T_3	39.969		
T_2	39.917		
T_1	39.991	*accepted value is 0.1000	
Reference PRT	39.9501		

i	$d = T_i' - T_i$	α	Δ		Predicted d	dev.
1	-0.008	1	6	$\Sigma d = -0.077$	-0.0071	-0.0009
2	-0.004	1	4	$\Sigma id = -0.033$	-0.0052	0.0012
3	-0.003	1	2	$2/k(k+1) = 1/10$	-0.0033	0.0003
4	-0.002	1	0	$3/k(k^2-1) = 1/20$	-0.0014	-0.0006

$$\Sigma dev^2 = 0.00000270$$

$$\hat{\alpha} = \frac{1}{10}\,[3(-0.033) - 5(-0.017)] = -0.00140$$

$$\hat{\Delta} = \frac{1}{20}\,[5(-0.017) - 2(-0.033)] = -0.00095$$

$$\hat{\sigma} = \sqrt{\Sigma dev^2/2} = \sqrt{0.00000135} = 0.00115$$

s.d. $(\hat{\alpha}) = \sigma\sqrt{\frac{7}{10}}$ s.d. (average T) $= \sigma/\sqrt{2}$

s.d. $(\hat{\Delta}) = \sigma\sqrt{\frac{1}{20}}$

APPENDIX B

Least Squares Analysis of Calibration Designs

In this appendix the least squares analysis is presented in matrix form for those wishing to prepare a general analysis. Each formal statement will be illustrated by its application to the calibration design of Section 6.

It is assumed that the expected value of the observations represented in vector form as $y' = (y_1\ y_2\ \ldots\ y_n)$ have expected values $E(y) = X\beta$ where β is the vector of parameters and X is the design matrix. It is also assumed that the errors of measurement are uncorrelated and have equal variance, i.e., $V(y) = \sigma^2 I$.

For the design of section 5,

$$X = \begin{pmatrix} 1 & -1 & 0 & 0 & -7 \\ -1 & 0 & 0 & 1 & -5 \\ 0 & 0 & 1 & -1 & -3 \\ 0 & 1 & -1 & 0 & -1 \\ 0 & 1 & 0 & -1 & 1 \\ -1 & 0 & 0 & 1 & 3 \\ 1 & 0 & -1 & 0 & 5 \\ 0 & -1 & 1 & 0 & 7 \end{pmatrix} \qquad \beta = \begin{pmatrix} A \\ B \\ C \\ D \\ \Delta \end{pmatrix}$$

The matrix of normal equations is given by $(X'X)\beta = X'y$ which for calibration designs is not of full rank.

$$X'X\beta = \begin{pmatrix} 4 & -1 & -1 & -2 & 0 \\ -1 & 4 & -2 & -1 & 0 \\ -1 & -2 & 4 & -1 & 0 \\ -2 & -1 & -1 & 4 & 0 \\ 0 & 0 & 0 & 0 & 168 \end{pmatrix} \qquad \beta = X'y = \begin{pmatrix} 1 & -1 & 0 & 0 & 0 & -1 & 1 & 0 \\ -1 & 0 & 0 & 1 & 1 & 0 & 0 & -1 \\ 0 & 0 & 1 & -1 & 0 & 0 & -1 & 1 \\ 0 & 1 & -1 & 0 & -1 & 1 & 0 & 0 \\ -7 & -5 & -3 & -1 & 1 & 3 & 5 & 7 \end{pmatrix} y$$

In order to solve this system, a restraint in the form $h'\beta = K$ is imposed leading to the augmented equations (see Zelen [4]),

$$\begin{pmatrix} X'X & h \\ h' & 0 \end{pmatrix} \begin{pmatrix} B \\ \lambda \end{pmatrix} = \begin{pmatrix} X'y \\ K \end{pmatrix}$$

For the design as used in the calibration of gage blocks the restraint is that A + B = K, giving h' = (1 1 0 0 0) and the augmented equations are

$$\begin{pmatrix} 4 & -1 & -1 & -2 & 0 & 1 \\ -1 & 4 & -2 & -1 & 0 & 1 \\ -1 & -2 & 4 & -1 & 0 & 0 \\ -2 & -1 & -1 & 4 & 0 & 0 \\ 0 & 0 & 0 & 0 & 168 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \end{pmatrix} \quad \beta = \begin{pmatrix} X'y \\ K \end{pmatrix}$$

The solution for the parameter values $\hat{\beta}$ are

$$\begin{pmatrix} \hat{\beta} \\ \lambda \end{pmatrix} = \begin{pmatrix} X'X & h \\ h' & 0 \end{pmatrix}^{-1} \begin{pmatrix} X'y \\ K \end{pmatrix} = \begin{pmatrix} C & g \\ g' & 0 \end{pmatrix} \begin{pmatrix} X'y \\ K \end{pmatrix}$$

where C is the indicated KxK matrix arising in the inversion process n. For the example

$$\begin{pmatrix} \hat{\beta} \\ \lambda \end{pmatrix} = \frac{1}{336} \begin{pmatrix} 35 & -35 & -7 & 7 & 0 & 168 \\ -35 & 35 & 7 & -7 & 0 & 168 \\ -7 & 7 & 91 & 21 & 0 & 168 \\ 7 & -7 & 21 & 91 & 0 & 168 \\ 0 & 0 & 0 & 0 & 2 & 0 \\ 168 & 168 & 168 & 168 & 0 & 0 \end{pmatrix} \begin{pmatrix} X'y \\ K \end{pmatrix} = \frac{1}{168} \begin{pmatrix} 35 & -14 & -7 & -14 & -21 & -14 & 21 & 14 & 84 \\ -35 & 14 & 7 & 14 & 21 & 14 & -21 & -14 & 84 \\ -7 & 14 & 35 & -42 & -7 & 14 & -49 & 42 & 84 \\ 7 & 42 & -35 & -14 & -49 & 42 & -7 & 14 & 84 \\ -7 & -5 & -3 & -1 & 1 & 3 & 5 & 7 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} y \\ K \end{pmatrix}$$

for which $C = \frac{1}{336} \begin{pmatrix} 35 & -35 & -7 & 7 & 0 \\ -35 & 35 & 7 & -7 & 0 \\ -7 & 7 & 91 & 21 & 0 \\ 7 & -7 & 21 & 91 & 0 \\ 2 & 0 & 0 & 0 & 2 \end{pmatrix}$

The variances of the parameters are given by $C_{ii}\sigma^2$ and of linear functions, $\ell'\beta$, the variance is $\ell'C\ell\sigma^2$. For the example

$$V(\hat{A}) = V(\hat{B}) = C_{11}\sigma^2 = C_{22}\sigma^2 = 35\sigma^2/336 = 5\sigma^2/48$$

$$V(\hat{C}) = V(\hat{D}) = C_{33}\sigma^2 = C_{44}\sigma^2 = 91\sigma^2/336 = 13\sigma^2/48$$

$$V(\hat{A-B}) = (C_{11} + C_{22} - 2C_{12})\sigma^2 = 140\sigma^2/336 = 5\sigma^2/12$$

$$V(\hat{A+B}) = 0$$

$$V(\hat{C+D}) = (C_{33} + C_{44} + 2C_{34})\sigma^2 = 224\sigma^2/336 = 2\sigma^2/3$$

$$V(\hat{C-D}) = (C_{33} + C_{44} - 2C_{34})\sigma^2 = 140\sigma^2/336 = 5\sigma^2/24$$

$$V(\hat{\Delta}) = C_{55}\sigma^2 = 2\sigma^2/336 = \sigma^2/168$$

FOOTNOTE 1

If one had assigned the two standards to positions B and C instead of to A and B as was done, then one would be repeating the measurement of the difference (B-C), and of the difference (A-D). These differences are internal to the pair of standards over the pair of unknowns and one might suspect that they add little to the transfer from standard to test item. This is confirmed by examination of the inverse of the matrix of normal equations,

$$\begin{pmatrix} 4 & -1 & -1 & -2 & 0 & 0 \\ -1 & 4 & -2 & -1 & 0 & 1 \\ -1 & -2 & 4 & -1 & 0 & 1 \\ -2 & -1 & -1 & -14 & 0 & 0 \\ 0 & 0 & 0 & 0 & 168 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 \end{pmatrix}^{-1} = \frac{1}{168} \begin{pmatrix} 56 & 0 & 0 & 28 & 0 & 84 \\ 0 & 14 & -14 & 0 & 0 & 84 \\ 0 & -14 & 14 & 0 & 0 & 84 \\ 28 & 0 & 0 & 56 & 0 & 84 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 84 & 84 & 84 & 84 & 0 & 0 \end{pmatrix}$$

The variances for the standards are

$$V(\hat{B}) = V(\hat{C}) = 14\sigma^2/168 = \sigma^2/12$$

$$V(\hat{B-C}) = 56\sigma^2/168 = \sigma^2/3$$

which is smaller than for the restraint A + B = K.

However, for the test items the variances are

$$V(\hat{A}) = V(\hat{D}) = 56\sigma^2/168 = \sigma^2/3$$

which is larger than that with the restraint A + B = K for which the corresponding variance is $13\sigma^2/48$.

The estimate for the test item, A, is

$$\hat{A} = \frac{1}{24}[8y_1 - 4y_2 - 4y_3 - 4y_5 - 4y_6 + 8y_7] + \frac{K}{2}$$

which does not involve y_4 and y_8 which are measurements of the difference between the two standards, i.e., of B-C. Thus, there is a gain in efficiency in the calibration of the test block by using positions A and B for the standards, the efficiency factor being $(\sigma^2/3)/(13\sigma^2/48) = 16/13$.

FOOTNOTE 2

If there were but a single standard, A, the inverse of the matrix of normal equations would be

$$\begin{pmatrix} 4 & -1 & -1 & -2 & 0 & 1 \\ -1 & 4 & -2 & -1 & 0 & 0 \\ -1 & -2 & 4 & -1 & 0 & 0 \\ -2 & -1 & -1 & 4 & 0 & 0 \\ 0 & 0 & 0 & 0 & 168 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}^{-1} = \frac{1}{168} \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 168 \\ 0 & 70 & 42 & 28 & 0 & 168 \\ 0 & 42 & 70 & 28 & 0 & 168 \\ 0 & 28 & 28 & 56 & 0 & 168 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 168 & 168 & 168 & 168 & 0 & 0 \end{pmatrix}$$

The variances of the test items are

$$V(\hat{B}) = V(\hat{C}) = 70\sigma^2/168 = 5\sigma^2/12$$

$$V(\hat{D}) = 56\sigma^2/168 = \sigma^2/3$$

FOOTNOTE 3

If the sum of all four were taken as the restraint, the inverse of the matrix of normal equations would be

$$\begin{pmatrix} 4 & -1 & -1 & -2 & 0 & 1 \\ -1 & 4 & -2 & -1 & 0 & 1 \\ -1 & -2 & 4 & -1 & 0 & 1 \\ -2 & -1 & -1 & 4 & 0 & 1 \\ 0 & 0 & 0 & 0 & 168 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 \end{pmatrix}^{-1} = \frac{1}{336} \begin{pmatrix} 49 & -21 & -21 & -7 & 0 & 84 \\ -21 & 49 & -7 & -21 & 0 & 84 \\ -21 & -7 & 49 & -21 & 0 & 84 \\ -7 & -21 & -21 & 49 & 0 & 84 \\ 0 & 0 & 0 & 0 & 2 & 0 \\ 84 & 84 & 84 & 84 & 0 & 0 \end{pmatrix}$$

The variances of all four test items are the same

$$V(\hat{A}) = V(\hat{B}) = V(\hat{C}) = V(\hat{D}) = 49\sigma^2/336 = 7\sigma^2/48$$

```
C**********GAGE  BLOCK  CALIBRATION**********
C     THIS PROGRAM COMPUTES THE VALUES OF THE UNKNOWN GAGE BLOCKS,
C     AND PERFORMS TWO STATISTICAL TESTS, THE F-TEST AND THE T-TEST,
C     TO DETERMINE IF THE PROCESS IS IN CONTROL.
C*****************************************************************
C                          INPUT
C
C     THIS PROGRAM CALLS FOR A USER CREATED DATA FILE WHICH CONSISTS
C     OF THE FOLLOWING:
C     (1) (CS(I),I=1,3) - DATE,OBSERVER,INSTRUMENT
C     (2) K  - VALUE OF RESTRAINT(MICROINCHES)
C     (3) NO - NOMINAL SIZE OF TEST BLOCKS(INCHES)
C     (4) S5 - ACCEPTED VALUE OF CHECK STANDARD(MICROINCHES)
C     (5) B5 - ACCEPTED S.D. OF THE INSTRUMENT(MICROINCHES)
C     (6) B6 - ACCEPTED TOTAL S.D.
C     (7) U2 - UNCERTAINTY IN THE RESTRAINT
C     (8) X(I),Y(I) - OBSERVED READINGS (EIGHT PAIRS)
C*****************************************************************
C
C
C*****************************************************************
C
C     DATA VALUES WHICH ARE DETERMINED BY THE CALIBRATION DESIGN,
C     AND WHICH ARE STORED WITHIN THIS PROGRAM IN DATA STATEMENTS,
C     ARE AS FOLLOWS:
C     (1) A(I,J) - LEAST SQUARES COEFF TO COMPUTE THE UNKNOWNS
C     (2) M(I,J) - LEAST SQUARES COEFF TO COMPUTE THE DEVIATION
C     (3) E(I)   - VARIANCE FACTOR
C     (4) F(I)   - DRIFT VECTOR
C     (5) C1     - MATRIX DIVISOR
C     (6) GS(I)  - BLOCK DESIGNATION
C
C     OTHER VARIABLES:
C     (1) N  - NO. OF BLOCKS IN THE CALIBRATION (N = 4)
C     (2) G1 - NO. OF OBSERVATIONS (G1 = 8)
C     (3) F4 - F RATIO (CRITICAL VALUE FOR P = .01); (F4 = 3.32)
C     (4) A4 - NO. OF DEGREES OF FREEDOM (A4 = 4)
C*****************************************************************
C
C
C*****************************************************************
C                          OUTPUT
C
C     (1)   DATE,OBSERVER,INSTRUMENT
C     (2)   COMPARATOR READINGS
C     (3)   OBSERVED DIFFERENCES
C     (4)   DEVIATIONS
C     (5)   VALUES OF THE UNKNOWNS
C     (6)   OBSERVED STANDARD DEVIATION
C     (7)   STATISTICAL TESTS
C     (8)   UNCERTAINTY STATEMENT
C*****************************************************************
```

```
250 DIM C$(3),B(4,9),M(8,8),A(9),X(8),Y(8),G$(4)
255 DATA 35,-14,-7,-14,-21,-14,21,14,84
260 DATA -35,14,7,14,21,14,-21,-14,84
265 DATA -7,14,35,-42,-7,14,-49,42,84
270 DATA 7,42,-35,-14,-49,42,-7,14,84
275 DATA 49,-7,-7,21,49,49,-7,21
280 DATA -7,87,13,-5,33,-41,53,35
285 DATA -7,13,89,25,-39,37,57,-7
290 DATA 21,-5,25,111,-27,3,-23,63
295 DATA 49,33,-39,-27,97,25,9,21
300 DATA 49,-41,37,3,25,103,13,-21
305 DATA -7,53,57,-23,9,13,73,-7
310 DATA  21,35,-7,63,21,-21,-7,63
312 DATA -7,-5,-3,-1,1,3,5,7
315 DATA .31250,.31250,.145833,.145833
320 DATA 168
322 DATA S.,S..,X,Y
330 N=4
335 G1=8
340 F4=3.32
345 A4=4
350 R1=G1+1
355 * READ COEFFICIENTS USED TO COMPUTE VALUES OF THE BLOCKS
360 FOR I = 1,N
365 FOR J1 = 1,R1
370 READ B(I,J1)
375 NEXT J1
380 NEXT I
385 * READ COEFFICIENTS USED TO COMPUTE THE DEVIATIONS
390 FOR I = 1,G1
395 FOR J1= 1,G1
400 READ M(I,J1)
405 NEXT J1
410 NEXT I
411 * READ DRIFT VECTOR
412 FOR I = 1,G1
413 READ F(I)
414 NEXT I
415 * READ VARIANCE VECTOR
420 FOR I = 1,N
425 READ E(I)
430 NEXT I
435 * READ MATRIX DIVISOR
440 READ C1
441 * READ BLOCK DESIGNATIONS
442 FOR I = 1,N
443 READ G$(I)
444 NEXT I
```

```
445 * DEFINE USER DATA FILE = DATA1
450 FILES DATA1
455 * READ ADMINISTRATIVE DATA AND PROCESS PARAMETERS
460 READ#1,C$(1),C$(2),C$(3)
465 READ#1,K,N0,S5,B5,B6,M2
470 * READ COMPARATOR READINGS AND COMPUTE THEIR DIFFERENCES
471 * ALSO, COMPUTE DRIFT =D1, AND S.D.(DRIFT) = S1
473 D1=0
475 FOR I = 1,G1
480 READ#1,X(I),Y(I)
485 A(I)=X(I)-Y(I)
487 D1=D1+A(I)*F(I)
490 NEXT I
492 D1=D1/C1
493 S1=B5*(1./C1)↑.5
495 * SET A(9)= RESTRAINT
500 A(9)=K
505 * COMPUTE VALUES = V(I), S.D. = Z(I), AND UNCERTAINTY = C(I)
510 FOR I = 1,N
515 Y1=0
520 FOR J1=1,R1
525 Y1=Y1+B(I,J1)*A(J1)
530 NEXT J1
535 V(I)=Y1/C1
540 Z(I)=(B6↑2-B5↑2*E(I))↑.5
545 C(I)=3*Z(I)+.5*M2
550 NEXT I
555 * COMPUTE CHECK STANDARD
560 C5=V(1)-V(2)
565 * COMPUTE  THE DEVIATIONS AND THE OBSERVED S.D.
570 S0=0
575 FOR I = 1,G1
580 D0=0
585 FOR J1=1,G1
590 D0=D0+M(I,J1)*A(J1)
595 NEXT J1
600 D(I)=D0/C1
605 S0=S0+D(I)↑2
610 NEXT I
615 S=(S0/A4)↑.5
620 * PERFORM STATISTICAL TESTS
625 F=(S/B5)↑2
630 T=(C5-S5)/B6
```

```
640 PRINT,906
645 PRINT DATE  ,CS(1)
650 PRINT OBS.  ,CS(2)
655 PRINT INSTR.,CS(3)
660 PRINT,905
665 PRINT OBSERVATIONS
670 FOR I = 1,G1
675 PRINT X(I),Y(I)
680 NEXT I
685 *PRINT OBSERVED DIFFERENCES AND DEVIATIONS
690 PRINT,695
695 FMT //,X18,A(I),X8, DEV(I)
700 FOR I = 1,G1
705 PRINT,710,A(I),D(I)
710 FMT X14,F8.3,X4,F9.3
715 NEXT I
720 * PRINT VALUES OF THE BLOCKS
722 PRINT,723
723 FMT //,X58,UNCERTAINTY
725 PRINT,730
730 FMT X19,NOM.,X8,CORR.,X10,S. D.,X4,3(S.D.)*.5(S.E.)
735 FOR I =1,N
740 PRINT,745,GS(I),NO,V(I),Z(I),C(I)
745 FMT F12.6,F11.2,X7,F8.5,F15.5
750 NEXT I
755 * PRINT STATISTICAL INFORMATION
760 PRINT,765
765 PRINT OBS. S.D,ACC. S.D.,F TEST,F RATIO,D.F.
770 PRINT,775,S,B5,F,F4,A4
775 FMT F7.4,X5,F9.5,X3,F8.3,F12.2,I10
780 IF F>F4 THEN 790
785 GO TO 800
790 PRINT,795
795 PRINT  **********S. D. IS NOT IN CONTROL**********
800 PRINT,805
805 PRINT OBS. CHECK,ACC. CHECK,T TEST
810 PRINT,815,C5,S5,T
815 FMT F10.5,F11.5,F12.5
820 IF ABS(T) > 3 THEN 830
825 GO TO 840
830 PRINT,835
835 PRINT **********CHECK STANDARD IS NOT IN CONTROL**********
840 PRINT,905
845* PRINT DRIFT AND S.D.(DRIFT)
885 PRINT,890,D1,S1
890 FMT  DRIFT        = ,F10.4/ S.D.(DRIFT) = F10.4
895 PRINT,900,K
900 FMT ///, RESTRAINT (S.*S..)                    = ,F8.2
901 PRINT,903,M2
902 PRINT,906
903 FMT  SYSTEMATIC ERROR(S.E.) IN RESTRAINT = ,F8.2
905 FMT //
906 FMT //////
910 STOP
915 END
```

INPUT = USER DATA FILE

```
10 MAY 28 1974,HOWELL,FEDERAL 1
20 6.4,.101,-.133,.32,.49,.20
30 52.0,52.5
40 45.2,52.1
50 50.0,45.1
60 53.1,50.0
70 52.3,45.2
80 45.1,52.0
90 52.0,50.1
100 50.1,52.3
```

OUTPUT

```
DATE          MAY 28 1974
OBS.          HOWELL
INSTR.        FEDERAL 1

OBSERVATIONS
     52            52.5
     45.2          52.1
     50            45.1
     53.1          50
     52.3          45.2
     45.1          52
     52            50.1
     50.1          52.3

                  A(I)          DEV(I)
                  -.500          .029
                 -6.900         -.046
                  4.900          .113
                  3.100          .571
                  7.100         -.237
                 -6.900         -.079
                  1.900         -.154
                 -2.200          .304
```

	MEAN.	CORR.	S. D.	UNCERTAINTY 3(S.D.)*.5(S.E.)
S.	.101000	2.95	.45618	1.46854
S..	.101000	3.45	.45613	1.46854
X	.101000	.92	.47452	1.52355
Y	.101000	-3.88	.47452	1.52355

OBS. S.D	ACC. S.D.	F TEST	F RATIO	D.F.
.3607	.32000	1.271	3.32	4

OBS. CHECK	ACC. CHECK	T TEST
-.50000	-.13300	-.74898

```
DRIFT          =      .0042
S.D.(DRIFT)    =      .0247

RESTRAINT (S.*S..)                        =     6.40
SYSTEMATIC ERROR(S.E.) IN RESTRAINT       =      .20
```

Metrologia 26, 107–113 (1989)

metrologia

An Extended Error Model for Comparison Calibration

C. Croarkin

Statistical Engineering Division, National Institute of Standards and Technology*, Gaithersburg, MD 20899, USA

Received: September 10, 1988 and in revised form November 18, 1988

Abstract

The usual error model for calibration experiments is extended to situations where there are both short-term and long-term random errors of measurement. Such error models are useful where short-term errors are related to instrumentation, and long-term errors are related to operating procedures, environmental factors or changes in the artifacts themselves. The concept of a check standard is advanced for estimating variability and maintaining statistical control of the measurement process.

Introduction

Comparison calibration relates a characteristic of an artifact or instrument to the defined unit for the quantity of interest. A reference standard, whose value has been independently established, is the basis for assigning a value to the unknown artifact. For calibrations at the highest accuracy levels, very precise comparators with linear responses over a small on-scale range are used to quantify small differences between artifacts of the same nominal value. We describe an error model and analysis where two unknowns are compared with two reference standards according to a specific design.

Calibration Model

In the simplest case, an unknown X with value X^*, yet to be determined, is assumed to be related to a reference standard R with known value R^* by

$$X^* = \Lambda + R^*$$

where Λ is small but not necessarily negligible.

Given a measurement x on the unknown and a measurement r on the reference standard, the responses are assumed to be of the form

$$x = \eta + X^* + e_x$$

and (1)

$$r = \eta + R^* + e_r$$

where η is instrumental offset and e_x and e_r are independent random errors which come from a distribution with mean zero and standard deviation σ.

The value of Λ is estimated[1] by the difference Λ where

$$\Lambda = x - r \tag{2}$$

and the value assigned to the unknown artifact is based on the known value of the reference standard, R^*, called the restraint, according to

$$\boldsymbol{X}^* = \Lambda + R^* . \tag{3}$$

The standard deviation of this estimate, $\sigma_{\boldsymbol{X}}$, depends on the error structure for $\boldsymbol{X}^*$ which is of the form

$$\boldsymbol{X}^* = X^* + e_x - e_r \tag{4}$$

so that

$$\sigma_{\boldsymbol{X}} = \sqrt{2}\sigma . \tag{5}$$

Calibration Designs

A more complicated case involves the calibration of several unknowns, such as a weight set of various denominations or a group of voltage cells in a temperature-controlled enclosure, relative to a single reference standard or group of standards. Any difference measurements which compare unknowns and reference standards with one another and each other are candidates for the calibration procedure.

* Formerly, the U.S. National Bureau of Standards

[1] Boldface type is used to denote a least-squares estimate from the data such as Λ

A calibration design is a subset of all candidate measurements which admits a least-squares solution for the unknowns. The design is constructed to be parsimonious so as, on one hand, to minimize the number of measurements and, on the other hand, to give estimates with reasonably high precision. We recognize that precision depends on the number of measurements, and Grabe [1] has shown how precision depends on the construction of the design. As we show in this paper, precision can also be limited by other factors.

In the earliest references to designs by Hayford and Benoit [2, 3], the term "weighing design" is used to describe a sequence of measurements for calibrating a weight set. In papers published in the 1960s and 1970s, Bose and Cameron [4, 5] and Chakravarti and Suryanarayana [6] extend the theory and application of designs; Cameron and Eicke [7] solve a problem peculiar to electrical circuits; and Cameron and Hailes [8] discuss the situation where there is drift in the measurement process. Recent publications [9–12] show that designs now enjoy general acceptance in the calibration laboratory and are routinely used for the calibration of mechanical and electrical units of measurement, as well as for mass measurements.

Expanded Calibration Model

Throughout this development, the one constant assumption has been that random errors of measurement are independent and come from a single error distribution (such as the normal distribution). With more precise measurement systems, we are now able to identify situations where these assumptions are called into question and a more realistic model is needed. We find that random errors of measurement for a single design, which takes at most a few hours' time, are not of the same magnitude as errors which afflict the measurement process over the course of several designs or days [2]. Thus, we are forced to admit two error distributions, one that arises in the short-term and one that arises in the long term.

It is convenient to think in terms of short-term instrumental variations and long-term artifact changes caused by environmental conditions and the like. The latter are assumed to vary randomly from design to design and to be constant for a single design. The model in (1) is expanded to include both types of errors so that

$$\begin{aligned} x &= \eta + \{X^* + \delta_X\} + e_x \\ r &= \eta + \{R^* + \delta_R\} + e_r \end{aligned} \tag{6}$$

where e_x and e_r are short-term errors of (1), and δ_X and δ_R, which represent long-term changes associated with X and R, come from a distribution with mean zero and standard deviation σ_b.

The error structure of the estimate, $\hat{X}^*$, given by

$$\hat{X}^* = X^* + \delta_X - \delta_R + e_x - e_r \tag{7}$$

now contains both types of error terms, and the standard deviation $\sigma_{\hat{X}}$ becomes

$$\sigma_{\hat{X}} = (2\sigma_b^2 + 2\sigma^2)^{1/2}.$$

Application to Designs

Standard deviations associated with solutions to a design depend upon the error structures of the model. We illustrate with an example where two unknown artifacts X_1 and X_2 with unknown values X_1^* and X_2^* are calibrated relative to two reference standards R_1 and R_2 with values R_1^* and R_2^*. All items have the same nominal value. A design consisting of the six comparisons $d_1, \ldots, d_6$ that can be made among the four items, two at a time, can be represented as:

Obs	R_1	R_2	X_1	X_2
d_1	1	−1		
d_2	1		−1	
d_3	1			−1
d_4		1	−1	
d_5		1		−1
d_6			1	−1

The model that follows from this design is:

$$\begin{aligned} d_1 &= \{R_1^* + \delta_{R_1}\} - \{R_2^* + \delta_{R_2}\} + \varepsilon_1 \\ d_2 &= \{R_1^* + \delta_{R_1}\} - \{X_1^* + \delta_{X_1}\} + \varepsilon_2 \\ d_3 &= \{R_1^* + \delta_{R_1}\} - \{X_2^* + \delta_{X_2}\} + \varepsilon_3 \\ d_4 &= \{R_2^* + \delta_{R_2}\} - \{X_1^* + \delta_{X_1}\} + \varepsilon_4 \\ d_5 &= \{R_2^* + \delta_{R_2}\} - \{X_2^* + \delta_{X_2}\} + \varepsilon_5 \\ d_6 &= \{X_1^* + \delta_{X_1}\} - \{X_2^* + \delta_{X_2}\} + \varepsilon_6 \end{aligned} \tag{8}$$

[2] The statistical term for this phenomenon is components of error with the errors sometimes referred to as within-time and between-time random errors

The terms $\varepsilon_1, \ldots, \varepsilon_6$ represent random errors of measurement and the terms $\delta_{R_1}, \delta_{R_2}, \delta_{X_1}$, and δ_{X_2} represent random changes in the artifacts. It is assumed that the ε terms come from a distribution with mean zero and standard deviation σ_w and that the δ terms come from a distribution with mean zero and standard deviation σ_b. All random errors are assumed to be mutually independent.

The solution to the design depends on the restraint. If the restraint is taken to be the average of the reference standards or

$$R^* = \tfrac{1}{2}(R_1^* + R_2^*),$$

then least-squares estimates (see, for example, Cameron et al. [13]) are as follows:

$$\begin{aligned}
R_1^* &= \tfrac{1}{8}(\ \ 2d_1 + d_2 + d_3 - d_4 - d_5) + R^* \\
R_2^* &= \tfrac{1}{8}(-2d_1 - d_2 - d_3 + d_4 + d_5) + R^* \\
X_1^* &= \tfrac{1}{8}(\ \ -3d_2 - d_3 - 3d_4 - d_5 + 2d_6) + R^* \\
X_2^* &= \tfrac{1}{8}(\ \ -d_2 - 3d_3 - d_4 - 3d_5 - 2d_6) + R^*
\end{aligned} \tag{9}$$

We rewrite the solutions in terms of model (8) and collect error terms to obtain

$$\begin{aligned}
R_1^* &= R_1^* + \tfrac{1}{8}(\ \ 4\delta_{R_1} - 4\delta_{R_2} + 2\varepsilon_1 + \varepsilon_2 + \varepsilon_3 - \varepsilon_4 - \varepsilon_5) \\
R_2^* &= R_2^* + \tfrac{1}{8}(-4\delta_{R_1} + 4\delta_{R_2} - 2\varepsilon_1 - \varepsilon_2 - \varepsilon_3 + \varepsilon_4 + \varepsilon_5) \\
X_1^* &= X_1^* + \tfrac{1}{8}(-4\delta_{R_1} - 4\delta_{R_2} + 8\delta_{X_1} - 3\varepsilon_2 - \varepsilon_3 - 3\varepsilon_4 - \varepsilon_5 + 2\varepsilon_6) \\
X_2^* &= X_2^* + \tfrac{1}{8}(-4\delta_{R_1} - 4\delta_{R_2} + 8\delta_{X_2} - \varepsilon_2 - 3\varepsilon_3 - \varepsilon_4 - 3\varepsilon_5 - 2\varepsilon_6)
\end{aligned} \tag{10}$$

Associated standard deviations are found from (10) as follows[3]:

$$\sigma_{R_1} = \sigma_{R_2} = (\tfrac{1}{2}\sigma_b^2 + \tfrac{1}{8}\sigma_w^2)^{1/2}$$

and (11)

$$\sigma_{X_1} = \sigma_{X_2} = (\tfrac{3}{2}\sigma_b^2 + \tfrac{3}{8}\sigma_w^2)^{1/2}.$$

The structure of (11) indicates how precision depends on the relationship between the components of error. For all four estimates, the contribution to the total variance from σ_b^2 is four times larger than the contribution from σ_w^2; thus, the size of σ_b relative to σ_w determines to what extent precision is affected by the number of design points.

Check Standard

The quantity σ_b can only be estimated from many designs involving the same artifact. Because calibrations are usually performed on a one-time basis, the prerequisite data for this analysis does not usually exist on the unknown itself. Thus, we designate a check standard for this purpose, and values of the check standard from many designs provide the basis for estimating σ_b.

For designs involving two reference standards, we create a check standard based on the difference between the two reference standards. For the design of (8), this difference

$$C = R_1^* - R_2^* \tag{12}$$

which is independent of the restraint, has an error structure of the form

$$C = R_1^* - R_2^* + \tfrac{1}{8}(8\delta_{R_1} - 8\delta_{R_2} + 4\varepsilon_1 + 2\varepsilon_2 + 2\varepsilon_3 - 2\varepsilon_4 - 2\varepsilon_5). \tag{13}$$

with associated standard deviation

$$\sigma_C = (2\sigma_b^2 + \tfrac{1}{2}\sigma_w^2)^{1/2}. \tag{14}$$

Hence

$$\sigma_b = (\tfrac{1}{2}\sigma_C^2 - \tfrac{1}{4}\sigma_w^2)^{1/2} \tag{15}$$

and (11) can be reduced to

$$\sigma_{R_1} = \sigma_{R_2} = \tfrac{1}{2}\sigma_C$$

and (16)

$$\sigma_{X_1} = \sigma_{X_2} = \frac{\sqrt{3}}{2}\sigma_C.$$

Estimates of Standard Deviations from the Data

Given n designs with check standard values $C_1, \ldots, C_n$, the quantity σ_C is estimated with $(n-1)$ degrees of freedom by

$$s_C = \left(\frac{1}{n-1}\sum_{i=1}^{n}(C_i - \bar{C})^2\right)^{1/2} \tag{17}$$

where $\bar{C}$ is the average of the check standard values[4].

The standard deviation, σ_w, is estimated from a single design with $(m-k+1)$ degrees of freedom where m is the number of comparisons in the design; k is the number of artifacts; and the additional degree of freedom comes from the known value of the restraint. For the design given by (8), the standard deviation σ_w is estimated with three degrees of freedom by

$$s_w = \left(\frac{1}{3}\sum_{i=1}^{6}(d_i - \bar{d}_i)^2\right)^{1/2} \tag{18}$$

[3] These equations are valid where R^* is known without random error; see the section headed, "A Matrix Approach", for the case where R^* is subject to random error

[4] This method of estimating the standard deviation assumes that the check standard is not drifting over time

where d_i is the predicted value for each difference measurement from the design; i.e.,

$$
\begin{aligned}
d_1 &= R_1^* - R_2^* \\
d_2 &= R_1^* \qquad - X_1^* \\
d_3 &= R_1^* \qquad\qquad - X_2^* \\
d_4 &= \quad R_2^* - X_1^* \\
d_5 &= \quad R_2^* \qquad - X_2^* \\
d_6 &= \qquad\quad X_1^* - X_2^*
\end{aligned}
$$

We can improve the estimate of σ_w by pooling the standard deviations $s_{w_1}, \ldots, s_{w_n}$ from the n designs. The pooled value s_p, which has 3 n degrees of freedom, is computed as

$$s_p = \left(\frac{1}{n} \sum_{i=1}^{n} s_{w_i}^2 \right)^{1/2}. \tag{19}$$

For the purpose of making statements of precision or uncertainty the population standard deviations σ_w, σ_b and σ_C are replaced by their respective estimates in the appropriate equations.

Process Control

Two aspects of statistical process control are relevant in the calibration process. Short-term control for measurements constituting a single design depends on σ_w, and long-term control for calibrations over time depends on σ_b via check standard measurements. The latter depends upon reliable estimates from historical data for the mean, $\bar{C}$, and the standard deviation, s_C. For any new calibration, the check standard value, C, is tested for agreement with past data by a t statistic where

$$t = \frac{|C - \bar{C}|}{s_C}.$$

The process is judged to be in control if

$$t \leq t_{\alpha/2}(\nu)$$

where $t_{\alpha/2}(\nu)$ is the upper $\alpha/2$ percentage point of Student's t distribution [14] with ν degrees of freedom. Otherwise, the calibration is discarded.

Short-term control for each design is exercised by comparing the standard deviation from the design, s_w, with a pooled value s_p from historical data. An F statistic is computed as

$$F = s_w^2 / s_p^2.$$

Short-term precision is regarded as being in control if

$$F \leq F_\alpha(\nu_1, \nu_2)$$

where $F_\alpha(\nu_1, \nu_2)$ is the upper α percentage point of Snedecor's F distribution [15] with ν_1 degrees of freedom in s_w and ν_2 degrees of freedom in s_p. Failure to meet this condition is taken as an indication that precision has deteriorated, and the current calibration results are discarded.

Case Study From Mass Calibration

The National Institute of Standards and Technology (NIST) maintains about thirty check standards for mass calibrations. These check standards, which cover a variety of designs, load levels, and balances, constitute the data base for constructing uncertainties associated with mass calibrations and for implementing statistical control of the calibration process.

The data base, which covers the last twenty years of calibration history at NIST, is reviewed on an annual basis to update uncertainty statements and to expose any trends or anomalies in the process. Standard deviations from the designs, s_w, are pooled by balance. Standard deviations for each check standard, s_C, are estimated by (17).

Analysis confirms that the long-term component of error, s_b, is negligible for the mass-calibration process except at the critical kilogram level. The majority of mass calibrations at NIST start at the kilogram level using the design of (8) with the restraint as the average of two reference kilograms and a check standard C as defined by (12). Standard deviations for this process are shown in the table below.

Standard Deviations at the Kilogram Level

Source	Notation	Eq.	Std. dev.
Kg balance	s_p	(19)	0.0316 mg
Check standard	s_C	(17)	0.0277 mg
Long-term change	s_b	(15)	0.0116 mg
Unknowns	s_{X_1}, s_{X_2}	(16)	0.0240 mg

Weights other than kilograms are related to the NIST unit of mass via a hierarchy of designs where the restraint for each design is taken from the solution to the previous design. For example, at the kilogram level, the unknown X_2 is a group of weights totaling a kilogram; the group constitutes the starting restraint for the next design in the series. Thus, any random error that influences the value assigned to X_2 is propagated to all other weights.

Application to Other Designs

The standard deviation associated with a measurement must be defined on a design-by-design basis. A

matrix approach is outlined in the next section; also see Croarkin [16, 17] for specific formulations for a design involving two reference standards and three unknowns and a design involving four reference standards and four unknowns.

The problem of definition can sometimes be avoided by judicious choice of a check standard. If one chooses a check standard with the same error structure as the artifacts being calibrated, then the standard deviation for the check standard also applies to the calibrated artifacts. For example, if we make all ten comparisons among five artifacts of the same nominal value, where one artifact is a designated check standard, then the check standard will have the same error structure as the unknowns.

A Matrix Approach

A matrix approach is outlined for estimating components of variance for any measurement design where there are both short-term random errors of measurement and long-term random changes in the artifacts. We also allow for the situation where the restraint has been estimated from a previous experiment, and the random errors associated with that measurement process are taken into account.

Given m difference measurements among k artifacts, where some artifacts are regarded as reference standards and some are regarded as test items or unknowns, the model for the measurement process

$$D = A[X^* + \delta] + \varepsilon \tag{20}$$

is shown in terms of matrix elements. The elements and their respective dimensions are defined as follows:

- $\underset{(m \times 1)}{D}$ a matrix of difference measurements
- $\underset{(m \times k)}{A}$ a matrix of zeroes and ones such that a plus or minus one in the j^{th} position indicates that the j^{th} artifact is involved in the i^{th} comparison and a zero indicates the converse
- $\underset{(k \times 1)}{X^*}$ a matrix of unknown values for the k artifacts
- $\underset{(k \times 1)}{\delta}$ a matrix of random errors with zero mean and standard deviation σ_b
- $\underset{(m \times 1)}{\varepsilon}$ a matrix of random errors with zero mean and standard deviation σ_w

Because the matrix A has rank $(k-1)$, a solution for an unknown X^*, as shown by Zelen [18], is achieved by imposing upon the system a restraint, or known value for a linear combination of the artifacts. Let the scalar R^* be the restraint, and let $\underset{(1 \times k)}{\mathscr{L}_R'}$ be a vector of zeroes and ones such that a one in the j^{th} position indicates that the j^{th} artifact is in the restraint and a zero indicates the converse.

For example, the vector [5]

$$\underset{(1 \times k)}{\mathscr{L}_R'} = (1 \quad 1 \quad 0 \ldots 0)$$

indicates that the *restraint* R^* is the *summation* for the first two artifacts.

Then a solution can be found from an augmented matrix B where

$$\underset{(k+2) \times (k+2)}{B} = \begin{pmatrix} \underset{(k \times k)}{A'A} & \underset{(k \times 1)}{\mathscr{L}_R} & \underset{(k \times 1)}{A'D} \\ \underset{(1 \times k)}{\mathscr{L}_R'} & \underset{(1 \times 1)}{0} & \underset{(1 \times 1)}{R^*} \\ \underset{(1 \times k)}{0} & \underset{(1 \times 1)}{0} & \underset{(1 \times 1)}{-I} \end{pmatrix}$$

has an inverse of the form

$$\underset{(k+2) \times (k+2)}{B^{-1}} = \begin{pmatrix} \underset{(k \times k)}{Q} & \underset{(k \times 1)}{h} & \underset{(k \times 1)}{X^*} \\ \underset{(1 \times k)}{h'} & \underset{(1 \times 1)}{0} & \underset{(1 \times 1)}{\bullet} \\ \underset{(1 \times k)}{\bullet} & \underset{(1 \times 1)}{\bullet} & \underset{(1 \times 1)}{\bullet} \end{pmatrix}$$

and Q is the covariance matrix; X^* is the vector of estimates for the unknowns; and other entries ($\bullet$) are irrelevant for this application.

The deviations from the fit are given by the vector ζ where

$$\underset{(1 \times m)}{\zeta'} = [D - AX^*]'$$

and the standard deviation for the design σ_w is estimated by

$$s_w = \left(\frac{\zeta'\zeta}{m-k+1}\right)^{1/2}$$

with $m-k+1$ degrees of freedom.

It is now assumed that a check standard C is tracked for many applications of the same design over time. The estimated value of C for any particular design is given by

$$C = \mathscr{L}_C'[X^*]$$

where, for example,

$$\underset{(1 \times k)}{\mathscr{L}_C'} = (1 \quad -1 \quad 0 \ldots 0)$$

indicates that the check standard is the computed difference between the first and second artifacts.

[5] The mark (') indicates the transpose of a matrix

The standard deviation σ_b can be estimated from the relationship

$$\sigma_b^2 = \frac{\sigma_C^2 - \mathscr{L}_C'[Q]\,\mathscr{L}_C\,\sigma_w^2}{\mathscr{L}_C'[Q\,A'\,A]\,\mathscr{L}_C}$$

where σ_C and σ_w should be estimated from the data of several designs [6].

Now consider a single unknown X_j whose estimated value is

$$X_j^* = [\mathscr{L}_{X_j}]'[X^*] \tag{21}$$

where, for example,

$$\underset{(1\times k)}{\mathscr{L}_{X_j}'} = (0 \quad 1 \quad 0 \ldots 0)$$

signifies that X_j refers to the second artifact in the design. Then the appropriate standard deviation for X_j^* is given by

$$\sigma_{X_j} = ([\mathscr{L}_{X_j}]'[Q\,A'\,A][\mathscr{L}_{X_j}]\,\sigma_b^2 + [\mathscr{L}_{X_j}]'[Q][\mathscr{L}_{X_j}]\,\sigma_w^2)^{1/2} \tag{22}$$

and the standard deviation associated with any linear combination of the unknowns is computed in a similar fashion. At this stage we assume that R^* is known without random error. Eq. (25) is appropriate if this assumption is not valid.

Mass calibration is a special case because values are assigned to sets of weights covering several denominations of mass. All values are related to a starting restraint, such as a kilogram reference standard, by a series of interrelated designs. The first series includes as an unknown, a single weight or a summation of weights, which becomes the restraint for the following series and so on throughout the entire weight set. Thus, we must account for imprecision associated with restraints after the first series.

Let $\mathscr{L}_\Sigma$ be a (k × 1) vector that defines the unknown whose value will be used as the restraint in the next series; this out-going restraint has value

$$X_\Sigma = [\mathscr{L}_\Sigma]'[X^*]. \tag{23}$$

The standard deviation associated with this restraint is computed as

$$\sigma_{X_\Sigma} = ([\mathscr{L}_\Sigma]'[Q\,A'\,A]\,[\mathscr{L}_\Sigma]\,\sigma_b^2 + [\mathscr{L}_\Sigma]'[Q]\,[\mathscr{L}_\Sigma]\,\sigma_w^2)^{1/2}. \tag{24}$$

[6] See the discussion under "Check Standard" and Eqs. (16) and (18)

To account for weights of various denominations, let W be a vector of nominal values for the k weights of the second series so that

$$\underset{(1\times k)}{W'} = (W_1, \ldots, W_k).$$

Now we redefine the design matrix A and the restraint vector $\mathscr{L}_R$ for the next series and let

$$R^* = X_\Sigma$$

and

$$\sigma_R = \sigma_{X_\Sigma}.$$

The matrix B and its inverse B^{-1} follow accordingly. The $\mathscr{L}_{X_j}$ vectors are also redefined for the weights in the series so that estimates can be computed according to (21). Then the appropriate standard deviation for the j^{th} weight, X_j, is given by (25) which follows:

$$\sigma_{X_j} = \left[[\mathscr{L}_{X_j}][Q\,A'\,A][\mathscr{L}_{X_j}]\,\sigma_b^2 + [\mathscr{L}_{X_j}]'[Q][\mathscr{L}_{X_j}]\,\sigma_w^2 + \left(\frac{[\mathscr{L}_{X_j}]'W}{[\mathscr{L}_R]'W}\right)^2 \sigma_R^2\right]^{1/2} \tag{25}$$

Standard deviations for the check standard for this series and other combinations of weights are computed similarly. It is noted that the process standard deviations, σ_w and σ_b, depend on the balance and the denominations of weights calibrated in the series; thus, they should be estimated separately for each series.

The process is extended to the next series by redefining the vector $\mathscr{L}_\Sigma$ so that it identifies the out-going restraint whose value is given by (23). Then the standard deviation for this restraint is given by (26) which follows:

$$\sigma_{X_\Sigma} = \left[[\mathscr{L}_\Sigma]'[Q\,A'\,A]\,[\mathscr{L}_\Sigma]\,\sigma_b^2 + [\mathscr{L}_\Sigma]'[Q]\,[\mathscr{L}_\Sigma]\,\sigma_w^2 + \left(\frac{[\mathscr{L}_\Sigma]'W}{[\mathscr{L}_R]'W}\right)^2 \sigma_R^2\right]^{1/2} \tag{26}$$

The standard deviations given by (22) and (24) are appropriate for values estimated in the initial series of weighings where the starting restraint is a known value. For values assigned by subsequent series of weighings, the imprecision of the estimated restraint contributes a component to the total standard deviation. Thus, (25) and (26) are appropriate.

Concluding Remarks

The proposed error model is especially enlightening where short-term errors are related to instrumentation. Then long-term errors are the result of operating procedures or environmental changes which affect the artifacts over time but are reasonably constant in the short-term so as not to affect the standard deviation from the design. Thus, there is motivation for isolating the long-term component in order to ascer-

tain whether precision can be improved given current instrumentation.

Other models may prove more useful or descriptive for other situations. For example, for mass calibrations which deal with weights of the same nominal mass, it is reasonable to assume that random changes in the weights can be characterized by a single error distribution. However, for weights which are not of the same nominal mass, we would allow for errors proportional to mass or, perhaps, to surface area.

Finally, the analysis of the design for four artifacts demonstrates that improved precision cannot always be attained by increasing the number of measurements in the design. The relative magnitudes of σ_w and σ_b and their contribution to the total variance must be understood before one can improve precision.

Acknowledgments. The author wishes to thank Dr. Richard Davis of the National Institute of Standards and Technology, who is responsible for maintaining the NIST unit of mass, for his interest and advice on the subject at hand.

References

1. M. Grabe: Metrologia **14**, 143–146 (1978)
2. J. A. Hayford: *On the least-square adjustment of weighings*, U.S. Coast and Geodetic Survey Report for 1892, Appendix 10
3. J.-R. Benoît: *L'étalonnage des séries de poids*, Trav. Mém. BIPM vol. **13**, 1907
4. R. C. Bose, J. M. Cameron: J. Res. Natl. Bur. Stand., Sect. **B: 79**, 323–332 (1965)
5. R. C. Bose, J. M. Cameron: J. Res. Natl. Bur. Stand., Sect. **B: 71**, 149–160 (1967)
6. I. M. Chakravarti, K. V. Suryanarayana: J. Combin. Theory **13**, 426–431 (1964)
7. J. M. Cameron, W. G. Eicke: *Designs for the Surveillance of the Volt Maintained by a Small Group of Saturated Standard Cells*, Natl. Bur. Stand. (U.S.) Tech. Note 430 (1967)
8. J. M. Cameron, G. E. Hailes: *Designs for the Calibration of Small Groups of Standards in the Presence of Drift*, Natl. Bur. Stand. (U.S.) Tech. Note 844 (1974)
9. C. P. Reeve: *The Calibration of Indexing Tables by Subdivision*, Natl. Bur. Stand. (U.S.) report NBSIR 75-750 (1975)
10. J. M. Cameron, M. C. Croarkin, R. C. Raybold: *Designs for the Calibration of Standards of Mass*, Natl. Bur. Stand (U.S.) Tech. Note 952 (1977)
11. C. P. Reeve: *The Calibration of a Roundness Standard*, Natl. Bur. Stand. (U.S.) report NBSIR 79-1758 (1979)
12. C. P. Reeve: *The Calibration of Angle Blocks by Intercomparison*, Natl. Bur. Stand. (U.S.) report NBSIR 80-1967 (1980)
13. J. M. Cameron, M. C. Croarkin, R. C. Raybold: Ref. 10 (above), pp. 10–11
14. E. S. Pearson, H. O. Hartley (eds.): *Biometrika Tables for Statisticians, Vol. I* (Cambridge University Press, Cambridge 1956), p. 138
15. Ibid., pp. 157–163
16. M. C. Croarkin: *Measurement Assurance Programs, Part II: Development and Implementation*, Natl. Bur. Stand. (U.S.) Spec. Publ. 676-II (1985), pp. 54–58
17. Ibid., pp. 66–75
18. M. Zelen: *Linear Estimation and Related Topics*, in: *Survey of Numerical Analysis*, J. Todd (ed.) (McGraw Hill, New York 1962), pp. 563–565

NBSIR 74-587

THE USE OF THE METHOD OF LEAST SQUARES IN CALIBRATION

J. M. Cameron

Institute for Basic Standards
National Bureau of Standards
Washington, D. C. 20234

September 1974

U. S. DEPARTMENT OF COMMERCE, Frederick B. Dent, Secretary
NATIONAL BUREAU OF STANDARDS, Richard W. Roberts, Director

THE USE OF THE METHOD OF LEAST SQUARES IN CALIBRATION

by

J. M. Cameron

1. Introduction

When more than one measurement is made on the same quantity, we are accustomed to taking an average and we have the feeling that the result is "better" than any single value that might be chosen from the set. Exactly why the average should be better needs some justification and the fundamental step toward a general approach to the problem of measurement was taken by Thomas Simpson in 1755. In showing the advantage of taking an average of values arising from a number of probability distributions, "he took the bold step of regarding errors, not as individual unrelated happenings, but as properties of the measurement process itself . . . He thus opened the way to a mathematical theory of measurement based on the mathematical theory of probability" [3, page 29].

The taking of an average is a special case of the method of least squares for which the original justification by Lengendre in 1805 did not involve any probability considerations but was advanced as a convenient method for the combination of observations. It was Gauss who recognized that one could not arrive at a "best" value unless the probability distribution of the measurement errors were known. In 1798 he showed the optimality of the least squares values when the underlying distribution is normal and in 1821 showed that the method of least squares leads to values of the parameters which have minimum variance among all possible unbiased linear functions* of the observations regardless of the underlying distribution. It is this property that gives the method of least squares its position of dominance among methods of combination of observations.

In this paper the statistical concepts needed for the method of least squares will be stated as a prelude to the usual modern version of the Gauss theorem. The formation of the observational equations and the derivation of the normal equations are illustrated for several situations arising in calibration. The role of restraints in the solution of systems which are not of full rank is discussed. The results are presented in a form designed to facilitate computation.

*An example of a nonlinear function with smaller variance than the average (the "best" linear estimator) is given by the midrange for the rectangular distribution. The midrange (average of the largest and smallest observation) has variance $1/[2(N+1)(N+2)]$ when based on n measurements, whereas the average has variance $1/12N$. Thus if $N \geq 3$, the midrange is to be preferred.

2. The Physical and Statistical Model of an Experiment

In physics, one is familiar with the construction and interpretation of the physical model of an experiment. One has a substantial body of theory on which to base such a model and one need only consider the determination of length by interferometric measurements to remind oneself of the various elements involved: a defined unit, the apparatus, the procedure, the corrections for environmental factors, etc. One realization of the experiment leads to values for the quantities of interest.

But one realizes that a repetition of the experiment will lead to different values--differences for which the physical model does not provide corrections. One is thus confronted with the need for a statistical model to account for the variations encountered in a sequence of measurements. In building the statistical model, one is first faced with the issue of what is meant by a repetition of the experiment--many readings within a few minutes or *ab initio* determinations a week apart.

The objective is to describe the output of the physical process not only in terms of the physical quantities involved but also in terms of the random variation and systematic influences due to environmental, procedural, or instrumental factors in the experiment.

3. Equation of Expected Values of the Observation

If one measured the same quantity again and again to obtain the sequence

$$y_1, y_2, \cdots y_n \cdots$$

then if the process that generates these numbers is "in control," the long run average or *limiting mean*, μ, will exist. By "in control" one means that the values of y behave as random variables from a probability distribution (for a discussion of this topic, see Eisenhart [1]). This limiting mean, μ, is usually called the *expected value* of y designated by the operator E( ) so that the statement becomes in symbols $E(y) = \mu$. Because y is regarded as a random variable one can represent it as

$$y = \mu + \varepsilon$$

where ε is the random component that follows some probability distribution with a limiting mean of zero, i.e., $E(\varepsilon) = 0$.

The quantity μ may involve one or more parameters. Consider the measurement of the difference in length of all distinct pairings of

four gage blocks, A, B, C, D. Denote the 6 measurements by $y_1, y_2, \ldots y_6$, then one may write

$$E(y_1) = A\text{-}B$$
$$E(y_2) = A\text{-}C$$
$$E(y_3) = A\text{-}D$$
$$E(y_4) = B\text{-}C$$
$$E(y_5) = B\text{-}D$$
$$E(y_6) = C\text{-}D$$

Other representations are useful.

Observation	Expected Value: E(y)	Matrix Form: $X\beta$
y_1	A - B	$\begin{bmatrix} 1 & -1 & 0 & 0 \\ 1 & 0 & -1 & 0 \\ 1 & 0 & 0 & -1 \\ 0 & 1 & -1 & 0 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} A \\ B \\ C \\ D \end{bmatrix}$
y_2	A - C	
y_3	A - D	
y_4	B - C	
y_5	B - D	
y_6	C - D	

Consider a sequence of measurements of the same quantity in the presence of a linear drift of Δ per observation. The expected values are thus:

Observation	Matrix Form: $X\beta$
$E(y_1) = \mu$	$\begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ \vdots & \vdots \\ 1 & (n-1) \end{bmatrix} \begin{bmatrix} \mu \\ \Delta \end{bmatrix}$
$E(y_2) = \mu + \Delta$	
$E(y_3) = \mu + 2\Delta$	
.	
.	
.	
$E(y_n) = \mu + (n-1)\Delta$	

There is an alternative representation that measures the drift from the central point of the experiment so that the drift is represented by . . . -3Δ, -2Δ, $-\Delta$, 0, Δ, 2Δ, 3Δ . . . for an odd number of observations and by . . . $\frac{-5\Delta}{2}$, $\frac{-3\Delta}{2}$, $\frac{-\Delta}{2}$, $\frac{\Delta}{2}$, $\frac{3\Delta}{2}$, $\frac{5\Delta}{2}$ . . . for an even number of observations.

If, as for example with some gage blocks, the value changes approximately linearly with time; then one can represent the observation as follows:

Expected Value $E(y)$	Matrix Form: $X\beta$
$E(y_1) = \alpha + \beta x_1$	$\begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix}$
$E(y_2) = \alpha + \beta x_2$	
$E(y_n) = \alpha + \beta x_n$	

The sequence of measurements for the intercomparison of 4 gage blocks is as follows:

Observation	Expected Value: $E(y)$	Matrix Form: $X\beta$
y_1	S. - S.. - $7\Delta/2$	$\begin{bmatrix} 1 & -1 & 0 & 0 & -7 \\ -1 & 0 & 0 & 1 & -5 \\ 0 & 0 & 1 & -1 & -3 \\ 0 & 1 & -1 & 0 & -1 \\ 0 & 1 & 0 & -1 & 1 \\ -1 & 0 & 0 & 1 & 3 \\ 1 & 0 & -1 & 0 & 5 \\ 0 & -1 & 1 & 0 & 7 \end{bmatrix} \begin{bmatrix} S. \\ S.. \\ X \\ Y \\ \Delta/2 \end{bmatrix}$
y_2	Y - S. - $5\Delta/2$	
y_3	X - Y - $3\Delta/2$	
y_4	S..- X - $\Delta/2$	
y_5	S..- Y + $\Delta/2$	
y_6	Y - S. + $3\Delta/2$	
y_7	S. - X + $5\Delta/2$	
y_8	X - S.. + $7\Delta/2$	

(Note that for simplicity, $\Delta/2$ is regarded as the parameter.)
For a detailed analysis of this and related experimental arrangements, see J. M. Cameron and G. E. Hailes [1]. The notation is that used in [1] where S. and S.. refer to reference standards and X and Y are the objects being calibrated.

If, as often occurs in the intercomparison of electrical standards, the comparator has a left-right polarity effect, this can be represented as an additive effect, α, as shown below for the intercomparison of 5 standards.

Observation	Expected Value: E(y)	Matrix Form: $X\beta$
y_1	$A - B + \alpha$	$\begin{bmatrix} 1 & -1 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} A \\ B \\ C \\ D \\ E \\ \alpha \end{bmatrix}$
y_2	$B - C + \alpha$	$\begin{bmatrix} 0 & 1 & -1 & 0 & 0 & 1 \end{bmatrix}$
y_3	$C - D + \alpha$	$\begin{bmatrix} 0 & 0 & 1 & -1 & 0 & 1 \end{bmatrix}$
y_4	$D - E + \alpha$	$\begin{bmatrix} 0 & 0 & 0 & 1 & -1 & 1 \end{bmatrix}$
y_5	$-A + E + \alpha$	$\begin{bmatrix} -1 & 0 & 0 & 0 & 1 & 1 \end{bmatrix}$
y_6	$-A + D + \alpha$	$\begin{bmatrix} -1 & 0 & 0 & -1 & 0 & 1 \end{bmatrix}$
y_7	$B - D + \alpha$	$\begin{bmatrix} 0 & 1 & 0 & -1 & 0 & 1 \end{bmatrix}$
y_8	$-B + E + \alpha$	$\begin{bmatrix} 0 & -1 & 0 & 0 & 1 & 1 \end{bmatrix}$
y_9	$C - E + \alpha$	$\begin{bmatrix} 0 & 0 & 1 & 0 & -1 & 1 \end{bmatrix}$
y_{10}	$A - C + \alpha$	$\begin{bmatrix} 1 & 0 & -1 & 0 & 0 & 1 \end{bmatrix}$

4. Statistical Independence

The sequence of differences from a zero measurement, y_0,

$$A: \quad y_1-y_0,\ y_2-y_0,\ y_3-y_0, \ldots, y_n-y_0, \ldots$$

are clearly dependent because an error in y_0 will be common to all. Similarly, the successive differences

$$B: \quad y_2-y_1,\ y_3-y_2, \ldots, y_n-y_{n-1}, \ldots$$

will be correlated in pairs because an error in y_n affects both the (n-1)st and n-th difference.

If it is assumed in both cases that each y_i has the form $\mu_i = \mu_i + \varepsilon_i$ where $E(\varepsilon_i) = 0$, $Var(\varepsilon_i) = \sigma^2$ and $cov(\varepsilon_i, \varepsilon_j) = 0$, then the variance of the differences for sequence A is, as one would expect,

$$V(y_i - y_0) = 2\sigma^2$$

and the covariance of two differences is

$$cov(y_i - y_0, y_j - y_0) = E[(\varepsilon_i - \varepsilon_0)(\varepsilon_j - \varepsilon_0)] = E(\varepsilon_0^2) = \sigma^2$$

because terms of the form $E(\varepsilon_i, \varepsilon_j) = 0$

For sequence B the variance is also $V(y_i - y_{i-1}) = 2\sigma^2$ and the covariance terms are

$$cov(y_i - y_{i-1}, y_j - y_{j-1}) = E[(\varepsilon_i - \varepsilon_{i-1})(\varepsilon_j - \varepsilon_{j-1})] = \begin{cases} 0 & \text{if } |i-j| \geq 2 \\ -\sigma^2 & \text{if } |i-j| = 1 \end{cases}$$

These variance-covariance relationships can be represented in matrix form:

Sequence A: $V = \begin{bmatrix} 2 & 1 & 1 & \cdots & 1 \\ 1 & 2 & 1 & \cdots & 1 \\ \vdots & & & & \\ 1 & 1 & 1 & \cdots & 2 \end{bmatrix} \sigma^2$ Sequence B: $V = \begin{bmatrix} 2 & -1 & 0 & 0 & \cdots & 0 \\ -1 & 2 & -1 & 0 & \cdots & 0 \\ \vdots & & & & & \\ 0 & 0 & 0 & 0 & \cdots & 2 \end{bmatrix} \sigma^2$

All are familiar with the phenomenon of much closer agreement among measurements taken immediately after each other when compared to a sequence of values taken days or weeks apart. The simplest statistical model for this case is that each day has its own limiting mean, $\mu_j = \mu + \delta_j$, where $E(\delta_j) = 0$, $Var(\delta_j) = \sigma_B^2$, $Cov(\delta_i, \delta_j) = 0$, and the successive values on each day have the form

$$y_{ij} = \mu_i + \varepsilon_{ij} = \mu + \delta_i + \varepsilon_{ij}$$

where $E(\varepsilon_{ij}) = 0$, $Var(\varepsilon_{ij}) = \sigma_w^2$, $Cov(\varepsilon_{ij}, \varepsilon_{k\ell}) = 0$, and $Cov(\varepsilon_{ij}, \delta_k) = 0$.

These three examples serve to illustrate the point that the physical conduct of the experiment is the essential element in dictating the appropriate statistical analysis. In all three cases the correlation among the variables vitiates the usual formula: standard deviation of the mean = $(1/\sqrt{n})$ standard deviation. (See Appendix, Section 1(b).)

It is in the physical conduct of the experiment that one has to build in the independence of the measurements. For Sequence A one could remeasure the zero setting each time or in Sequence B, make an independent duplicate measurement. Ordinarily this is too much of an expense to pay to achieve uncorrelated variables just for a simpler analysis.

Statistical independence is to be desired in the sense that if the successive measurements are highly correlated, then many measurements are only slightly better than a single one. The really important issue is that the proper statistical model be used so that the results are valid.

5. Normal Equations For the Method of Least Squares (independent random variables)

When there are more observations than parameters, the "best" (in the sense of minimum variance) linear unbiased estimates for the parameters are given by the so-called least squares estimators. For example, assume one has the problem of deriving values for A, B, C, and D from the following measurements.

Measurements	Expected Value: E(y)	Matrix Form: $X\beta$
y_1	A	$\begin{bmatrix}1&0&0&0\end{bmatrix}\begin{bmatrix}A\\B\\C\\D\end{bmatrix}$
y_2	B	$\begin{bmatrix}0&1&0&0\end{bmatrix}$
y_3	C	$\begin{bmatrix}0&0&1&0\end{bmatrix}$
y_4	D	$\begin{bmatrix}0&0&0&1\end{bmatrix}$
y_5	A + B	$\begin{bmatrix}1&1&0&0\end{bmatrix}$
y_6	B + C	$\begin{bmatrix}0&1&1&0\end{bmatrix}$
y_7	C + D	$\begin{bmatrix}0&0&1&1\end{bmatrix}$
y_8	D + A	$\begin{bmatrix}1&0&0&1\end{bmatrix}$

An obvious estimator, $\tilde{A}$, is the average of the three values,

	Expected Value
y_1	A
y_5-y_2	(A+B)-B
y_8-y_4	(A+D)-D

so that, assuming independent measurements with variance, σ^2,

$$\tilde{A} = \frac{1}{3}(y_1 + y_5 - y_2 + y_8 - y_4)$$

$$\text{Var}(\tilde{A}) = \frac{5}{9}\sigma^2$$

The least squares estimator is obtained by forming the normal equations (see Appendix, Section 2).

$$3A + B + \quad + D = y_1 + y_5 + y_8$$

$$A + 3B + C \quad = y_2 + y_6 + y_5$$

$$B + 3C + D = y_3 + y_7 + y_6$$

$$A \quad + C + 3D = y_4 + y_8 + y_7$$

The solution gives the following estimators for the parameters.

$$\hat{A} = (7y_1 - 3y_2 + 2y_3 - 3y_4 + 4y_5 - y_6 - y_7 + 4y_8)/15$$

$$\hat{B} = (-3y_1 + 7y_2 - 3y_3 + 2y_4 + 4y_5 + 4y_6 - y_7 - y_8)/15$$

$$\hat{C} = (2y_1 - 3y_2 + 7y_3 - 3y_4 - y_5 + 4y_6 + 4y_7 - y_8)/15$$

$$\hat{D} = (-3y_1 + 2y_2 - 3y_3 + 7y_4 - y_5 - y_6 + 4y_7 + 4y_8)/15$$

Using formula (1.11) of Appendix, gives

$$\text{Var}(\hat{A}) = 105\sigma^2/225 = 21\sigma^2/45 = 7\sigma^2/15$$

which can be compared to the variance of $\tilde{A}$ which was $25\sigma^2/45$. The Gauss theorem on least squares guarantees that no other linear unbiased estimator will have smaller variance.

In matrix form one has

$$(X'X)\hat{\beta} = \begin{bmatrix} 3 & 1 & 0 & 1 \\ 1 & 3 & 1 & 0 \\ 0 & 1 & 3 & 1 \\ 1 & 0 & 1 & 3 \end{bmatrix} \begin{bmatrix} A \\ B \\ C \\ D \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ y_8 \end{bmatrix}$$

$$\hat{\beta} = \frac{1}{15} \begin{bmatrix} 7 & -3 & 2 & -3 \\ -3 & 7 & -3 & 2 \\ 2 & -3 & 7 & -3 \\ -3 & 2 & -3 & 7 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 \end{bmatrix} y$$

$$\hat{\beta} = \frac{1}{15} \begin{bmatrix} 7 & -3 & 2 & -3 & 4 & -1 & -1 & 4 \\ -3 & 7 & -3 & 2 & 4 & 4 & -1 & -1 \\ 2 & -3 & 7 & -3 & -1 & 4 & 4 & -1 \\ -3 & 2 & -3 & 7 & -1 & -1 & 4 & 4 \end{bmatrix} y$$

When only differences among a group of objects (such as gage blocks, voltage cells, etc.) are measured the normal equation will not be of full rank so that a unique solution will not exist. For the design involving differences between all distinct pairings of objects the normal equations are, for the case of 4 objects discussed in Section 3,

$$3A - B - C - D = y_1 + y_2 + y_3 = q_1$$

$$-A + 3B - C - D = -y_1 + y_4 + y_5 = q_2$$

$$-A - B + 3C - D = -y_2 - y_4 + y_6 = q_3$$

$$-A - B - C + 3D = -y_3 - y_5 - y_6 = q_4$$

Or in matrix form:

$$X'X\hat{\beta} = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 1 & 1 & 0 \\ 0 & -1 & 0 & -1 & 0 & 1 \\ 0 & 0 & -1 & 0 & -1 & -1 \end{bmatrix} \begin{bmatrix} 1 & -1 & 0 & 0 \\ 1 & 0 & -1 & 0 \\ 1 & 0 & 0 & -1 \\ 0 & 1 & -1 & 0 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & -1 \end{bmatrix} \beta = \begin{bmatrix} 3 & -1 & -1 & -1 \\ -1 & 3 & -1 & -1 \\ -1 & -1 & 3 & -1 \\ -1 & -1 & -1 & 3 \end{bmatrix} \beta = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 1 & 1 & 0 \\ 0 & -1 & 0 & -1 & 0 & 1 \\ 0 & 0 & -1 & 0 & -1 & -1 \end{bmatrix} y$$

which can be seen not to be of full rank because the sum of the four equations is zero.

One needs a baseline to which the differences can be referred--a restraint to bring the system of equations up to full rank. If one of the objects were designated as the standard, or if a number (or all) of them were regarded as a reference group whose value was known, values for the items could be obtained.

If the restraint $A = K_o$ is invoked, the normal equations become (using the methods of Appendix, Section 3)

$$\begin{aligned} 3A - B - C - D + \lambda &= q_1 \\ -A + 3B - C - D &= q_2 \\ -A - B + 3C - D &= q_3 \\ -A - B - C + 3D &= q_4 \\ A &= K_o \end{aligned} \qquad \begin{bmatrix} 3 & -1 & -1 & -1 & 1 \\ -1 & 3 & -1 & -1 & 0 \\ -1 & -1 & 3 & -1 & 0 \\ -1 & -1 & -1 & 3 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{\lambda} \end{bmatrix} = \begin{bmatrix} X'y \\ K_o \end{bmatrix}$$

The solution is given by

$$\begin{aligned} \hat{A} &= K \\ \hat{B} &= K+(-2y_1-y_2-y_3+y_4+y_5)/4 \\ \hat{C} &= K+(-y_1-2y_2-y_3-y_4+y_6)/4 \\ \hat{D} &= K+(-y_1-y_2-2y_3-y_5-y_6)/4 \\ \lambda &= 0 \end{aligned} \qquad \begin{bmatrix} \hat{\beta} \\ \hat{\lambda} \end{bmatrix} = \frac{1}{4} \begin{bmatrix} 0 & 0 & 0 & 0 & 4 \\ 0 & 2 & 1 & 1 & 4 \\ 0 & 1 & 2 & 1 & 4 \\ 0 & 1 & 1 & 2 & 4 \\ 4 & 4 & 4 & 4 & 0 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & -1 & 0 & -1 & 0 & 1 & 0 \\ 0 & 0 & -1 & 0 & -1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} y \\ K_o \end{bmatrix}$$

$$\begin{bmatrix}\hat{\beta}\\ \hat{\lambda}\end{bmatrix} = \frac{1}{4}\begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 4\\ -2 & -1 & -1 & 1 & 1 & 0 & 4\\ -1 & -2 & -1 & -1 & 0 & 1 & 4\\ -1 & -1 & -2 & 0 & -1 & -1 & 4\\ 0 & 0 & 0 & 0 & 0 & 0 & 0\end{bmatrix}\begin{bmatrix} y\\ K_o\end{bmatrix}$$

The variances of the values are $V(\hat{A}) = 0$; $V(\hat{B}) = V(\hat{C}) = V(\hat{D}) = \sigma^2/2$.

If the restraint $A + B + C + D = K_1$ is invoked, the normal equations become

$$\begin{aligned} 3A - B - C - D + \lambda &= q_1\\ -A + 3B - C - D + \lambda &= q_2\\ -A - B + 3C - D + \lambda &= q_3\\ -A - B - C + 3D + \lambda &= q_4\\ A + B + C + D &= K_1 \end{aligned} \qquad \begin{bmatrix} 3 & -1 & -1 & -1 & 1\\ -1 & 3 & -1 & -1 & 1\\ -1 & -1 & 3 & -1 & 1\\ -1 & -1 & -1 & 3 & 1\\ 1 & 1 & 1 & 1 & 0\end{bmatrix}\begin{bmatrix}\hat{\beta}\\ \hat{\lambda}\end{bmatrix} = \begin{bmatrix} X'y\\ K_1\end{bmatrix}$$

and the solution is given by

$$\hat{A} = (y_1+y_2+y_3+K_1)/4$$
$$\hat{B} = (-y_1+y_4+y_5+K_1)/4$$
$$\hat{C} = (-y_2-y_4+y_6+K_1)/4$$
$$\hat{D} = (-y_3-y_5-y_6+K_1)/4$$
$$\lambda = 0$$

$$\begin{bmatrix}\hat{\beta}\\ \hat{\lambda}\end{bmatrix} = \frac{1}{16}\left[\begin{bmatrix} 3 & -1 & -1 & -1 & 4\\ -1 & 3 & -1 & -1 & 4\\ -1 & -1 & 3 & -1 & 4\\ -1 & -1 & -1 & 3 & 4\\ 4 & 4 & 4 & 4 & 0\end{bmatrix}\right]\begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0\\ -1 & 0 & 0 & 1 & 1 & 0 & 0\\ 0 & -1 & 0 & -1 & 0 & 1 & 0\\ 0 & 0 & -1 & 0 & -1 & -1 & 0\\ 0 & 0 & 0 & 0 & 0 & 0 & 1\end{bmatrix}\begin{bmatrix} y\\ K_1\end{bmatrix}$$

$$= \frac{1}{16}\begin{bmatrix} 4 & 4 & 4 & 0 & 0 & 0 & 4\\ -4 & 0 & 0 & 4 & 4 & 0 & 4\\ 0 & -4 & 0 & -4 & 0 & 4 & 4\\ 0 & 0 & -4 & 0 & -4 & -4 & 4\\ 0 & 0 & 0 & 0 & 0 & 0 & 0\end{bmatrix}\begin{bmatrix} y\\ K_1\end{bmatrix}$$

The variances of the values are $V(\hat{A}) = V(\hat{B}) = V(\hat{C}) = V(\hat{D}) = 3\sigma^2/16$.

Although it is a simple matter to change the reference point for the parameters (i.e., change the restraint) after one solution has been found, the corresponding change of variances for the parameter values should not be ignored. These variances are given by the diagonal terms of the inverse of the matrix of normal equation, the inverse being indicated by double brackets in these examples. The difference in variance for $\hat{B}$ in the last example, arises from the fact that in the first case one is concerned only with the difference between A (the standard) and B, whereas in the second case it is the difference between B and the average of the others that is involved.

For completeness, the matrices of normal equations and their inverses for the examples of Section 3 are shown below.

Linear Drift

$$X = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ \cdot & \\ \cdot & \\ 1 & (n-1) \end{bmatrix} \qquad X'X = \begin{bmatrix} n & n(n-1)/2 \\ n(n-1)/2 & n(n-1)(2n-1)/6 \end{bmatrix}$$

$$(X'X)^{-1} = \frac{12}{n^2(n^2-1)} \begin{bmatrix} n(n-1)(2n-1)/6 & -n(n-1)/2 \\ -n(n-1)/2 & n \end{bmatrix}$$

y a linear function of x

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \cdot & \\ \cdot & \\ 1 & x_n \end{bmatrix} \qquad X'X = \begin{bmatrix} n & \Sigma x \\ \Sigma x & \Sigma x^2 \end{bmatrix} \qquad (X'X)^{-1} = \frac{1}{n\Sigma x^2 - (\Sigma x)^2} \begin{bmatrix} \Sigma x^2 & -\Sigma x \\ -\Sigma x & n \end{bmatrix}$$

Gage block design

$$X = \begin{bmatrix} 1 & -1 & 0 & 0 & -7 \\ -1 & 0 & 0 & 1 & -5 \\ 0 & 0 & 1 & -1 & -3 \\ 0 & 1 & -1 & 0 & -1 \\ 0 & 1 & 0 & -1 & 1 \\ -1 & 0 & 0 & 1 & 3 \\ 1 & 0 & -1 & 0 & 5 \\ 0 & -1 & 1 & 0 & 7 \end{bmatrix}$$

$$\begin{bmatrix} X'X & B \\ B' & 0 \end{bmatrix} = \begin{bmatrix} 4 & -1 & -1 & -2 & 0 & 1 \\ -1 & 4 & -2 & -1 & 0 & 1 \\ -1 & -2 & 4 & -1 & 0 & 0 \\ -2 & -1 & -1 & 4 & 0 & 0 \\ 0 & 0 & 0 & 0 & 168 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$\begin{bmatrix} X'X & B \\ B' & 0 \end{bmatrix}^{-1} = \frac{1}{336} \begin{bmatrix} 35 & -35 & -7 & 7 & 0 & 168 \\ -35 & 35 & 7 & -7 & 0 & 168 \\ -7 & 7 & 91 & 21 & 0 & 168 \\ 7 & -7 & 21 & 91 & 0 & 168 \\ 0 & 0 & 0 & 0 & 2 & 0 \\ 168 & 168 & 168 & 168 & 0 & 0 \end{bmatrix}$$

Intercomparison of 5 standards (Sum of all used as restraint)

$$X = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 & 1 \\ 0 & 1 & -1 & 0 & 0 & 1 \\ 0 & 0 & 1 & -1 & 0 & 1 \\ 0 & 0 & 0 & 1 & -1 & 1 \\ -1 & 0 & 0 & 0 & 1 & 1 \\ -1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & -1 & 0 & 1 \\ 0 & -1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & -1 & 1 \\ 1 & 0 & -1 & 0 & 0 & 1 \end{bmatrix}$$

$$\begin{bmatrix} X'X & B \\ B' & 0 \end{bmatrix} = \begin{bmatrix} 4 & -1 & -1 & -1 & -1 & 0 & 1 \\ -1 & 4 & -1 & -1 & -1 & 0 & 1 \\ -1 & -1 & 4 & -1 & -1 & 0 & 1 \\ -1 & -1 & -1 & 4 & -1 & 0 & 1 \\ -1 & -1 & -1 & -1 & 4 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 10 & 0 \\ 1 & 1 & 1 & 1 & 1 & 0 & 0 \end{bmatrix}$$

$$\begin{bmatrix} X'X & B \\ B' & 0 \end{bmatrix}^{-1} = \frac{1}{25}\begin{bmatrix} 4 & -1 & -1 & -1 & 1 & 0 & 5 \\ -1 & 4 & -1 & -1 & -1 & 0 & 5 \\ -1 & -1 & 4 & -1 & -1 & 0 & 5 \\ -1 & -1 & -1 & 4 & -1 & 0 & 5 \\ -1 & -1 & -1 & -1 & 4 & 0 & 5 \\ 0 & 0 & 0 & 0 & 0 & 5/2 & 0 \\ 5 & 5 & 5 & 5 & 5 & 0 & 0 \end{bmatrix}$$

6. Standard Deviation

By substituting the computed values for the parameters into the equations of expected values for the observation, one has a *predicted value* to compare to the actual observation. The difference, d, between the observed and predicted value is called the *deviation* and is used to determine an estimate, s, of the standard deviation, σ, of the process

$$s = \sqrt{\frac{\Sigma d_i^2}{n-k+m}}$$

where n is the number of measurements, k is the number of parameters and m is the number of restraints.

Ordinarily one has available a sequence of values of the standard deviation say $s_1, s_2, s_3, \ldots, s_n$ based on $\nu_1, \nu_2, \nu_3, \ldots, \nu_n$ degrees of freedom. One forms the estimate of σ by combining these in quadrature

$$\hat{\sigma} = \sqrt{\frac{\nu_1 s_1^2 + \nu_2 s_2^2 + \ldots + \nu_n s_n^2}{\nu_1 + \nu_2 + \ldots + \nu_n}}$$

with degrees of freedom $N = \Sigma\nu$. In assigning a standard deviation to the parameters or linear combinations of them, the value $\hat{\sigma}$ is used rather than the value of s from a single experiment.

The variance of the sums of two parameter values is given by adding the corresponding diagonal terms (variances) in the inverse of the matrix of normal equations and the appropriate off diagonal terms (covariances) and multiplying by σ^2. For the case of the intercomparison of 5 standards given at the end of Section 5:

$$\text{s.d. } (\hat{A}+\hat{B}) = \sqrt{\sigma_A^2 + \sigma_B^2 + 2\sigma_{AB}} = \frac{\sigma}{\sqrt{25}}\sqrt{[4+4+2(-1)]} = \frac{\sigma\sqrt{6}}{5}$$

For the variance of the difference, the covariance terms enter negatively so that for the same example

$$\text{s.d.}(\hat{A}-\hat{B}) = \sqrt{\sigma_A^2 + \sigma_B^2 - 2\sigma_{AB}} = \frac{\sigma}{\sqrt{25}}\sqrt{[4+4-2(-1)]} = \frac{\sigma\sqrt{10}}{5}$$

For other linear combinations, formula 1.10-M of the Appendix would be used.

For the linear function example, the predicted value of y for x_o is $\hat{y}_o = \hat{\alpha} + \hat{\beta}x_o$ which has a variance of

$$\begin{bmatrix} 1 & x_o \end{bmatrix} \begin{bmatrix} C_{11} & C_{12} \\ C_{12} & C_{22} \end{bmatrix} \begin{bmatrix} 1 \\ x_o \end{bmatrix} \sigma^2 = (C_{11} + x_o^2 C_{22} + 2x_o C_{12})\sigma^2$$

where the terms C_{11}, C_{12}, C_{22} are the elements of $(X'X)^{-1}$ given in Section 5 for the case of y as a linear function of x.

7. Correlated Measurements

In the previous section it was assumed that the observations were uncorrelated, i.e., that $V(y_i) = \sigma^2$, $\text{cov}(y_i, y_j) = 0$ or in matrix form $V = \text{Var}(y) = \sigma^2 I$ where I is the identity matrix. Section 4 of the Appendix discusses the general case where one knows the matrix, V, of variances and covariances for the observations.

Quite often a transformation of variables can be achieved to obtain variables that are uncorrelated. A simple example is provided by the case of cummulative errors, i.e., in the case where

$$y_1 = \mu_1 + \varepsilon_1$$

$$y_2 = \mu_2 + \varepsilon_1 + \varepsilon_2$$

$$y_3 = \mu_3 + \varepsilon_1 + \varepsilon_2 + \varepsilon_3$$

The variance covariance matrix of the y's assuming $E(\varepsilon_i) = 0$, $\text{Var}(\varepsilon) = \sigma^2$, $\text{cov}(\varepsilon_i\varepsilon_j) = 0$ is given by

$$V = \sigma^2 \begin{bmatrix} 1 & 1 & 1 & . & . & 1 \\ 1 & 2 & 2 & & & 2 \\ 1 & 2 & 3 & & & 3 \\ . & & & & & \\ . & & & & & \\ 1 & 2 & 3 & . & . & n \end{bmatrix}$$

If one transforms to variables x_i where

$$x_1 = y_1 \qquad = \mu_1 + \varepsilon_1$$

$$x_2 = y_2 - y_1 \quad = \mu_2 - \mu_1 + \varepsilon_2$$

$$x_3 = y_3 - y_2 \quad = \mu_3 - \mu_2 + \varepsilon_3$$

$$\cdot$$

$$\cdot$$

$$x_n = y_n - y_{n-1} = \mu_n - \mu_{n-1} + \varepsilon_n$$

The expected values and variances become

$$E(X) = \begin{bmatrix} \mu_1 \\ \mu_2 - \mu_1 \\ \cdot \\ \cdot \\ \mu_n - \mu_{n-1} \end{bmatrix} \qquad V(X) = \begin{bmatrix} \sigma^2 & 0 & . & . & 0 \\ 0 & \sigma^2 & & & 0 \\ \cdot & & & & \\ \cdot & & & & \\ 0 & 0 & & & \sigma^2 \end{bmatrix}$$

In matrix form $X = Ty$ where $T = \begin{bmatrix} 1 & 0 & 0 & . & . & 0 & 0 \\ -1 & 1 & 0 & & & 0 & 0 \\ 0 & -1 & 1 & & & 0 & 0 \\ \cdot & & & & & & \\ \cdot & & & & & & \\ 0 & 0 & 0 & & & -1 & 1 \end{bmatrix}$

and if one computes Var(Ty) = TVT', one gets

$$\text{Var}(Ty) = \begin{bmatrix} 1 & 0 & 0 & . & . \\ -1 & 1 & 0 & & \\ 0 & -1 & 1 & & \\ . & & & & \\ . & & & & \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 & . & . & . & 1 \\ 1 & 2 & 2 & & & & 2 \\ 1 & 2 & 3 & & & & 3 \\ . & & & & & & \\ . & & & & & & \end{bmatrix} \begin{bmatrix} 1 & -1 & 0 & . & . \\ 0 & 1 & -1 & & \\ 0 & 0 & 1 & & \\ . & & & & \\ . & & & & \end{bmatrix} \quad \sigma^2 = \sigma^2 I$$

REFERENCES

(1) Cameron, J. M., and Hailes, G. E., "Designs for the Calibration of Small Groups of Standards in the Presence of Instrumental Drift," NBS Technical Note 844, Aug. 1974.

(2) Eisenhart, C., "Realistic Evaluation of the Precision and Accuracy of Instrument Calibration Systems," NBS J. Research, 67C (1963), 161-187.

(3) Eisenhart, C., "The Meaning of 'Least' in Least Squares," J. Wash. Acad. Sci., 54 (1964), 24-33.

(4) Goldman, A. J., and Zelen, M., "Weak Generalized Inverses and Minimum Variances Linear Unbiased Estimation," NBS J. Research 68B (1964) 151-172.

(5) Zelen, M., "Linear Estimation and Related Topics," Chapter 17 of Survey of Numerical Analysis, edited by J. Todd, McGraw Hill, New York, 1957, 558-584.

APPENDIX: FORMULAS FROM STATISTICS

1. Background and Notation

(a) Expected Value

The expected value, μ, of a random variable, y, will be written

$$E(y) = \mu$$

The mean μ may represent a linear function of some basic parameters $\beta_1, \beta_2, \ldots \beta_k$ with known coefficients $x_1, x_2, \ldots, x_k$

$$E(y) = \mu = x_1\beta_1 + x_2\beta_2 + \ldots + x_k\beta_k$$

The expected value of n observed values $y_1, y_2, \ldots, y_n$ can then be written

$$E(y_1) = x_{11}\beta_1 + x_{12}\beta_2 + \ldots + x_{1k}\beta_k \qquad (1.1)$$

$$E(y_2) = x_{21}\beta_1 + x_{22}\beta_2 + \ldots + x_{2k}\beta_k$$

$$\vdots$$

$$E(y_n) = x_{n1}\beta_1 + x_{n2}\beta_2 + \ldots + x_{nk}\beta_k$$

This may be written in matrix notation as

$$\begin{bmatrix} E(y_1) \\ E(y_2) \\ \vdots \\ E(y_n) \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} \cdots x_{1k} \\ x_{21} & x_{22} \cdots x_{2k} \\ & \\ x_{n1} & x_{n2} \cdots x_{nk} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \\ \beta_k \end{bmatrix} \qquad (1.1\text{-M})$$

or as $E(y) = X\beta$

where the vectors y and β and the matrix, X, are easily identified.

(b) Variance, Covariance

The variance, σ_i^2, of a random variable, y_i, is defined as

$$\sigma_i^2 = E\{(y_i - \mu_i)^2\} = E(y_i^2) - 2\mu_i E(y_i) + \mu_i^2 = E(y_i^2) - \mu_i^2 \qquad (1.2)$$

and the covariance σ_{ij} of the variables y_i and y_j by

$$\sigma_{ij} = E\{(y_i - \mu_i)(y_j - \mu_j)\} = E(y_i y_j) - \mu_i \mu_j \qquad (1.3)$$

The variance of cy where c is some constant is

$$\text{Var}(cy) = E\{(cy - c\mu)^2\} = c^2\sigma^2 \qquad (1.4)$$

The variance of a sum of two variables

$$\text{Var}(y_1 + y_2) = E\{[y_1 + y_2 - (\mu_1 + \mu_2)]^2\} = E\{[(y_1 - \mu_1) + (y_2 - \mu_2)]^2\}$$

$$= E(y_1 - \mu_1)^2 + E(y_2 - \mu_2)^2 + 2E\{(y_1 - \mu_1)(y_2 - \mu_2)\}$$

$$= \sigma_1^2 + \sigma_2^2 + 2\sigma_{12} \qquad (1.5)$$

which we may write as

$$\sigma_1^2 + \sigma_2^2 + 2\sigma_{12} = [1 \quad 1] \begin{bmatrix} \sigma_1^2 + \sigma_{12} \\ \sigma_{12} + \sigma_2^2 \end{bmatrix} = [1 \quad 1] \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \qquad (1.5\text{-M})$$

For independent random variables $\sigma_{ij} = 0$ and

$$\text{Var}(\Sigma y_i) = \Sigma \sigma_i^2 \qquad (1.6)$$

EXAMPLE:

$$\text{Var}(ay_1 + by_2 + cy_3) = E\{[(ay_1 - a\mu_1) + (by_2 - b\mu_2) + (cy_3 - c\mu_3)]^2\}$$

$$= a^2\sigma_1^2 + b^2\sigma_2^2 + c^2\sigma_3^2 + 2ab\sigma_{12} + 2ac\sigma_{13} + 2bc\sigma_{23} \qquad (1.7)$$

which may be written as

$$[a \; b \; c] \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_2^2 & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_3^2 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} \qquad (1.7\text{-M})$$

(c) Linear Function of Random Variables

A linear function

$$L = a_1y_1 + a_2y_2 + \ldots + a_ny_n \tag{1.8}$$

has expected value

$$E(L) = a_1E(y_1) + a_2E(y_2) + \ldots + a_nE(y_n) \tag{1.9}$$

or in matrix notation

$$E(L) = (a_1\ a_2\ \ldots\ a_n)\begin{bmatrix} E(y_1) \\ E(y_2) \\ E(y_n) \end{bmatrix} = a'\mu \tag{1.9-M}$$

The variance is given, by analogy with (1.7) by

$$V(L) = [a_1\ a_2\ \ldots\ a_n]\begin{bmatrix} \sigma_1^2 & \sigma_{12} & \ldots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \ldots & \sigma_{2n} \\ \cdot & & & \\ \cdot & & & \\ \sigma_{n1} & \sigma_{n2} & \ldots & \sigma_n^2 \end{bmatrix}\begin{bmatrix} a_1 \\ a_2 \\ \cdot \\ \cdot \\ a_n \end{bmatrix} \tag{1.10-M}$$

which reduces to the usual formula

$$V(\Sigma a_iy_i) = \Sigma a_i^2\sigma_i^2 \tag{1.11}$$

if $\sigma_{ij} = 0$.

For two linear functions L_1 and L_2 the covariance term is given by

$$E\{[a_1(y_1-\mu_1) + \ldots\ldots + a_n(y_n-\mu_n)][b_1(y_1-\mu_1) + \ldots\ldots + b_n(y_n-\mu_n)]\}$$

$$= a_1b_1E(y_1-\mu_1)^2 + a_2b_2E(y_2-\mu_2)^2 + \ldots + a_nb_nE(y_n-\mu_n)^2$$

$$+ (a_1b_2 + a_2b_1)E(y_1-\mu_1)(y_2-\mu_2) + (a_1b_3 + a_3b_1)E(y_1-\mu_1)(y_3-\mu_3)$$

$$+ (a_2b_3+a_3b_2)E(y_2-\mu_2)(y_3-\mu_3) + \ldots$$

This reduces to the usual formulas:

$$\text{If } \sigma_{ij} = 0 \qquad \text{then } \operatorname{Cov}(L_1, L_2) = \Sigma a_i b_i \sigma_i^2$$
$$\text{If } \sigma_i = \sigma \qquad \text{then } \operatorname{Cov}(L_1, L_2) = \sigma^2 \Sigma a_i b_i \tag{1.12}$$

For the case of $L_1 = a_1y_1 + a_2y_2 + a_3y_3$ and $L_2 = b_1y_1 + b_2y_2 + b_3y_3$, the covariance can be written:

$$(a_1\ a_2\ a_3)\begin{bmatrix} b_1\sigma_1^2 + b_2\sigma_{12} + b_3\sigma_{13} \\ b_2\sigma_2^2 + b_1\sigma_{12} + b_3\sigma_{23} \\ b_3\sigma_3^2 + b_1\sigma_{13} + b_2\sigma_{23} \end{bmatrix} = (a_1\ a_2\ a_3)\begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_2^2 & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_3 \end{bmatrix}\begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} \tag{1.12-M}$$

giving the general formula for the variance and covariance of two linear functions

$$\begin{bmatrix} a_1\ a_2 \cdots a_n \\ b_1\ b_2 \cdots b_n \end{bmatrix}\begin{bmatrix} \sigma_1^2 & \sigma_{12} \cdots \sigma_{1n} \\ \sigma_{12} & \sigma_2^2 \cdots \sigma_{2n} \\ \vdots & \\ \sigma_{1n} & \sigma_{2n} \cdots \sigma_n^2 \end{bmatrix}\begin{bmatrix} a_1 & b_1 \\ a_2 & b_2 \\ \vdots & \vdots \\ a_n & b_n \end{bmatrix} \tag{1.13-M}$$

or in general for p such function, i.e., for a pxn matrix A

$$\operatorname{Var}(AY) = AVA' \tag{1.14-M}$$

(d) Quadratic Forms in Random Variables

We have from (1.2)

$$E(y^2) = \sigma^2 + \mu^2 \tag{1.15}$$

We wish to extend this to include the case of a more general quadratic expression in the y's, consider for example

$$E[(ay_1 + by_2)^2] = Ea^2y_1^2 + Eb^2y_2 + 2abE(y_1y_2)$$

$$= a^2\sigma_1^2 + a^2\mu_1^2 + b^2\sigma_2^2 + b^2\mu_2^2 + 2ab\mu_1\mu_2 + 2ab\sigma_{12}$$

which may be displayed as a matrix product as follows:

$$[y_1\ y_2]\begin{bmatrix} a^2 & ab \\ ab & b^2 \end{bmatrix}\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = [\mu_1\ \mu_2]\begin{bmatrix} a^2 & ab \\ ab & b^2 \end{bmatrix}\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} + [a\ \ b]\begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}\begin{bmatrix} a \\ b \end{bmatrix}$$

This example illustrates the general formula:

$$E[(a_1y_1 + \ldots + a_ny_n)^2] = E\left\{[y_1\ y_2 \ldots y_n]\begin{bmatrix} a_1^2 & a_1a_2 & \ldots & a_1a_n \\ a_2a_1 & a_2^2 & \ldots & a_2a_n \\ \cdot & & & \\ \cdot & & & \\ a_na_1 & a_na_2 & \ldots & a_n^2 \end{bmatrix}\begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ y_n \end{bmatrix}\right\}$$

$$= [\mu_1 \ldots \mu_n]\begin{bmatrix} a_1^2 & \ldots & a_{1n} \\ \cdot & & \cdot \\ \cdot & & \cdot \\ a_{1n} & \ldots & a_n^2 \end{bmatrix}\begin{bmatrix} \mu_1 \\ \cdot \\ \cdot \\ \mu_n \end{bmatrix} + [a_1 \ldots a_n]\begin{bmatrix} \sigma_1^2 & \sigma_{12} & \ldots & \sigma_{1n} \\ \cdot & & & \\ \cdot & & & \\ \sigma_{1n} & \sigma_{2n} & \ldots & \sigma_n^2 \end{bmatrix}\begin{bmatrix} a_1 \\ \cdot \\ \cdot \\ a_n \end{bmatrix}$$

or

$$E\{y'Ay\} = \mu'A\mu + a'Va \qquad (1.16\text{-M})$$

where $A = \begin{bmatrix} a_1^2 & a_1a_2 & \ldots \\ a_1a_2 & a_2^2 & \ldots \\ \cdot & & \\ \cdot & & \end{bmatrix}$ and $V = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \ldots \\ \sigma_{12} & \sigma_2^2 & \ldots \\ \cdot & & \\ \cdot & & \end{bmatrix}$

The last term can be replaced by the trace of AV so that we have

$$E(Y'AY) = \mu'A\mu + \text{Trace}(AV) \qquad (1.17\text{-M})$$

For an excellent treatment of these statistical topics one should consult Zelen [5].

2. The Gauss Theorem on Least Squares (Independent, Equal Variance, Full Rank)

Let the n observations $y_1, y_2, \ldots, y_n$ have expected values

$$\begin{aligned} E(y_1) &= x_{11}\beta_1 + x_{12}\beta_2 \cdots x_{1k}\beta_k \\ E(y_2) &= x_{21}\beta_1 + x_{22}\beta_2 \cdots x_{2k}\beta_k \\ &\cdot \\ &\cdot \\ &\cdot \\ E(y_n) &= x_{n1}\beta_1 + x_{n2}\beta_2 \cdots x_{nk}\beta_k \end{aligned} \qquad (2.1)$$

and be statistically independent with common variance, σ^2. These two conditions can be expressed in matrix form as follows:

$$E(y) = \begin{bmatrix} E\,y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & & x_{1k} \\ x_{21} & x_{22} & & x_{2k} \\ \cdot & \cdot & & \\ \cdot & \cdot & & \\ \cdot & \cdot & & \\ x_{n1} & x_{n2} & & x_{nk} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \cdot \\ \cdot \\ \beta_k \end{bmatrix} = X\beta \qquad (2.1\text{-M})$$

$$V(y) = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix} = \sigma^2 I$$

The Gauss theorem states that the minimum variance unbiased linear estimator of any linear function, L, of the parameters, $\beta_1\ \beta_2 \ldots \beta_k$, say

$$L = a_1\beta_1 + a_2\beta_2 + \ldots a_k\beta_k$$

is given by substituting the values of β_i which minimize

$$Q = \Sigma[y_i - (x_{i1}\beta_1 + \ldots + x_{ik}\beta_k)]^2 \qquad (2.2)$$

considered as a function of the β_i. These values, $\hat{\beta}_1$, $\hat{\beta}_2 \ldots \hat{\beta}_k$ are the solutions to the k equations, called the *normal equations*.

$$\begin{aligned}
&\Sigma x_{i1}^2\hat{\beta}_1 + \Sigma x_{i1}x_{i2}\hat{\beta}_2 + \ldots + \Sigma x_{i1}x_{ik}\hat{\beta}_k = \Sigma x_{i1}y_i \\
&\Sigma x_{i2}x_{i1}\hat{\beta}_1 + \Sigma x_{i2}^2\hat{\beta}_2 + \ldots + \Sigma x_{i2}x_{ik}\hat{\beta}_k = \Sigma x_{i2}y_i \\
&\cdot \\
&\cdot \\
&\cdot \\
&\Sigma x_{ik}x_{i1}\hat{\beta}_1 + \Sigma x_{ik}x_{i2}\hat{\beta}_2 + \ldots + \Sigma x_{ik}^2\hat{\beta}_k = \Sigma x_{ik}y_i
\end{aligned} \qquad (2.3)$$

or in matrix form

$$(X'X)\hat{\beta} = X'y \qquad (2.3\text{-M})$$

The solution to these equations can be written as

$$\hat{\beta} = (X'X)^{-1}X'y \qquad (2.4\text{-M})$$

because X was assumed to be of rank k. The matrix $(X'X)^{-1}$ is the *inverse of the matrix of normal equations* and plays an important role in least squares analysis. Let its elements be denoted by c_{ij} so that

$$(X'X)^{-1} = \begin{bmatrix} c_{11} & c_{12} \ldots c_{1k} \\ c_{21} & c_{22} \ldots c_{2k} \\ \cdot & \\ \cdot & \\ \cdot & \\ c_{k1} & c_{k2} \ldots c_{kk} \end{bmatrix} \qquad (2.5\text{-M})$$

The standard deviation, σ, is estimated from the deviations d_i, where

$$d_i = y_i - (x_{i1}\hat{\beta}_1 + x_{i2}\hat{\beta}_2 + \ldots + x_{ik}\hat{\beta}_k) \qquad (2.6)$$

by the quantity, s,

$$s = \sqrt{\frac{\Sigma d^2}{n-k}} \tag{2.7}$$

and is said to have n-k *degrees of freedom.*

The standard deviation of the values for the coefficients $\hat{\beta}_i$ are given by

$$s.d.(\hat{\beta}_i) = \sigma\sqrt{c_{ii}} \tag{2.8}$$

and for a linear function $L = a_1\hat{\beta}_1 + a_2\hat{\beta}_2 \ldots a_k\hat{\beta}_k$ is [see equation (1.10-M)]

$$s.d.(L) = \sigma\left\{[a_1 \ldots a_k]\begin{bmatrix} c_{11} & \ldots & c_{1n} \\ \cdot & & \\ \cdot & & \\ \cdot & & \\ c_{n1} & \ldots & c_{nn} \end{bmatrix}\begin{bmatrix} a_1 \\ \cdot \\ \cdot \\ \cdot \\ a_k \end{bmatrix}\right\}^{1/2} \tag{2.9-M}$$

3. The Gauss Theorem on Least Squares (Independent, Equal Variance, With Restraints)

If the parameters, β_i, are required to satisfy the m linear equations

$$\left\{\begin{array}{l} \psi_1 = b_{11}\beta_1 + b_{12}\beta_2 \ldots b_{1k}\beta_k = K_1 \\ \quad \cdot \\ \quad \cdot \\ \psi_m = b_{m1}\beta_1 + b_{m2}\beta_2 \ldots b_{mk}\beta_k = K_m \end{array}\right. \tag{3.1}$$

or in matrix form

$$B'\beta = K \tag{3.1-M}$$

then using the method of Lagrangian multipliers, it turns out that the minimum variance unbiased linear estimators are given by minimizing

$$F = Q + 2\lambda_1 (\psi_1 - K_1) + 2\lambda_2(\psi_2 - K_2) + \ldots + 2\lambda_m(\psi_m - K_m) \tag{3.2}$$

considered as a function of the β's and λ's. ($2\lambda_i$ is chosen rather than just λ_i so that in setting $\partial F/\partial\beta_i = 0$, a common factor of 2 can be divided out.)

This leads to the normal equations

$$\left\{\begin{array}{l} \Sigma x_1^2\beta_1 + \ldots + \Sigma x_1 x_k \beta_k + b_{11}\lambda_1 + \ldots + b_{m1}\lambda_m = \Sigma x_1 y \\ \quad \cdot \\ \quad \cdot \\ \quad \cdot \\ \Sigma x_k x_1 \beta_1 + \ldots + \Sigma x_k^2 \beta_k + b_{1k}\lambda_1 + \ldots + b_{mk}\lambda_m = \Sigma x_k y \\ b_{11}\beta_1 + \ldots + \beta_{1k}\beta_k = K_1 \\ \quad \cdot \\ \quad \cdot \\ \quad \cdot \\ b_{m1}\beta_1 + \ldots + \beta_{mk}\beta_k = K_m \end{array}\right. \tag{3.3}$$

or in matrix form

$$\begin{bmatrix} X'X & B \\ B' & 0 \end{bmatrix} \begin{bmatrix} \beta \\ \lambda \end{bmatrix} = \begin{bmatrix} X'y \\ K \end{bmatrix} \qquad (3.3\text{-M})$$

and the solution is given by

$$\begin{bmatrix} \hat{\beta} \\ \hat{\lambda} \end{bmatrix} = \begin{bmatrix} X'X & B \\ B' & 0 \end{bmatrix}^{-1} \begin{bmatrix} X'y \\ K \end{bmatrix} \qquad (3.4\text{-M})$$

If X'X was already of full rank, then B must be of rank m for the inverse to exist. If X'X is of rank (k-m) and B' consists of m rows, then the indicated inverse will exist if B is orthogonal to X'X, i.e. that (X'X)B = 0, and B is of rank m. Also if B is a combination of such an orthogonal set of restraints, denoted by H, and the vectors of X'X, then the inverse exists if the mxm matrix B'H is of rank m, i.e., the determinant $|B'H| \neq 0$.

EXAMPLE: If the differences A-B, B-C, C-D, D-E, E-A are measured, then the 5 measurements y_1, y_2, y_3, y_4, y_5 (assumed independent with equal variance) can be represented as

$$E(y) = \begin{bmatrix} A-B \\ B-C \\ C-D \\ D-E \\ -A+E \end{bmatrix} = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 1 & -1 \\ -1 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} A \\ B \\ C \\ D \\ E \end{bmatrix} = X\beta$$

$$X'X = \begin{bmatrix} 2 & -1 & 0 & 0 & -1 \\ -1 & 2 & -1 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 \\ 0 & 0 & -1 & 2 & -1 \\ -1 & 0 & 0 & -1 & 2 \end{bmatrix} \qquad \text{rank of } X'X \text{ is } 4$$

The restraint $A+B+C+D+E = [1\ 1\ 1\ 1\ 1] \begin{bmatrix} A \\ B \\ C \\ D \\ E \end{bmatrix} = H' \begin{bmatrix} A \\ B \\ C \\ D \\ E \end{bmatrix} = K$

is orthogonal to X'X because H'(X'X) = (1 1 1 1 1) (X'X) = [0 0 0 0 0]. If the given restraint were A + B = K_0, then B' = (1 1 0 0 0) and |B'H| = 2 ≠ 0 so that the restraint is sufficient to produce a solution.

The standard deviation estimate is changed from that given in formula (2.7) to become

$$s = \sqrt{\frac{\Sigma d^2}{n-k+m}} \qquad \text{degrees of freedom} = (n-k+m) \tag{3.5}$$

where m is the number of restraints.

Formulas (2.8) and (2.9) still apply for the standard deviation of the parameter values and of linear combinations of them.

4. The Gauss Theorem on Least Squares (General Case)

If the observed values $y_1\ y_2\ \ldots\ y_n$ have variances σ_i^2 and covariances σ_{ij} so that

$$\text{Var}\ (y) = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \ldots & \sigma_{1n} \\ \sigma_{12} & \sigma_2^2 & & \sigma_{2n} \\ \cdot & & & \\ \sigma_{1n} & \sigma_{2n} & & \sigma_n^2 \end{bmatrix} = V \tag{4.1-M}$$

and the parameters are subject to the m restraints

$$\begin{cases} b_{11}\beta_1 + \ldots + b_{1k}\beta_k = K_1 \\ \cdot \\ b_{m1}\beta_1 + \ldots + b_{mk}\beta_k = K_m \end{cases} \tag{4.2}$$

or in matrix form

$$B'\beta = K \tag{4.2-M}$$

Then the least squares estimators for β are given by

$$\begin{bmatrix} \hat{\beta} \\ \hat{\lambda} \end{bmatrix} = \begin{bmatrix} X'V^{-1}X & B \\ B' & 0 \end{bmatrix}^{-1} \begin{bmatrix} X'V^{-1}y \\ K \end{bmatrix}$$

where as before $\hat{\lambda}' = [\hat{\lambda}_1\ \hat{\lambda}_2\ \ldots\ \hat{\lambda}_m]$ is a vector of Lagrangian multipliers entering into the minimization process.

For a discussion of this general case, the reader is referred to the Goldman-Zelen article [4].

PRACTICAL REPRESENTATION OF THE MASS UNIT
EFFECTIVE JANUARY 1990

Volume 95, Number 1, January-February 1990
Journal of Research of the National Institute of Standards and Technology

[J. Res. Natl. Inst. Stand. Technol. 95, 79 (1990)]

New Assignment of Mass Values and Uncertainties to NIST Working Standards

Volume 95 Number 1 January-February 1990

Richard S. Davis

National Institute of Standards and Technology,
Gaithersburg, MD 20899

For some time it had been suspected that values assigned to NIST working standards of mass were some 0.17 mg/kg larger than mass values based on artifacts representing mass in the International System of Units (SI). This relatively small offset, now confirmed, has had minimal scientific or technological significance. The discrepancy was removed on January 1, 1990. We document the history of the discrepancy, the studies which allow its removal, and the methods in place to limit its effect and prevent its recurrence. For routine calibrations, we believe that our working standards now have a long-term stability of 0.033 mg/kg (3σ) with respect to the national prototype kilograms of the United States. We provisionally admit an additional uncertainty of 0.09 mg/kg (3σ), systematic to all NIST mass measurements, which represents the possible offset of our primary standards from standards maintained by the Bureau International des Poids et Mesures (BIPM). This systematic uncertainty may be significantly reduced after analysis of results from the 3rd verification of national prototype kilograms, which is now underway.

Key words: calibration; international standards; kilogram; mass; national standards; SI; standards.

Accepted: January 6, 1990

1. Introduction

The kilogram (kg) is one of the seven base units which form the foundation of the Système International d'Unités or International System of Units, abbreviated SI. Used world wide to express the results of physical measurements, the SI specifies that the kilogram is the unit of mass and that the mass of the International Prototype Kilogram exactly equals 1 kg. The International Prototype referred to in the definition is a cylinder made of an alloy of platinum and iridium and stored at the International Bureau of Weights and Measures (BIPM) in France. The kilogram is thus the only remaining base unit of the SI to rely on an artifact for its definition.

When the SI was established, replicas of the International Prototype were manufactured by the BIPM for use as national prototype kilograms. At long intervals, the national prototypes are returned to the BIPM where their assigned mass is verified by measurements directly traceable to the International Prototype [1]. It was intended by the founders of the SI that the national prototype kilograms would be the primary mass standards within each country. There are, however, several practical difficulties with this scheme. The following discusses the reasons for these difficulties and the steps we have taken to overcome them.

In order for the kilogram unit to be useful, methods must exist to measure multiples and submultiples of 1-kg standards. These methods, when successful, rely on good equipment and sound experimental practice. In addition to these, a calibra-

tion service requires rigorous tests to maintain statistical control of the measurement process. At NIST, statistical rigor was introduced in the 1960s through the pioneering work of Pontius and Cameron [2]. Present methods are simply refinements of the system which they established.

The uncertainty of a 1-kg standard, expressed as a dimensionless ratio, propagates directly to mass values of multiples and submultiples derived from the standard. For example, if a kilogram standard has a relative uncertainty of 1 ppm[1], all multiples and submultiples derived from the standard will have an uncertainty component of 1 ppm propagated from the standard. In the field of precision measurement, uncertainty is usually reported at an estimated level of 1 standard deviation. All uncertainties are combined by the root-sum-square (RSS) method according to guidelines recommended by the International Committee for Weights and Measures (CIPM) [3]. In NIST calibration reports, on the other hand, uncertainties are estimated at a level of 3 standard deviations. Furthermore, any uncertainty deemed "systematic" to a series of measurements is added directly to the "random" uncertainties, which are combined by RSS. However, in the rest of this paper, we follow the CIPM recommendations unless otherwise noted.

In addition to the SI, the United States recognizes the U.S. Customary System of units for legal metrology. In this system, the avoirdupois pound (lb) is the unit of mass. It is, by definition, exactly equal to 0.45359237 kg.

2. History of NIST Mass Standards Before 1980

2.1 Primary Mass Standards of Platinum-Iridium

Kilograms K20 and K4 are the two national prototypes of the United States. Kilogram K20 has historically been considered the primary U.S. kilogram standard with K4 being relegated to use as a "check standard." The history of these two artifacts through 1985 has already been documented in a previous report [1]. One important question which remained open in [1] is whether the mass values assigned by BIPM to their working standards have been consistent with the SI definition of mass. The cause for concern was that the embodiment of the SI definition, the International Prototype Kilogram, had not been used since 1946. This situation has changed within the past year as BIPM embarked on only the third calibration of national prototype kilograms since 1889. Preliminary results obtained by BIPM as a part of the 3rd verification confirm the long-term stability of their working standards to within required limits [4].

2.2 Secondary Mass Standards

Platinum-iridium alloy (approximate density 21,500 $kg \cdot m^{-3}$) is too expensive a material for widespread use. At present, stable alloys of non-magnetic stainless steel (approximate density 8,000 $kg \cdot m^{-3}$) are usually specified for use as secondary standards. Before such alloys were available, practical standards were typically made of plated brass (approximate density 8,400 $kg \cdot m^{-3}$). The densities of these alloys assume importance because mass metrology is almost always performed in the ambient air (density ca. 1.2 $kg \cdot m^{-3}$) using balances which are, in essence, force or torque transducers. The effect of air buoyancy thus becomes a confounding influence which must be removed by correction.

The size of the necessary buoyancy correction relative to the mass of interest is given by:

$$(1-\rho_a/\rho_s)/(1-\rho_a/\rho_x)-1 \approx \rho_a(1/\rho_x-1/\rho_s), \quad (1)$$

where ρ_a = ambient air density
ρ_s = density of the known standard
ρ_x = density of the unknown secondary standard.

Equation (1) makes clear that, when comparing weights of nearly equal density, the importance of the correction is relatively small. Buoyancy corrections are typically 10 ppm between alloys of stainless steel and brass; corrections of less than 5 ppm are typical for comparisons between various alloys of non-magnetic stainless steel. (Specifications for the highest quality analytical weights limit the alloy density to within a narrow range in order to ensure that buoyancy corrections between nominally equal weights will be small.)

By contrast, the buoyancy correction between (i) primary standards of platinum-iridium alloy and (ii) secondary standards of brass or stainless steel typically ranges from 87-97 ppm. In our laboratory, the densities of secondary kilogram standards are determined by hydrostatic weighing. The density of ambient air is now determined from the CIPM-1981 equation-of-state for moist air [5]. The latter requires knowledge of ambient temperature, barometric pressure, relative humidity, and carbon-

[1] 1 ppm = 1 part per million = 1×10^{-6}.

dioxide level. A discussion of the accuracy which can be expected from buoyancy corrections in our laboratory is given in [1].

The above considerations dictate that calibrations carried out by NIST on a routine basis be performed with secondary standards having a density near to that of the unknown weight.

2.2.1 N_1 and N_2 Two weights, designated N_1 and N_2, have served as NIST secondary standards of mass since 1965. The weights were fabricated in 1948 of a nickel-chromium alloy having a nominal density of 8,340 $kg \cdot m^{-3}$, which is close to that of the brass weights which were then in common use. These weights were given an initial calibration in terms of a platinum-iridium prototype (K4) in 1948. They were recalibrated against both K20 and K4 in 1958. The newer calibration gave mass values which were systematically higher by about 0.06 mg/kg. There is no indication in the existing records what, if any, uncertainty was assigned to either calibration. When, in 1965, N_1 and N_2 were placed in service as secondary mass standards, the mass assigned to them was based on selected data from the 1958 series of measurements. Presumably, this decision was made because the 1958 measurements were performed by remote control on a two-pan, Rueprecht balance having a standard deviation below 0.02 mg. By 1965, this device had been replaced by a single-pan balance which was much more convenient to use but which had an inferior standard deviation of about 0.15 mg. Further, remote weighing was not possible on the single-pan balance.

Based on the 1958 measurements, the mass of N_1 and N_2 taken together was calculated to be:

$$R = 2 \text{ kg} - 10.059 \text{ mg}.$$

The difference in mass between N_1 and N_2 was calculated by pooling a large amount of data:

$$C = -19.476 \text{ mg}.$$

These two numbers, R and C, fix the individual values of each kilogram. The uncertainty in C is largely statistical in nature. It depends almost entirely on the standard deviation of the balance used to compare the mass of N_1 with N_2. Thus its uncertainty could be rigorously assigned. In addition, the significance of any measured change in C could also be determined.

The uncertainty of R was much more problematic. The statistical component of this uncertainty resulting from the balance used in the measurements may, of course, be calculated. There are at least two additional components which increase the uncertainty of R (but not of C):

1. The uncertainty in the accepted mass of K20 with respect to the International Prototype Kilogram.
2. The accuracy of the correction for air buoyancy between the platinum-iridium and the nichrome kilograms.

Rather than base an estimate of these uncertainties on what was considered insufficient metrological data, calibration reports prior to January 1, 1990 state:

"It is assumed that the present 'accepted values' of the two NIST standards at the 1 kilogram level, designated N_1 and N_2, are without error. Estimates of the uncertainty of the accepted values of the NIST standards relative to the International Prototype Kilogram can be provided on request. However, these estimates have no real meaning in either national or international comparison. This is because of the lack of sufficient data to provide a realistic estimate of the uncertainty in the values assigned to the prototype kilograms K20 and K4, particularly in regard to long term, or between-run variability. Changes in the accepted values for the NIST standards at the kilogram level, as and when they occur, will be reported in the scientific papers of the Bureau and will be given wide distribution..."

Except for the change in name of the institution, the above wording had been in place at least since 1967. The reports of that time (and well beyond) also referenced a technical note entitled "The Accepted Values of the NBS Standards at the 1 kg Level and Associated Uncertainty Estimates," to be published at a future date. Unfortunately, this note was never produced. Section 3 of the present paper may therefore be regarded as fulfilling a promise of long standing.

In looking over calibration documentation extending back 25 years, it seems that the original intention was to reserve N_1 and N_2 for calibration of other working standards of similar density. These working standards would be used in routine calibration work and thereby would spare N_1 and N_2 from excessive wear. But the calibration of working standards of 1 kg could only be done on the single-pan balance mentioned above. Thus working standards would be assigned an uncer-

tainty which was large relative to the precision of commercially available balances unless the calibration were based on the average of many measurements. But the latter strategy would no longer spare N_1 and N_2 from excessive use.

Faced with this problem, N_1 and N_2 began to be used as working standards themselves in routine calibrations. They were never cleaned (except for gentle dusting with a brush) in order to prevent discontinuous changes in their mass. It was, of course, recognized that checks must be established to ensure the constancy of the mass assigned to the summation of N_1 and N_2. Two criteria were routinely used.

The first criterion was the constancy of C. A measurement of C was available each time N_1 and N_2 were used. In time, a newer balance of similar design was obtained. This device, which is still in use, has a standard deviation of about 0.035 mg. If values of C were seen to change significantly with time, it would mean that the summation mass of N_1 and N_2 had deviated from its accepted value. This test is effective in checking whether one or the other kilograms has suffered damage since its last use. However, the test fails to detect changes common to both artifacts. Because N_1 and N_2 are virtually identical and receive identical use, such changes cannot be ruled out *a priori*. Thus the constancy of C is not a sufficient test to rule out a change in the summation mass of the two kilograms. A control chart showing values of C over time is given in figure 1. The second criterion is discussed below in section 2.2.2.

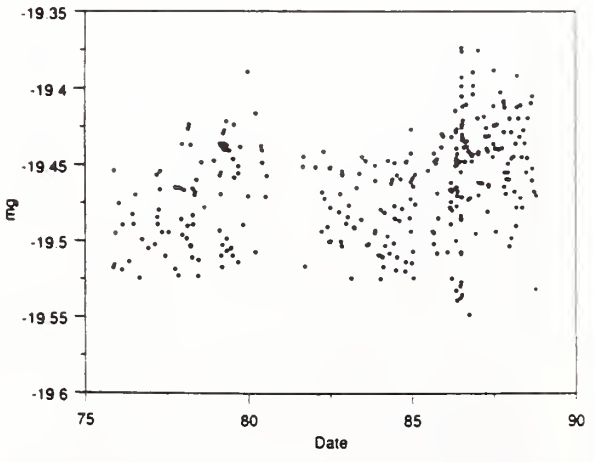


Figure 1. Mass values of $N_1 - N_2$ as a function of time. Measurements were taken on a balance having a standard deviation of 0.035 mg for a single reading.

In 1969, the masses of N_1 and of N_2 were redetermined 10 times with respect to K20 and K4. Measurements were made on a one-pan balance having a standard deviation of 0.14 mg for a single observation. The results of these measurements indicated that N_1 and N_2 were an average of 0.09 mg/kg below their accepted value. However, because the uncertainties propagated from the prototype kilograms and from the correction for air buoyancy could still not be assessed, these data were not used.

2.2.2 100-g Check Standards The second criterion used to monitor the constancy in mass of N_1 and N_2 was the evolution in time of two 100-g "check" standards. A measurement of one or the other of these standards in terms of N_1 and N_2 was obtained each time a routine calibration was performed on a set of weights from 1 kg to 100 g. Such measurements are carried out dozens of times each year. If the mass of the 100-g check standards was seen to change over time, it would be evidence that either their mass or that of N_1 and N_2 was changing. It is unlikely that the mass of the 100-g check standards would change in exact proportion to the mass of the 1-kg working standards. This test suffers, however, from low precision. The statistical precision in the assignment of mass to a 100-g standard is about ten-times lower than the relative precision of mass assigned to 1-kg weights. The reason is simply that all mass comparisons between 1 kg and 100 g are performed on the same balance. One would need to average about 10^2 mass determinations of a 100-g check standard in order to have the same relative precision as one single mass determination of a 1-kg standard.

The 100-g check standard suffers from an additional problem. Since it receives heavy use, its mass can reasonably be expected to decrease with time due to wear. Control charts showing mass values obtained over time for our 100-g check standards, JMC-1 and JMC-2, are given in figure 2. The apparent rapid loss in mass early in the service life of JMC-1 is not unusual. Such behavior is also seen, for instance, in our 1-g check standard where there can be no possibility that the source is instability in 1-kg working standards. Thus the 100-g check standards, while essential to guard against measurement blunders and catastrophic changes in working standards, are themselves susceptible to long-term instability.

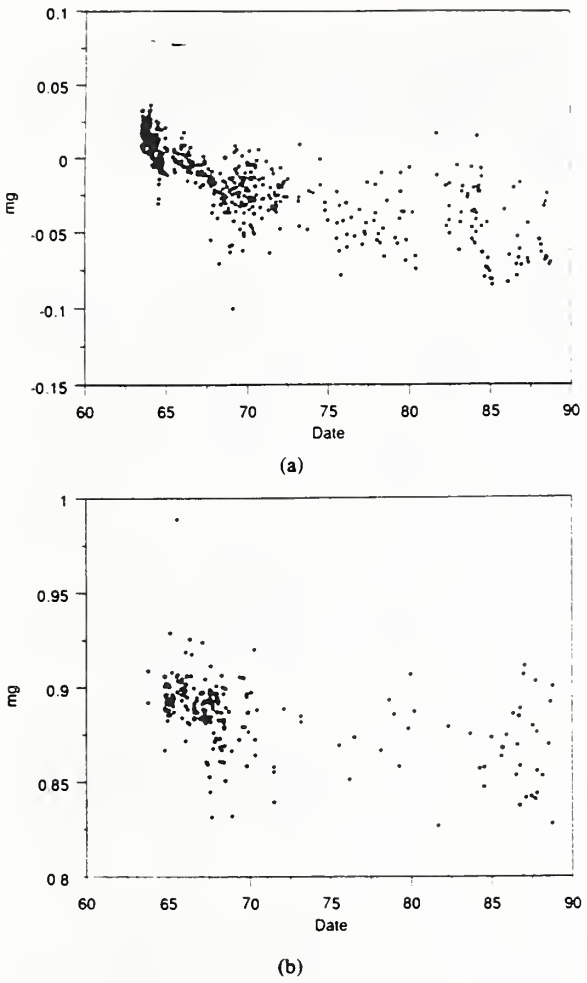


Figure 2. Mass values of 100-g check standards JMC-1(a) and JMC-2(b). These values are based on the accepted mass of N_1 and N_2 prior to January 1, 1990.

2.2.3 State Laboratories Each state within the United States maintains a well-equipped laboratory for primary mass metrology, typically placed administratively within the State Department of Agriculture. Training of personnel and many aspects of quality control are coordinated through the NIST Office of Weights and Measures (OWM). The OWM organizes regional round-robin measurements involving State mass-standards of various nominal values. These round-robins also include standards recently calibrated by NIST. An examination of round-robin results for 1-kg masses does not reveal systematic differences between NIST and the States developing over time. But the precision of these comparisons limits conclusions to about 0.5 mg/kg.

2.3 Fundamental Measurements

Some fundamental constants offer a check on the constancy of mass standards. During the 1970s, measurements of the Avogadro constant N_A [6] and the Faraday constant F relied directly on mass values maintained at NIST. These measurements can be compared with related measurements at other laboratories as is done during periodic CODATA adjustments of the fundamental constants [8].

In the case of the NIST determination of the Faraday constant, routine mass calibrations of a 5-g and 3-g working standard were used. It was estimated that the uncertainty in these calibrations was 0.5 ppm (standard error). This estimate contributed less than 10 percent of the combined experimental uncertainty. The Faraday constant has, therefore, little bearing on the present discussion.

This is not true in the case of the Avogadro constant. In order to have their mass values directly traceable to national standards, the experimenters made direct use of K20 and K4. Several calibrations at the 1-kg level were carried out on the newly developed NBS-2 balance [9]. This balance operates under remote control and, at that time, had a standard deviation of less than 0.005 mg. (After initial testing at NBS, the balance was transferred to the BIPM where improved conditions have reduced its standard deviation five-fold.) Unfortunately, N_1 and N_2 were not measured during the experiments, although several stainless-steel kilograms were calibrated in terms of K20 and K4. Two of these kilograms had also been measured against K20 and K4 in 1969 as part of the series of mass determinations which included N_1 and N_2 (see sec. 2.2.1, above). These results were completely consistent with the 1969 measurements and thus raise the question of whether the mass values for N_1 and N_2 dating from 1958 were still appropriate.

3. History of Mass Standards after 1980

About 10 years ago, NIST began a program to tie the mass values disseminated by its calibration services with international standards. It was foreseen that improvements in commercial balance technology and improved precision in measuring critical fundamental constants would soon make this step necessary. In addition, questions of international compatibility of national standards began

to be raised at this time. In order to assess the presently accepted values of NIST secondary standards with respect to the SI, four major areas had to be addressed:

1. A meaningful calibration of K20 and K4 with respect to accepted representations of SI standards.
2. A reliable method for making corrections for air buoyancy between primary standards of platinum-iridium and secondary standards of nichrome or stainless steel.
3. A balance which could compare kilogram masses with a precision no worse than 0.005 mg.
4. Demonstration that primary standards could indeed be used periodically to calibrate secondary standards and that mass values so determined did not suffer from serious, unexplainable discontinuities.

We now briefly describe efforts made in these four areas.

3.1 Tie to International Standards

As mentioned in section 2.2.1 above, the main reason given in the past for not basing mass calibrations on routine comparisons with K20 was that the long-term stability of platinum-iridium prototype kilograms had not been rigorously established. One reason for this apparent lack of understanding is the infrequency with which the International Prototype Kilogram is used. The BIPM faces this same problem because it is their job to recertify national prototype kilograms upon request and to provide new national prototype kilograms when required. These activities must be carried out during the long intervals when the International Prototype Kilogram is not accessible.

As described in [1], the BIPM has set in place the following system in which all the mass standards involved are made of platinum-iridium alloy.

Two working standards are used in the calibration of an unknown prototype. The measured difference in mass between the two working standards is used to check that neither has suffered a catastrophic change in mass. The working standards are cleaned at about 15-year intervals. Within these intervals, however, their mass is redetermined periodically against a third kilogram which is reserved for just this use. This third kilogram is cleaned just prior to its use in recalibrating the working standards. Based on the history of the last 40 years, it appears that the BIPM representation of the SI unit of mass is stable to within about 0.02 mg (0.02 ppm). Therefore, it seems a reasonable goal to achieve compatibility with the mass representation currently maintained at the BIPM. These measurements are reported in detail in [1].

3.2 Corrections for Air Buoyancy

In eq (1), the quantity ρ_a is typically determined from an equation-of-state for moist air. The inputs to this equation are temperature, barometric pressure, relative humidity, and ambient level of carbon dioxide. The last of these has relatively little effect. It is obvious that errors in measuring the required experimental input parameters will propagate to the final result. In the 1970s, however, it was appreciated that the equation-of-state itself has great importance and that several such equations were in wide use. Furthermore, it had not yet been demonstrated experimentally that any of the equations-of-state in use were adequate for actual mass comparisons.

At NIST, Jones derived a semi-empirical equation-of-state based on up-to-date data [10]. This equation, with minor changes, was endorsed for use in mass metrology by the CIPM in 1981 [5]. The equation given in [5] is now referred to as the CIPM-81 equation-of-state for moist air and is used for mass metrology by most national laboratories. The NIST began using this equation for international work in 1981. Use of CIPM-81 instead of its predecessor [11] makes a negligible change to routine mass calibrations. As of January 1, 1990, however, CIPM-81 has been adopted for use in all calibration software.

In order to test the efficacy of CIPM-81, it is necessary to determine the mass difference between two nominally equal weights with and without reliance on the equation-of-state. The latter measurement is typically done in vacuum. This type of comparison was done at the Physikalisch-Technische Bundesanstalt (PTB) [12]. Results agreed to within the expected uncertainty, 1×10^{-4} in ρ_a.

It is also necessary to measure the input parameters with sufficient accuracy. In general, this requires the use of transducers whose calibration is checked at frequent intervals by defining instruments. Our capabilities as they existed in 1985 are described in [1]. Since that time, we have improved the accuracy of our measurements of barometric pressure and of relative humidity.

3.3 Improved Balance

The balance used for primary mass metrology must operate by remote control in order to ensure that the weights being compared remain in sufficient equilibrium with the air of the weighing chamber. Schoonover and Keller have demonstrated that severe systematic errors may intrude if the equilibrium constraints are violated [13]. In addition, the balance itself must have sufficiently high precision. We consider the balance to be suitable when either of the two following conditions is met: 1. The contribution of the balance imprecision to the uncertainty of working standards is negligible compared to the imprecision of routine mass calibrations. 2. The imprecision of the balance is negligible compared to typical instabilities of mass standards.

In [1], we described modifications made to an existing balance which allowed it to fulfill the first criterion. Although working reasonably well, we wanted to improve efficiency by fully automating it. In order to make the job of automation more straight forward, the balance was fitted with an electro-magnetic servocontrol system [14]. Introduction of the servocontrol also resulted in a modest improvement in precision [15].

3.4 Stability of Mass Values

It remains to demonstrate that the work undertaken since 1980 has led to an improved representation of the SI unit of mass.

3.4.1 K20 and K4 The most recent mass value for kilogram K20 results from the 1984 calibration at the BIPM [1]. As discussed in [1], the cleaning process at the BIPM removed significant amounts of surface pollution from the two prototypes. (The kilograms had also been cleaned at NIST but by a less effective technique). Since 1984, NIST has adopted the BIPM cleaning method. Values obtained for the difference in mass between K20 and K4 are shown in figure 3. These have standard deviation of 0.0019 mg. We would expect a standard deviation of 0.0013 mg based solely on the observed standard deviation of the balance which was used. The difference is negligible.

3.4.2 N_1 and N_2 Throughout the last 10 years, N_1 and N_2 continued to be used as working standards for routine mass calibrations. In 1982, they were measured against K20 and K4 prior to sending the latter two weights to BIPM for recalibration. The results, calculated after receiving the new

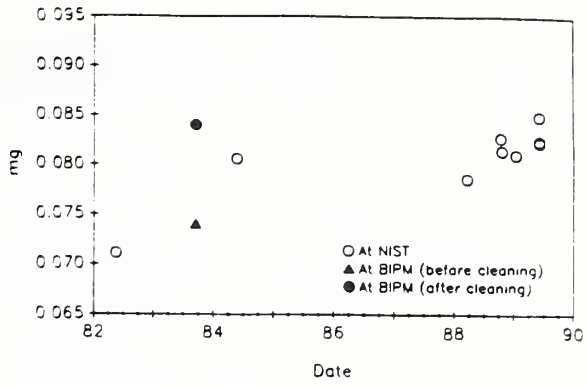


Figure 3. Mass values of K20−K4 as a function of time. Measurements were taken on a balance having a standard deviation of 0.0018 mg for a single reading.

BIPM certificate, indicate that the value of $R/2$ was 0.103 mg±0.025 mg below that accepted. The uncertainty is at an estimated level of one standard deviation and is dominated by problems with auxiliary equipment used in measuring air buoyancy. The value of C was found to be −19.474 mg±0.003 mg, consistent with the control chart data shown in figure 2.

From 1986 to 1988, mass values of N_1 and N_2 were determined three times against K20 and K4 in a more careful series of measurements. Several other stainless-steel kilograms were also involved in the measurements. These are discussed in section 3.4.3, below. It is sufficient to mention at this point that this series of measurements was consistent with the long-term measurements of the other kilograms involved. The results of the 1986–1988 measurements are summarized in table 1. The uncertainty types and the rules for combining uncertainty conform to recommendations of the CIPM [3]. (This reference defines Type A and Type B uncertainties.) Components 2, 4, and 5 will be discussed in more detail in section 3.4.3. In assessing whether the observed change in $R/2$ after 1986 is significant, one must not include Type B components, which we believe to be systematic to all measurements in table 1. It is interesting to note that the observed change in $R/2$ after 1986 is three times greater than the change in C. It is also interesting that the data of figure 1 show a statistically significant variation with time. A linear fit to the data predicts that the value of C in April 1988 was −19.454±0.0028 mg (1 standard deviation), in satisfactory agreement with the measurement shown in table 1.

Table 1. Recent determinations of the masses of kilograms N_1 and N_2 with respect to secondary standards calibrated against K20. The values of C are subject to a measurement uncertainty of 0.0013 mg (1 standard deviation)

Date	$R/2$ $[(N_1+N_2)/2]$	C $[N_1-N_2]$
1986 Aug	1 kg −5.159 mg	−19.440 mg
1987 Nov	1 kg −5.192 mg	−19.451 mg
1988 Apr	1 kg −5.193 mg	−19.447 mg
Accepted values:	1 kg −5.0295 mg	−19.476 mg (1965)
		−19.454 mg (Apr. 1988)

Uncertainty (1 standard deviation or 68 percent confidence level) for measured values of $R/2$

Component	Type A	Type B
1. Instability of K20 since 1985 BIPM calibration	included in 4	nil
2. Calibration precision of secondary standard	0.0035 mg	nil
3. Correction of secondary standards for air buoyancy	included in 2	0.01 mg
4. Instability of secondary standards	0.0036 mg	nil
5. Calibration precision of $(N_1+N_2)/2$	0.0005 mg	nil
6. Correction of N_1 and N_2 for air buoyancy	nil	0.001 mg
RSS	0.0050 mg	0.010 mg
Combined Type A and B: 0.011 mg		

During this period, several kilograms which were submitted to NIST for calibration were measured against N_1 and N_2 using routine calibration procedures. The test kilograms were also measured against stainless-steel kilograms which are discussed in the next section using our best 1-kg balance. The results were, in all cases, consistent with table 1.

There was now good evidence that the accepted value of $R/2$ was 0.164 mg below the accepted value. Less certain evidence suggests that more than half of this difference had been present since at least 1969 (see sec. 2.2.1). This computes to an average change of order −0.004 ppm/yr.

The standards N_1 and N_2 were again checked in 1989. Although these measurements were not as extensive, they show that the average mass had dropped by another 0.05±0.013 mg (1 standard deviation) with respect to four stainless-steel kilograms reserved for special use. This change thus appears to be real and serves as a warning that N_1 and N_2 are now losing mass at a greatly increased rate. The value of C measured during these measurements had returned to within 0.012 mg of the accepted value.

3.4.3 New Secondary and Working Standards of Mass Kilograms N_1 and N_2 have served as both secondary standards—artifacts of practical density which most accurately represent mass as specified in the SI; and working standards—artifacts of practical density used as standards in routine calibration work. Our intention was to separate these roles by acquisition of new standards, all made of non-magnetic stainless steel. The choice of alloy simply reflects the fact that the highest quality 1-kg weights which are commercially available are now made of stainless steel. Several stainless-steel kilograms were already on hand for use as secondardy standards. Three of these, designated D2, E1, and E2 are about 25 years old. The physical characteristics of all three kilograms are similar; D2 was described in some detail in [1]. We also made use of a newer kilogram, designated CH-1, whose characteristics are also described in [1]. The four artifacts were grouped in pairs: CH-1 and D2 formed one pair while E1 and E2 formed the second pair. When not in use, the pairs were stored in separate containers of different design. The pair E1, E2 was never subjected to any type of cleaning except for gentle dusting with a soft brush. The pair CH-1, D2 was cleaned on various occasions.

The pair CH-1, D2 was compared eight times against primary standards K20 and K4. The mass values of CH-1 resulting from these measurements are shown in figure 4(a). Figure 4(b) shows measurements of the mass difference between CH-1 and D2. Note that results displayed in figure 4(a) include a buoyancy correction of approximately 95 mg while the correction for air buoyancy needed for the results in figure 4(b) was less than 3 mg. Figure 5 shows similar data for the pair E1, E2. In this case, however, the pair CH-1, D2 was used as the standard. The mass value assigned to the standard was the same for all the data shown. Pertinent statistical parameters are summarized in table 2. The outlying point in the mass difference of CH-1 and D2 was repeatable. Because the difference returned to its previous values upon recleaning the two kilograms, we assume the outlying value was due to some type of surface contamination. At any

Table 2. Statistical parameters inferred from measurements of secondary standards

Mass of:	s_{total}	DF	s_w	s_b	DF
CH-1	0.0085 mg	8	0.0013 mg		
CH-1 − D2	0.0052 mg	18	0.0013 mg	0.0036 mg	16.9
E1	0.0015 mg	4	0.0011 mg	0.0011 mg	2.5
E1 − E2	0.0016 mg	9	0.0013 mg	0.0009 mg	4.2

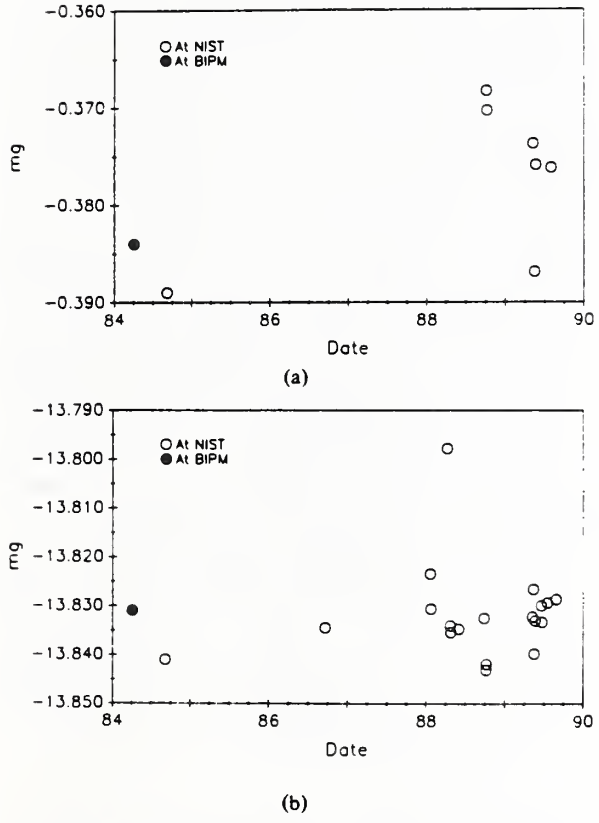


Figure 4. (a) Mass of CH-1 as a function of time. These values are based on direct comparison with K20. The balance used has a standard deviation of 0.0018 mg for a single reading. (b) Mass of CH-1 − D2 as a function of time using the same balance as in (a). There is one outlying point which has been excluded in the data analysis.

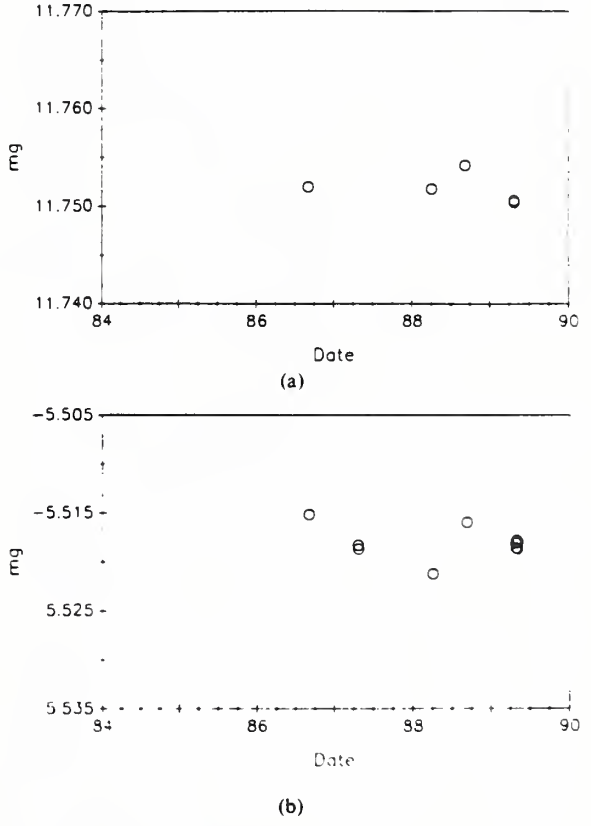


Figure 5. (a) Mass of E1 as a function of time. These values are based on direct comparison with CH-1 and D2. The balance used has a standard deviation of 0.0018 mg for a single reading. (b) Mass of E1 − E2 as a function of time using the same balance as in (a).

rate, the outlying point is not included in the calculations for table 2.

In table 2, s_{total} is the estimated standard deviation of the data shown in figures 4 and 5. The number of degrees of freedom in this estimate is given in the next column. The quantity s_w refers to the "within-group" standard deviation—that component of the observed standard deviation which can be attributed to the balance precision. This number is pooled from a great many measurements and thus has a large number of degrees of freedom. The "between-group" standard deviation, s_b, is a measure of increased variability seen over long time periods. This quantity is calculated from the others in the table. The estimated number of degrees of freedom [16] in s_b is given in the last column. A full discussion of these parameters as well as their treatment in the context of mass calibrations has been given by Croarkin [17]. It is interesting to note that

the data of figure 2, when subjected to the same analysis, indicate that $\bar{s}_b$ for these measurements is 0.0116 mg [17].

The Croarkin model is not sufficient to model direct comparisons of CH-1 and D2 with K20 and K4. This is because uncertainties in buoyancy corrections have little effect on measured differences between weights of the same density but have large effects on measured differences between weights of different density. While the transducers used to measure the parameters of temperature, pressure, relative humidity, and carbon-dioxide level have excellent short-term precision, slow drifting between recalibration leads to an additional "between group" uncertainty. If the error model of [17] is extended to include buoyancy effects, the data of table 2 can be used to compute an additional parameter $s_\rho = 0.007$ mg (DF = 3.9). This parameter characterizes daily variability in the measured mass difference between a kilogram of platinum-iridium and a kilogram of stainless steel due solely to measurement of the air buoyancy correction.

Although based on somewhat limited data, it seems that E1 and E2, kilograms of the identical alloy and which are never cleaned, have a more stable mass than CH-1 and D2. This is a curious result in the sense that mass values for E1 and E2 are based on direct comparison with CH-1 and D2. In these comparisons, it is assumed that the summation mass of CH-1 and D2 is the average of all recent measurements which are in statistical control. The evidence thus suggests that this average is a better estimate of the mass of CH-1 and D2 than, for instance, the most recently obtained values.

As mentioned in the introduction to this section, it was envisioned that use of N_1 and N_2 as working standards would be superseded by stainless-steel kilograms. These would have a nominal density of 8000 kg·m^{-3}. In 1985, six such kilograms, identical to CH-1, were obtained for this purpose. They are marked 1,2,...,6 but for purposes of discussion we shall refer to them as C1, C2,...,C6. Until January 1988, these six kilograms were used extensively for various cleaning studies. Now, however, they will be used as working standards as described below in section 4.

4. Summary of the Change on January 1, 1990

Beginning on January 1, 1990, the mass values assigned to working standards of the NIST calibration service are based on a calibration chain which starts with mass values assigned to NIST primary standards K20 and K4 by the BIPM, continues with mass values assigned to secondary standards CH-1 and D2 with direct reference to K20 and K4, and finally to working standards C1, C2,...,C6 by direct reference to CH-1 and D2.

4.1 Effect on Industry and Technology

An Ad Hoc Committee of the National Conference of Standards Laboratories (NCSL) was formed in order to help assess industrial and technological implications of the actions contemplated for January 1, 1990. Members of the Committee include representatives from civilian and military standards laboratories, balance manufacturers, and weight manufacturers. All were asked to estimate the impact which a change of roughly 0.15 mg/kg would have on their programs. The members could not identify a single instance where such a change would affect a manufactured product or a critical measurement. Virtually all concerned, however, recognized that a change of this magnitude could be noticeable within their metrology laboratory. This is not surprising since typical NIST calibrations give an uncertainty of about 0.075 mg (3 standard deviations) for calibrations of 1-kg standards and users of these standards often have balances of comparable precision to our own.

In recent years, calibrations for primary national laboratories of other countries have been carried out using secondary standards CH-1 and D2 with assigned values based directly on measurements against K20. These measurements are not, therefore, in need of correction.

4.2 Implementation

Based on the data shown in section 3.4.2, it is clear that, by 1988, mass values assigned to NIST working standards were some 0.164 mg/kg higher than our best estimate of their actual value (that is, the value directly traceable to the representation of the SI unit of mass). At the beginning of the decade, the discrepancy was about 0.10 mg/kg. There is evidence that, between 1988 and 1989, the discrepancy grew still greater.

In early 1988, and based on the data available to that point, it was decided to assign new mass values to NIST working standards on January 1, 1990. On the same date, the new quality-control procedures designed to keep mass values assigned to NIST working standards closely tied to the SI representation of mass would be in place. Various standards

organizations were informed of these intentions by letter. The letter also stated that the new mass values would be of the order of 0.15 ppm lower than the present values. Also in 1988, the NCSL Ad Hoc Committee was established to help in the implementation of the change. The target date of January 1, 1990 was chosen to coincide with the date on which international changes in the representations of the SI volt, ohm, and kelvin would be implemented. Guidelines developed by the Ad Hoc Committee are given in the Appendix.

These guidelines treat the discrepancy between the accepted mass of NIST working standards and the mass traceable to SI representations as equal in magnitude to 0.17 mg/kg (0.17 ppm) throughout the decade from 1980 through 1989. Based on data presented above, we see that this is an oversimplification. Our best data, taken between 1986 and 1988, give the discrepancy as 0.164 mg/kg. Less accurate data, however, suggest that the discrepancy grew slowly throughout the decade and then increased rapidly in the last year. A time-dependent correction algorithm with time-dependent uncertainty could, of course, be devised based on these data. The complexity of applying such an algorithm combined with its trivial scientific or technological benefit made this course unwise. Instead we recommend correction of −0.17 mg/kg made to NIST calibration certificates dated during the 1980s. This, we believe, will provide sufficient continuity with certificates issued after January 1, 1990.

The BIPM is conducting the 3rd verification of national prototype kilograms. When this exercise is completed (perhaps in 2 years) we will have a much better idea of the internal stability of BIPM standards and the stability of these standards with respect to the national prototype kilograms. For the present, we estimate that the mass values used by NIST in its calibrations represent SI values as maintained by the BIPM to within 0.03 mg/kg or 0.03 ppm (1 standard deviation). This uncertainty will not be included in NIST calibration reports except to say that it is systematic to all mass measurements.

5. Future Plans

We plan to participate in the 3rd verification of national prototype kilograms being organized by the BIPM. Consequently, in early 1990, we will send our national prototype (K20) to BIPM for a lengthy set of comparisons.

We plan to recalibrate our working standards in terms of secondary standards CH-1 and D2 at approximately 6-month intervals. The working standards will not, initially, be cleaned although the secondary standards will. We foresee calibrating the secondary standards in terms of our primary standards K20 and K4 at about 2-year intervals. Based on the data presented above, we believe this procedure will permit us to know the mass ratio between our working standards and our primary standards to within 0.01 ppm (1 standard deviation) at all times. As noted at the end of the previous section, this uncertainty does not include possible discrepancies between NIST standards and those of the BIPM. We tentatively set the latter uncertainty at 0.03 ppm (1 standard deviation).

It would be helpful to have a balance of 1-kg capacity and a standard deviation of order 0.005 mg for use in routine calibration work. Such a device would help compensate for the fact that, since January 1, 1990, we are formally recognizing that our working standards are subject to uncertainty.

A major goal of the new quality-control system is to improve international compatibility regarding practical mass standards. We are, therefore, seeking to promote international comparisons of stainless-steel mass standards in order to ascertain the degree of compatibility among various industrialized countries.

In conclusion, we note that a system of metrology ultimately based on an artifact standard will necessarily have shortcomings. Over a long enough period of time, mass differences between any two artifact standards will be unstable; the estimated standard deviation based on the complete data record will diverge. If the mass of one of the artifacts is arbitrarily assumed to be constant, its actual instability will in time be revealed by measurements of true physical constants. While there has as yet been no such revelation [18], modern technology may soon be expected to put the present definition of the SI kilogram to a severe test.

6. Appendix. Notice of Change in the Unit of Mass Traceable to The National Institute of Standards and Technology

On January 1, 1990 the unit of mass as disseminated by the National Institute of Standards and Technology (NIST) will shift by 0.17 mg/kg (0.17

ppm). This small shift will bring the unit of mass traceable to NIST into better agreement with international standards. Since the avoirdupois pound is defined as 0.45359237 kg, pound masses traceable to NIST will also be affected to the same extent (0.17 μlb/lb, or 0.17 ppm).

Most people will be unaffected by this small change so that continued traceability to NIST can be maintained without taking any action. Unaffected users will be those whose mass standards are assigned an uncertainty greater than 1 mg/kg or 1μlb/lb (1 ppm). Included in the *un*affected list are:

1. Analytical weights certified to be within any of the tolerances prescribed by NIST/NBS or ASTM/ANSI or to any OIML tolerance except E_1.
2. Direct-reading balances and scales.
3. Any analytical weights which have been assigned an uncertainty greater than 1 mg/kg or 1 μlb/lb (1 ppm). [This will typically include all weights greater than 2 kg or less than 20 g which were calibrated by NIST/NBS (see table A1). In some special cases, however, NIST calibrations at weight denominations other than those shown in table A1 may have an uncertainty lower than 1 mg/kg.]

Traceability to NIST of the above three categories is unaffected by the change which will take effect on January 1, 1990. No action need be taken. In addition, any calibration certificate dated January 1, 1990 or later already has any necessary changes incorporated.

Table A1. Typically, action need be taken only for these nominal values of weights *and* only if the assigned uncertainty is below the value given. Although this table shows weight denominations most likely to require correction, denominations which may require correction are not necessarily limited to those shown

Nominal mass	Uncertainty	Nominal mass	Uncertainty
2 kg	2.00 mg	50 lb	50 μlb
1 kg	1.00 mg	30 lb	30 μlb
		20 lb	20 μlb
500 g	0.50 mg	10 lb	10 μlb
300 g	0.30 mg		
200 g	0.20 mg	5 lb	5 μlb
100 g	0.10 mg	3 lb	3 μlb
		2 lb	2 μlb
50 g	0.05 mg	1 lb	1 μlb
30 g	0.03 mg		
20 g	0.02 mg	0.5 lb	0.5 μlb
		0.3 lb	0.3 μlb
		0.2 lb	0.2 μlb

Weights which will be affected by the change which will take effect on January 1, 1990 are all those which do not fall into category 3 above and, in addition, whose calibration certificate bears a date before January 1, 1990. Affected weights are those which have an assigned calibration uncertainty of less than 1 mg/kg (1 ppm). Based on typical NIST calibration reports, these will generally be weights with denominations between 2 kg and 20 g or 5 lb and 0.2 lb. Other denominations may be affected in special cases, however.

The following actions will be necessary in order to maintain traceability to NIST for the affected weights:

a. Weights whose calibration certificate bears a date after January 1, 1980 and before January 1, 1990.

   After January 1, 1990 the mass of each affected weight should be *reduced* by 0.17 mg/kg (0.17 ppm) as shown in table A2. This applies both to the true mass and the apparent mass. The uncertainty stated in the report remains the same.

   (Alternatively, the mass values stated in the calibration certificate may remain uncorrected provided the stated uncertainty is increased by 0.17 mg/kg).

b. Weight sets whose calibration certificate bears a date before January 1, 1980 but which have been subjected to a surveillance test within the 10 years preceding January 1, 1990. (An example of a surveillance report

Table A2. Corrections to apply to calibrations dated between January 1, 1980 and January 1, 1990. The denominations shown are those of table A1

Nominal mass	Correction	Nominal mass	Correction
2 kg	−0.3400 mg	50 lb	−8.500 μlb
1 kg	−0.1700 mg	30 lb	−5.100 μlb
		20 lb	−3.400 μlb
500 g	−0.0850 mg	10 lb	−1.700 μlb
300 g	−0.0510 mg		
200 g	−0.0340 mg	5 lb	−0.850 μlb
100 g	−0.0170 mg	3 lb	−0.510 μlb
		2 lb	−0.340 μlb
50 g	−0.0085 mg	1 lb	−0.170 μlb
30 g	−0.0051 mg		
20 g	−0.0034 mg	0.5 lb	−0.085 μlb
		0.3 lb	−0.051 μlb
		0.2 lb	−0.034 μlb

issued by NIST is shown at the end of this Appendix.

After January 1, 1990 the mass of each affected weight should be *reduced* by 0.17 mg/kg (0.17 ppm) as shown in table A2. This applies both to the true mass and the apparent mass. The assertions of the surveillance report will remain in effect.

(Alternatively, the mass values stated in the calibration certificate may remain uncorrected provided the stated uncertainty is increased by 0.17 mg/kg).

c. Weights whose calibration certificate bears a date before January 1, 1980 and which have had no surveillance test subsequent to January 1, 1980.

After January 1, 1990 the uncertainty assigned to each affected weight should be increased by 0.17 mg/kg (0.17 ppm) until a new calibration or surveillance test is performed.

Sample Surveillance Report Issued by NIST

April 1, 1987

In reply refer to: 731/12345

Company XYZ
1 Metrology Blvd.
Grovers Corner, NJ 00000
Attention: J. Doe

Subject: Recalibration of Mass Standards previously calibrated under NBS Test No. 00/G00000 (copy attached)

Items: Nine (9) Mass Standards: 100 g – 1 g

The above items have been intercompared in sums. The differences as measured have been compared with the differences computed from the value under G00000. One or more of the items have been checked against national standards. The results of this test indicate that there is no significant change since the last calibration. This test assures the continuing accuracy of the values under G00000.

Sincerely,

Richard S. Davis
Group Leader, Mass Group
Center for Manufacturing Engineering

Attachment

7. Acknowledgments

Many colleagues at NIST have assisted in one or more areas of the above work. Dr. Joe D. Simmons first directed that the work be done. Mr. Randall Schoonover and Mr. Jerry Keller provided historical information and much useful advice. Mr. Henry Oppermann of the Office of Weights and Measures shared historical data with the author. Mrs. Ruth Varner and Mrs. M. Carroll Croarkin provided computational help and welcome advice on statistical questions.

The National Conference of Standards Laboratories aided materially by organizing an Ad Hoc Committee under their aegis. Committee members helped assess the technological implications of the changes discussed above and recommended methods of implementation of those changes.

The staff of the BIPM provided calibrations of NIST mass standards and cooperated fully in detailed explanations of their calibration process.

About the author: Richard S. Davis is a physicist in the NIST Center for Manufacturing Engineering.

8. References

[1] Davis, R. S., J. Res. Natl. Bur. Stand. (U.S.) **90**, 263 (1985).
[2] See for instance, Pontius, P. E., and Cameron, J. M., Realistic Uncertainties and the Mass Measurement Process, Natl. Bur. Stand. (U.S.) Monograph 103, August 1967.
[3] Giacomo, P., Metrologia **17**, 73 (1981).
[4] Comité International des Poids et Mesures, Document CIPM/89-9.
[5] Giacomo, P., Metrologia **18**, 33 (1982).
[6] Deslattes, R. D., in Proceedings of course LXVIII Metrology and Fundamental Constants, Summer School of Physics—Enrico Fermi, Varenna Italy (1976), Soc. Italiana di Fisica, Bologna (1980) p. 38.
[7] Bower, V. E., Davis, R. S., Murphy, T. J., Paulsen, P. J., Gramlich, J. W., and Powell, L. J., J. Res. Natl. Bur. Stand. (U.S.) **87**, 21 (1982).
[8] Cohen, E. R., and Taylor, B. N., Rev. Mod. Phys. **59**, 1121 (1987).
[9] Almer, H. E., J. Res. Natl. Bur. Stand. (U.S.) **76C**, 1 (1972).
[10] Jones, F. E., J. Res. Natl. Bur. Stand. (U.S.) **83**, 419 (1978).
[11] Varner, R. N., and Raybold, R. C., National Bureau of Standards Mass Calibration Computer Software, Natl. Bur. Stand. (U.S.) Tech. Note 1127, July 1980.
[12] Balhorn, R., PTB-Mitteilungen **93**, 303 (1983).
[13] Schoonover, R. M., and Keller, J., in Report of the 68th National Conference on Weights and Measures 1983, Natl. Bur. Stand. (U.S.) Special Publ. 663 (1983) p. 39.
[14] The basic design of the servo system is given in: Schoonover, R. M., and Taylor, J. E., An Investigation of a user-operated mass calibration package, Natl. Inst. Stand. Technol. Report NISTIR 88-3876 (1988).
[15] Davis, R. S., Comité Consultatif pour la Masse et les Grandeurs Apparentées, Document CCM/88-8 (1988).
[16] Brownlee, K. A., Statistical Theory and Methodology in Science and Engineering, John Wiley and Sons, Inc., New York (1965) p. 300.
[17] Croarkin, C., Metrologia **26**, 107 (1989).
[18] Davis, R. S., Metrologia **26**, 75 (1989).

NIST-114A
(REV. 3-89)

U.S. DEPARTMENT OF COMMERCE
NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY

BIBLIOGRAPHIC DATA SHEET

1. PUBLICATION OR REPORT NUMBER
NISTIR 4941

2. PERFORMING ORGANIZATION REPORT NUMBER

3. PUBLICATION DATE
NOVEMBER 1992

4. TITLE AND SUBTITLE

Selected Publications for the Advanced Mass Measurements Workshop

5. AUTHOR(S)

Georgia L. Harris, editor

6. PERFORMING ORGANIZATION (IF JOINT OR OTHER THAN NIST, SEE INSTRUCTIONS)

U.S. DEPARTMENT OF COMMERCE
NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY
GAITHERSBURG, MD 20899

7. CONTRACT/GRANT NUMBER

8. TYPE OF REPORT AND PERIOD COVERED

9. SPONSORING ORGANIZATION NAME AND COMPLETE ADDRESS (STREET, CITY, STATE, ZIP)

10. SUPPLEMENTARY NOTES

[ ] DOCUMENT DESCRIBES A COMPUTER PROGRAM; SF-185, FIPS SOFTWARE SUMMARY, IS ATTACHED.

11. ABSTRACT (A 200-WORD OR LESS FACTUAL SUMMARY OF MOST SIGNIFICANT INFORMATION. IF DOCUMENT INCLUDES A SIGNIFICANT BIBLIOGRAPHY OR LITERATURE SURVEY, MENTION IT HERE.)

This volume is a collection of NBS/NIST publications that have been found essential and useful in the Advanced Mass Measurements Workshop sponsored by the Office of Weights and Measures. Publications in this volume originally appeared as articles and in numerous publications. Some of the original publications are out of print. This volume also contains information relevant to the change in the mass unit of 1990. This publication, which combines many relevant publications edited for the workshop, will serve as a useful text and reference for all professionals interested in mass calibrations.

12. KEY WORDS (6 TO 12 ENTRIES; ALPHABETICAL ORDER; CAPITALIZE ONLY PROPER NAMES; AND SEPARATE KEY WORDS BY SEMICOLONS)

calibration designs; mass measurement; measurement assurance, statistical analysis, statistical control

13. AVAILABILITY

[ ] UNLIMITED
[X] FOR OFFICIAL DISTRIBUTION. DO NOT RELEASE TO NATIONAL TECHNICAL INFORMATION SERVICE (NTIS).
[ ] ORDER FROM SUPERINTENDENT OF DOCUMENTS, U.S. GOVERNMENT PRINTING OFFICE, WASHINGTON, DC 20402.
[ ] ORDER FROM NATIONAL TECHNICAL INFORMATION SERVICE (NTIS), SPRINGFIELD, VA 22161.

14. NUMBER OF PRINTED PAGES
662

15. PRICE
A99

ELECTRONIC FORM