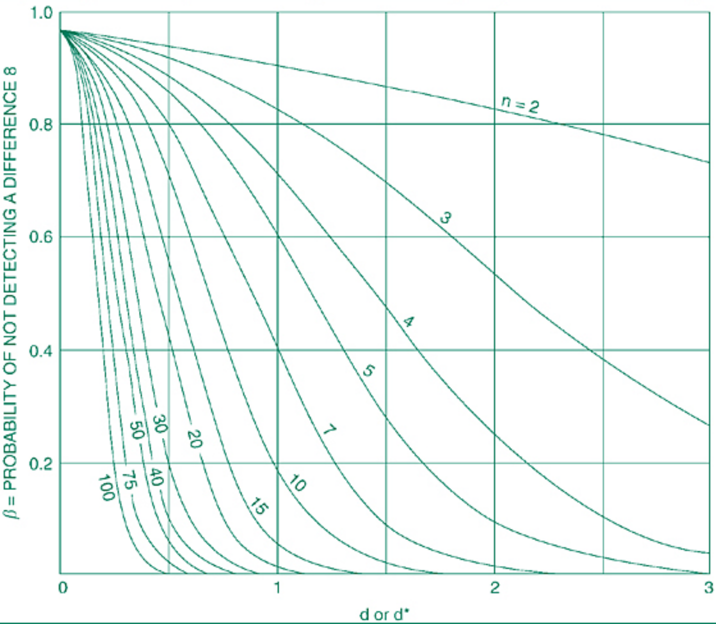
SECOND EDITION

Statistical Techniques for Data Analysis



John K. Taylor Ph.D. and Cheryl Cihon Ph.D.

CHAPMAN & HALL/CRC

Statistical Techniques

for

Data Analysis

Second Edition

Statistical Techniques

for

Data Analysis

Second Edition

John K. Taylor Ph.D.
Formerly of the National Institute of
Standards and Technology

and

Cheryl Cihon Ph.D.
Bayer HealthCare, Pharmaceuticals

CHAPMAN & HALL/CRC

A CRC Press Company
Boca Raton London New York Washington, D.C.

Library of Congress Cataloging-in-Publication Data

Cihon, Cheryl.
Statistical techniques for data analysis / Cheryl Cihon, John K. Taylor.—2nd. ed.
p. cm.
Includes bibliographical references and index.
ISBN 1-58488-385-5 (alk. paper)
1. Mathematical statistics. I. Taylor, John K. (John Keenan), 1912-II. Title.

QA276.C4835 2004
519.5—dc22 2003062744

Direct all inquiries to CRC Press LLC, 2000 N.W. Corporate Blvd., Boca Raton, Florida 33431.

Visit the CRC Press Web site at www.crcpress.com

International Standard Book Number 1-58488-385-5
Library of Congress Card Number 2003062744
Printed in the United States of America 1 2 3 4 5 6 7 8 9 0
Printed on acid-free paper

Preface

Data are the products of measurement. Quality measurements are only achievable if measurement processes are planned and operated in a state of statistical control. Statistics has been defined as the branch of mathematics that deals with all aspects of the science of decision making in the face of uncertainty. Unfortunately, there is great variability in the level of understanding of basic statistics by both producers and users of data.

The computer has come to the assistance of the modern experimenter and data analyst by providing techniques for the sophisticated treatment of data that were unavailable to professional statisticians two decades ago. The days of laborious calculations with the ever-present threat of numerical errors when applying statistics of measurements are over. Unfortunately, this advance often results in the application of statistics with little comprehension of meaning and justification. Clearly, there is a need for greater statistical literacy in modern applied science and technology.

There is no dearth of statistics books these days. There are many journals devoted to the publication of research papers in this field. One may ask the purpose of this particular book. The need for the present book has been emphasized to the authors during their teaching experience. While an understanding of basic statistics is essential for planning measurement programs and for analyzing and interpreting data, it has been observed that many students have less than good comprehension of statistics, and do not feel comfortable when making simple statistically based decisions. One reason for this deficiency is that most of the numerous works devoted to statistics are written for statistically informed readers.

To overcome this problem, this book is not a statistics textbook in any sense of the word. It contains no theory and no derivation of the procedures presented and presumes little or no previous knowledge of statistics on the part of the reader. Because of the many books devoted to such matters, a theoretical presentation is deemed to be unnecessary, However, the author urges the reader who wants more than a working knowledge of statistical techniques to consult such books. It is modestly hoped that the present book will not only encourage many readers to study statistics further, but will provide a practical background which will give increased meaning to the pursuit of statistical knowledge.

This book is written for those who make measurements and interpret experimental data. The book begins with a general discussion of the kinds of data and how to obtain meaningful measurements. General statistical principles are then described, followed by a chapter on basic statistical calculations. A number of the most frequently used statistical techniques are described. The techniques are arranged for presentation according to decision situations frequently encountered in measurement or data analysis. Each area of application and corresponding technique is explained in general terms yet in a correct scientific context. A chapter follows that is devoted to management of data sets. Ways to present data by means of tables, charts, graphs, and mathematical expressions are next considered. Types of data that are not continuous and appropriate analysis techniques are then discussed. The book concludes with a chapter containing a number of special techniques that are used less frequently than the ones described earlier, but which have importance in certain situations.

Numerous examples are interspersed in the text to make the various procedures clear. The use of computer software with step-by-step procedures and output are presented. Relevant exercises are appended to each chapter to assist in the learning process.

The material is presented informally and in logical progression to enhance readability. While intended for self-study, the book could provide the basis for a short course on introduction to statistical analysis or be used as a supplement to both undergraduate and graduate studies for majors in the physical sciences and engineering.

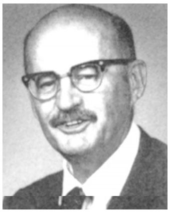
The work is not designed to be comprehensive but rather selective in the subject matter that is covered. The material should pertain to most everyday decisions relating to the production and use of data.

Acknowledgments

The second author would like to express her gratitude to all the teachers of statistics who, over the years, encouraged her development in the area and gave her the tools to undertake such a project.

Dedication

This book is dedicated to the husband, son and family of Cheryl A. Cihon, and to the memory of John K. Taylor.



The late John K. Taylor was an analytical chemist of many years of varied experience. All of his professional life was spent at the National Bureau of Standards, now the National Institute of Standards and Technology, from which he retired after 57 years of service.

Dr. Taylor received his BS degree from George Washington University and MS and PhD degrees from the University of Maryland. At the National Bureau of Standards, he served first as a research chemist, and then managed research and development programs in general analytical chemistry, electrochemical analysis, microchemical analysis, and air, water, and particulate analysis. He coordinated the NBS Center for Analytical Chemistry's Program in quality assurance, and conducted research activities to develop advanced concepts to improve and assure measurement reliability. He provided advisory services to other government agencies as part of his official duties as well as consulting services to government and industry in analytical and measurement programs.

Dr. Taylor authored four books, and wrote over 220 research papers in analytical chemistry. Dr. Taylor received several awards for his accomplishments in analytical chemistry, including the Department of Commence Silver and Gold Medal Awards. He served as past chairman of the Washington Academy of Sciences, the ACS Analytical Chemistry Division, and the ASTM Committee D 22 on Sampling and Analysis of Atmospheres.

Cheryl A. Cihon is currently a biostatistician in the pharmaceutical industry where she works on drug development projects relating to the statistical aspects of clinical trial design and analysis.

Dr. Cihon received her BS degree in Mathematics from McMaster University, Ontario, Canada as well as her MS degree in Statistics. Her PhD degree was granted from the University of Western Ontario, Canada in the field of Biostatistics. At the Canadian Center for Inland Waters, she was involved in the analysis of environmental data, specifically related to toxin levels in major lakes and rivers throughout North America. Dr. Cihon also worked as a statistician at the University of Guelph, Canada, where she was involved with analyses pertaining to population medicine. Dr. Cihon has taught many courses in advanced statistics throughout her career and served as a statistical consultant on numerous projects.

Dr. Cihon has authored one other book, and has written many papers for statistical and pharmaceutical journals. Dr. Cihon is the recipient of several awards for her accomplishments in statistics, including the National Sciences and Engineering Research Council award.

Table of Contents

List of Figures

List of Tables

CHAPTER 1

What are Data?

Data may be considered to be one of the vital fluids of modern civilization. Data are used to make decisions, to support decisions already made, to provide reasons why certain events happen, and to make predictions on events to come. This opening chapter describes the kinds of data used most frequently in the sciences and engineering and describes some of their important characteristics.

DEFINITION OF DATA

The word data is defined as things known, or assumed facts and figures, from which conclusions can be inferred. Broadly, data is raw information and this can be qualitative as well as quantitative. The source can be anything from hearsay to the result of elegant and painstaking research and investigation. The terms of reporting can be descriptive, numerical, or various combinations of both. The transition from data to knowledge may be considered to consist of the hierarchal sequence

$$\text{Data} \xrightarrow{\text{analysis}} \text{Information} \xrightarrow{\text{model}} \text{Knowledge}$$

Ordinarily, some kind of analysis is required to convert data into information. The techniques described later in this book often will be found useful for this purpose. A model is typically required to interpret numerical information to provide knowledge about a specific subject of interest. Also, data may be acquired, analyzed, and used to test a model of a particular problem.

Data often are obtained to provide a basis for decision, or to support a decision that may have been made already. An objective decision requires unbiased data but this should never be assumed. A process used for the latter purpose may be more biased than one for the former purpose, to the extent that the collection, accumulation, or production process may be biased, which is to say it may ignore other possible bits of information. Bias may be accidental or intentional. Preassumptions and even prior misleading data can be responsible for intentional bias, which may be justified. Unfortunately, many compilations of data provide little if any information about intentional biases or modifying circumstances that could affect decisions based upon them, and certainly nothing about unidentified bias.

Data producers have the obligation to present all pertinent information that would impact on the use of it, to the extent possible. Often, they are in the best position to provide such background information, and they may be the only source of information on these matters. When they cannot do so, it may be a condemnation of their competence as metrologists. Of course, every possible use of data cannot be envisioned when it is produced, but the details of its production, its limitations, and quantitative estimates of its reliability always can be presented. Without such, data can hardly be classified as useful information.

Users of data cannot be held blameless for any misuse of it, whether or not they may have been misled by its producer. No data should be used for any purpose unless their reliability is verified. No matter how attractive it may be, unevaluated data are virtually worthless and the temptation to use them should be resisted. Data users must be able to evaluate all data that they utilize or depend on reliable sources to provide such information to them.

It is the purpose of this book to provide insight into data evaluation processes and to provide guidance and even direction in some situations. However, the book is not intended and cannot hope to be used as a “cook book” for the mechanical evaluation of numerical information.

KINDS OF DATA

Some data may be classified as “soft” which usually is qualitative and often makes use of words in the form of labels, descriptors, or category assignments as the primary mode of conveying information. Opinion polls provide soft data, although the results may be described numerically. Numerical data may be classified as “hard” data, but one should be aware, as already mentioned, that such can have a soft underbelly. While recognizing the importance of soft data in many situations, the chapters that follow will be concerned with the evaluation of numerical data. That is to say, they will be concerned with quantitative, instead of qualitative data.

Natural Data

For the purposes of the present discussion, natural data is defined as that describing natural phenomena, as contrasted with that arising from experimentation. Observations of natural phenomena have provided the background for scientific theory and principles and the desire to obtain better and more accurate observations has been the stimulus for advances in scientific instrumentation and improved methodology. Physical science is indebted to natural science which stimulated the development of the science of statistics to better understand the variability of nature. Experimental studies of natural processes provided the impetus for the development of the science of experimental design and planning. The boundary between physical and natural science hardly exists anymore, and the latter now makes extensive use of physical measuring techniques, many of which are amenable to the data evaluation procedures described later.

Studies to evaluate environmental problems may be considered to be studies of natural phenomena in that the observer plays essentially a passive role. However, the observer can have control of the sampling aspects and should exercise it, judiciously, to obtain meaningful data.

Experimental Data

Experimental data result from a measurement process in which some property is measured for characterization purposes. The data obtained consist of numbers that often provide a basis for decision. This can range anywhere from discarding the data, modifying it by exclusion of some point or points, or using it alone or in connection with other data in a decision process. Several kinds of data may be obtained as will be described below.

Counting Data and Enumeration

Some data consist of the results of counting. Provided no blunders are involved, the number obtained is exact. Thus several observers would be expected to obtain the same result. Exceptions would occur when some judgment is involved as to what to count and what constitutes a valid event or an object that should be counted. The optical identification and counting of asbestos fibers is an example of the case in point. Training of observers can minimize variability in such cases and is often required if consistency of data is to be achieved. Training is best done on a direct basis, since written instructions can be subject to variable interpretation. Training often reflects the biases of the trainer. Accordingly, serial training (training some one who trains another who, in turn, trains others) should be avoided. Perceptions can change with time, in which case training may need to be a continuing process. Any process involving counting should not be called measurement but rather enumeration.

Counting of radioactive disintegrations is a special and widely practiced area of counting. The events counted (e.g., disintegrations) follow statistical principles that are well understood and used by the practitioners, so will not be discussed here. Experimental factors such as geometric relations of samples to counters and the efficiency of detectors can influence the results, as well. These, together with sampling, introduce variability and sources of bias into the data in much the same way as happens for other types of measurement and thus can be evaluated using the principles and practices discussed here.

Discrete Data

Discrete data describes numbers that have a finite possible range with only certain individual values encountered within this range. Thus, the faces on a die can be numbered, one to six, and no other value can be recorded when a certain face appears.

Numerical quantities can result from mathematical operations or from measurements. The rules of significant figures apply to the former and statistical significance applies to the latter. Trigonometric functions, logarithms, and the value of π, for example, have discrete values but may be rounded off to any number of figures for computational or tabulation purposes. The uncertainty of such numbers is due to rounding alone, and is quite a different matter from measurement uncertainty. Discrete numbers should be used in computation, rounded consistent with the experimental data to which they relate, so that the rounding does not introduce significant error in a calculated result.

Continuous Data

Measurement processes usually provide continuous data. The final digit observed is not the result of rounding, in the true sense of the word, but rather to observational limitations. It is possible to have a weight that has a value of 1.000050...0 grams but not likely. A value of 1.000050 can be uncertain in the last place due to measurement uncertainty and also to rounding. The value for the kilogram (the world's standard of mass) residing in the International Bureau in Paris is 1.000...0 kg by definition; all other mass standards will have an uncertainty for their assigned value.

VARIABILITY

Variability is inevitable in a measurement process. The operation of a measurement process does not produce one number but a variety of numbers. Each time it is applied to a measurement situation it can be expected to produce a slightly different number or sets of numbers. The means of sets of numbers will differ among themselves, but to a lesser degree than the individual values.

One must distinguish between natural variability and instability. Gross instability can arise from many sources, including lack of control of the process [1]. Failure to control steps that introduce bias also can introduce variability. Thus, any variability in calibration, done to minimize bias, can produce variability of measured values.

A good measurement process results from a conscious effort to control sources of bias and variability. By diligent and systematic effort, measurement processes have been known to improve dramatically. Conversely, negligence and only sporadic attention to detail can lead to deterioration of precision and accuracy. Measurement must entail practical considerations, with the result that precision and accuracy that is merely "good enough", due to cost-benefit considerations, is all that can be obtained, in all but rare cases. The advancement of the state-of-the-art of chemical analysis provides better precision and accuracy and the related performance characteristics of selectivity, sensitivity, and detection [1].

The inevitability of variability complicates the evaluation and use of data. It must be recognized that many uses require data quality that may be difficult to achieve. There are minimum quality standards required for every measurement situation (sometimes called data quality objectives). These standards should be established in advance and both the producer and the user must be able to determine whether they have been met. The only way that this can be accomplished is to attain statistical control of the measurement process [1] and to apply valid statistical procedures in the analysis of the data.

POPULATIONS AND SAMPLES

In considering measurement data, one must be familiar with the concepts and distinguish between (1) a population and (2) a sample. Population means all of an object, material, or area, for example, that is under investigation or whose properties need to be determined. Sample means a portion of a population. Unless the population is simple and small, it may not be possible to examine it in its entirety. In that case, measurements are often made on samples believed to be representative of the population of interest.

Measurement data can be variable due to variability of the population and to all aspects of the process of obtaining a sample from it. Biases can result for the same reasons, as well. Both kinds of sample-related uncertainty – variability and bias – can be present in measurement data in addition to the uncertainty of the measurement process itself. Each kind of uncertainty must be treated somewhat differently (see Chapter 5), but this treatment may not be possible unless a proper statistical design is used for the measurement program. In fact, a poorly designed (or missing) measurement program could make the logical interpretation of data practically impossible.

IMPORTANCE OF RELIABILITY

The term reliability is used here to indicate quality that can be documented, evaluated, and believed. If any one of these factors is deficient in the case of any data, the reliability and hence the confidence that can be placed in any decisions based on the data is diminished.

Reliability considerations are important in practically every data situation but they are especially important when data compilations are made and when data produced by several sources must be used together. The latter situation gives rise to the concept of data compatibility which is becoming a prime requirement for environmental data [1,2]. Data compatibility is a complex concept, involving both statistical quality specification and adequacy of all components of the measurement system, including the model, the measurement plan, calibration, sampling, and the quality assurance procedures that are followed [1].

A key procedure for assuring reliability of measurement data is peer review of all aspects of the system. No one person can possibly think of everything that could cause measurement problems in the complex situations so often encountered. Peer review in the planning stage will broaden the base of planning and minimize problems in most cases. In large measurement programs, critical review at various stages can verify control or identify incipient problems.

Choosing appropriate reviewers is an important aspect of the operation of a measurement program. Good reviewers must have both detailed and general knowledge of the subject matter in which their services are utilized. Too many reviewers misunderstand their function and look too closely at the details while ignoring the generalities. Unless specifically named for that purpose, editorial matters should be deferred to those with redactive expertise. This is not to say that glaring editorial trespasses should be ignored, but rather the technical aspects of review should be given the highest priority.

The ethical problems of peer review have come into focus in recent months. Reviews should be conducted with the highest standards of objectivity. Moreover, reviewers should consider the subject matter reviewed as privileged information. Conflicts of interest can arise as the current work of a reviewer parallels too closely that of the subject under review. Under such circumstances, it may be best to abstain.

In small projects or tasks, supervisory control is a parallel activity to peer review. Peer review of the data and the conclusions drawn from it can increase the reliability of programs and should be done. Supervisory control on the release of data is necessary for reliable individual measurement results. Statistics and statistically based judgments are key features of reviews of all kinds and at all levels.

METROLOGY

The science of measurement is called metrology and it is fast becoming a recognized field in itself. Special branches of metrology include engineering metrology, physical metrology, chemical metrology, and biometrology. Those learned in and practitioners of metrology may be called metrologists and even by the name of their specialization. Thus, it is becoming common to hear of physical metrologists. Most analytical chemists prefer to be so called but they also may be called chemical metrologists. The distinguishing feature of all metrologists is their pursuit of excellence in measurement as a profession.

Metrologists do research to advance the science of measurement in various ways. They develop measurement systems, evaluate their performance, and validate their

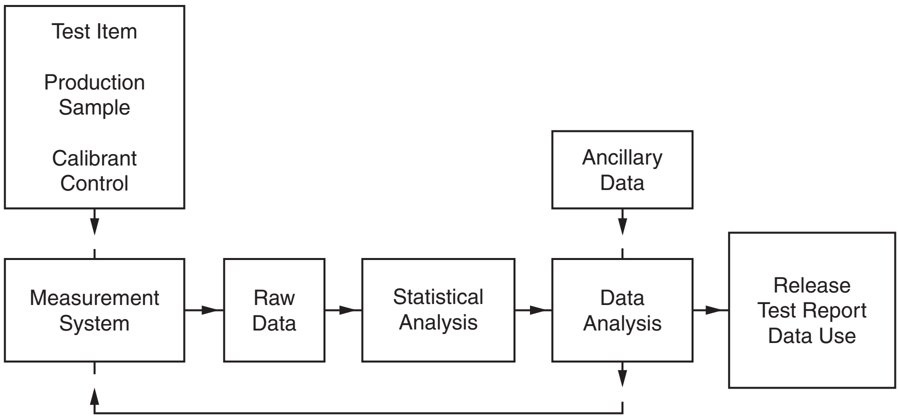


Figure 1.1. Role of statistics in metrology.

applicability to various special situations. Metrologists develop measurement plans that are cost effective, including ways to evaluate and assess data quality.

Statistics play a major role in all aspects of metrology since metrologists must contend with and understand variability.

The role of statistics is especially important in practical measurement situations as indicated in Figure 1.1. The figure indicates the central place of statistical analysis in data analysis which is or should be a requirement for release of data in every laboratory. When the right kinds of control data are obtained, its statistical analysis can be used to monitor the performance of the measurement system as indicated by the feedback loop in the figure. Statistical techniques provide the basis for design of measurement programs including the number of samples, the calibration procedures and the frequency of their application, and the frequency of control sample measurement. All of this is discussed in books on quality assurance such as that of the present author [1].

COMPUTER ASSISTED STATISTICAL ANALYSES

It should be clear from the above discussion that an understanding and working facility with statistical techniques is virtually a necessity for the modern metrologist. Modern computers can lessen the labor of utilizing statistics but a sound understanding of principles is necessary for their rational application. When modern computers are available they should be used, by all means. Furthermore, when data are accumulated in a rapid manner, computer assisted data analysis may be the only feasible way to achieve real-time evaluation of the performance of a measurement system and to analyze data outputs.

Part of the process involved in computer assisted data analysis is selecting a software package to be used. Many types of statistical software are available, with capabilities ranging from basic statistics to advanced macro programming features. The examples in the forthcoming chapters highlight MINITAB™ [3] statistical software for calculations. MINITAB has been selected for its ease of use and wide variety of analyses available, making it highly suitable for metrologists.

The principles discussed in the ensuing chapters and the computer techniques described should be helpful to both the casual as well as the constant user of statistical techniques.

EXERCISES

1-1. Discuss the hierarchy: Data $\longrightarrow$ information $\longrightarrow$ knowledge.

1-2. Compare "hard" and "soft" data.

1-3. What are the similarities and differences of natural and experimental data?

1-4. Discuss discrete, continuous, and enumerative data, giving examples.

1-5. Why is an understanding of variability essential to the scientist, the data user, and the general public?

1-6. Discuss the function of peer review in the production of reliable data and in its evaluation.

REFERENCES

[1] Taylor, J.K. *Quality Assurance of Chemical Measurements,* (Chelsea, MI: Lewis Publishers, 1987).

[2] Stanley, T.W., and S.S. Verner. "The U.S. Environmental Protection Agency's Quality Assurance Program," in *Quality Assurance of Environmental Measurements,* ASTM STP 967, J.K. Taylor and T.W. Stanley, Eds., (Philadelphia: ASTM, 1985), p. 12.

[3] *Meet MINITAB,* Release 14 for Windows (Minitab Inc. 2003).