DATA ANALYSIS FOR SCIENTISTS AND ENGINEERS

STUART L. MEYER

DATA ANALYSIS FOR SCIENTISTS AND ENGINEERS

STUART L. MEYER
Northwestern University

John Wiley & Sons, Inc., New York · London · Sydney · Toronto

"Now this is the peculiarity of scientific method, that when once it has become a habit of mind, that mind converts all facts whatsoever into science. The field of science is unlimited; its material is endless, every group of natural phenomena, every phase of social life, every stage of past or present development is material for science. *The unity of all science consists alone in its method, not in its material.* The man who classifies facts of any kind whatever, who sees their mutual relation and describes their consequences, is applying the scientific method and is a man of science. The facts may belong to the past history of mankind, to the social statistics of our great cities, to the atmosphere of the most distant stars, to the digestive organs of a worm, or to the life of a scarcely visible bacillus. It is not the facts themselves which form science, but the methods by which they are dealt with."

Karl Pearson, *The Grammar of Science*

Library of Congress Cataloging in Publication Data

Meyer, Stuart L 1937–
Data analysis for scientists and engineers.

Bibliography: p.
1. Mathematical statistics. 2. Probabilities.
I. Title.
QA276.M437 519.2 74–8873
ISBN 0-471-59995-6

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

PREFACE

This book evolved from a personal need: the need to have in one place, with a consistent style and notation, practically all that an experimental scientist *needs* to know to deal with data and *wants* to know to satisfy his or her intellectual curiosity.

For many years there has really been no discipline in which to teach data analysis *for experimental scientists.* The usual medium in science departments for imparting knowledge in this area has been the teaching laboratory. I believe that an appreciation of the means of extracting and evaluating information from experimental data is one of the basic goals of any laboratory course. Moreover, this is a skill and appreciation that should be developed as early as possible in a student's career, regardless of whether he or she intends to be an experimental or theoretical scientist. Therefore, although this book can be used in departments that have formal courses on the subject, I have assumed that the student's first acquaintance with data analysis will come through the science laboratory and that mature knowledge will come largely from self-study.

The early parts of the book are written to accompany a typical first-course science laboratory, and versions of Parts I and II have been used for several years at Northwestern University as a supplement to the beginning physics laboratory (and as a guide for the graduate teaching assistants). In addition, a version of Parts III and IV has been used in an integrated program of mathematics, physics, and chemistry for engineering students. Despite this history, the book is intended primarily for self-study and reference by the student, and much of the material goes beyond the usual content of a first course to provide the core of what is needed by a professional scientific investigator. Therefore, I have kept the discussion as complete and lucid as possible with many worked-out examples and cases that are of direct interest to the science student.

The wide availability of digital computer capability has made traditional, formerly cumbersome analysis methods more convenient and accurate and also has permitted the use of more sophisticated techniques that are not practical with slide rules, pencils, and paper alone. Some of these techniques are introduced here because they will become utilized more and more in the near future. I have not discussed programming or specific computer programs, since this material is readily available elsewhere. Also, the use of individual programs is often determined by the availability of library routines in a particular computer center, and it was more important to present the *methods* underlying applicable library programs instead of the programs themselves. Finally, the advances in technology have made available to *individuals* computing power that required a large facility only a decade ago. All of the numerical discussions and worked-out examples included here may be *followed* by the reader without *any* computational assistance and may be *duplicated* with a pocket calculator of the "electronic slide rule" type.

The level of this book requires only the rudiments of calculus as a preliminary. Although the early discussions are on a level that is understandable to the beginning student, the subject is developed so that it is fairly complete and self-contained, and it ends at a level of sophistication that is suitable for the professional scientist. This will enable the beginning student to "grow into" the book. The volume also may be used as a general reference and guide as the reader becomes more sophisticated in the design and analysis of experiments. Many discussions, formulas, tables, and graphs are included for convenience, completeness, and review. I hope that the book will find a useful place on the permanent bookshelf of the working scientist.

I wrote this work from the viewpoint of the *practitioner* of data analysis, that is, the scientist or engineer *using* the techniques, rather than from the viewpoint of the mathematician who develops new ones. Nevertheless, it is important for the practitioner to *understand* the bases of the techniques that he uses. For this reason, a fair amount of time is spent in developing the *ideas* of data analysis instead of merely presenting a series of recipes. Few problems encountered in the real world fit any one recipe exactly. It is important, therefore, to understand data-analysis concepts and methods thoroughly, and to have an appreciation for their spheres of usefulness.

The presentation is designed to be self-contained, eschewing the elegant obscurity of sophisticated methods of proof for the tedious clarity of more straightforward ones. All available sources have been utilized, and my major efforts are directed toward making the concepts palatable at the level of the reader and making them form a comprehensible, coherent, and useful whole. If I have erred, it has been on the side of too many explanation instead of too few, more steps in the mathematics rather than not enough steps; although this might please the experts less, it will help the students more. There is a famous explanation of why a classical work of J. Willard Gibbs was not widely known: "It is a little book which is little read because it is a little hard." I have tried to make this book easy to understand, even at the expense of greater than minimum length. However, there is sufficient depth in this single volume to reward careful and continued study.

Even the simplest problem often has subtleties that require more nuances of thought than the casual reader may consider at first glance. Some of the ramifications are not needed early in the book, and we sometimes return to the same problem again when it is appropriate to discuss other aspects of it. In addition, considerable material has been added because of my personal (possibly idiosyncratic) tastes. This material is usually set in special type or otherwise identified as not being absolutely essential to the main line of discussion.

I am pleased to acknowledge the support of the U.S. Army Research Office—Durham under Grant DA-ARO-D-31-124-71-G175 whose assistance facilitated the completion of the manuscript. I thank Publishers—Hall Syndicate and Cartoon Features for arranging permission to use cartoons and to Trendline Inc. for permission to use a graph. My sincere appreciation goes to Professors A. J. F. Siegert, Meyer Dwass, Robert Eisberg, and Hugh Young for reading parts of the manuscript and for helpful suggestions and to Mr. King L. Leung for able assistance with the calculation of the tables and graphs.

Stuart L. Meyer

Evanston, Illinois

CONTENTS

PART I

Introduction to Scientific Measurement

PART II

Introduction to Graphical Techniques and Curve Fitting

PART III
Probability

PART IV
Some Probability Distributions and Applications

PART V
Statistical Inference

APPENDICES

PART I

Introduction to Scientific Measurement

1

THE MEANING OF MEASUREMENT

The distinction is sometimes made between the "exact" sciences and the other sciences. In the first category are usually put the physical sciences such as physics and chemistry; in the second category are, broadly speaking, biology, psychology, etc. This terminology is unfortunate since it seems to imply that physical science makes statements that are true whereas the others do not. In fact, all scientists deal with the truth. The distinction is that the exact sciences can more easily measure or at least assign a value to the amount of truth in a scientific statement. It is preferable to call the physical sciences "quantitative" disciplines, since one can discuss in numerical fashion the amount of truth in any given statement of physical fact.

The alert student will note at this point that we have not yet defined truth. We must ask what it means to say that a measurement is "true" and what we mean when we say that a given statement about the physical world is "true." We make measurements in the physical sciences, and we perform experiments to test hypotheses about the physical world. There is no clear distinction between experiment and measurements, and the former of necessity involve the latter; that is, an experiment inescapably requires the measurement of *something*.

A Definition of Measurement

For the purposes of this discussion we shall consider that a measurement is the quantitative determination of the value of some *fixed* physical constant that is characteristic of a physical object or system or of a parameter that is needed for the description of a *reproducible* physical situation. Note carefully the adjectives *fixed* and *reproducible*. Both carry the necessary idea that the measurement can be repeated either tangibly or conceptually. The important point is that an *independent* determination of the value of the measured quantity can be made.

SIGNIFICANCE

Examples of quantities subject to measurement are legion: the velocity of light in vacuum; the mass of a proton; the length of an inch scale in centimeters. When the measurement is made, we assign a value to the result. If one looks in a handbook that is modern, one finds the following values: the velocity of light in vacuum is 2.997925×10^{10} cm/sec; the mass of the proton in MeV/c^2 (millions of electron volts divided by the square of the velocity of light) is 938.256 ± 0.005; and the length of an inch in centimeters is 2.54 *exactly*.

What do these numbers mean? It is intuitively easy to understand that one makes a measurement and gets a result. We have three results, however, and they appear to be in three different forms. The first number consists of seven digits with a decimal point as well as an exponent. By writing down seven digits, we are expressing a measure of the reliability of this number. We tacitly mean that the next to last digit is truly meaningful and that the last number is our best guess. We say that the number has seven significant figures. In this case the velocity of light is most likely to be halfway between the numbers 2.99792 and 2.99793. The use of the exponent is a standard way of separating the number of digits needed to specify the *magnitude* of a number from the number of digits needed to convey its *significance*.

The second number represents another way of conveying the significance of a numerical quantity. The mass of the proton is believed likely to be between the values 938.251 and 938.261 MeV. We shall discuss what "likely" means at a later time. At that point we shall be able to discuss whether two numbers that represent the same quantity are *consistent* with each other.

The last is the strangest one of all: 2.54 *exactly*. For one of the few times to be encountered, one's first impression is correct in this instance. This is a unique quantity known as a *defined constant*. This is the only kind of quantity that is exact. The metric and English systems of units arose independently, and for many years the conversion between them was a measured quantity like all the (undefined) quantities we shall encounter. If you find an old mathematics or physics book, you may even find the conversion constant written as 2.54001. (The U.S. Coast and Geodetic Survey still uses 2.54001.) Recently, however, it was decided to *fix* the inch as the distance that is equal to 2.54 of the units known as the centimeter. This was defined in terms of 1/100 of the distance between two ruled lines inscribed on a platinum-iridium bar kept under standard conditions in Sevres, France. Since 1960, the standard centimeter has been defined by international agreement in terms of the wavelength of the orange spectral line of the light emitted by the pure isotope krypton-86. The official centimeter is now defined to be 16,507.6373 wavelengths.

DIMENSIONS

The basic dimensions that we deal in are length (L), time (T), and mass (M). All else may be expressed in terms of these using various physical relations, since the two sides of any equation must have the same dimensions.

Thus, force F is defined in our basic set of dimensions by the relation between the acceleration produced on a mass and the force acting on it:

$$F = ma$$

We shall denote the equality of dimensions by $\stackrel{D}{=}$. Thus

$$F \stackrel{D}{=} MLT^{-2}$$

represents the dimensions of force.

The gravitational force equation tells us that the gravitational force between two masses M_1 and M_2 in isolation is proportional to the product of the masses and to the inverse of the square of the distance between them:

$$F = \frac{GM_1M_2}{r^2}$$

where G, the gravitational constant, has dimensions

$$G \overset{D}{=} M^{-2}L^{+2}F = M^{-2}L^{+2}MLT^{-2}$$
$$\overset{D}{=} L^{3}M^{-1}T^{-2}$$

To determine the dimensions of any quantity we need only recall the equations relating the quantity of interest to quantities whose dimensions we know and treat the dimensions as ordinary algebraic symbols.

The arguments (x) of various functions such as $\exp(x)$ and $\sin(x)$ must be dimensionless.

UNITS

It is desirable to write all equations so that they are independent of the system of units. Since our systems of units are arbitrary, it would be absurd to do otherwise. It should be obvious that the left side of any equation must have the same units as the right side. This provides a convenient check on any equations we write down (a necessary but not sufficient condition for the correctness of the equation).

Except for dimensionless constants, all quantities used must be regarded in some system of units, for example, centimeter-gram-second (CGS) system, meter-kilogram-second (MKS) system, foot-pound-second (English) system, etc. As with dimensions, the units of the left-hand side of an equation must be the same as the right-hand side. The units should be continually checked through any calculation since errors arising from discrepancies in units are most common! All units used must be rationalized to be consistent, since combinations of mixed units are myriad and hence inconvenient.

For example, an English moat (it must be English) is to be filled from a reservoir. How many acre-feet of water must be drawn from the reservoir to fill the moat to a depth of 2.00 fathoms if the moat is 0.500 furlongs in length and 900 barleycorns wide?

$$\text{Volume } V = 900 \text{ barleycorn-furlong-fathoms}$$

We may always multiply anything by unity. We recognize that

$$1 \text{ barleycorn} = \frac{1}{3} \text{ inch}$$
$$1 \text{ fathom} = 6 \text{ feet}$$
$$1 \text{ furlong} = \frac{1}{8} \text{ mile}$$
$$1 \text{ acre} = 43{,}560 \text{ square feet}$$

Therefore

$$V = 900 \text{ barleycorn} \frac{1}{3} \frac{\text{inch}}{\text{barleycorn}} \times \frac{1}{12} \frac{\text{foot}}{\text{inch}}$$
$$\times \text{furlong} \frac{1}{8} \frac{\text{mile}}{\text{furlong}} 5280 \frac{\text{feet}}{\text{mile}}$$
$$\times \text{fathoms} \cdot 6 \frac{\text{feet}}{\text{fathom}}$$
$$= (900)(1/3)(1/12)(1/8)(5280)(6) \text{ foot}^3$$
$$= 99{,}000 \text{ ft}^3 = 99{,}000 \text{ ft}^2 \frac{1 \text{ acre}}{43560 \text{ ft}^2}$$
$$= 2.27\,27(\dot{2}7) \text{ acre-ft}$$

The advantages of the metric system should be obvious!

Units and standards of both the metric and the U.S. Customary Systems are discussed in detail in Appendix I.

Dimensional Analysis

Given a physical situation describable by physical variables $x_1, x_2, \ldots$, we can sometimes deduce from dimensional analysis certain limitations on the form of any possible relationship among the variables. Dimensional analysis is not capable of completely determining the unknown functional relationship, but it can delimit the possibilities and, in some simple cases, it can give the complete relationship to within a constant of proportionality. Appendix II contains a discussion of this subject.