MODERN BUSINESS STATISTICS

JOHN E. FREUND
FRANK J. WILLIAMS

Modern Business Statistics

JOHN E. FREUND

Professor of Mathematics
Arizona State College, Tempe

FRANK J. WILLIAMS

Associate Professor of Statistics
University of Tennessee

Englewood Cliffs, N.J.

PRENTICE-HALL, INC.

1958

LIBRARY OF CONGRESS
CATALOG CARD NO.: 58-8938

PRINTED IN THE UNITED STATES OF AMERICA

58956

Preface

To keep pace with the ever-increasing demand of business for college-trained personnel there has been, over the past few years, a continuous and general modernizing of approaches to the various areas of Business Administration. Old courses have been revised; new courses and entirely new departments have been added. As new problems have arisen in various fields of business and new methods of attacking them have appeared, the work of the colleges has been brought forward accordingly. The necessary adjustments have been relatively slow in the field of statistics, and until a few years ago many statistics courses were in almost every respect the same as those taught twenty or thirty years earlier. However, in the past few years there has been evidence—the appearance and acceptance of recently produced "modern" textbooks—that the college version of business statistics is about to catch up with the "statistics of business."

We would not like to be misunderstood: we do not suggest that the traditional, the so-called descriptive, methods no longer have a place in modern business statistics. One glance at the table of contents of this book should dispel any doubts as to our attitude in this respect. Descriptive statistics is being used in today's business, it will probably always be used, and we would be careless, indeed, if we were to underestimate its importance. However, our idea, and we are sure that of many of our colleagues, is that there has been a poor balance between the more traditional topics of purely descriptive statistics and the newer topics of statistical inference. It is well known today that many problems never before thought amenable to quantitative analysis (for example, predicting responses to advertising campaigns, locating oil wells, or finding customers) can be attacked by the methods of inductive statistics, and it is hard to see how the logical foundation underlying the statistical solutions to such problems can be ignored in an up-to-date treatment of business statistics.

It is the opinion of the authors that the concepts of modern statistics definitely are not beyond the understanding of any reasonably diligent college student. Indeed, this book has been keyed to the lowest possible level of mathematical rigor at which it is felt that modern statistics can effectively be taught. The mathematical training assumed of the reader does not go beyond a course in college algebra or equivalent preparation.

The order and emphasis of the material covered in this book follows the current trend in the teaching of business statistics. Although the first five chapters deal primarily with descriptive methods, some of the basic ideas of statistical inference are introduced at a very early stage. Having learned to distinguish between samples and populations and being aware of the fact that they require different methods of analysis, the reader is prepared, for example, for the subsequent material on the standard deviation. The early distinction between σ and s in Chapter 4 was decided upon by the authors after a careful survey of professional statisticians, among them many teachers of statistics.

Since few instructors teach introductory statistics in exactly the same fashion, the authors would like to point out that there is room for considerable flexibility in the use of this text. For instance, those who may want to devote more time to graphical presentation can take up Sections 17.9 and 17.10, dealing with the charting of time series, and Appendix I, dealing with pictorial presentation, immediately after Section 2.3. Those who may want to devote more time to inductive methods can take up Chapter 11, an optional and somewhat more advanced chapter on tests of hypotheses, and Appendix II, a brief treatment of Quality Control, immediately after Chapter 10. Furthermore, the index number chapters or the time series chapters can be taken up at any time before the material on statistical inference, although the time series chapters should profitably be preceded by the chapter on linear regression.

Most of the examples used in the exercises and illustrations are based on actual experiments, surveys, and other kinds of studies, although in some instances the numbers were simplified and the sample sizes reduced, to avoid calculations which would serve only to distract the reader. The authors are greatly indebted to their many friends and colleagues who so generously provided problem material. The authors would also like to express their appreciation to their colleagues and students, whose helpful suggestions contributed greatly to the final version of the manuscript.

The authors are greatly indebted to the editorial staff of Prentice-Hall, Inc., for their courteous co-operation in the production of the book; to Miss Marianne Byrd, Mrs. Alice Kegan, Miss Gladys Brackney, and Miss Betty Prater for typing parts of various drafts of the manuscript; and above all to their wives for their cheerful encouragement in spite of the demands made on their husbands' time by the writing of this book.

Finally, the authors would like to express their appreciation and indebtedness to Professor R. A. Fisher and Messrs. Oliver and Boyd, Ltd., Edinburg, for permission to reprint parts of Tables III and IV

from their book *Statistical Methods for Research Workers;* to Professor E. S. Pearson and the Biometrika trustees for permission to reproduce the material in Tables IV and V; to Henry Holt and Company for permission to reprint Table I; to Professor E. S. Pearson and the Cambridge University Press for permission to reproduce the page of random numbers (page 171); to the American Society for Testing Materials for permission to reproduce the material in Table VII; and to The Macmillan Company for permission to reprint Table IX.

John E. Freund
Frank J. Williams

Contents

Chapter 1. INTRODUCTION 1

1.1 Introduction, 1
1.2 Descriptive and Inductive Statistics, 3
1.3 Sources of Data, 6
1.4 The Direct Collection of Data, 7
1.5 Published Data, 8
1.6 Mathematical Prerequisites, 12
1.7 Subscripts and Summations, 13

Chapter 2. FREQUENCY DISTRIBUTIONS 17

2.1 Introduction, 17
2.2 The Construction of Numerical Distributions, 19
2.3 Graphical Presentations, 32
2.4 The Construction of Categorical Distributions, 38
2.5 Tabular Presentations, 40

Chapter 3. MEASURES OF LOCATION 43

3.1 Introduction, 43
3.2 The Arithmetic Mean, Ungrouped Data, 46
3.3 The Arithmetic Mean, Grouped Data, 51
3.4 The Median, Ungrouped Data, 57
3.5 The Median, Grouped Data, 60
3.6 The Mode, 63
3.7 The Geometric and Harmonic Means, 66
3.8 The Weighted Mean, 69
3.9 Quartiles, Deciles, and Percentiles, 72
3.10 Further Comparisons, 75

Chapter 4. MEASURES OF VARIATION 77

4.1 Introduction, 77
4.2 The Range, 80
4.3 The Average Deviation, 81
4.4 The Standard Deviation, Ungrouped Data, 84
4.5 The Standard Deviation, Grouped Data, 91

Chapter 4. MEASURES OF VARIATION (Continued)

4.6 Further Measures of Variation, 95
4.7 Measures of Relative Variation, 95
4.8 Further Remarks About the Standard Deviation, 97

Chapter 5. FURTHER DESCRIPTIONS 100

5.1 Introduction, 100
5.2 Measures of Symmetry and Skewness, 104
5.3 Measures of Peakedness, 107

Chapter 6. PROBABILITY 110

6.1 The Meaning of Probability, 110
6.2 Some Rules of Probability, 113
6.3 Probabilities in Games of Chance, 120
6.4 Mathematical Expectation, 123

Chapter 7. THEORETICAL DISTRIBUTIONS 127

7.1 Introduction, 127
7.2 The Binomial Distribution, 131
7.3 The Mean and Standard Deviation of the Binomial Distribution, 138
7.4 Continuous Distributions, 142
7.5 The Normal Curve, 145
7.6 Some Applications, 152
7.7 The Binomial Distribution and the Normal Curve, 156
7.8 Fitting a Normal Curve to Observed Data, 160
7.9 Further Theoretical Distributions, 165

Chapter 8. SAMPLING DISTRIBUTIONS 169

8.1 Random Sampling, 169
8.2 Sampling Distributions, 172

Chapter 9. PROBLEMS OF ESTIMATION 186

9.1 Introduction, 186
9.2 The Estimation of Means (Large Samples), 188
9.3 The Estimation of Means (Small Samples), 195
9.4 The Estimation of Proportions, 199
9.5 Standard Errors and Probable Errors, 206
9.6 Sampling from Small Populations, 208

Chapter 10. TESTS OF HYPOTHESES 212

10.1 Introduction, 212
10.2 Type I and Type II Errors, 217
10.3 Null Hypotheses and Significance Tests, 221
10.4 Tests Concerning Proportions, 225
10.5 Differences Between Proportions, 228
10.6 Tests Concerning Means, 233
10.7 Differences Between Means (Large Samples), 237
10.8 Differences Between Means (Small Samples), 240

Chapter 11. FURTHER TESTS OF HYPOTHESES 245

11.1 Tests Concerning k Proportions, 245
11.2 The Analysis of an r by k Table, 253
11.3 Tests of "Goodness of Fit," 257
11.4 Tests Concerning k Means, 261
11.5 Analysis of Variance, 268

Chapter 12. PROBLEMS OF SAMPLING 270

12.1 Introduction, 270
12.2 A Test of Randomness, 272
12.3 Runs Above and Below the Median, 277
12.4 Sample Designs, 279
12.5 Double, Multiple, and Sequential Sampling, 283

Chapter 13. LINEAR REGRESSION 286

13.1 Introduction, 286
13.2 The Method of Least Squares, 288
13.3 Linear Regression, 296
13.4 Limits of Prediction, 299
13.5 Multiple Linear Regression, 304

Chapter 14. CORRELATION 307

14.1 The Coefficient of Correlation, 307
14.2 The Interpretation of r, 314
14.3 A Significance Test for r, 317
14.4 The Calculation of r from Grouped Data, 320
14.5 Rank Correlation, 327
14.6 Multiple and Partial Correlation, 330
14.7 The Correlation of Qualitative Data, 333

Chapter 15. INDEX NUMBERS: BASIC CONCEPTS 338

15.1 Introduction, 338
15.2 Purpose of the Index, 340

Chapter 15. INDEX NUMBERS: BASIC CONCEPTS (Continued)

15.3 Availability and Comparability of Data, 341
15.4 Selection of the Items, 344
15.5 Choice of the Base Period, 345
15.6 Choice of the Weights, 347
15.7 Methods of Construction, 347

Chapter 16. INDEX NUMBERS: THEORY AND APPLICATION 350

16.1 Introduction, 350
16.2 Unweighted Index Numbers, 350
16.3 Weighted Index Numbers, 356
16.4 Weighted Averages of Price Relatives, 362
16.5 Chain Index Numbers, 365
16.6 Shifting the Base, 372
16.7 The Use of Index Numbers in Deflating, 373
16.8 Mathematical Properties of Index Numbers, 376
16.9 Current Problems, 378

Chapter 17. TIMES SERIES ANALYSIS: BASIC CONCEPTS 381

17.1 Introduction, 381
17.2 The Behavior of Time Series, 384
17.3 Secular Trend, 384
17.4 Seasonal Variation, 387
17.5 Cyclical Variation, 388
17.6 Irregular Variation, 389
17.7 A Word About the Classical Approach, 390
17.8 The Preliminary Adjustment of Time Series, 391
17.9 Graphical Presentations of Time Series, 393
17.10 Logarithmic Line Charts, 395

Chapter 18. TIME SERIES ANALYSIS: SECULAR TREND 400

18.1 Introduction, 400
18.2 Linear Trends: Semi-Averages, 402
18.3 Linear Trends: Least Squares, 404
18.4 Modified Trend Equations, 408
18.5 Parabolic Trends, 412
18.6 Exponential Trends, 416
18.7 Other Trend Curves, 419
18.8 The Smoothing of Time Series, 423
18.9 Forecasting, 426

Chapter 19. TIME SERIES ANALYSIS: SEASONAL AND CYCLICAL VARIATION 429

19.1 Seasonal Variation, 429
19.2 Some Preliminary Considerations, 431
19.3 The Method of Simple Averages, 432
19.4 The Ratio-to-Trend Method, 436
19.5 The Ratio-to-Moving Average Method, 440
19.6 Deseasonalized Data, 447
19.7 The Use of Seasonal Indexes in Forecasting, 449
19.8 Cyclical Variation, 453

Appendix I. PICTORIAL PRESENTATIONS 461

I.1 Introduction, 461
I.2 Bar Charts, 462
I.3 Pie Charts and Component Bar Charts, 466
I.4 Statistical Maps, 471

Appendix II. QUALITY CONTROL 474

II.1 Introduction, 474
II.2 The Control Chart, 476
II.3 $\bar{X}$ and R Charts, 478
II.4 $\bar{X}$ and σ Charts, 482
II.5 Control Charts for Attributes, 484
II.6 Acceptance Sampling, 489

Appendix III. CALCULATIONS WITH ROUNDED NUMBERS 492

III.1 Rounded Numbers, 492
III.2 Calculations with Rounded Numbers, 493

Appendix IV. THE USE OF LOGARITHM AND SQUARE ROOT TABLES 495

IV.1 The Use of Logarithm Tables, 495
IV.2 The Use of Square Root Tables, 497

STATISTICAL TABLES 499

ANSWERS TO ODD EXERCISES 521

INDEX 529

CHAPTER 1

Introduction

1.1 Introduction

In the last few decades, the growth of statistical ideas and statistical methods has made itself felt in almost every phase of human activity. In business, it has brought about drastic changes in production, in the efficient use of materials, in marketing, in various phases of business research, and in management. Here, statistical data and statistical techniques have become a vital factor in the decisions, analyses, and forecasts of the modern businessman.

There can be no doubt that it is impossible to understand the meaning and implications of most work done in business and economic research without having at least a speaking acquaintance with the subject of statistics. Clearly, numerical data derived from surveys, experiments, and other sources form the raw material on which analyses and forecasts are based, so it is essential to know how to squeeze usable information from such data. This, in fact, is the major objective of statistics.

Numerous textbooks have been written on business statistics, psychological statistics, educational statistics, agricultural statistics, medical statistics, and other specific areas of application. It is true, of course, that these diversified fields demand somewhat different and specialized techniques in particular problems; yet the fundamental principles that underlie all the various methods are identical regardless of the field of application. This will become evident to the reader once he realizes that *statistical methods in general are nothing but a refinement of everyday thinking.*

The approach we shall use in this elementary treatment of statistics is keynoted by the above statement; it is our goal to introduce the beginning student in business and economics to the ideas and the concepts which are fundamental to the understanding of modern statistics. Although the examples and exercises used in this text will deal primarily with subjects closely related to problems of business and economics, we shall repeatedly remind the reader that the various ideas and techniques are applicable also to other social sciences as well as the natural sciences. It is hoped that the approach used in this text will not only provide the reader with a sound understanding of the scientific principles used in business or economic research and in subsequent planning and operations, but that it will also enable him to gain a better understanding of the scope and limitations of empirical (scientific) knowledge in general.

As we have said before, the study of statistics may be directed toward applications in particular fields of inquiry. Furthermore, statistics may also be presented in varying degrees of mathematical refinement and in almost any balance between theory and application. Because it is, in our opinion, much more important to understand the *meaning and implications* of a few basic concepts than it is to memorize a large assortment of impressive sounding formulas, we shall have to sacrifice some of the mathematical detail that is sometimes covered in an elementary course in statistics. This is unfortunate in some respects, but it will avoid our getting lost in an excessive amount of detail which might easily obscure the most important issues. By stressing ideas over computational skills we hope to avoid the dangerous effect that often results from the indiscriminate application of statistical methods without a thorough understanding of the fundamental logical concepts that are involved.

Traditionally, business statistics has been looked upon as applying mainly to "front office" operations; hence, until recently, most elementary textbooks stressed the collection and sources of data and the presentation of numerical data in tables, charts, pictures, and maps. Although the importance of statistics in this phase of business cannot be denied, there has been a pronounced change in attitude toward the subject: business statistics is now viewed as providing quantitative bases for arriving at well-informed decisions with respect to *all* matters connected with the operation of a business. *In short, it is now recognized that, in its broadest sense, business statistics includes also the methods and inferences needed to provide an adequate flow of quality raw materials, to select and evaluate the performance of both machines and personnel, to design products and maintain their quality, and to evaluate new methods of production, advertising, and selling the goods and services of industry, both in the immediate present and in the near and distant future.* In view of all this, our goal is to introduce the reader to some of the methods of statistics, the traditional as well as the new, which are playing an ever increasing role in, as we said, *all matters connected with the operation of a business.*

1.2 Descriptive and Inductive Statistics

Everything dealing even remotely with the collection, analysis, interpretation, and presentation of numerical data may be classified as statistics. It includes even such diversified tasks as the computation of a ballplayer's batting average by a *team statistician;* the collection and presentation of data on births, marriages, and deaths as *vital statistics;* and the study of laws governing the behavior of atomic particles by a specialist in *statistical mechanics* or *quantum statistics.*

The word *statistics* itself can be given a variety of interpretations. For example, it is used in the plural to denote simply a collection of numerical data. Such statistics may be found in the *Economic Almanac,* the financial pages of newspapers, U.S. Census Reports, the *Statistical Abstract of the United States,* the records of county clerks, or wherever numerical data are collected and recorded. The second meaning, also in the plural, is that of the totality of methods that are employed in the collection and analysis of numerical data. In this sense, statistics is a branch of applied mathematics, and it is this field of mathematics which we shall study in this book. In order to complete this linguistic study of the word *statistics,* we might also mention that the term *statistic,* in the singular, is used to denote a particular quantity such as an average, an index number, or a coefficient of correlation which one calculates on the basis of a given set of data.

The problems that we shall discuss in this book are essentially of two types, belonging either to the field of *descriptive statistics* or that of *inductive statistics.* Although the term *descriptive statistics* is often used to denote merely the tabular or graphical presentation of data, we shall use it in a much wider sense. By descriptive statistics we shall understand any treatment of numerical data which does not involve generalizations. In contrast, we shall speak of *inductive statistics* the very moment that we make generalizations, predictions, or estimations.

To clarify this distinction with an example, let us consider a consumers' rating service which wants to compare 60 watt electric light bulbs produced by two different manufacturers. Taking *five* light bulbs of Brand A, they find that these bulbs burn out after 985, 863, 1024, 972, and 746 hours of continuous use. Repeating this test with *five* light bulbs of Brand B, they find that these bulbs burn out after 892, 1071, 993, 929, and 785 hours of continuous use. On the basis of this information we can say that the five Brand A light bulbs had an *average* lifetime of

$$\frac{985 + 863 + 1024 + 972 + 746}{5} = 918 \text{ hours}$$

while those of Brand B had an *average* lifetime of

$$\frac{892 + 1071 + 993 + 929 + 785}{5} = 934 \text{ hours}$$

What we have done so far belongs to the domain of descriptive statistics. We followed simple arithmetical rules in calculating the two averages which are, indeed, descriptive of the two sets of figures obtained in the test. However, if we concluded from this experiment that Brand B is superior to Brand A, that is, that the average lifetime of the light bulbs of Brand B is *in general* greater than the average lifetime of those of Brand A, our reasoning would have to go far beyond the information with which we were supplied and we would find ourselves in the domain of inductive statistics.

So long as we merely calculated the two averages, we did not add anything to the information with which we were supplied; we merely rearranged it in a different and possibly more useful form; this is characteristic of descriptive statistics. As soon as we generalized, we said more than we were given in the original data and this, in turn, is characteristic of inductive statistics.

In the given example it does not follow by any means that *in general* Brand B is superior to Brand A. As can be seen from the experiment, there are considerable differences even among the light bulbs manufactured by the same company. Hence, it is quite conceivable that Brand B showed up better in the test simply because we were "lucky" in picking Brand B light bulbs that happened to be of particularly good quality. It is equally conceivable that if we repeated the experiment, the result might easily be reversed if *by chance* we picked for our sample Brand B light bulbs of slightly inferior quality. Before reaching any conclusions we shall, therefore, have to investigate whether the difference of $934 - 918 = 16$ hours in the two averages might not be due to the variability (lack of uniformity) of the two products, together with chance factors involved in the selection of the particular samples used in the test.

If we decided on the basis of the above experiment that Brand B is in general superior to Brand A, we would be making a generalization that may or may not be correct and we would, consequently, be taking a risk. The careful evaluation, analysis, and control of the chances that must be taken when we make such a generalization is one of the main tasks of inductive statistics.

There are many questions that immediately come to one's mind when thinking about the above example. For instance, if the difference between the averages of the two samples had been *very large* we might have been inclined to say that one brand is superior to the other without making a detailed statistical analysis. Similarly, if the difference had been *very small* we might have been inclined to say that the products are of about equal quality. *But what if the difference were neither "very small" nor "very large"—where do we draw the line?* Also, how do we know that five light bulbs of each kind will provide us with enough information to reach any conclusion whatsoever? Would it not have been better, perhaps, to test 10 light bulbs of each kind, or possibly 20 or more? These are important questions, and they are the kind of questions we shall be able to answer with the methods of inductive statistics.

There is one other question that is extremely important whenever we wish to make a generalization like the one concerning the two brands of light bulbs, and that question is: *Was the experiment (or survey) conducted in such a way that a generalization is at all possible?* For instance, if the tests of the two kinds of light bulbs had been conducted at different voltages, the observed difference in the average lifetimes might easily have been caused by differences in voltage rather than by differences in quality. Although it would seem rather silly to conduct the experiment of our illustration by using different voltages, it is amazing how many surveys or experiments are conducted in such a way that no sensible statistical analysis and, of course, no generalizations are possible.

We have made this last point to emphasize that the statistical treatment of a problem does not merely consist of looking at a set of data, performing some calculations, and reaching a conclusion. The questions as to how the data were collected and how the whole experiment or survey was planned are of prime importance. Unless proper care is taken in the planning and design of an experiment (survey or other kind of investigation), we may not be able to reach any valid conclusion whatsoever. Generally speaking, poorly designed experiments cannot be salvaged by fancy mathematical or statistical techniques.

To summarize, we shall use the terms *descriptive statistics* and *inductive statistics* with reference to the kinds of problems we wish to solve—not with reference to a particular formula or statistic we may choose to employ. For example, we may calculate an average solely for the purpose of describing a set of data or we may use it to make generalizations or predictions.

1.3 Sources of Data

Because numerical data are the raw material of statistical investigations, one of the first steps in any statistical study must be the collection of suitable data. The only step that ordinarily precedes the collecting of data is the careful planning of the study and the precise formulation of its purpose, scope, and objectives.

Data needed for studies arising in the operation of a business organization will of necessity often have to come from the records of the organization itself. Data thus taken by a firm from its accounting records, payrolls, inventories, sales vouchers, and the like, for its own use in statistical investigations are called *internal data.* Although internal data are usually of considerable importance, in many situations they must be supplemented (or replaced) by numerical information from sources outside the firm. Data obtained from such sources as local, state, and national governments, private reporting organizations, trade associations, and trade publications are correspondingly referred to as *external data.*

Having made this distinction, we should not fail to note that some studies require both kinds of data. For instance, if a steel company wants to prepare a chart showing its stockholders how its accident record compares with that of the entire industry, data on the company's own accident rate will come from its internal records while the figures for the entire industry must be sought externally—perhaps, from a trade publication.

It is theoretically within the power of a firm to set up and maintain its own records in such a way as to provide it with whatever data are thought to be useful or necessary. Although this may not always be done and there may sometimes be regret for not having kept proper records, in principle, the collection of internal data generally does not pose any serious problems. The collection of data from external sources is another matter, however, and it is often quite troublesome. Sometimes data available from two or more sources do not agree, and sometimes data considered to be relevant to the solution of a problem are not available from any source, at least within limits imposed by practical considerations of cost and time.

External data are often classified as *primary*, meaning that they are published by the same organization by which they were collected, and *secondary*, meaning that they are published by an organization other than the one by which they were collected. For example, the Bureau of Labor Statistics collects the data needed for its *Consumer Price Index*, performs the necessary calculations, and publishes the index in its *Monthly Labor Review.* This is thus the primary source