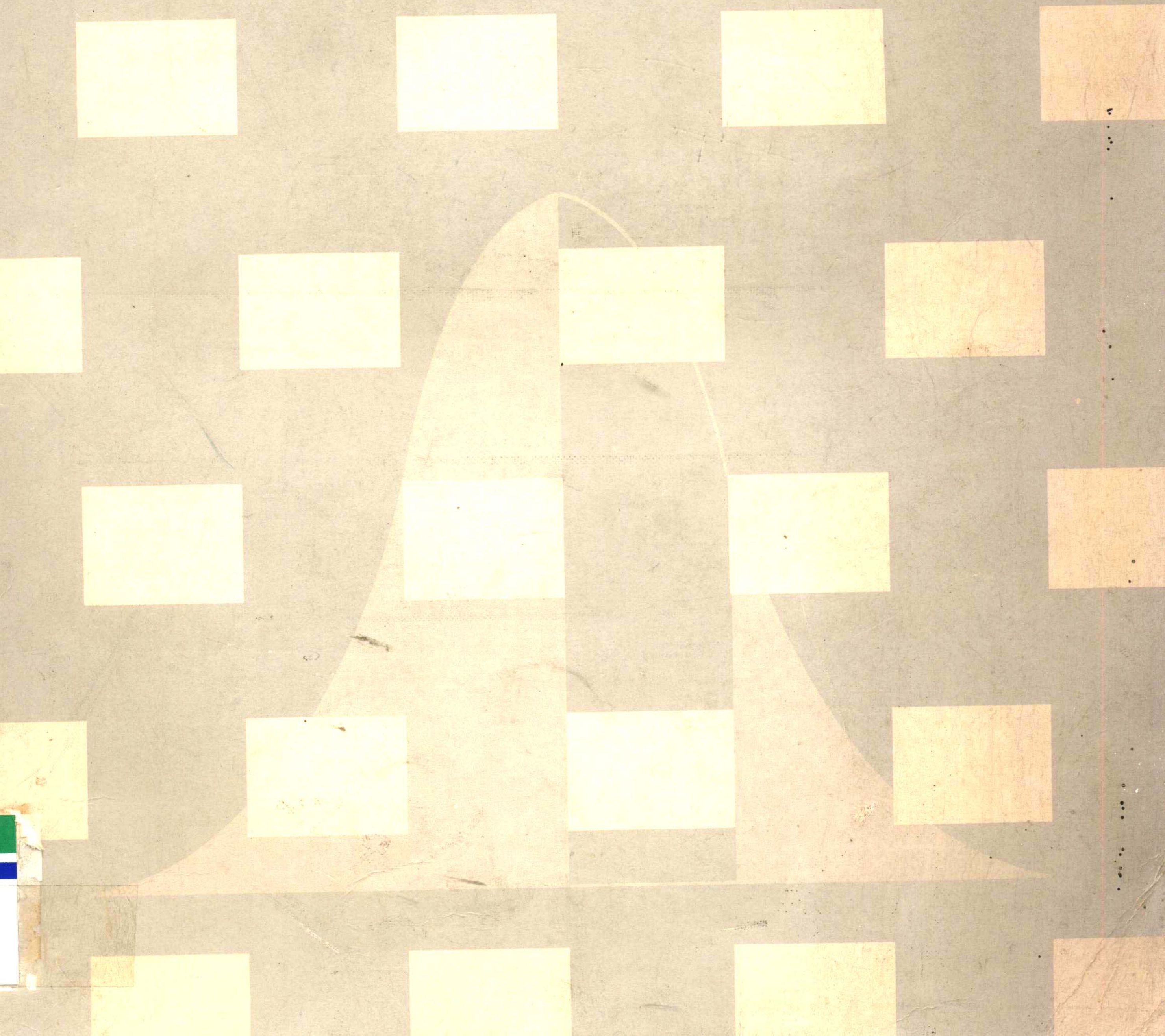
a computer-assisted approach to elementary statistics

examples and problems

BULGREN



A Computer-Assisted Approach to Elementary Statistics
Examples and Problems

William Bulgren

University of Kansas

Wadsworth Publishing Company, Inc.

Belmont, California

L. C. Cat. Card No.: 78-137147

Printed in the United States of America

1 2 3 4 5 6 7 8 9 10—75 74 73 72 71

A Computer-Assisted Approach to Elementary Statistics
Examples and Problems

Preface

This book is designed for classroom and laboratory use in the study of elementary statistics. It provides a logical, graded, tested series of computer-oriented exercises that are intended to encourage thinking in terms of the meaning of statistics, the statistical methods, and the results of statistical analysis.

A belief that the best learning situations involve the student personally and subjectively in the subject matter has guided the preparation of this book. Programmed texts, computer-assisted learning, and other learning media are now used to accomplish this involvement. Such techniques demand that students do more than passively observe a series of lectures on their subject matter. Introductory courses in statistics lend themselves particularly well to laboratory work. Surprisingly, however, those offered by most colleges today are organized on a lecture basis—to the dissatisfaction, I believe, of student and teacher alike. Students become dissatisfied with traditional approaches to elementary statistics, because the important concepts of statistics are not easily learned by hearing someone talk about them or by reading about them in a textbook. One learns these things best, it seems, by doing the work.

The instructor will find this book so arranged that it may be used with most elementary statistics textbooks in either a one-semester or a full-year course.

I appreciate the helpful suggestions and criticism offered by many students, especially Kenneth Norland of the University of Kansas; and I am indebted to my wife, Janis, for her help in preparing the manuscript and to Gayle Johnson for typing it. I am also indebted to Earl J. Schweppe, Chairman of the Computer Science Department, and to Paul Wolfe, Director of the Computation Center, University of Kansas. Lastly, Alexander Kugushev, Mathematical Sciences Editor of Wadsworth Publishing Company deserves the credit for getting me started on such an enterprise as this book in the first place. I have received from many and return this book to those who find an interest or need.

Contents

1

Introduction

The use of the computer in education has stimulated a great deal of interest and perhaps some unjustified anxiety concerning the computer's effect on education. This book outlines a set of exercises for use in a noncalculus elementary statistics course. For those who still have doubts about the ultimate use of computers in statistical education or who fear the effect of these techniques on the teaching of statistics, it is hoped that the following chapters will show the potential savings in time, increase in efficiency, and a new type of freedom which these new tools offer.

The book consists of eight chapters plus four appendices. Each chapter discusses important concepts and principles of statistics. Each also includes a discussion of the prerequisites for the chapter, a brief introduction with one or more examples for illustration, and a list of exercises to provide greater insight and understanding of the process being illustrated.

There are ample exercises for courses of one to four semester hours. Although the student probably will be unable to write more than two or three programs from each chapter in the allotted time, he will learn a great deal by reading and thinking about the Statement of the Problem and the Discussion and Questions of each exercise.

Four appendices are included to aid you, the student, in successfully completing this book. To avoid objections arising from the use of a particular programming language in the text, Appendix I and Appendix IV are included. Appendix I gives the essential parts of FORTRAN programming. Your competence will increase as you continue to practice this skill. Although much of FORTRAN is not explained in Appendix I, you will be able to resolve the remaining mysteries by reading any complete FORTRAN manual. The following is a list of some *typical* FORTRAN programming books.

1. Anderson, Decima M., *Computer Programming: FORTRAN IV.* Appleton-Century-Crofts, 1966.
2. Dimitry, Donald L., and Mott, Thomas H. Jr., *Introduction to FORTRAN IV Programming.* Holt, Rinehart and Winston, 1966.
3. McCracken, Daniel D., *A Guide to FORTRAN IV Programming.* John Wiley & Sons, Inc., 1965.
4. Organick, Elliot I., *FORTRAN IV Primer.* Addison-Wesley Publishing Co., Inc., 1966.

It is also suggested that manufacturers' manuals on FORTRAN be consulted, along with any supplementary material pertaining to the particular installation you plan to use, especially the program library.

Appendix IV includes programs written by students that give answers to selected examples and exercises in FORTRAN. In practice, you will find that once a program has been written, you will modify and expand it several times to account for various alternatives. Many of the exercises add to exercises that have already been completed earlier. This program growth exists from exercise to exercise within each chapter, and occasionally from chapter to chapter; hence, all programs should be saved until you have finished the book.

Appendix II is concerned with generating pseudorandom numbers. These numbers represent experimental data with which you will perform computer experiments as described in Chapters 2-8. The appendix summarizes a technique to generate pseudorandom numbers and presents a FORTRAN function as well as a table consisting of 1000 random digits.

Appendix III is concerned with generating normal random numbers. These numbers represent experimental data with which you will perform computer experiments as described in Chapters 2-8. The appendix summarizes a technique to generate normal random numbers and presents a FORTRAN function as well as a table consisting of 1000 normally distributed random numbers.

Since this book will be used as a supplement to an introductory textbook in elementary statistics, the following table illustrates which chapters of this book correspond to chapters of some typical elementary statistics textbooks. Since all elementary textbooks contain the basic elements included in this book, the table is only a guideline and by no means lists all the excellent textbooks available on the market.

Bulgren	[1] Alder & Roessler	[2] Freund	[3] Hoel	[4] Huntsberger	[5] Mendenhall
2	2, 4	2, 3, 4	2	2, 3	3
3	5	5	3	4	4
4	6, 7, 8	6, 7, 8	4, 5	5, 6	5, 6, 7
5	9	9, 10	6, 7	6, 7, 8, 9	8
6	10, 13, 16	9, 10	6, 7	7, 8, 9	9
7	12	14	9	11	10
8	17	12	12	12	13

For example, Chapter 2 of this book (Bulgren) would be utilized with Chapter 3 of Mendenhall [5], Chapters 2 and 4 of Alder and Roessler [1], and so forth.

Typical Elementary Textbooks

1. Alder, Henry L., and Roessler, Edward B., *Introduction to Probability and Statistics,* 4th ed. W. H. Freeman and Co., 1968.
2. Freund, John E., *Modern Elementary Statistics*, 3rd ed. Prentice-Hall, Inc., 1967.
3. Hoel, Paul G., *Elementary Statistics*, 2nd ed. John Wiley & Sons, Inc., 1966.
4. Huntsberger, David V., *Elements of Statistical Inference.* Allyn and Bacon, Inc., 1967.
5. Mendenhall, William, *Introduction to Probability and Statistics,* 3rd ed. Wadsworth Publishing Co., Inc., 1971.

2

Describing Distributions of Measurements

Prerequisites

The exercises of this chapter require a basic understanding of methods for describing distributions of measurements. Since the computer is also being utilized, you will be studying a technique which forces you to organize your thoughts concerning a particular problem and to carefully plan a method for solving a problem before writing the actual computer program for solution.

Introduction

You have already seen the two basic methods for describing sets of measurements: graphical and numerical. Examples of graphical methods are frequency histograms and relative frequency histograms. Numerical descriptive methods involve such things as measures of central tendency (for instance, the mean) and measures of variability (for instance, the variance).

Example 2.1. The customary way to organize data derived from observations is to display them as a frequency histogram, which shows the number of times the variable falls in different intervals or within certain limits. Table 2-1 illustrates a frequency table of pseudorandom numbers contained in the interval from zero (0)

Table 2-1 Frequency Table of Pseudorandom Numbers

i	Interval	Midpoint (X)	Frequency (f_i)	Relative Frequency (f_i/n)	Cumulative Relative Frequency
1	0.0-0.1	.05	5	0.10	0.10
2	0.1-0.2	.15	2	0.04	0.14
3	0.2-0.3	.25	6	0.12	0.26
4	0.3-0.4	.35	7	0.14	0.40
5	0.4-0.5	.45	2	0.04	0.44
6	0.5-0.6	.55	5	0.10	0.54
7	0.6-0.7	.65	4	0.08	0.62
8	0.7-0.8	.75	9	0.18	0.80
9	0.8-0.9	.85	6	0.12	0.92
10	0.9-1.0	.95	4	0.08	1.00
			50		

to one (1), which is denoted by U(0,1). By adding all the frequencies contained in Table 2-1, one finds the total number of observations to be 50. Now the end points of the classes (0.0, 0.1, 0.2, 0.3, and so forth, in Table 2-1) are called *class limits* or *class boundaries.* The midpoint between the upper and lower class boundaries is called the *class mark* or *class midpoint.* The difference between the largest and smallest class boundaries is called the range of the table. (In Table 2-1, the range is 1.0-0.0 = 1.0.) The number of observations (variates) falling within a given class interval is called the *class frequency,* or simply, *frequency.* (In Table 2-1, the frequency for class 1 is 5, or $f_1 = 5$.) The set of frequencies with their respective classes or class midpoints is called a *frequency distribution* (Table 2-1 is such a distribution.) Experience shows that in classifying data in a frequency table, a number between 10 and 20 classes is desirable. When the number of classes is less than 10, much accuracy is lost, and when the number of classes is greater than 20, the obtained summary is too extensive.

The number of observations (variates) falling within a given class interval divided by the total number of observations is called the *relative class frequency* or *relative frequency.* (In column 5 of Table 2-1, the relative frequency for class 1 is .10, or $f_1/50 = 5/50 = .10$.) The set of relative frequencies with their respective classes or class midpoints is called a *relative frequency distribution.* The *cumulative frequency (cumulative relative frequency)* for a given class is the sum of frequencies (relative frequencies) of all classes up to and including the given class, assuming that the class with the smallest class limits is written first. Table 2-1 illustrates only the cumulative relative frequency.

Histograms for the data of Table 2-1 are shown in Figures 2-1, 2-2, and 2-3. In Figure 2-1, called a *frequency histogram,* the frequency corresponding to a class is represented by the height of a rectangle whose base is the class interval. In Figure 2-2, called a *relative frequency histogram,* the relative frequency corresponding to a class is represented by the height of a rectangle whose base is the class interval. In Figure 2-3, called a *cumulative relative frequency histogram,* the

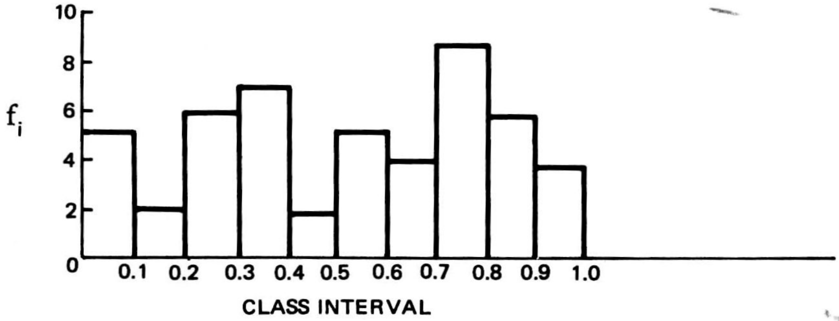


Figure 2-1 Frequency Histogram

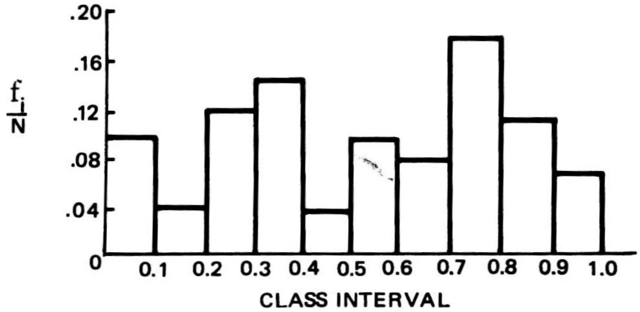


Figure 2-2 Relative Frequency Histogram

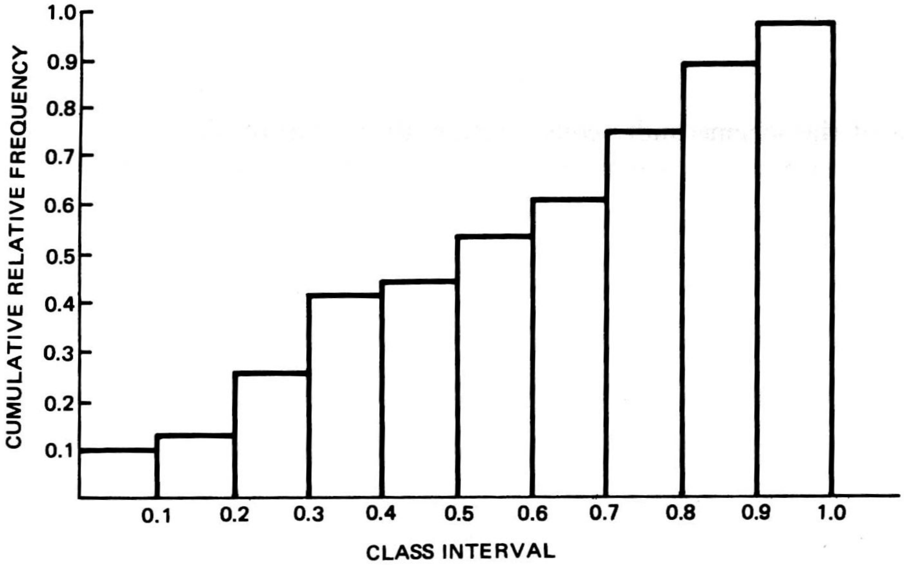


Figure 2-3 Cumulative Relative Frequency Histogram

cumulative relative frequency corresponding to a class is represented by the height of a rectangle whose base is the class interval.

Another way to represent a set of observations is by using a frequency polygon. A broken-line graph having class midpoints represented on the horizontal axis and frequencies on the vertical axis is called a *frequency polygon.* A frequency polygon for the data of Table 2-1 is shown in Figure 2-4.

If the number of class intervals increases, while at the same time the total number of observations is increased, the histogram may approach a curve, called a *frequency curve*. This curve is approximated by fitting a smooth curve to the points of a frequency polygon.

Next consider a scheme for representing any frequency distribution. Such a generalized scheme must be capable of representing a frequency distribution with any number of intervals and involving any number of values. The score value corresponding to the midpoint of a given interval will be represented by X, and the identification number for that class is affixed as a subscript, that is, X_i. It should be clear that

$$n = \sum_{i=1}^{c} f_i = f_1 + f_2 + \cdots + f_c ,$$

where c denotes the number of classes. In our example, $50 = 5 + 2 + \cdots + 4$. Note that

$$\sum_{i=1}^{c} \frac{f_i}{n} = 1.$$

In terms of the scheme under consideration, the sum of the f_1 scores in interval 1 is $f_1 X_1$, the sum of the f_2 scores in interval 2 is $f_2 X_2$, and so forth. Thus, the sum of

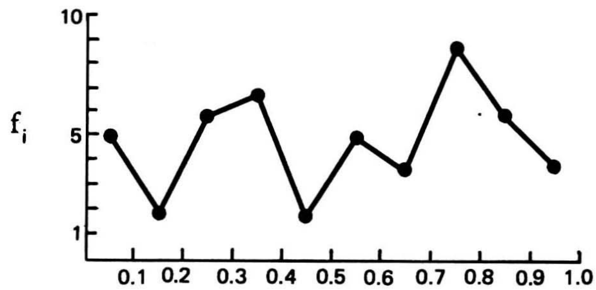


Figure 2-4 Frequency Polygon