ARNOLD NAIMAN
ROBERT ROSENFELD
GENE ZIRKEL

Understanding Statistics

THIRD EDITION

THIRD EDITION

UNDERSTANDING STATISTICS

ARNOLD NAIMAN
Late Professor
Nassau Community College

ROBERT ROSENFELD
Nassau Community College

GENE ZIRKEL
Nassau Community College

McGRAW-HILL BOOK COMPANY

New York St. Louis San Francisco Auckland
Bogotá Hamburg Johannesburg London Madrid
Mexico Montreal New Delhi Panama Paris
São Paulo Singapore Sydney Tokyo Toronto

UNDERSTANDING STATISTICS

2 3 4 5 6 7 8 9 0 DOCDOC 8 9 8 7 6 5 4 3

ISBN 0-07-045863-4

This book was set in Helvetica by Waldman Graphics, Inc.
The editors were John J. Corrigan,
Stephen Wagley, and Jo Satloff;
the designer was Nicholas Krenitsky;
the production supervisor was Charles Hess.
New drawings were done by J & R Services, Inc.
R. R. Donnelley & Sons Company was printer and binder.

Library of Congress Cataloging in Publication Data

Naiman, Arnold.
Understanding statistics.

Includes index.
1. Statistics. I. Rosenfeld, Robert.
II. Zirkel, Gene. III. Title.
QA276.12.N34 1983 519.5 82-17183
ISBN 0-07-045863-4

In memory of our colleague
DR. ARNOLD NAIMAN

Preface

We wrote the first edition of *Understanding Statistics* with our late colleague Dr. Arnold Naiman for students with little mathematical background, such as those we had all worked with at Nassau Community College for a number of years. We wanted a text that would be elementary enough to reach these students and still be mathematically sound and suitable for a one-semester college-level course. We are pleased that the text has "worked" for so many people. Over a period of 10 years, through two editions, many students and professors who have used the book have given us their comments on it. We have learned from them that its major strengths are its readability, its often humorous approach, and its problem sets. In putting together this third edition, we have tried to maintain these strengths while responding to users who have recommended changes.

The objective of this book is to show readers how statistics is used, not to train them to be statisticians. Students using it will gain an appreciation of the proper use of statistics and statistical terms that confront them in textbooks, newspapers, magazines, and on TV and radio. Our major emphasis is on understanding sampling and hypothesis testing.

In this edition we have included a class survey at the beginning of the book so that students will have real data *about themselves* to use. We then pose problems based on these data in special sets of questions at the end of most exercise sets. We have found this option particularly useful during the early part of the course.

We feel that the best way to introduce inferential statistics is through probability theory. Therefore, after a brief discussion of descriptive statistics, probability is treated intuitively. This leads into the binomial distribution, and the normal distribution is then introduced as an approximation to the binomial distribution. Chapter 8 then brings everything together and discusses the method of statistical hypothesis testing. One-sample binomial tests are used to introduce this important idea.

The next few chapters discuss other types of hypothesis testing. Two-sample binomial, one- and two-sample tests of sample means with both large samples (z scores) and small samples (t scores), chi-square tests, and tests about population variance (including a light introduction to analysis of variance) are included. A chapter on correlation and prediction and one on nonparametric tests conclude the text. Once students have mastered the basic material through Chapter 8, the instructor can select from the remaining chapters those topics appropriate to the needs of his or her students.

No formal proofs are presented. When feasible, theorems are motivated by an appeal to common sense. While this presentation is not mathematically rigorous, care is taken that the material is at all times mathematically accurate. Topics are introduced informally by questions and examples that lead naturally to the development of pertinent ideas. Notation is kept as simple as possible, and illustrations are used throughout for clarification.

Numerous examples and exercises are provided from various fields, including biology, medicine, business, psychology, education, and political science. Ranging from the frivolous to the serious, they have been chosen carefully to arouse student interest. They are not just lists of numerical exercises. In this edition we have added many new exercises including a nonroutine, thought-provoking question at the beginning of most problem sets.

A glossary of new words, symbols, and formulas is given at the end of each chapter, and answers to odd-numbered exercises are given at the end of the book.

Appendix A contains a selection of typical arithmetic problems that illustrate the mathematical skills needed for the material in the book. We strongly recommend that each student do these problems at the beginning of the course. Students should be able to handle signed numbers, but no manipulative skills from algebra are needed.

A number of changes have been made for this edition. In Chapter 2 we have added some elementary material on rates, especially birth rates and mortality rates. We rewrote Chapter 3 to focus on the histogram as a picture of a frequency table, and put less emphasis on details of intervals, boundaries, and so on. In Chapter 5 we now include a section on solving binomial problems with the aid of binomial probability tables, which we include in the appendix of tables.

In Chapter 13 we have expanded the section on 2 by 2 contingency tables, and we point out the relation between such problems and two-sample binomial problems and give a simplified way to compute X^2 for 2 by 2 tables. In Chapter 16 we have added the Mann-Whitney U test for comparing two samples. We have restored the appendix on probability written for the first edition by Dr. Naiman.

There have been a few notation changes. In Chapter 12 the notation on confidence intervals has been simplified, with less use of inequality symbols. We continue to mention both $\overline{X}$ and m for sample means, and although we use the symbol m most of the time, we do use $\overline{X}$ more than in the past. We have changed the symbol for the statistic in the runs test from U to R, and now reserve the U for the new material on the Mann-Whitney U test.

Since calculators and computers have become commonplace, we deleted most of the material on coding that is fast becoming obsolete. What is left is relegated to a few exercises that focus on the properties of the mean and the standard deviation. Similarly, the square root table was deleted, as was the random numbers table, since most teachers reported that they do not use them. Also in connection with the use of calculators, we have included some material in Chapter 1 on rounding off and working with approximate numbers.

Amid all these changes, we hope that you still find the essentials that have made our text popular over the years: simplicity, accuracy, and a blend of humorous examples with real-world problems. These are elements that have motivated students with little math background and interest.

We wish to thank our colleagues at Nassau Community College for their encouragement and helpful suggestions, particularly Professors Frank Avenoso, James Baldwin, Eli Berlinger, Alice Berridge, Mauro Cassano, Dennis Christy, Jerry Kornbluth, George Miller, Aaron Schein, Michael Steuer, and Abraham Weinstein. Thanks also to Roy McLeod of LaGuardia Community College.

For helpful detailed criticism of the manuscript we would like to thank Professors Daniel Brunk; Wilfrid Dixon, UCLA; James Edmondson, Santa Barbara City College; and Paul Kroll, formerly of William Paterson College of New Jersey. For this edition in particular we wish to thank Sister Mary Erwin Baker, Saint Mary's College; Professors Donald Evans, Polk Community College; David M. Crystal, Rochester Institute of Technology; John S. Mowbray, Shippensburg State College; Norman Neff, Trenton State College; Charles A. Oprian, Western Illinois University; Ronald E. Pierce, Eastern Kentucky University; Maxine D. Reed, Tennessee State Technical Institute at Memphis; William M. Self, Pittsburg State University; William Scott, Ocean County College; and Patricia L. Smith, Old Dominion University.

We are grateful to all who helped and whose suggestions made these improvements possible.

ROBERT ROSENFELD
GENE ZIRKEL

Contents

Introduction

1

When you applied to college, you probably filled out a form designed to establish your "financial need." Maybe you took the SAT or ACT exam. Then some stranger used these numbers to help decide what college you got into and how much money you had to pay. To a large degree you were being treated as a collection of numerical values, a collection of statistics.

You should understand how statistics are used because decisions that affect you personally are based on statistics. "Your grade-point average is only 1.47; sorry kid, we'll have to put you on probation. I know that you just had a bad time with your parents, but let's face it, everybody's got problems." That hurts.

There are probably times when you have mixed feelings about being treated as a source of statistics.

DEAR FRIEND:

ØUR RELIABLE CØMPUTERIZED MATCHING SYSTEM HAS PRØCESSED YØUR VITAL STATISTICS AND WE ARE PLEASED TØ ENCLØSE HEREWITH THE NAMES, ADDRESSES (WITH ZIP CØDES) AND PHØNE NUMBERS (WITH AREA CØDES) ØF SIX IDEAL MATCHES. WE ARE HØPEFUL THAT YØU CAN ESTABLISH A LASTING RELATIØNSHIP WITH AT LEAST ØNE ØF THEM.

WE REMIND YØU THAT YØUR FEE IS NØT REFUNDABLE.

SINCERELY,
CØMPUTERIZED DATING SERVICE

Products are sold to you all the time with numbers thrown at you: (1) "I used Grit toothpaste, and now I have 20 percent fewer cavities." (Fewer than what?) (2) "Hey, kids! Start your day with Daystart Cereal! It has twice as much iron as a delicious slice of toast and more vitamin C than two slices of bacon!" (So who said that bread was a good source of iron in the first place, or that bacon has a lot of vitamin C?)

Doctors prescribe medicine and treatment for you, basing their judgment on statistical information. (1) Use of this pill will cause deleterious side effects in 1.4 percent of its users. (Is the risk worth taking?) (2) There is a 40 percent chance that an adult suffering from a herniated spinal disk will recover spontaneously. (Should we go ahead with the operation?)

TWO USES OF STATISTICS

Consider the following example. Maybe you have participated in this kind of survey if you live near a big shopping mall.

> A random sample of adults was taken in a large shopping plaza in the city of Niles. Of those questioned, 15 percent used NoCav brand toothpaste. Subsequently a concentrated advertising campaign was undertaken to sell NoCav to the public. A second survey taken 3 weeks after this campaign showed that 19 percent of those questioned used NoCav toothpaste.

Are we correct in assuming that the rise from 15 percent in the first sample to 19 percent in the second sample is due to advertising? If we have doubts that the advertising caused a substantial increase in the use of NoCav toothpaste, what questions should we ask concerning the data presented? What about the data that were omitted from the presentation?

In this example we see numbers used in two different ways. The number 15 percent is used to describe the fraction of people in the first sample who used NoCav. As such it summarizes with conciseness and clarity the unreported fact that of 140 persons interviewed, 21 used NoCav. This is an example of descriptive statistics. **Descriptive statistics** is the use of numbers to summarize information which is known about some situation. In contrast to this use of numbers, if we use this sample to imply that approximately 15 percent of *all* the adults in Niles used NoCav, then we are using the number to infer something about a larger population for which we do not have complete information. This is an example of statistical inference. **Statistical inference** is the use of numbers to give numerical information about larger groups than those from which the original raw data were taken.

In characterizing a large amount of data by a few descriptive statistics we gain clarity and compactness, but we lose detail. The following statistics, by summarizing information, describe in some way the populations from which they were taken.

1. The average IQ at Nostrum College is 109.
2. The marks on the last exam ranged from 51 to 98.
3. Nielsen reports that 25 percent of those who were interviewed watched the President's news conference last Sunday night.

The following are examples of statistical inference. We might infer from appropriate samples that

1. Between 20 and 25 percent of American college students are married.

2. Cholesterol level and heart trouble are related.
3. 25 percent of all television viewers watched the President's news conference last Sunday.

Here is a more detailed example of statistical inference.

> Suppose there were a disease in which three-fourths of the patients recovered without treatment within 3 months of contracting the disease. Suppose also that a doctor claims to have discovered a new drug to cure this disease. We shall administer the drug to 100 patients. Even if the drug were useless, we would still expect about 75 (three-fourths) of these people to recover. Due to chance variations more or less than 75 may recover.

One of the problems of statistical inference for the example given above is to decide how many must recover before we are willing to accept the drug as a cure. Certainly if all 100 recovered, we would be enthusiastic about the drug's potential. But how about 95 or 90 or 80 recoveries? Where should we draw the line?

The job of deciding where to draw the line is an important one for the statistician. It is one of the main skills we hope you will develop from this book. Can we confidently say that the new drug saves lives, or is it likely that this result occurred by chance? Even if all 100 recovered, it is possible (though very unlikely) that they would have recovered anyway. Perhaps, just by luck, this particular group of 100 patients was unusually resistant to the disease.

It is important for the statistician to pick the sample in an impartial way. If by chance we happened to test the drug on only mild cases, our results would be misleading. We would hope that the sample is truly a mirror of the population we want to learn about (in this case all victims of this disease).

Sample surveys, polls, and statistical tests have become a part of our way of life. Every day, people present figures to prove or disprove some claim: Does a certain food additive cause cancer? Does smoking marijuana lead to heroin use? In this book we will study some of the tests that statisticians use when making claims. We hope to show you how such tests should be done properly and how to interpret the "proof" of such claims.

SOME STATISTICAL TERMS

If we are listing the ages of students in a certain school, then each age is called a **raw score.** In general, a raw score is any number as it originally appears in an experiment. A collection of such scores about one particular thing is frequently referred to as a **distribution** of scores. If we consider the grades that your class gets on the first test in this course, then you will be very interested not only in the entire distribution of the class's grades, but also in one particular raw score, namely, your own grade on the test.

Often we collect raw scores about several different things. We may, for instance, collect information on heights, weights, and ages of the people who belong to a particular organization. All the information we have is called our **data.** By the way, "data" is a plural noun. For example, you should say, "The data *show* that X is more popular than Y." It is incorrect to say "The data *shows* that . . . ." The singular of data is datum.

The word **population** is used to refer to *all* the persons, objects, scores, or measurements under consideration. The word **sample** refers to any portion of the population. A population may be large or small.

Suppose a scientist is trying to determine the average weight of all 1-year-old male white rabbits which are raised in laboratories using a certain diet. It is impossible for her to weigh every rabbit in the population because the population never exists completely at any one time. If she selects 50 rabbits and determines their average weight, these 50 would be referred to as a sample from the population.

We used the word **random** in examples at the beginning of this chapter. No word is more important to the theory of inferential statistics than this word. An item is chosen "at random" from a population if in the selection every item in the population has the same probability or chance of being selected; the *process* of selection does not favor any particular item either intentionally or inadvertently. A sample of items in which each item is chosen this way is called a **random sample.** Entire textbooks have been written describing procedures for selecting random samples, and the process can become quite technical. In this text we prefer to leave the idea to your intuition. It will be sufficient to think of a random sample as one that has been picked "fairly"—without prejudicing the chances of any member of the population to be chosen. For example, if we want to pick a random sample of 20 people from some population, then every possible grouping of 20 people should have an equal chance of being selected as the sample. The practice of putting paper slips into a large drum, mixing them well, and then picking one without looking is a simple model of random selection.

Statistical testing is frequently based upon the assumption that the sample was picked randomly. If it turns out that the sample was not random, the results may not be useful. Hidden, unsuspected bias can completely destroy the usefulness of statistical information and statistical inferences made from such information. For example, if random phone calls were made at 1 P.M. to sample the population of all voters, many people with full-time day jobs would be missed.

Note: The word *random* describes the *process* by which the sample was chosen. This does *not guarantee* that the sample will turn out to be representative.

ARITHMETIC, CALCULATORS, AND ROUNDING OFF

During this course you will have to do a lot of problems, and most likely you will use a calculator for some of the arithmetic. One of the characteristics of calculators is that they routinely display a lot of digits, often more than make sense in a particular problem. This means that you will want to round off your results, to some convenient *approximate* value. This section gives you some informal rules for reasonable rounding off.

BASIC RULE FOR ROUNDING OFF AT A GIVEN DIGIT

Look at the digit following the one that is to be rounded. If it is 4 or less, simply drop it; but if it is 5 or more, then add one to the digit that is being rounded.

EXAMPLE 1-1 Round these numbers off to two decimal places: 16.837, 8.00319, 9.105, and 10.1349.

SOLUTION

original data	next digit	rule	result
16.837	7	add 1 to 3	16.84
8.00319	3	drop	8.00
9.105	5	add 1 to 0	9.11
10.1349	4	drop	10.13

Notice that in rounding the last number we did *not* round off twice. A common *mistake* is to first change 10.1349 to 10.135 and then to 10.14. This is not correct. We look *only* at the next digit when rounding off. ■

EXAMPLE 1-2 Round off these numbers to the nearest hundred: 5826, 9084, 163.7, and 4041.

SOLUTION

original data	next digit	rule	result
5826	2	drop	5800
9084	8	add 1 to 0	9100
163.7	6	add 1 to 1	200
4041	4	drop	4000

Note two things here. The first answer is 5800 and not 58. We need the zeros to indicate that the number is large, in the thousands. Note also that in the last answer the first zero is accurate and the last two are not. When it is important to indicate this, we place a bar over the last accurate digit. In this case we would write $4\bar{0}00$.

In the same vein we only write final zeros after a decimal point if needed. The numbers 6.0, 6, and 6.00 are all different if they are approximate values. ■

CALCULATING WITH APPROXIMATE NUMBERS

If we find the average of the eight numbers 6, 9, 9, 0, 7, 7, 5, and 4, we obtain 47/8 = 5.875. Now if these eight numbers represent the number of people in eight families, then they are **exact numbers.** No family had *about* 7 people. Therefore we can say that the average is exactly 5.875.

However, if the data represent eight measurements, then they are **approximate numbers.** For example, the numbers could be the weight in tons of feed that a large ranch possessed at the end of each of the last eight months, or the number of quarts of strawberries eight people picked on a picnic. In the case of measurements the data are not exact, but approximate. They were rounded off to the nearest whole number. For us to claim to know the average correct to the nearest thousandth is silly. Since our data were only given correct to the nearest whole number, we can only approximate the average to the nearest whole number. Thus to say that the average is "about 6" is more sensible than to give the false impression that we know this figure correct to three decimal places.

> Our basic rule of thumb is this:
>
> The final result of calculations with approximate numbers should be in agreement with the data you started with, so round them off appropriately.

There are other more complicated rules, and in fact the whole science of numerical analysis deals with the accuracy and precision of numbers after calculations. However, this simple rule of thumb will suffice for our purposes.

INTERMEDIATE RESULTS

When reporting intermediate results, before the final answer, it is not unusual to keep one or more extra digits. Thus to find the average of the approximate numbers 19.62, 18.3, 17.064, and 16.21, we first find the sum, 71.184. Dividing this sum by 4 yields an average of 17.796. Since the weakest item in our data, 18.3, was correct to only one decimal place, we round off the average to 17.8. It would not be uncommon to record the sum of the numbers, which is an intermediate result, to two decimal places as 71.18.

Different calculators may yield different numbers of digits in their answers, and if you use pencil and paper, you will probably use less digits than a calculator. This can sometimes lead to a slightly different final answer.

Which answer is wrong? Neither one. Remember we are dealing with approximate numbers. Do not be concerned if your answer differs by 1 in the least accurate digit from the answer obtained by your neighbor, by your instructor, or from an answer key.

WARNING

You should be careful *not to round off too early in your calculations.* Always carry at least one extra digit in your intermediate results until you obtain your final answer.

Because many people do their arithmetic on calculators or even on computers, we must be careful. For example, if the sum of 17 approximate numbers is 68.23, their average, 68.23/17, may be displayed on a calculator as 4.013 529 411 764. Of course, you know that this figure is wrong and that the answer is better expressed as 4.01. Years ago, when we all used pencil and paper, there was little danger that students would carry out a long division to 13 digits. Today, however, it is important to realize that calculators often give senseless and misleading information when dealing with approximate numbers. It is up to us to use these machines intelligently. The following quotation from *Statistical Method in Biological Assay* by D. J. Finney underscores our point.

> Bioassays are seldom sufficiently precise to warrant quotation of results to more than 4 significant digits: a statement that a test preparation is estimated to have a potency of 35.71685 units per mg is both stupid and confusing.

One calculator says that the square root of 34.26 is 5.832 042 506 64. If 34.26 is an exact value, this is fine, but if 34.26 represents an approximate number, then, depending on the original data, 5.8 might be a better answer. Remember, it is your responsibility to round off calculator results reasonably.

Throughout this book we will usually assume that all data in any one problem are given to the same accuracy or precision. This is in keeping with ordinary usage. Statisticians do not measure some people to the nearest half inch and others to the nearest inch. Thus if some data are given as 17, 21, 19, and 30, we presume that 30 is also correct to the nearest unit. Similarly, 18.02, 191, 19.61, and 10 imply that the 10 and the 191 are correct to two decimal places. It would be better to write them as 10.00 and 191.00 to indicate this.

STUDY AIDS

VOCABULARY

1. Descriptive statistics	**2.** Statistical inference	**3.** Raw score
4. Distribution	**5.** Data	**6.** Population
7. Sample	**8.** Random	**9.** Random sample
10. Exact number	**11.** Approximate number	

EXERCISES

1-1 Michelangelo tosses a fair coin 4 times and obtains 4 heads. Is this sample of the results of a coin-tossing experiment a *random* sample?

1-2 What is the point of the NoCav example in this chapter? What questions would you want to ask to help you decide if the advertising was effective? What are some other possible explanations for the increase in users of NoCav?

1-3 "There are three kinds of lies: lies, damned lies, and statistics."—Benjamin Disraeli, Prime Minister of England (1804–1881). Why do people both admire and fear statistics? What are some of the advantages to the use of statistics? Are there any disadvantages?

1-4 95 percent of the people who use heroin started out using marijuana regularly. Therefore, using marijuana regularly leads to using heroin. Comment.

1-5 98 percent of the people who use marijuana first drank milk on a regular basis. Therefore, drinking milk on a regular basis leads to using marijuana. Comment.

1-6 Classify each of the following as either statistical inference or descriptive statistics.

(a) Walter Krankrite predicts the results of an election after looking at the votes in 15 of 100 districts.

(b) Dr. Bea Kareful, an ecologist, says that the flesh of fish in a certain lake contains an average of 400 units of mercury.

(c) At Webelo Normal High School last year the average SAT score was 528.

(d) The safety councils of Pessam and Mystic counties predict 600 automobile accidents for the next July 4 weekend.

(e) Last year 72 percent of the workers in Scrooge and Marley's accounting firm missed at least 1 day of work.

1-7 For each of the following statements describe the *population* or *populations* that should have been sampled to get this information. If necessary, clarify the question until it is clear what population is meant.

(a) 30 percent of all suicides are widows.

(b) Malignant tumors were found in 80 percent of the rats injected with 10 ml of chemical X.

(c) English majors at Hudson University have higher grade-point averages than chemistry majors.
(d) Too much cholesterol is bad for your heart.
(e) Girls learn to speak before boys do.

1-8 Have someone in the class secretly mix in a large bag any amount of dried yellow split peas with any amount of dried green split peas. Without counting or even seeing *all* the peas, discuss any methods that could be used to estimate what fraction of the peas in the bag are green. Test your methods. Were they successful? In this experiment, what is the population? What is the sample?

1-9 Find some uses of statistics (sample, average, percentile, etc.) in texts that you use in other courses. Can you classify them as either descriptive or inferential?

1-10 Find some uses of statistics in current magazines and newspapers. Classify them as descriptive or inferential.

1-11 An advertisement states that three-fourths of doctors interviewed recommended Brand X. What is your reaction?

1-12 Answer (a), (b), or (c).
(a) Find some reference to the 1936 survey by *Literary Digest* which predicted that Alf Landon would easily win the United States presidential election (e.g., Huff, *How to Lie with Statistics*).
(b) Find some reference to the polls on the June 18, 1970, British election. (Check newspapers of that week.)
(c) Find some references which discuss the randomness of the December 1, 1969, draft lottery. (Check newspapers of that week, or see the book, *A Sampler on Sampling* by Bill Williams.)

1-13 Round off as indicated.
(a) 16.43 (tenths)
(b) 50,631 (hundreds)
(c) 40,538 (tens)
(d) 18.062 (tenths)
(e) 40,100 (thousands)
(f) 19.8963 (hundredths)

1-14 Estimate these square roots if the original data were given as indicated.
(a) $\sqrt{3.120}$ inches (tenth of an inch)
(b) $\sqrt{0.0196}$ tons (hundredths of a ton)
(c) $\sqrt{800}$ degrees (nearest 10 degrees)
(d) $\sqrt{89}$ volts (nearest volt)
(e) $\sqrt{26{,}000{,}000}$ people (nearest 1000 people)

1-15 Average these expenditures: \$16,000, \$120,000, \$4$\bar{0}$0,000.

1-16 Find the sum of 0.00160, 0.00058, and 0.002098.

1-17 (a) Calculate $z = \dfrac{90.3 - 46.12}{20.3}$ if the original data were given to tenths of an inch.
(b) Calculate $z = \dfrac{1031 - 982.8}{2.41}$ if the original data were given to the nearest fathom.

1-18 We measure the left thumbs of eight people and obtain the following measurements: 2.30, 1.92, 2.10, 2.41, 1.88, 1.70, 2.00, and 1.80 inches. Using your rule of thumb, of course, find the average length of left thumbs.

1-19 (a) Multiply 0.18422 × 1.9 and round to the nearest tenth.