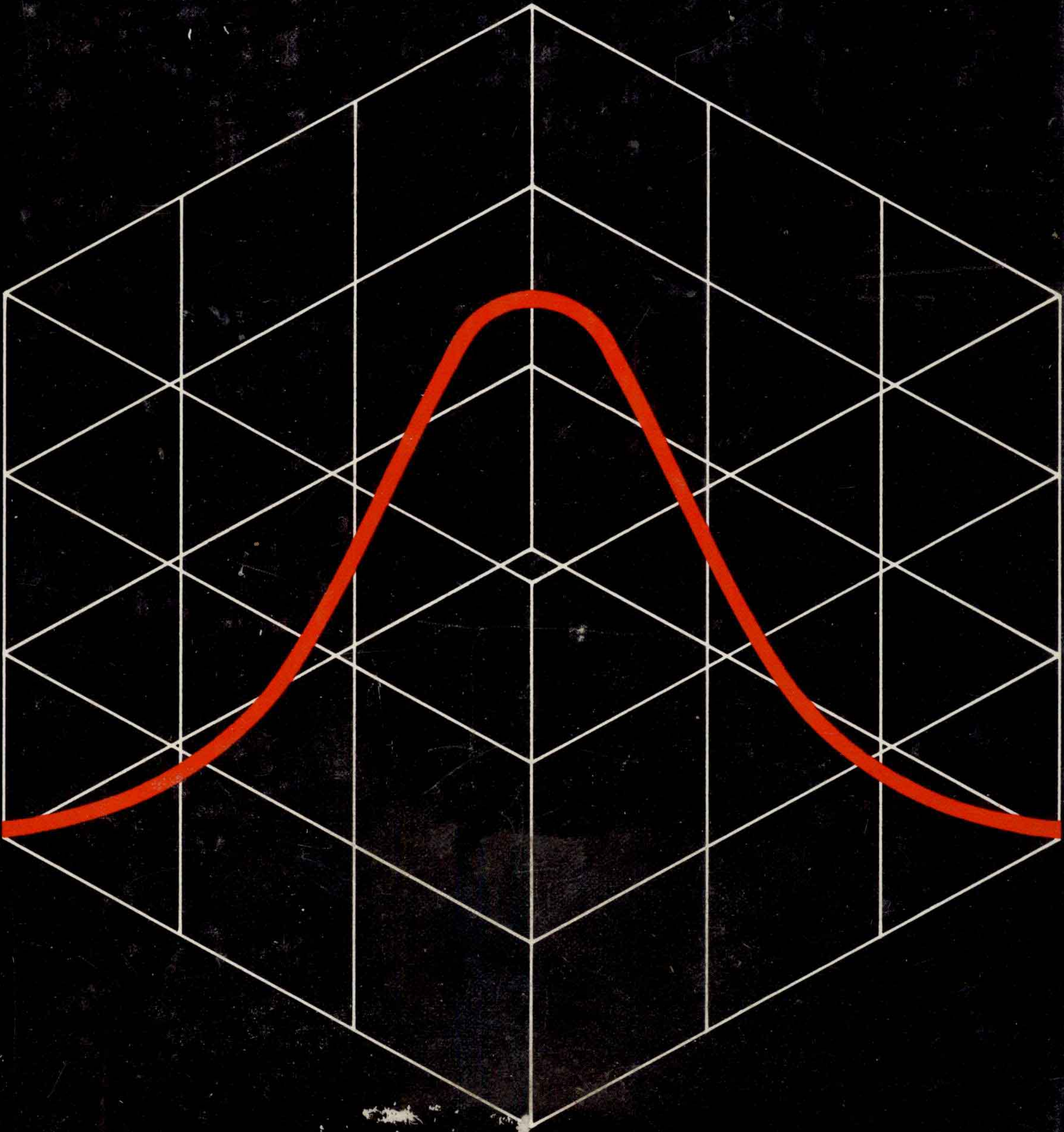
# STATISTICS

Discovering
Its
Power

Wonnacott
and
Wonnacott



# STATISTICS
# Discovering Its Power

**Ronald J. Wonnacott**
Department of Economics
University of Western Ontario

**Thomas H. Wonnacott**
Department of Statistics
University of Western Ontario

175 YEARS OF PUBLISHING
1807 1982

**John Wiley & Sons**
**New York Chichester Brisbane Toronto Singapore**

**Library of Congress Cataloging in Publication Data:**

Wonnacott, Ronald J
Statistics: discovering its power.

(Wiley series probability and mathematical
statistics: Applied probability and statistics
section ISSN 0271-6356)
Includes index.
1. Statistics. I. Wonnacott, Thomas H., 1935-
joint author. II. Title.
QA276.12.W66 519.5 80-26507
ISBN 0-471-01412-5

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

# Preface

This is an introduction to statistics at an easy and applied level. Written for a one- or two-semester course, the topics are carefully ordered so that students who use the book for only one semester can nevertheless be confident that they have covered the most important subjects.

## TO THE STUDENT

Statistics is the intriguing study of how you can describe an unknown world by opening a few windows on it. You will discover the excitement of thinking in a way you have never thought before.

This book is not a novel, and it cannot be read that way. Whenever you come to a numbered example in the text, try first to answer it yourself. Only after you have given it hard thought and, we hope, solved it, should you consult the solution we provide. The same advice holds for the exercise problems at the end of each section. These problems have been kept computationally as simple as possible, so that you can concentrate on insight rather than arithmetic. At the same time, we have tried to make them realistic by the frequent use of real data.,

The more challenging problems and sections are indicated by a star (*). For example, in Chapters 8 and 9 we give some problems that are best answered by using a computer package: we want students who like computers to see their power; but at the same time we keep these exercises optional, so that other students can fully master the text without using a computer.

Brief answers to all odd-numbered problems are given in the back of the book. Their completely worked-out solutions are available in the student's manual.

## TO THE INSTRUCTOR

Our previous textbooks (*Introductory Statistics* and *Introductory Statistics for Business and Economics*) have been aimed at the middle to high end of the first-course market; this one is aimed at the more elementary level. It is substantially easier than our earlier books in two ways: the standard statistical topics covered here are, whenever possible, treated in a simpler way; and the more demanding topics have been dropped altogether. Yet our previous books remain very helpful as references; those students who occasionally want a more advanced treatment, for example, can refer to our *Introductory Statistics* with no roadblocks—it has similar notation, and follows the same order of topics.

With so many books already designed for the elementary course, why another? We felt there are still too few texts that are both intuitively appealing and cover the important topics. In taking on this challenge, we have introduced two distinctive features:

*1. Teaching by example.* New concepts are introduced and illustrated with examples. Many of these examples, formally numbered and set off in blue for easy reference, are formed as questions for students to answer themselves. This type of learning is a real pleasure, in both the classroom and individual reading; students find that working out the answers can be as enjoyable as doing a recreational puzzle.

In a sense, it's the Socratic method: we try to ask the right question that will start students thinking on their own. In our own classrooms, we have often found the ensuing discussion teaches us as much as the students, and sharing this learning is a very rewarding experience.

*2. Highlighting of important topics.* As you skim the Table of Contents, you will see the topics we feel are important. Some deserve special note here. First, we regard regression as the most powerful tool that can be learned in a basic course; so we get to it early and deal with it thoroughly. We include multiple regression, for example, and emphasize its value in reducing bias in observational studies.

We stress confidence intervals and $p$ values, because we feel these are much more teachable than classical hypothesis testing. Of course, we do examine classical tests—emphasizing tail probabilities (prob-values) more than testing at a fixed level such as 5%. But the text is structured so you can spend as much or as little time on this subject as you choose. Our subtle shift in emphasis from testing to estimation manifests itself in several ways. For example, nearly all statistical techniques (such as the difference of two means and regression slopes) are first introduced with confidence intervals; tests are only undertaken later. And to further stress estimation, we have a short chapter on shrinkage estimates of the Bayesian or James-Stein type. Such modern topics are not only very useful and interesting, but are also easily learned—they supplement common sense rather than contradict it.

In a one-semester course, the material up to multiple regression in Chapter 8 can easily be covered. Alternatively, the course can cover the basics as far as confidence intervals in Chapter 5, followed by a smorgasbord chosen from the remaining chapters. In a two-semester course, it should be possible to teach all of the book at a relaxed and thoughtful pace.

## ACKNOWLEDGMENTS

So many people have contributed so much to this book that it is impossible to thank them adequately. However, we must express our special thanks to Jon Baskerville, Robin Carter, L. K. Chan, A. M. Chaudhry, John Koval, and especially to the students of Stats 137 at the University of Western Ontario during 1976 to 1980, who gave many helpful suggestions on the manuscript as it developed.

London, Ontario, Canada, 1981

**Thomas H. Wonnacott**
**Ronald J. Wonnacott**

# Contents

## PART III RELATING TWO OR MORE VARIABLES

## PART IV FURTHER TOPICS

# PART I

# BASIC STATISTICS

# CHAPTER 1

## The Nature of Statistics

*Life is the art of drawing sufficient conclusions from insufficient premises.*

Samuel Butler

Statistics, like life, is an art—the art of making wise decisions in the face of uncertainty. Many people think of statistics as simply collecting numbers. Indeed, this was its original meaning: State-istics was the collection of population and economic information vital to the state. But statistics is now much more than this. It has developed into a scientific method of analysis widely applied in business and all the social and natural sciences. To get an idea of what modern statistics is, we will examine a couple of typical applications—a political poll, and an experimental surgical technique.

### 1-1 RANDOM SAMPLING

Before a presidential election, the Gallup poll tries to pick the winner. It also tries to predict how much support each candidate will get from men and women, whites and blacks, Protestants and Catholics, and so on. To be concrete, consider the problem of predicting the proportion of the population that will support the Democratic candidate in the next

U.S. presidential election. Clearly, canvassing the entire population would be an unrealistic task. All we can do is to take a sample, in the hope that the sample porportion will provide a good estimate of the population proportion.

Just how should the sample be chosen? Some interesting lessons can be learned from history. In 1936, for example, when polling was in its infancy, the *Literary Digest* tried to predict the U.S. vote in the presidential election. They mailed questionnaires to ten million voters chosen from lists such as telephone books and club memberships—lists that tended to be more heavily Republican than the voting population at large. Only a quarter responded—and, as it turned out, they tended to be much more Republican than the nonrespondents. This sample was so mismanaged ("biased") that it pointed to a Republican majority. Election day produced a rude surprise: Less than 40% of the voter population were Republicans, and the Democratic incumbent, Roosevelt, was elected with an historic majority.

Other examples of biased samples are easy to find. Informal polls of people on the street are often biased because the interviewer may select people that seem civil and well dressed; a surly worker or harassed mother is overlooked. Members of congress cannot rely on their mail as an unbiased sample of their constituency, since mail is a sample of people with strong opinions and includes an inordinate number of cranks and members of pressure groups.

From such bitter experience important lessons have been learned: To avoid bias, *every* voter must have a chance to be counted. And to avoid slighting any voter, even unintentionally, the sample should be selected *randomly*. There are various ways of doing this, but the simplest to visualize is the following: Put each voter's name on a chip, stir the chips thoroughly in a large bowl, and draw out a sample of, say, a thousand chips. This gives the names of the thousand voters who make up what is called a *simple random sample* of size $n = 1000$.

Unfortunately, in practice simple random sampling is often very slow and expensive. For example, in polling the population of American voters, it would be very difficult to track down the many isolated voters who would turn up in the sample. Much more efficient is *multistage sampling:* From the nation as a whole, take a random sample of a few cities (and counties); within each of these cities, take a random sample of a few wards; finally, within each ward take a random sample of several individuals. While methods like this are frequently used, in this book we will assume simple random sampling (as in drawing chips from a bowl)—leaving the sophisticated variations to an advanced textbook.

Simple random samples will not reflect the population perfectly, of course. If only a few voters are drawn at random, the luck of the draw will be a factor. For example, how might a sample of just 10 voters turn out, from a population of voters split 50-50 Democrat and Republican? The likeliest result is a sample of 5 Democrats, but the luck of the draw might produce 8 or 9 Democrats—just as 10 flips of a fair coin might produce 8 or 9 heads. That is, the sample proportion of Democrats might be 80 or 90%—a far cry from the population proportion of 50%.

In larger samples, the sample proportion $P$ will be a more reliable estimate of the population proportion of Democrats (which we denote by $\pi$, the Greek equivalent of our $P$.) In fact, the easiest way to show how well $\pi$ is estimated by $P$ is to give a so-called *confidence interval:*

$$\pi = P \pm \text{a small error} \tag{1-1}$$

with crucial questions being, "How small is this error?" and "How sure are we that we are right?" Since this typifies the very core of the book, we state the answer more precisely, in the language of Chapter 5 (where you will find it fully derived):

*For simple random sampling, we can state with 95% confidence that*

$$\boxed{\pi = P \pm 1.96\sqrt{\frac{P(1-P)}{n}}} \tag{1-2}$$

*where $\pi$ and $P$ are the population and sample proportions, and $n$ is the sample size.*

Before we illustrate this confidence interval, we repeat the warning that we gave in the preface: Every numbered example in this text is an exercise that you should actively work out yourself, rather than passively read. We therefore put each example in the form of a question for you to answer; if you get stuck, then you may read the solution. But in all cases remember that *statistics is not a spectator sport.* You cannot learn it by watching, any more than you can learn to ride a bike by watching. You have to jump on and take a few spills.

**Example 1-1**

The Gallup poll is a combination of multistage and other kinds of random sampling that provides about the same accuracy as simple random sampling. Throughout this book, therefore, we will do little damage in assuming it actually *is* a simple random sample for purposes of applying equation (1-2).

Just before the 1980 presidential election, a Gallup poll of 1500 voters showed 720 for Carter and the remaining 780 for Reagan (ignoring third-party candidates). Calculate the 95% confidence interval for the population proportion $\pi$ of all voters who were for Carter.

**Solution**

The sample size is $n = 1500$ and the sample proportion is

$$P = \frac{720}{1500} = .48$$

Substitute these into equation (1-2):

$$\begin{aligned} \pi &= .48 \pm 1.96\sqrt{\frac{.48(.52)}{1500}} \\ \pi &= .48 \pm .03 \end{aligned} \qquad (1\text{-}3)$$

That is, with 95% confidence, the proportion for Carter in the whole population of voters was between 45% and 51%.

One of the major objectives of this book will be to construct confidence intervals like equation (1-3)—or, as we will hereafter abbreviate references to equations, "like (1-3)." Another related objective is to *test hypotheses.* For example, suppose a claim is made that only 40% of the population supports Carter. In mathematical terms, this hypothesis may be written $\pi = .40$. On the basis of the information in (1-3), we would reject this hypothesis, of course. In general, there is always this kind of close association between confidence intervals and hypothesis tests.

We can make several other crucial observations about (1-2):

1. The estimate is *not* made with certainty; we are only 95% confident. We must concede the 5% possibility that the draw turned

up a misleading sample—just as in flipping a coin 10 times, it is possible that 8 or 9 heads will occur.

2. As sample size $n$ increases, we note that the error allowance in (1-2) shrinks. For example, if we increased our sample to 15,000 voters, and continued to observe a proportion of .48 for Carter, the 95% confidence interval would shrink to the more precise value:

$$\pi = .48 \pm .01 \tag{1-4}$$

This is also intuitively correct: A larger sample contains more information, and hence allows a more precise conclusion.

In conclusion, what does random sampling accomplish? It allows us to make an *unbiased* estimate of the unknown population—including a confidence interval that shows the uncertainty involved.

**PROBLEMS**

1-1 Ten days before the 1980 presidential election, a Gallup poll showed the following percentages supporting Carter. (As we mentioned already, treat the sample from each group as a random sample.)

| | | | |
|---|---|---|---|
| men | 49% ($n$ = 600) | under 30 | 48% ($n$ = 200) |
| women | 58% ($n$ = 600) | over 30 | 55% ($n$ = 1000) |

(a) For each group, calculate a 95% confidence interval for the population proportion supporting Carter.

(b) Star each case where you can conclude with 95% confidence that Carter had a majority (or minority).

1-2 Project yourself back in time to six recent U.S. presidential elections. In parentheses we give the results of the Gallup pre-election poll of 1500 voters (ignoring third-party candidates, as usual).

| *Year* | *Democrat* | *Republican* |
|---|---|---|
| 1960 | Kennedy (51%) | Nixon (49%) |
| 1964 | Johnson (64%) | Goldwater (36%) |
| 1968 | Humphrey (50%) | Nixon (50%) |
| 1972 | McGovern (38%) | Nixon (62%) |
| 1976 | Carter (51%) | Ford (49%) |
| 1980 | Carter (48%) | Reagan (52%) |

(a) In each case, construct a 95% confidence interval for the propor-