Elementary Statistics

DONALD R. BURLESON

Elementary Statistics

Library of Congress Cataloging in Publication Data

Burleson, Donald R
Elementary statistics.

Bibliographies.
Includes index.
1. Statistics. I. Title.
QA276.12.B87 519.5 79-19885
ISBN 0-87626-213-2

Design by David Ford

17 Dunster Street, Cambridge, Massachusetts 02138

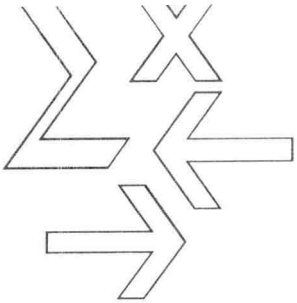
10 9 8 7 6 5 4 3 2 1

Exercises and Examples in Particular Areas of Application
(P = "Ponderables" Problems; E = Examples)

Chapter, Section	Natural Sciences	Social Sciences	Business-Economics	Education	General Applications
1 A	P8		P2	9	6
B	3, 4				E1, E2
C					1, 4
2 A	17		10	9, 18, 20, E6	1, 2, 3, 4, 5, 16, E7, E13
B		P10		6, 9	E7, E13
C	E5, E8	3, 9, E7	P1		4, 12, P2
3 A	5	7, 10, P2	8, P7		9, 11, 12, P1 P3, E8, E9, E10
B	5	3, E1	1		3, E3
4 A	8, 11, P4	1, 7, 14, P6, P8	3, 9, 13, E6		5, P9, E4, E8 E9
B	1, 5, 6	P5, E1, E3	3, 4, E2, E4		2, 7, 8, P2
C	5	8, 13, E3, E4	1, P1	10, 11	E1, E2, E5, E6
5 A	5, 7, 9, P1, P3	14, 15, P2	1, 11, 13, 16, 17, 19, 20, 21, 22, E2, E3, E7, E8		3, E1, E4, E5
B	P4, P7	11, 12, 13	1, 2, 7, 8, 9, 10, 15, 17, P1, P9		3, 4, 5, 6, 19, 20, P3, P5, E1, E2, E3, E4, E5, E6, E7
C	14, E3, E4	7, 9, 12, 13, P1, P4	1, 2, 16, 18, P2, P6	10, 11	3, 4, 5, 6, 20, E1, E2, E5, E6
D		3, 8	1, 2, 10		4, 5, 9, E1, E2, E3
6 A	2, 11	1, 12, 14, P1, P2, P3	5, 6, 7, 9, 16	15, E1	3, 8, 10, E2
B	12, 14	10, 13, 15, 16, 17, P3, E3	6, 7, 8, 18, E4	3, 11	1, 2, 4, P4, E1
7 A	5, 9	6, 7, 8, 11, E1	4, E3	1, 3	2, 10, E4
B	1, 3, 4, 5, 6, 7, 10, 11, P1, P2	8, 9, E1		2	E2
8 A		8		3, 6, 7, E1, E2	1, 2, 5
B	10	E1	1, 2	5	3, 6, 7, 9, E2
C	6	4	3, E1		1, 2, 5, 7, 9
D		6, 10	1, 2		3, 4, 5, 7, 8, 9, E2
E	6		8	1, 2, 7	3, 4, 5, 9, 10, E1, E2

Topics Introduced in "Ponderables"

1	A	Geometric mean, harmonic mean, midrange, quartiles, midquartile
	B	Square root interpolation, mean deviation, semi-interquartile range (Q)
	C	Frequency distribution standard deviation, frequency distribution mode
2	B	Odds, conditional probability, independent events
	C	Hypergeometric distribution
3	A	Tchebycheff's Theorem
5	A	Small-sample t-test
	B	Two-sample t-test
	C	Differences between means
	D	Double interpolation in F tables
6	A	Factorial design, interaction, two-way ANOVA
	B	Yates' correction
8	A	Sign test for the median



Preface

This book is designed to serve a standard one-semester general statistics course at the college freshman level, for students with perhaps only a very modest background in mathematics.

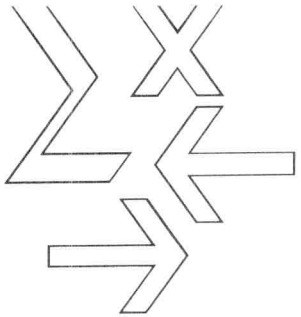
Having written a number of mathematics textbooks and other books, and having taught statistics and a great variety of other courses for many years, I have again employed here an approach that I have long found effective both in writing and in teaching: an informal and conversational exposition, with examples worked out in detail at every turn to illustrate what has just been introduced and explained. I have made every effort to avoid "skipped steps" and "gaping holes" in the discussions, so that the student has the best chance of following the examples smoothly. With this discussion-and-examples format the student should be encouraged to go through all the examples carefully, because they have been designed to bring out various important aspects of the topics in question, as have the exercises. Many examples and exercises are object lessons in such matters as the effect of sample size on the computations.

I have sought to make the book flexible, both in terms of course content and level. The number of topics available should make it easy for the instructor to select whatever is desired, and the exercise sets are plentiful enough and the discussion inclusive enough to make it possible to stress any particular topic as needed. In fact, the exercises themselves provide considerable flexibility in course level, because each section includes a regular exercise set and special problems called "Ponderables." These addititional problems offer either more challenging questions about the section or an extension of the topics covered, or both. The instructor may assign these or not as preferred, or may use them for classroom discussion.

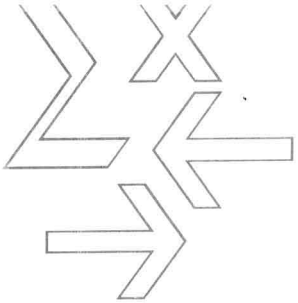
For their kind assistance I wish to thank all those at Winthrop Publishers who have worked with me on the book's production, as well as the reviewers who have made useful suggestions in manuscript. I want to express appreciation to my colleagues and students at Middlesex Community College for their words of encouragement. Finally, but especially, I want to thank my wife Sue and my sons Bruce and Brian for the patience and understanding that they have

shown once again during a long and difficult time; I well know that when a community college instructor writes a textbook and keeps up with a heavy teaching load, the demands on the family are inordinate.

Donald R. Burleson



Contents



1 Statistics: A Beginning

Well begun is half done.—Horace

A MEASURING CENTRAL TENDENCY

Probably it would surprise no one to say that we live in an increasingly complex world—and a world increasingly *numerical*. We are surrounded, in our daily lives, by numerical information. Advertisers hurl conflicting figures and claims at us from all sides, parading data before us to sell us their products. Our banking facilities and our insurance policies are founded on stunning amounts of numerical analysis. Our jobs often call for struggling through mazes of numerical information. Our physical health depends to no small degree upon the applications of numerical methods to medical science, to the production and distribution of food, to the preservation of the environment. Every day, weighty decisions are made in business and government—from the village store to the giant corporation, from the town level to the international level—by considerations of numerical data. Our cities flourish because of intelligent numerical planning and decision making, or crumble because of the lack of it. Our industries, our governments, all facets of our lives are shaped by the manipulation of numerical information.

It is therefore natural that the science of *statistics* has evolved as one of the most important of all human endeavors. In this book we will examine some of the concepts and procedures belonging to this far-reaching subject, to gain an acquaintance with the fundamentals of its applications to human life.

Statistics has so many aspects and so many uses that a brief definition of it, although necessary, is by itself unlikely to convey any true picture of its real nature. However, we may start by defining **statistics** as

the science of organizing and analyzing numerical data for the purposes of description and decision making.

That is, statistics concerns itself with certain numerical *descriptions* of the world, or parts of the world, and with the making of *decisions* on numerical

grounds—intelligent decisions that often must be made earlier than we would have liked, and on the basis of only partial information.

Let us look at a few preliminary terms and ideas.

First, the total body of information that we can be concerned about in any given situation—the "sea of information" on whose shores we stand looking and wondering—is called a **population.**

In statistics a population may or may not have anything to do with people. For example, a population *may* be all the people in Los Angeles, or all the people in California, or in the United States, or in the world. Or it may be all the catsup bottles that come off a conveyor belt in a given year, or the heights of all the trees in a certain forest, or all the members of a certain rare family of butterflies, or all the possible readings on a meter for a chemical experiment. In short, a population may be a lot of things—it is, in any given situation, the *total* body of information that one would theoretically have to deal with.

The trouble with populations is that they tend to be *large.* Sometimes they are enormous, in fact. If the population in question, for example, is composed of all the actual weights of the schoolchildren in a given country, that population may contain millions of numbers. (A population, for that matter, may even be *infinite.* For example, if a population consists of all the *possible* weights of schoolchildren, then it literally contains infinitely many numbers.)

For this reason, statisticians concern themselves with *sampling.*

A **sample,** as one would expect, is a selection of items taken from a population. For example, a sample could consist of 100 schoolchildren selected from among *all* the schoolchildren in Massachusetts (a population consisting of many thousands). Or a sample could consist of 250 observations of an experiment that can be performed any number of times: a sample of size 250 taken from an infinite population.

The reason for sampling is clear: a population, typically, is very large, so that the statistician cannot possibly hope to handle *all* the information about it, but must be content with looking only at *some* of the information (a sample).

A statistician hopes that the sample studied will paint some reasonably accurate and enlightening picture of the population from which it was taken. This is one of the ways in which statistics makes it possible to make informed decisions based upon partial information. But if this approach is to work sufficiently well, the sample taken should (if possible) be of good size, and should be *random.*

A sample is **random** if it is drawn from the population in such a way that no one could have predicted which items would be selected and which would not, each item having the same chance of being selected. This selection, needless to say, must be done with some care. For example, if out of a population of accounts in file folders we wish to select one folder at random (to audit that account, say), we may assign numbers to the folders and then take a number from a table of *random digits,* and select that corresponding folder (or keep taking random digits until one set of digits matches a folder number). This way, no one could say, "You picked my folder because you don't like me!" because no one could have predicted which folder we would select. On the

other hand, it would *not* be a random procedure simply to close our eyes and reach into the drawer containing the alphabetized folders and pick one, because (for example) we would be more likely to choose a folder starting with M in the middle of the drawer than to pick one starting with A or Z.

If a sample is selected in such a way that it is *not* random, then we may say that it is **biased.** For a very obvious example, suppose we are sampling people and measuring their heights to gain an idea of the public's heights in general. If we were to take our sample by stopping basketball players on their way to and from the locker room, our sample would be so thoroughly biased that it would provide no real idea of most people's heights.

Now, suppose that from some population, a random sample of numbers is taken.

One aspect of the sample which we may wish to measure is **central tendency:** We may wish to determine, in some sense, where the "middle" of the data is. And to do so, we should compute some sort of *statistic* that measures central tendency. In popular terminology, we want to compute some sort of "average" for the sample numbers. It turns out that there are a number of different ways to do this. (In computing an "average" of some sort for a sample of numbers drawn from a population, we become involved with **descriptive statistics:** computations which simply *describe* some aspect of a sample or a population. This descriptive aspect of statistics we will look at in various forms before pursuing **inferential statistics:** the science of drawing probable inferences or conclusions about populations by analyzing samples. More about this later.)

The most common kind of "average"—the better term being "measure of central tendency"—is called the **mean** (also sometimes called the **arithmetic mean** to distinguish it from certain other means).

Given a set of sample numbers

$$\{x_1, x_2, x_3, \ldots, x_n\}$$

(where n is the number of items in the sample), the *mean* of the sample is denoted $\bar{x}$ and is defined to be

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n}.$$

That is, we simply add the numbers and divide by the number of things added. (This is what most people naturally think of when the term "average" is used.) Because it is a little awkward to have to write the sum of the x's "strung out" as shown, we may more compactly write

$$x_1 + x_2 + x_3 + \ldots + x_n = \sum x$$

where the $\sum$ (Greek capital letter *sigma*) simply means "the sum of . . . ." Thus

$\sum x$ (read "sigma x") means "the sum of all the x's." Thus we can write the formula for the sample mean $\bar{x}$ more simply as

$$\bar{x} = \frac{\sum x}{n}.$$

EXAMPLES

1. Given the eleven sample numbers

{52, 57, 47, 50, 40, 41, 52, 42, 40, 59, 47}

the sample mean is computed as

$$\bar{x} = \frac{\sum x}{n} = \frac{\sum x}{11}$$

$$= \frac{52+57+47+50+40+41+52+42+40+59+47}{11}$$

$$= \frac{527}{11} = 47.909090\ldots \approx 47.9$$

(where $\approx$ means "is approximately equal to"). We have retained the mean to the nearest tenth here in keeping with a general principle:

Always compute a sample statistic correct to one *more* decimal place than the given data.

(Thus, having given data in integer form calls for a mean to the nearest tenth.)

2. Given the eight sample numbers

{23.09, 38.20, 29.71, 24.58, 31.34, 28.12, 34.50, 22.76}

the sample mean is

$$\bar{x} = \frac{\sum x}{n} = \frac{232.30}{8} = 29.0375,$$

which must be rounded off. Because the given data are in hundredths, we should round this result off to the nearest thousandth; again, to one more decimal place than the given numbers. The question here is whether to round off to 29.037 or 29.038 (because 29.0375 is exactly halfway between them). The general practice, when a number occurs halfway like this, is to round the

number off to an *even* digit. Thus we would conclude

$\bar{x} \approx 29.038$.

(But, for instance, 29.0765 would have been rounded to 29.076.)

There are other ways, besides the mean, to measure central tendency. One such way is called the **median**, denoted $\tilde{x}$. To compute the median $\tilde{x}$ for a set of numbers, we first arrange the numbers *in order*, say smallest to largest, then count to the physical center of the list and take the number encountered there. (If *two* numbers are encountered there, we take the mean of those two.) Thus, in the ordered list, just as many numbers will precede the median value as will follow it. Formally, the median stands in position $\frac{n+1}{2}$ in the list, where n as usual is the number of numbers involved. Thus in a list of 25 numbers the median is the number standing in position $\frac{25+1}{2} = 13$ in the list, that is, the thirteenth number. In a list of 30 numbers, the median stands in position $\frac{30+1}{2} = 15.5$, that is, halfway between the fifteenth number and the sixteenth.

$\tilde{x}$ = the score which divides the ordered list of data into two equal parts.

EXAMPLES

3. Given the 13 sample numbers

{81, 70, 67, 76, 88, 84, 67, 83, 78, 81, 76, 73, 74}

to find the median $\tilde{x}$ we first arrange the numbers in order:

{67, 67, 70, 73, 74, 76, 76, 78, 81, 81, 83, 84, 88}

and, counting to the middle and finding the seventh number $\left(\text{the number in position } \frac{n+1}{2} = \frac{13+1}{2} = 7\right)$ to be 76, we conclude that

$\tilde{x} = 76.0$.

Again, we have expressed a sample statistic to one more decimal place than the given data.

4. Given the 14 sample numbers

{35.6, 38.1, 31.7, 25.4, 38.5, 25.9, 28.1, 34.7, 35.8, 37.2, 30.5, 28.1, 35.3, 28.1}

to find the median $\tilde{x}$ we order the numbers:

{25.4, 25.9, 28.1, 28.1, 30.5, 31.7, 33.6, 34.7, 35.3, 35.6, 35.8, 37.2, 38.1, 38.5}

and count to the middle, finding the *two* numbers

33.6, 34.7

so that we are obliged to take their mean. The desired median is

$$\tilde{x} = \frac{33.6+34.7}{2} = \frac{68.3}{2} = 34.15.$$

We express this statistic in hundredths, one more decimal place than the given data, which were expressed in tenths.

Thus the not very informative word "average" can signify at least two different things: mean and median. There are others. Unfortunately, we are sometimes not told, when the word "average" is used, *which* meaning was intended, and there is even a possibility of deliberate manipulation (e.g., reporting the median instead of the mean, or vice-versa, just because it makes the speaker's case sound better). An honest commentator will specify *what* measure of central tendency is intended if the word "average" is used. After all, $\bar{x}$ and $\tilde{x}$ do not have to turn out equal to each other. For example, for the numbers

{75, 91, 84, 60, 48, 82, 78, 71, 77, 71}

the mean is

$$\bar{x} = \frac{737}{10} = 73.7$$

but the median is

$$\tilde{x} = \frac{75+77}{2} = 76.0.$$

In some cases, the difference between the two may be quite considerable. For example, suppose we measure the diameters of the tree trunks in a certain little grove and find them to be (in centimeters)

{32, 36, 34, 29, 39, 38, 40, 31, 27, 37}.

The mean is $\bar{x} = \frac{343}{10} = 34.3$, and the median is $\tilde{x} = \frac{34+36}{2} = 35.0$, very nearly the

same. *But:* Suppose, standing at the edge of the grove, there is an enormous tree whose trunk diameter is 215 centimeters (or 2.15 meters). If this new tree were added to the sample as an eleventh item, then the new mean would be

$$\bar{x} = \frac{558}{11} \approx 50.7 \text{ centimeters}$$

(a considerable change from 34.3), but because the numbers would then be arranged

{27, 29, 31, 32, 34, 36, 37, 38, 39, 40, 215}

the median $\tilde{x}$ would only be shifted from 35.0 to 36.0. The introduction of the "atypical" number 215 has distorted the original mean far more than it has distorted the original median, and in this situation it may well be that the median paints a more informative picture for us (concerning what *most* of the trees look like) than the mean.

However, the mean is often preferred over the median in statistical analysis because of its stability—its *reliability.* That is, in certain circumstances we can rely on it to have certain kinds of values. If we repeatedly take different random samples (at a fixed sample size) from the same population and compute their sample means, these means will tend to fluctuate by chance much less than the medians would. For that reason, sample means, together with other information, may be put to good use for estimating certain characteristics of the population itself. More of this later.

One further measure of central tendency which we may examine here is the statistic called the **mode,** which will be denoted $\hat{x}$. The mode of a set of numbers is simply that number (if any) which occurs more often than any of the others. The mode may or may not exist, and when it exists may or may not be unique.

EXAMPLES

5. For the set of sample numbers

{38, 53, 49, 38, 47, 38, 50, 47, 38, 45}

the mode is

$$\hat{x} = 38.0$$

because 38 occurs four times and none of the other numbers occurs as many as four times.