Elementary Statistical Methods

Elementary Statistical Methods

Gregory Quenell

St. Regis Workshop

Published by St. Regis Workshop, Paul Smiths, New York.
Printed in the United States of America.

ISBN 978-0-578-08705-4

Preface

This book follows the syllabus of the Introductory Statistics course we teach at Plattsburgh State University.

The course is intended to introduce students to some basic statistical ideas: data representation, the construction and interpretation of confidence intervals, hypothesis testing, and simple linear regression. It also includes some basic probability theory, and a brief look at random variables.

The present book is intended to introduce these topics in a clear and economical way, by means of examples and exercises. A student who masters the material in this small volume will have met all the course expectations there are for Introductory Statistics.

Most of the students taking Introductory Statistics at Plattsburgh are doing so in order to satisfy a college-wide quantitative reasoning requirement; many will take this as their one and only college-level mathematics course. The mathematical prerequisites are therefore minimal, and the approach throughout emphasizes method, rather than theory. Examples, worked out in some detail, make up most of the text. By studying these examples and then imitating them in the exercise sets, the student will learn how to approach questions about probability, how to apply a variety of statistical techniques, and how to interpret certain kinds of statistical data. Formulas and methods are presented mostly without proofs or careful derivations. Questions about where the formulas come from and why they work belong in another course with a stronger mathematical prerequisite.

There is little or no real-world data in this book. Most of the numbers in the examples and exercises are made up or come from a computer random-number generator. There are two reasons for staying away from real data. First, real data are dated, and a book that makes

a big deal out of using real data will have to be reissued (with a higher price tag) every year or two. Second, real data are messy. A working statistician will, indeed, have to deal with outliers, tainted data sets, poor sampling techniques, and numerous other ills whose resolution calls for the kind of judgment that can come only from experience. This book is meant for use in an introductory course in basic statistical methods, and the (made-up) data presented here are intended to illustrate those methods clearly, so that they may be learned easily, without the distractions that inevitably tarnish real data. An introductory lesson in, say, driving a standard-transmission car would surely take place on level ground and would not involve parallel parking on a steep hill in a snowstorm. At night.

Most of the names of places, people, businesses, and even drugs and diseases in the examples and exercises are also made up. M&M's are real, of course (and the name is a registered trademark of Mars, Incorporated), as are Canada and the various U.S. states that are mentioned from time to time. Centerville, Statsburgh, and Fiction County are not meant to represent any real places, and the orchards, bakeries, pin factories, and transmission shops that populate them are just as imaginary.

Contents

Chapter 1: Preview and Review

1.1 Terminology

We begin by looking at some of the basic terminology that's used in the study and practice of statistics. The course includes a dozen or so statistical methods, each with its own set of formulas, but they all share some basic vocabulary, which we will introduce by means of a simple example.

> An advertising specialist in the city of Videopolis wants to know the average number hours that Videopolis residents spend watching television over the course of a weekend. The specialist randomly selects the names of 150 residents of Videopolis and sends each one a note asking "How many hours did you spend watching television last weekend?" Miraculously, every one of the 150 residents replies, so the specialist now has a list of 150 numbers. He takes the average of those 150 numbers, and figures that's a pretty good estimate of the average number of weekend television-watching hours for the whole city.

This little story describes a very common use of statistics: making an inference about a large group (all the residents of a city) based on information collected from a smaller group (the 150 people who were surveyed). We use the story to introduce four important terms that we will use again and again. They are *population*, *parameter*, *sample*, and *statistic*.

- The *population* is the entire group under study. In the story above, the population is the set of all residents of Videopolis. The specialist is trying to estimate an average that involves every one of them.
- The *sample* is the subset of the population that's actually surveyed (or measured, or interrogated, or experimented on). In our story, the sample is the set of 150 people who were asked about their television-watching habits.
- A *parameter* is a general term for any number that describes some aspect of the population. In our story, the parameter of interest is the average number of hours that a resident of Videopolis spends watching television over the course of a weekend, where

the average is taken over *all* residents of the city. In many cases, we have no way to determine the exact value of a parameter. In our story, for example, it would be entirely impractical to collect information on the television-watching habits of every resident of Videopolis. We can't find the exact value of this parameter; the best we can do is to estimate it.

- A *statistic* is a general term for any number that is calculated using data gleaned from a sample. In our story, the statistic is the average of the 150 numbers that the specialist collects. Typically, a statistic is used to estimate the value of an unknown parameter.

Example 1.1 Carefully describe the population and the sample in the following situation. What parameter is being estimated? What is the statistic?

> A biologist wants to estimate the incidence rate of a certain parasite among the trout in Dreary Lake. She uses a gill net to capture 25 Dreary Lake trout, and finds that the parasite is present in 10 of those 25. She estimates that the parasite is present in 40% of all the trout in Dreary Lake.

Solution:

The population is the set of all the trout in Dreary Lake.
The sample is the 25 trout that were caught in the gill net.
The parameter is the incidence rate of the parasite among all the trout in Dreary Lake.
The statistic is the percentage of the trout in the sample in which the parasite was present. The value of the statistic in this example is 40%.

■

From a description such as the one in Example 1.1, it is usually easiest to identify the sample and the statistic first, because we are told that some set of objects (the sample) is being interrogated or measured, and some number (the statistic) is being calculated. Once we have a description of the sample, we can usually identify the population by asking (and answering) the question "What larger group does the sample represent?" Similarly, once we know the statistic, we can usually identify the parameter: it's the number that is approximated by the statistic.

This strategy—identifying the sample and the statistic first, and then using them to determine the population and parameter—will be helpful in the following examples.

Example 1.2 Describe the sample, the statistic, and the population in the following situation. What is the parameter that's being estimated?

> A coin enthusiast uses a micrometer to measure the thickness of 60 U.S. dimes. He calculates the average of the thicknesses of the 60 dimes to be 41.55 thousandths of an inch. He estimates that the average thickness of a U.S. dime is 41.55 thousandths of an inch.

Solution:

The sample is the set of 60 dimes that the enthusiast measures.
The statistic is the average of the thicknesses of the dimes in the sample. Its value in this instance is 41.55 thousandths of an inch.
The population is probably meant to be the set of all U.S. dimes.
The parameter of interest is the average thickness of all U.S. dimes.

■

Example 1.3 Describe the sample, the statistic, and the population in the following situation. What is the parameter that's being estimated?

> A teacher acquires ten one-pound bags of M&Ms and counts the number of blue M&Ms in each bag. The average of these counts turns out to be 88.4.

Solution:

The sample is the ten bags of M&Ms.
The statistic is the average number of blue M&Ms in the ten bags in the sample. The value of the statistic in this instance is 88.4.
The population is not made explicit, but it's probably meant to be the set of all one-pound bags of M&Ms, everywhere.
The parameter would then be the average number of blue M&Ms in a one-pound bag, with the average taken over all one-pound bags of M&Ms.

■

Example 1.4 Describe the sample, the statistic, and the population in the following situation. What is the parameter that's being estimated?

> A quality-control inspector selects ten tires from a newly-arrived shipment and measures the tread depth on each of the ten tires. The average of the tread depths on the selected tires is 8.5 mm, so the inspector concludes that the whole shipment is acceptable.

Solution:

The sample is the set of ten tires that the inspector measured.
The statistic is the average tread depth for the tires in the sample. Its value in this instance is 8.5 mm.
The population is the whole shipment of tires.
The parameter is probably the average tread depth of the tires in the shipment.

■

Example 1.5 Describe the sample, the statistic, and the population in the following situation. What is the parameter that's being estimated?

> A physician conducts a study to determine the effectiveness of a certain weight-loss program. He has fifty patients who complete the program; their average weight loss is 8.3 pounds. The physician reports that the program results in an average weight loss of 8.3 pounds, with a margin of error of 2.7 pounds (at the $\alpha = 0.05$ significance level).

Solution:

The sample is clearly the fifty patients who complete the program.
The statistic is the average amount of weight lost by the fifty patients in the program. The value of the statistic in this instance is 8.3 pounds.
The population is rather abstract. The physician seems to be making a claim about what the program *would* do for anyone who might try it. So the population is the set of all *potential* participants in the weight-loss program.
The parameter, even more abstract, is the average amount of weight that would be lost, with the average taken over all the people who might try the program.

■

Example 1.6 Describe the sample, the statistic, and the population in the following situation. What is the parameter that's being estimated?

> An agent suspects that a particular pair of dice may be loaded. She rolls the dice 200 times, and notes that double-sixes come up 12 times out of 200, or 6% of the time.

Solution:

> The sample is the set of 200 rolls that the agent performs.
> The statistic is the percentage (or proportion) of the sample rolls that result in double-sixes. Its value in this instance is 6% (or the ratio 12/200).
> The population (abstract again) is the totality of all rolls that have ever been or will ever be made with this particular pair of dice.
> The parameter is the percentage of all the rolls of this pair of dice that result in double-sixes.

■

Two more terms will show up throughout the course, though not as frequently as the four above. *Survey* and *census* are almost synonymous in casual use. In statistics, however, there is a small but important difference between the meanings of these two words.

- A *survey* is the process of collecting data from the members of a sample. The results of a survey are generally used to find the value of a statistic.
- A *census* is the process of collecting data from every member of a population. In a case where a genuine census can be carried out—that is, where we can collect data from every single member of the group under study—techniques of statistical estimation are not needed. The parameter of interest can be calculated directly. If our population consists of all the students at a particular university, for example, and the datum of interest is something we can easily find in the registrar's records, then we can conduct a census.

It is only in special situations that we can determine the exact value of a parameter using the equivalent of a registrar's records. Most of the time, conducting a census of a large population is either impossible or too expensive to be practical. In order to answer a question about some aspect of a large population, we select a representative sample from the population, conduct a survey of the sample, and use the results of the survey to give an approximate answer to the question.

1.2 Percentages

We'll need to handle two kinds of percentage problems. The first type has the form "What is p percent of N?" where the percentage p and the number N are given. The second has the form "What percentage of N is x?" (This could also be phrased "What percentage is x of N?") Problems of each type can be handled by using a memorized formula, or by setting up a proportion and solving for an unknown variable.

Finding a given percentage of a given number

When the question has the form "What is p percent of N?" we can interpret the word "of" to mean "times" and the word "percent" to mean "divided by 100." Then "p percent of N" means just $\frac{p}{100} \times N$.

Example 1.7 What is 15% of 1200?

Solution: Using the idea above, we rewrite 15% of 1200 as

$$\frac{15}{100} \times 1200 \;=\; 180.$$

Fifteen percent of 1200 is 180.

Finding the percentage when two numbers are given

To find a general formula for dealing with problems of the form "What percentage of N is x?" we use a little algebra. Letting p denote the unknown percentage, we write the question "What percentage of N is x?" as "What value of p satisfies $\frac{p}{100} \times N = x$?" Solving this equation for p, we get the general formula

$$p \;=\; 100 \times \frac{x}{N}. \tag{1.1}$$

Example 1.8 What percentage of 650 is 130?

Solution: To answer this question, we use Formula 1.1 with $N = 650$ and $x = 130$. We get

$$p \;=\; 100 \times \frac{130}{650} \;=\; 20.$$

One hundred thirty is 20 percent of 650.

Using proportions to solve percentage problems

Many students prefer to use proportions to handle all percentage problems. The usual memory device is "part over whole equals part over whole." One of the two part-over-wholes is $\frac{p}{100}$, where p is the percentage. The other is $\frac{x}{N}$, where x is often the size of some subset (such as a sample) of a larger set (such as a population) with size N.

Example 1.9 Use the proportions method to find 15% of 1200.

Solution: We let x be the unknown number (that is, x denotes 15% of 1200), and form the equation

$$\frac{x}{1200} = \frac{15}{100}.$$

To solve for x, we multiply both sides of the equation by 1200, getting

$$\cancel{1200} \times \frac{x}{\cancel{1200}} = 1200 \times \frac{15}{100}.$$

The two 1200s on the left side of the equation cancel, and we get

$$x = 1200 \times \frac{15}{100} = 180.$$

Fifteen percent of 1200 is 180.

■

Example 1.10 Use the proportions method to answer the question "What percentage of 650 is 130?"

Solution: The unknown in this example is the percentage p, so one of our part-over-wholes will be $\frac{p}{100}$. For the other part-over-whole, we note that the size of the whole set is 650, and the part that we're interested in has size 130, so we have $x = 130$ and $N = 650$. We get the equation

$$\frac{130}{650} = \frac{p}{100}.$$

In this case, we solve for p. We'll interchange the two sides of the equation and then multiply both sides by 100. We get

$$\cancel{100} \times \frac{p}{\cancel{100}} = 100 \times \frac{130}{650}$$

The 100s cancel, and we get

$$p \;=\; 100 \times \frac{130}{650} \;=\; 20.$$

One hundred thirty is 20 percent of 650.

■

Here are some more examples of both kinds of percentage problems.

Example 1.11 What is 52% of 829?

Solution 1: By the first method, we get that 52% of 829 is $\frac{52}{100} \times 829$, which is 431.08.

Solution 2: To use proportions, we set up the equation $\frac{x}{829} = \frac{52}{100}$. Solving this for x, we get

$$x \;=\; 829 \times \frac{52}{100} \;=\; 431.08.$$

The two methods, of course, give the same answer.

■

Example 1.12 In a survey of 655 people, 81.5% said they opposed needless violence. How many people in the sample said they opposed needless violence?

Solution: Using either the percentage formula or the proportions, we find that 81.5% of 655 is

$$81.5 \times \frac{655}{100} \;=\; 533.825.$$

It doesn't make sense, however, to say that 533.825 people were opposed to needless violence. Because of the way the question is posed, the answer must be a whole number. We round to the nearest whole number of people. Of the 655 people who were surveyed, 534 said they are opposed to needless violence.

■

Example 1.13 The Spindle College Engineering Club has 25 members, of whom six will be graduating this year. What percentage of the membership of the Spindle College Engineering Club will be graduating this year?