Chapter 2: Estimating Population Parameters

Vocabulary

Population: The collection of all people or objects to be studied.

Census: Collecting data from everyone in a population.

Sample: Collecting data from a small subgroup of the population.

Statistic: A number calculated from sample data in order to understand the characteristics of the data. For example, a sample mean average, a sample standard deviation, or a sample percentage.

Parameter: A number that describes the characteristics of a population like a population mean or a population percentage. Can be calculated from an unbiased census, but is often just a guess about the population.

Sampling Distribution: Take many random samples from a population, calculate a sample statistic like a mean or percent from each sample and graph all of the sample statistics on the same graph. The center of the sampling distribution is a good estimate of the population parameter.

Sampling Variability: Random samples values and sample statistics are usually different from each other and usually different from the population parameter.

Point Estimate: When someone takes a sample statistic and then claims that it is the population parameter.

Margin of Error: Total distance that a sample statistic might be from the population parameter. For normal sampling distributions and a 95% confidence interval, the margin of error is approximately twice as large as the standard error.

Standard Error: The standard deviation of a sampling distribution. The distance that typical sample statistics are from the center of the sampling distribution. Since the center of the sampling distributions is usually close to the population parameter, the standard error tells us how far typical sample statistics are from the population parameter.

Confidence Interval: Two numbers that we think a population parameter is in between. Can be calculated by either a bootstrap distribution or by adding and subtracting the sample statistic and the margin of error.

95% Confident: 95% of confidence intervals contain the population value and 5% of confidence intervals do not contain the population value.

90% Confident: 90% of confidence intervals contain the population value and 10% of confidence intervals do not contain the population value.

99% Confident: 99% of confidence intervals contain the population value and 1% of confidence intervals do not contain the population value.

Bootstrapping: Taking many random samples values from one original real random sample with replacement.

Bootstrap Sample: A simulated sample created by taking many random samples values from one original real random sample with replacement.

Bootstrap Statistic: A statistic calculated from a bootstrap sample.

Bootstrap Distribution: Putting many bootstrap statistics on the same graph in order to simulate the sampling variability in a population, calculate standard error, and create a confidence interval. The center of the bootstrap distribution is the original real sample statistic.

Introduction: The goal of learning Statistics or Data Science is to be able to analyze data to learn about populations in the world around us. The best way to understand a population is collect and analyze unbiased data from that population, namely a census. The trouble is we rarely have an unbiased census. It is sometimes impossible to collect data from everyone in a population. We have to rely on samples, small subgroups of the population. The next few chapters deal with the subject of using samples to understand populations. This is sometimes called "inferential statistics". We will start by trying to distinguishing between population parameters from sample statistics.

Section 2A – Statistics and Parameters

Vocabulary

Population: The collection of all people or objects to be studied.

Census: Collecting data from everyone in a population.

Sample: Collecting data from a small subgroup of the population.

Bias: When data does not represent the population.

The goal of collecting and analyzing data is to understand the world around us. To this end, our goal is understand populations. The population is all of the people or objects you plan to study. A population can be large (like all people living in Brazil) or small (like all students in a particular statistics class). It goes without saying that the larger the population the more difficult it is to understand.

The best data for representing populations is an unbiased census. A census is an attempt to collect data from everyone in a population. A census is easier if we have a small population like the people in a particular statistics class. The advantage of collecting an unbiased census is that we can calculate population values (parameters) directly with reasonable certainty. Governments may sometimes attempt to do a census and collect data on all of the people living in a particular country. It should be noted that though they attempt to get data on everyone, they rarely succeed. There will always be some people fall through the cracks and are not represented in the census. An unbiased census of a large population still represents a high percentage of the people, so is generally better than a small sample of people.

A data scientist rarely has the ability to collect a census unless the population is relatively small. People that work in statistics and data science usually rely on collecting samples. Remember a sample is a small subgroup of the population. It is usually less than 10% of the population and is often significantly less than 10%. If the sample is unbiased, we then try to analyze the sample data and make guesses as to what is happening at the population level. Therefore, a data scientist or statistician needs to be able to use sample values (statistics) to figure out approximate population value (parameters).

Statistic: A number calculated from sample data in order to understand the characteristics of the data.

Parameter: A population value. It can be calculated from an unbiased census, but is often just a guess about what someone thinks the population value might be.

It is very important to note that statistics and parameters are not the same thing. A statistic calculated from 250 people in a sample will often be very different from the actual population parameter from millions of people. The question that is important to ask is how far off is the sample statistic from the population parameter? That is sometimes called "margin of error" and is a key topic in this chapter.

Common Statistics

$\bar{x}$: ("x-bar") Sample mean average

s : Sample standard deviation (typical distance from the sample mean)

s^2: Sample variance (sample standard deviation squared)

$\hat{p}$: ("p-hat") Sample proportion (sample percentage)

n : Sample size or frequency (number of people or objects in the sample)

r : Sample correlation coefficient (measures quantitative relationships between samples)

b_1 : Sample slope (The slope of a regression line calculated from sample data.)

b_0 : Sample Y-intercept (The Y-intercept of a regression line calculated from sample data.)

Common Parameters

μ : ("mu") Population mean average

σ : ("sigma") Population standard deviation (typical distance from the population mean)

σ^2: Population variance (population standard deviation squared)

π: ("pi") Population proportion (population percentage) *(Some people use "p" for population proportion.)*

N : Population size or frequency (number of people or objects in the population)

ρ : ("rho") Population correlation coefficient (measures quantitative relationships between populations. Note this is not a "p". It is the Greek letter "rho".)

β_1 : Population slope (The slope of the population regression line. Used when studying quantitative relationships between populations.)

β_0 : Population Y-intercept (The Y-intercept of the population regression line. Used when studying quantitative relationships between populations.)

Let us look at some examples of using statistics and parameters. It is important to be able to identify if a number used is a statistic or a parameter and what letter we might use in the computer program.

Example

"We think the mean average ACT score for all high school students is about 22. The mean average ACT score for a random sample of 85 high school students was 21.493"

μ = 22 (parameter)
n = 85 (statistic)
$\bar{x}$ = 21.493 (statistic)

Example

"A random sample showed that 13.2% of adults were infected, but this indicates that the population percentage could be 17%". *(Note: Computer programs often require you to convert the percentages into decimal proportions.)*

$\hat{p}$ = 0.132 (statistic)
π = 0.17 (parameter)

Example

The standard deviation for the heights of all women is thought to be about 2.5 inches. A random sample of women heights had a standard deviation of 2.618 inches.

σ = 2.5 (parameter)
s = 2.618 (statistic)

Example

"Sample data indicated that the correlation coefficient was 0.239 and the slope was 47.3 dollars per pound. Let's compare these to the population claims that the correlation coefficient is zero and the slope is about 50 dollars per pound."

r = 0.239 (statistic)
b_1 = 47.3 (statistic)
ρ = 0 (parameter)
β_1 = 50 (parameter)
