Chapter 3 (Part 1): What do Samples Tell Us?

Overview: In Chapter 2, we talked about how to select which individuals will be in a sample. In this section, we will discuss the size of the sample and the variability of your results depending on the sample size and the population that the sample comes from. We will discuss the difference between a parameter and a statistic as well as the statistical meaning of the word confidence.

Motivating Example: Suppose a chef makes a large pot of soup. He wants to check if the soup is over-salted so he takes one spoonful of the soup and tastes it. Suppose an editor wants to double check a 600-page novel for typos so they spot check 35 pages. Suppose a psychologist wants to determine the affects of an anti-anxiety drug on patients, she analyzes the results of 500 patient files. In these examples there is a sample of size 1, a sample of size 35, and a sample of size 500. How should you decide how large a sample size needs to be?

We begin with some definitions:

Definition of a Parameter and a Statistic: Suppose we have a population of individuals

- A **parameter** (for population)

  is a number that describes the population
  e.g. population average, population variance

- A **statistic** is any formula (or role) that uses values from th sample

  A statistic is used to estimate the population parameter
  e.g.
  the sample size is 3, suppose x1, x2, and x3
  (x1+x2-5)/9-->This is the statistics

  The number or value resulted from the statistic is called observed value for example: suppose 11, 13, 9
  (11+13-5)/9=2.11-->This is the observed value

Practice Question: Suppose we are interested in the average height of all current NHL players and we currently have access to players on the Vancouver Canucks.

1. The population is:

   (A) Current NHL players (B) Current Canucks players (C) All hockey players

2. The sample is:

   (A) Current NHL players (B) Current Canucks players (C) All hockey players

3. What is the parameter we are interested in?

   (A) The average height of all current NHL players.
   (B) The average height of current Canuck's players.

4. The average height of current Canuck's players is 6ft 1in. What does this value represent?

   Select all that apply.

   (A) The population parameter.
   (B) The statistic.
   (C) The observed value of a statistic.

Examples of desired Parameters and the Statistic used to estimate them:

1. Parameter:

population mean

Denoted by: $\mu(mu)$

$$\mu = \frac{\Sigma X}{N}$$

Statistic:

sample mean

Denoted by: $\bar{x}(xbar)$

$$\bar{x} = \frac{x1 + x2 + ... + xn}{n}$$ -> for sample size n

2. Parameter:

population proportion

Denoted by: $p(smallp)$ (pi in professional books)

$$p = \frac{X}{N}$$

No. of individuals in the population with the character of interest / No. of individuals in the population

Statistic:

sample propotion

Denoted by: $\hat{p}(phat)$

$$\hat{p} = \frac{x}{n}$$

x individuals of the sample size n with the character of interest / No. of individuals in the sample

3. Parameter:

population median

Denoted by: $\tilde{\mu}$

Statistic:

sample median

Denoted by: $\tilde{x}$

Order the numbers and then take the "middle " number

4. Parameter:

population varience

Denoted by: $\sigma^2(sigma\ square)$

$$\sigma^2 = \frac{1}{N}\Sigma(X - \mu)^2$$

Statistic:

sample variance

Denoted by: $s^2(ssquare)$

$$S^2 = \frac{1}{n-1}\Sigma_{i=1}^{n}(xi - \bar{x})^2$$

Note: A statistic doesn't necessarily do a good job estimating the population parameter. For example, you could use a statistic that just always equals 6. Unless you are trying to estimate a parameter that has the value 6, this would be a terrible statistic.

We usually use statistics that involves computing the same quantity that is desired in the population just on the available sample instead.

Sampling Bias vs Statistic Bias:

Bias is a consistent, repeated deviation of the sample statistic from the population parameter in the same direction when we take many samples.

Bias is in the sample if

Bias is in the statistic if

Note: We will not be exploring bias in a statistic in this course but if you take other stats courses this is a very important topic.

Definition: Another thing which can greatly impact a sample is **variability**.

Variability describes

When we are sampling, we would like there to be little or no sampling bias and small variability. A good analogy is that of a dart board. You can think of the population parameter as the bullseye and the darts being thrown as the statistic estimating the parameter. Sampling bias results in the dart being consistently thrown to the same part of the board, away from the bullseye. Variability results in darts being thrown all over the board, not very close together.

Question: We already saw that if our sample is not random, it will be biased. Thus, to reduce bias in a study, we need to select our sample randomly. If the population we are sampling from does not have small variability, what can we do to make sure our sample accurately reflects the population?

Answer: Increase the sample size. The larger the sample, the lower something called the **variance of the sampling distribution** is (we will define this later on). Basically, a large sample can better capture the population.

Note: Samples don't always need to be large. If the population does not have large variability then the sample can be small.

Consider the 3 scenarios discussed in the motivating example:

(a) A chef takes one spoonful of soup as a sample for the entire pot. Why is this an appropriate sample size?

(b) An editor spot-checks 35 pages of a 600 page novel for typos. Why is this an appropriate sample size?

(c) A psychologist analyzes 500 patient files in order to determine the affects of an anti-anxiety drug on patients. Why is her sample so large?