Introduction to Statistics I

Textbook: Elementary Statistics (4th Edition, by Navidi and Monk), and Workshop Statistics (4th Edition, by Rossman and Chance).

What's the most mysterious or interesting part of our universe?
What if you wanted to learn something new about it, what would you do?

Statistics is the study of the procedures for

- collecting information (sampling),
- describing information, and
- drawing conclusions from information.

§1.1: Sampling

We get this info by **sampling** from a population.
Each person, place, or thing in our sample is called an **observational unit** (obv-unit).

The info (e.g., height) we gather from each obv-unit is called a **variable** (var).
Why? Because they **vary** from obv-unit to obv-unit.

Sample vs. Population

A **population** is the whole collection of obv-units we're interested in.

A **sample** is a subgroup of obv-units from whom data is collected.

A **sample size** (denoted n) is the number of obv-units in the sample.

Example: If we're interested in the height of 20 yr olds, we might ask 300 of them their height. ...

Obv-units: 20 yr olds.
Population: **ALL** 20 yr olds.
Sample: the 300 we collected info from
Sample Size: $n = 300$
Variable: Height

Statistic vs. Parameter

A **statistic** is a number that *summarizes* our sample data (e.g., the sample average).

A **parameter** is a number that summarizes the *entire population* (e.g., the population average).

From our Example: A statistic is the average height from our sample. A parameter is the average height of *ALL* 20 yr olds.

Research Question (**RQ**): What's the average length of a word in the Gettysburg Address?

Population/Parameter?

Poll:

bit.ly/introstatsdata

Poll: Gettysburg Address 1

Population: all words in Address. **Parameter**: average length of *ALL* words in Address.

Activity: 1.1a

Obv-unit: words in Address. **Variable**: word length.

Sampling

Let's limit our sample to ten words from the Address. What's the statistic?

Statistic: average length of words in our ten word sample from Address.

Four score and seven years ago our fathers brought forth on this continent a new nation, conceived in Liberty, and dedicated to the proposition that all men are created equal. Now we are engaged in a great civil war, testing whether that nation or any nation so conceived and so dedicated, can long endure. We are met on a great battle-field of that war.

We have come to dedicate a portion of that field, as a final resting place for those who here gave their lives that that nation might live. It is altogether fitting and proper that we should do this. But, in a larger sense, we can not dedicate—we can not consecrate—we can not hallow—this ground. The brave men, living and dead, who struggled here, have consecrated it, far above our poor power to add or detract.

The world will little note, nor long remember what we say here, but it can never forget what they did here. It is for us the living, rather, to be dedicated here to the unfinished work which they who fought here have thus far so nobly advanced. It is rather for us to be here dedicated to the great task remaining before us—that from these honored dead we take increased devotion to that cause for which they gave the last full measure of devotion—that we here highly resolve that these dead shall not have died in vain—that this nation, under God, shall have a new birth of freedom—and that government of the people, by the people, for the people, shall not perish from the earth.

Each student: select 10 words from above address that appear to be of "typical" length.
Then, report average length of your ten words.

bit.ly/introstatsdata

Poll: Gettysburg Address 2

Our sampling method is **biased**. On average, we ended up with statistics that were consistently high (actual average is 4.29). We **overestimated** the parameter.

How might we sample better?

What if we closed our eyes and picked 10 words that way?

Alternative Sampling Method

Assign each word in the Address a number.

Then, use a computer to sample ten random numbers. Use those ten words.
(check "Show Sampling Options", then "Draw Samples")

bit.ly/introstatsdata

Applets: Sampling Words

Simple Random Sampling (**SRS**): sampling methodology in which all members of a population have an *equal chance of being selected* for a sample. SRS is an *unbiased sampling method*.

What leads to biased samples?

- Sampling from only a (nonrepresentative) part of the population

- Convenience Samples (lazy sampling)

- Voluntary Samples (self-selection)

- Allowing non-response

Most Straitforward SRS: Accomplished by numbering ALL the members of the population using a **sampling frame** (list of all the obv-units, often not available).

A sample of convenience may be acceptable if it's reasonable to believe that there's **no systematic difference** between the sample and the population.

Note - the Gettysburg activity illustrated the concept of **sampling variability**: that statistics vary from sample-to-sample.

Other sampling methods described in the book: **Stratified**, **Cluster**, **Systematic**.

What did we learn?

- Basic idea of stats
- Simple random sampling (SRS),
- Sampling variability
- We want unbiased/consistent sampling method