Sampling Principles (part 1)

Ryan Miller

Introduction

- Understanding how data are organized, summarized, and displayed is important, but arguably most important is how they are collected

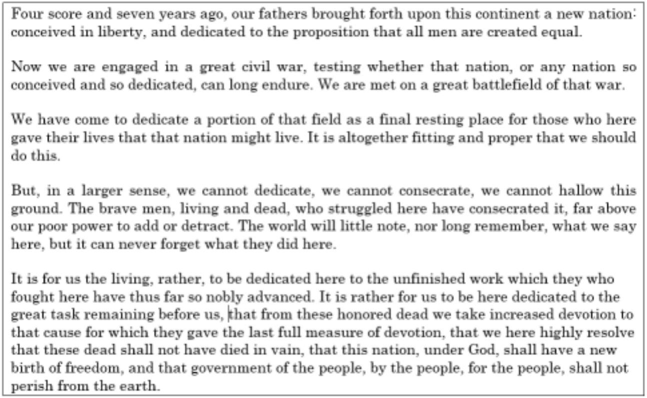
Introduction

- Understanding how data are organized, summarized, and displayed is important, but arguably most important is how they are collected
- All research questions pertain to some target **population**, or comprehensive group of cases
    - In most circumstances it is impossible to collect data on the entire population, instead we must rely on a **sample** or subset of cases
    - A sample must be **representative** of the population in order to produce *reliable conclusions*

An Example

- Earlier this semester I showed by MATH-256 students the text of the Gettysburg Address (target population), asking each of them to come up with a representative sample of 5 words
    - I then had them summarize their sample using the sample mean (ie: I asked them to report their average word length)
    - If there samples were truly representative, we'd expect them to have sample means that cluster around the population mean

The Gettysburg Address



Four score and seven years ago, our fathers brought forth upon 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 battlefield of that war.

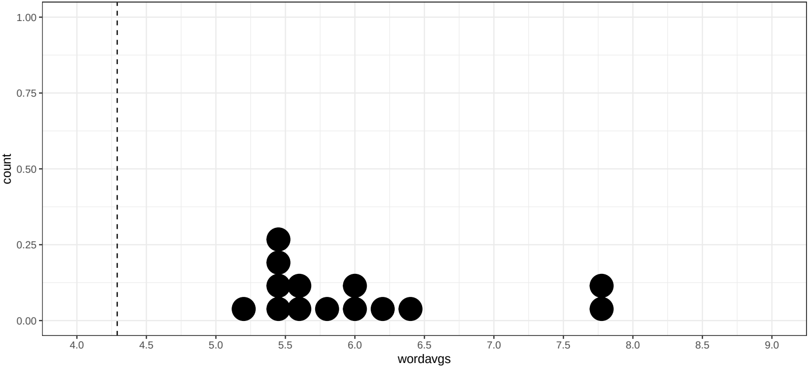
We have come to dedicate a portion of that field as a final resting place for those who here gave their lives that that nation might live. It is altogether fitting and proper that we should do this.

But, in a larger sense, we cannot dedicate, we cannot consecrate, we cannot 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.

Results

- The population's mean word length, in *statistical notation*, is $\mu = 4.295$
  - Why were the MATH-256 students' samples so far off?



Sampling Bias

- The answer boils down to **sampling bias**, which is a *systematic tendency* for some cases from the population to be more likely to make it into a sample than others
  - In the Gettysburg Address, students tended to pick longer words
  - For example, the longer words stood out more prominently, or the students felt awkward selecting multiple two-letter words in their sample

Simple Random Sampling

- The *ideal sampling procedure* is **simple random sampling**, a protocol where each case in the target population has an identical chance of ending up in the sample
- Do you think it would be easy or hard to collect a simple random sample of Xavier students?

Simple Random Sampling

- The *ideal sampling procedure* is **simple random sampling**, a protocol where each case in the target population has an identical chance of ending up in the sample
- Do you think it would be easy or hard to collect a simple random sample of Xavier students?
  - It would actually be quite hard since the University is unlikely to give you a list of all enrolled students
  - Often times we attempt to get representative samples in other ways

Other Sampling Protocols

- A **convenience sample** is exactly what the name suggests, a sample that is easily collected (ie: low monetary or time costs)
  - Convenience samples are not random, but they can be representative if carefully selected
  - You might be able to get a representative sample by standing near the center of campus on a typical day and stopping people who walked by

Other Sampling Protocols

- A **convenience sample** is exactly what the name suggests, a sample that is easily collected (ie: low monetary or time costs)
  - Convenience samples are not random, but they can be representative if carefully selected
  - You might be able to get a representative sample by standing near the center of campus on a typical day and stopping people who walked by
- A **stratified sample** is a more complex scheme where the population is broken into similar subcategories, which are sampled separately (typically simple random sampling)
  - The analysis methods we cover will need extensions in order to apply to this type of data
  - Nevertheless, we should be able to recognize these sampling schemes for precisely that reason

Practice

For each scenario, determine whether it describes a *population* or a *sample*, as well as whether or not the sample is biased.

1. To estimate the size of trout in a lake, an angler records the weight of the 12 trout he catches over a weekend
2. A subscription based music website tracks the listening history of its active users
3. The Department of Transportation announces that of the 250 million registered cars in the US, 2.1% are hybrids
4. A car rental company installs an experimental data collection device on the first 20 vehicles from an alphabetized list of license plates

Practice (solution)

1. This is a sample and it's biased, there is no way that the angler has an equal chance of catching every trout in the lake
2. This is a population because it includes all of the website's users
3. This is a population because it's safe to assume the DOT has registration on the overwhelming majority of cars in the US
4. This is a sample and it's unbiased, there is no reason to believe that license plate numbers are related to anything meaningful about each vehicle