Sampling from a Population

Ryan Miller

The Candy Activity

Today I've brought with me a bag containing 100 pieces of candy, it is your job to correctly determine the weight of the bag

With your group, you will:

1. Sample 5 candy pieces from the bag
2. Weigh your sample
3. Multiply your sample's weight by 20 to estimate the entire bag's weight
4. Return your sample to the bag

The group whose estimate is closest to the bag's weight will be given the entire bag to consume or distribute as they see fit

How Accurate is an Estimate?

Today we will discuss sampling from a population. A lot of effort in statistics is devoted to analyzing data, but how the data are collected is *absolutely critical* - often *much more important* than the analysis techniques used

Today we will discuss the following concepts:

- **Populations** versus **samples**
- **Statistical inference**
- **Simple random samples**
- **Bias** and **Variability**
- Sources of bias

Populations vs. Samples

Every statistical analysis begins with a question - ie: How much does the bag of candy weigh?

- The best approach is to weigh the entire bag
- But what if your access to the bag is limited?
- In our example, the 100 pieces of candy in the bag represent a **population** - *all of the cases* we want to learn about
- I didn't allow you access to the entire population, but rather a **sample** - *a subset of cases* from the population

We denote the size of a sample using n, ie: $n = 5$

Practice

In a study on hand washing, researchers in several cities across the United States pretended to comb their hair in public restrooms while observing whether or not people washed their hands after going to the bathroom. They found that 85% of the 6,000 individuals they observed washed their hands.

What is the population? What is the sample?

- We could say the population is all people in the US that use public restrooms
- But people are likely to behave differently when someone else is in the restroom with them
- It would be wise to restrict the population to people in the US using a restroom *with another occupant*

Statistical Inference

- A *fundamental goal* of statistics is to use information from a sample to make *reliable* statements about a population
- This idea is called **statistical inference**

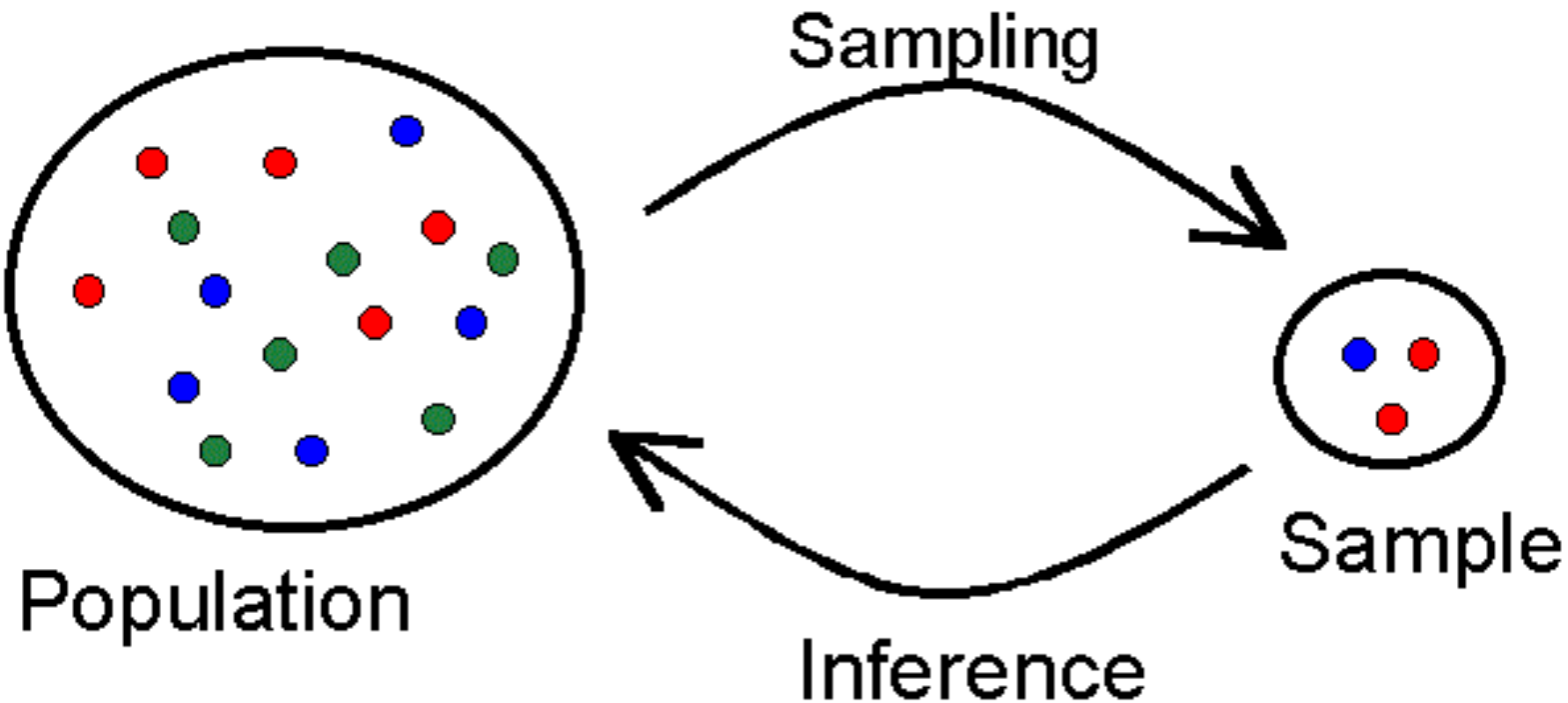


Image credit: http://testofhypothesis.blogspot.com/2014/09/the-sample.html

Statistical Inference - Notation

Statisticians use different notation to distinguish *population parameters* (things we want to know) from *estimates* (things derived from a sample). For a few common measures, this notation is summarized below:

Statistic	Population Parameter	Estimate (from sample)
Mean	μ	$\bar{x}$
Standard Deviation	σ	s
Proportion	p	$\hat{p}$
Correlation	ρ	r

For example, μ is the mean of the population, while $\bar{x}$ is the mean of the cases that ended up in the sample.

Now Let's Weigh the Bag

I didn't know what your estimates would be when I prepared these slides... but I predict that **all** of them are way **too high**!

Simple Random Samples

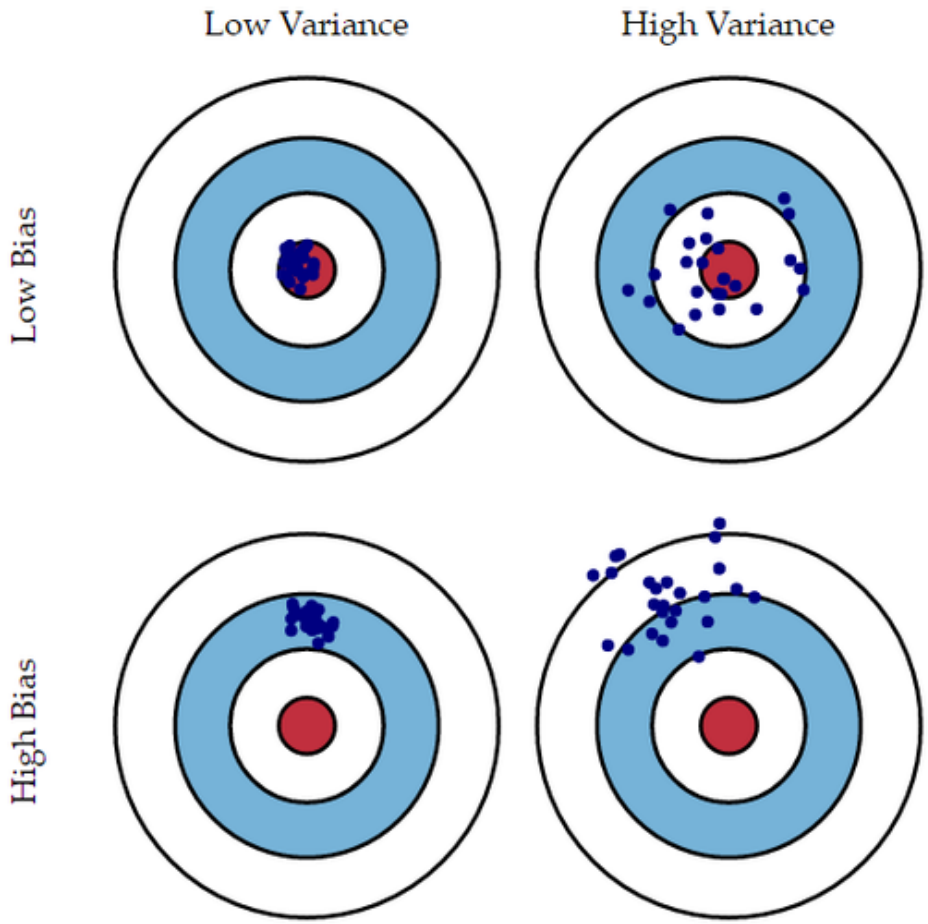
- It is *essential* for a sample to be **representative** of the population in order for statistical inference to be valid
- The simplest way to get representative samples is to use **random sampling**
  - Random is NOT the same as haphazard! Was your sample of candy random? How could you make it random?
- It has been well-established that humans are surprisingly bad at randomness
  - Algorithms can easily differentiate between human generated "random" sequences and actual random sequences

Randomness and Variability

- Any given sample, regardless of how it was collected, only contains a subset of cases from the population
- This introduces **variability** when trying to use the sample to estimate a population parameter
  - Just by random chance, some samples will yield more accurate estimates than other samples, even if an ideal sampling protocol is used
  - Next week we'll approach the goal of trying to understand this variability, today we'll continue learning about sampling

Bias and Variability

To summarize, there are *two reasons* why an estimate might not accurately represent a population parameter, **bias** and **variability**:



Variance decreases with larger sample sizes

Bias is not improved by a larger sample

Case Study - The 1936 President Election

- In 1936, Franklin Roosevelt was up for re-election versus Republican candidate Alfred Landon
- The country was in the midst of the Great Depression, with nearly 20% of the country unemployed and real income at roughly two-thirds of what it was in 1929 before the depression
- Roosevelt and Landon had very different views regarding the role of government in bringing the United States out of the depression

Case Study - The 1936 President Election

- Since 1916, the *Literary Digest* magazine had correctly predicted the winner of 5 straight presidential elections
- Prior to the 1936 election, the *Literary Digest* sampled *2.4 million* people and predicted a landslide victory for Landon: 57% - 43%
- In the actual election, Roosevelt won by a landslide: 62% - 38%

How could the Digest have been so far off?

- Take a minute to discuss this with your group
- Consider whether the inaccurate estimate could be due to **bias** or **variability**

Case Study - The 1936 President Election

Selection Bias

- The *Literary Digest* sent 10 million questionnaires to addresses gathered from telephone books and club memberships
- This disproportionately screened out the poor; Only 1 in 4 households owned a telephone at the time, and club members tended to be upper class
- Selection bias resulted in a non-representative sample

Non-response Bias

- Of the 10 million questionnaires, only 2.4 million were returned
- Responders tend to be different from non-responders
- The 2.4 million respondents likely weren't even representative of the 10 million people polled

That was 1936, surely today we understand the importance of representative samples . . . right?

Case Study - CTE and Football

Chronic traumatic encephalopathy (CTE) is degenerative brain disease found in individuals with a history of repetitive brain trauma. In July 2017 a paper published in JAMA by researchers at Boston University generated a lot of media buzz:

- *"CTE in 99% of former NFL player's brains in new study"* - Sports Illustrated
- *"111 NFL Brains. All But One Had CTE"* - NY Times
- *"CTE found in 99% of studied brains from deceased NFL players"* - CNN
- *"99% of Deceased NFL Players in One Study Had CTE"* - Forbes
- *"Brain disease affects 99% of NFL players in study"* - BBC News

Discussion

With your group take a look at the NY Times article: https://www.nytimes.com/interactive/2017/07/25/sports/football/nfl-cte.html

1. What is the actual population of the study?
2. What is the population most people jump to conclusions about when they see the headlines on the previous slide?

Case Study - CTE and Football

Article Link: "I'm a brain scientist and I let my son play football"

The study population in the most recent CTE paper represents a biased sample, as stated by the authors themselves. This means only the brains of self-selecting people who displayed neurological symptoms while living were studied. This is important because this sample was not a reflection of the general football population. The study was based on 202 brains out of the millions of people who have played football all of which are former NFL players.

So, when you hear 99 percent of football players had CTE, that doesn't mean that almost every football player will get CTE, and it doesn't mean your child has a 99-percent chance of developing CTE if he or she plays football. It means 99 percent of a specifically selected study sample had some degree of CTE; not 99 percent of the general football population. This is an important distinction.

Examples of Sampling Bias - CTE and Football (cont.)

Because of this sampling bias, we cannot estimate the prevalence or incidence of CTE (meaning the total number of cases and the number of new cases expected each year in football players); nor can we establish risk or a cause-effect relationship between head injury and development of CTE. To do that you need a randomly selected population comprised of people with the disease and people without the disease.

Some Other Sources of Bias

When collecting data it is *crucial* to be aware of potential sources of bias, some examples include:

1. Social Desirability Bias - Respondents tend to answer questions in ways that portray themselves in a positive light Link
2. Habituation Bias - Respondents tend to provide similar answers for similarly worded or structured questions (the brain going on autopilot) Link
3. Leading Questions - The wording of a question impacts how people respond, great examples in the textbook
4. Cultural Bias - Questions are often to be constructed with one's own culture in mind, they might not even make sense to people from other cultures.

This isn't a complete list, there are countless reasons for data not being representative of the population of interest

Practice

With your group, discuss whether each of the following are a **sample** or a **population**. If the data are a sample, describe the target population and whether 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. An online poll seeking to learn about adult workers asks: "What do you think of having an everyday uniform for work, like what Steve Jobs did?" 24% of people said they loved the idea

Practice - Solutions

1. This is a sample, the population is all trout in the lake. It is a biased sample because the angler isn't randomly catching fish, he is likely fishing in a single spot and is more likely to catch certain sizes of trout
2. This is a population, the website has data on all of its active users.
3. This is a population, the DoT has information on all registered cars
4. This is a sample, the population is all adult workers. It is a biased sample because the of the social desirability typically associated with Steve Jobs.

Conclusion

Right now you should:

1. Be able to the describe the population represented by a sample
2. Understand the importance of random sampling
3. Be able to recognize sources of bias
4. Know that the phrase *statistical inference* refers to the process of using a sample to learn about a population

If you want more information:

- Read Ch 1.2