Sampling from a Population

Ryan Miller

Grinnell College
Statistics

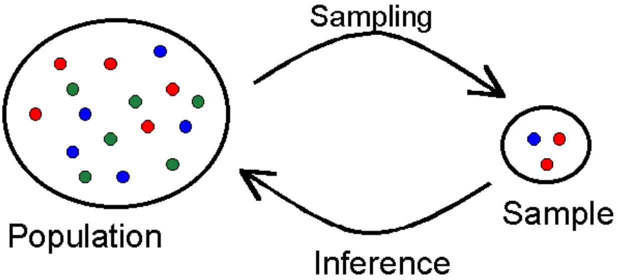
Introduction

Suppose a biologist wants to learn about the species of fish that reside within a particular lake

1) Do they need to capture and study *every* fish in this lake in order to achieve their goal?
2) What trade-offs are involved in collecting data on only *some* of the fish rather than *all* of them?

Sampling from a population

The data we collect is typically a **sample**, or a subset of cases, from a broader **population**, the collection of *all* cases we might be interested in:



We'll denote the number of cases in our sample as n and the size of the population as N (which is often unknown)

Sampling from a population

- **Inference** addresses the statistical question: “how reliably will trends in a sample reflect what is true of the population?”

Sampling from a population

- **Inference** addresses the statistical question: "how reliably will trends in a sample reflect what is true of the population?"
    - For example, if two variables, X and Y, have a correlation of $r = 0.71$ *in the sample data*, how do you think these variables are related *in the population*?

Sampling from a population

- **Inference** addresses the statistical question: "how reliably will trends in a sample reflect what is true of the population?"
    - For example, if two variables, X and Y, have a correlation of $r = 0.71$ *in the sample data*, how do you think these variables are related *in the population*?
- As a starting point, we might use the sample correlation as an **point estimate** of the correlation in the population
    - If the sample data are **representative**, this estimate should be *close* to the population-level correlation

Notation for estimates and population parameters

Statisticians use notation to distinguish *population parameters* (things we want to know) from *estimates* (things derived from a sample):

	Population Parameter	Estimate (from sample)
Mean	μ	$\overline{x}$
Standard Deviation	σ	s
Proportion	p	$\hat{p}$
Correlation	ρ	r
Regression	b_0, b_1	$\hat{b}_0, \hat{b}_1$

Sampling and estimation

- The Gettysburg Address is a famous speech made by Abraham Lincoln during the American Civil War
  - The speech contains 268 words, which we'll consider to be the members of our target population
- Your aim is to estimate the *average word length* of the speech
- Using the text of the Gettysburg Address on the next slide:
  1. Select a sample of 5 words
  2. Count the number of letters in each word
  3. Find the average number of letters in your sample
     - If your word lengths are 1, 2, 3, 4, and 5 you can use the `R` command `mean(c(1,2,3,4,5))`

Sampling from 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.

Sources of sampling error

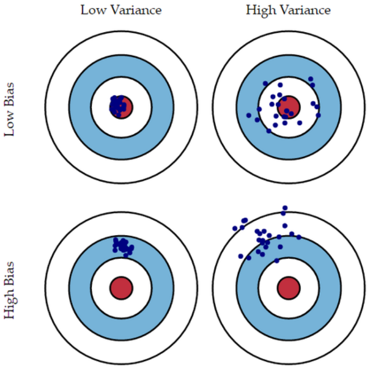
There are 2 main reasons for a sample estimate to differ from what's true of the population:

1) **Sampling Bias** - a systematic flaw in the way cases were selected that leads to certain types of cases being disproportionately represented in the sample data
2) **Sampling Variability** - since a sample doesn't include all of the population, any individual sample might differ from the population due to *random chance* (ie: "the luck of the draw")

Having a larger sized sample will *reduce sampling variability* but *will not fix sampling bias*.

Sampling error

Consider 4 different sampling procedures:



Here each “dot” represent an estimate from a *different sample*.

Practice

- In 1936, Franklin Roosevelt was up for re-election versus Republican candidate Alfred Landon
- The *Literary Digest* sampled 2.4 million voters and predicted a landslide victory for Landon: 57% to 43%
    - The *Literary Digest*'s poll had correctly predicted all 5 elections since it began in 1916

Practice

- In 1936, Franklin Roosevelt was up for re-election versus Republican candidate Alfred Landon
- The *Literary Digest* sampled 2.4 million voters and predicted a landslide victory for Landon: 57% to 43%
  - The *Literary Digest*'s poll had correctly predicted all 5 elections since it began in 1916
  - However, Roosevelt won the actual election by a landslide: 62% to 38%

1) What is the *population* and what is the *sample*?
2) What is the *population parameter* and what is the *sample estimate*?
3) Was the *Digest*'s inaccurate estimate likely due to *sampling bias* or *sampling variability* (or both)?

Practice (solution)

1) The population is all of the people who voted in the 1936 election. The sample is the 2.4 million voters contacted by the *Literary Digest*.
2) The population parameter is the proportion who voted for Roosevelt (or Landon since either proportion would tell you the other). The sample estimate would then be 43% (the proportion of those sampled by the *Digest* who said they'd vote for Roosevelt)
3) Sampling bias - the sample size was enormous (making sampling variability a non-issue). However, the sample was biased by disproportionately including wealthier individuals.

Types of bias in the *Literary Digest* example

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

Types of bias in the *Literary Digest* example

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
- Respondents tend to be different from non-respondents
- The 2.4 million respondents likely weren't even representative of the 10 million people polled

Three common sampling approaches

- **Convenience sampling** - select all cases from the target population that are easily accessible
  - Pros: data is easy to collect
  - Cons: high potential for sampling bias (though not guaranteed)

Three common sampling approaches

- **Convenience sampling** - select all cases from the target population that are easily accessible
  - Pros: data is easy to collect
  - Cons: high potential for sampling bias (though not guaranteed)
- **Simple random sampling** - randomly select cases from the target population
  - Pros: eliminates sampling bias, all sampling error is due to sampling variability
  - Cons: can be difficult to execute

Three common sampling approaches

- **Convenience sampling** - select all cases from the target population that are easily accessible
  - Pros: data is easy to collect
  - Cons: high potential for sampling bias (though not guaranteed)
- **Simple random sampling** - randomly select cases from the target population
  - Pros: eliminates sampling bias, all sampling error is due to sampling variability
  - Cons: can be difficult to execute
- **Stratified or cluster random sampling** - randomly select cases separately from different population segments, potentially using different strategies for each segment
  - Pros: low potential for sampling bias, more flexibility than simple random sampling
  - Cons: data analysis is complicated (sampling weights, etc.)

Conclusion

- **Inference** is the process of using an estimate *from a sample* to describe a characteristic of a *population*
    - We're almost always interested in inference when we are performing a data analysis
- Estimates from a sample can deviate from the truth about a population for two reasons:
    - **Sampling bias** and **sampling variability**
- Sampling bias is a product of *how we collect our sample* (ie: simple random sampling, convenience sampling, etc.)
- Sampling variability is multifaceted, primarily involving *sample size* and *variability within the population*
    - We'll explore sampling variability in greater detail in today's lab