Sampling from a Population

Ryan Miller

Outline

1. Samples and populations
   - Definitions, examples, and notation
2. Sources of sampling error
   - Sampling bias and variability
3. Sampling methods
   - Convenience sampling, simple random sampling, and other approaches

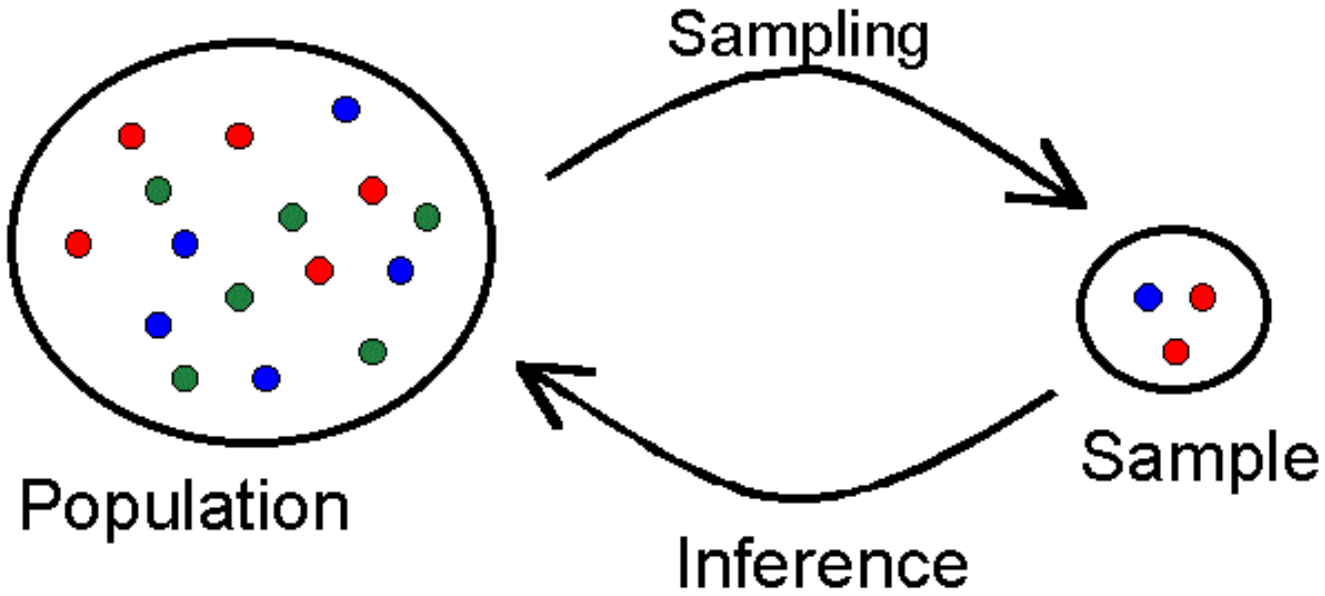
Introduction

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:



Note: We'll denote the number of cases in our sample as n and the size of the population as N (which is sometimes 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 a sample, 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 a sample, how do you think these variables are related in the population?
- As a starting point, we might use the sample correlation as an **estimate** of the population-level correlation
  - 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	$\bar{x}$
Standard Deviation	σ	s
Proportion	p	$\hat{p}$
Correlation	ρ	r
Regression	β_0, β_1	b_0, b_1

Two sources of sampling error

There are two main reasons why trends observed in the sample data might differ from those in 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

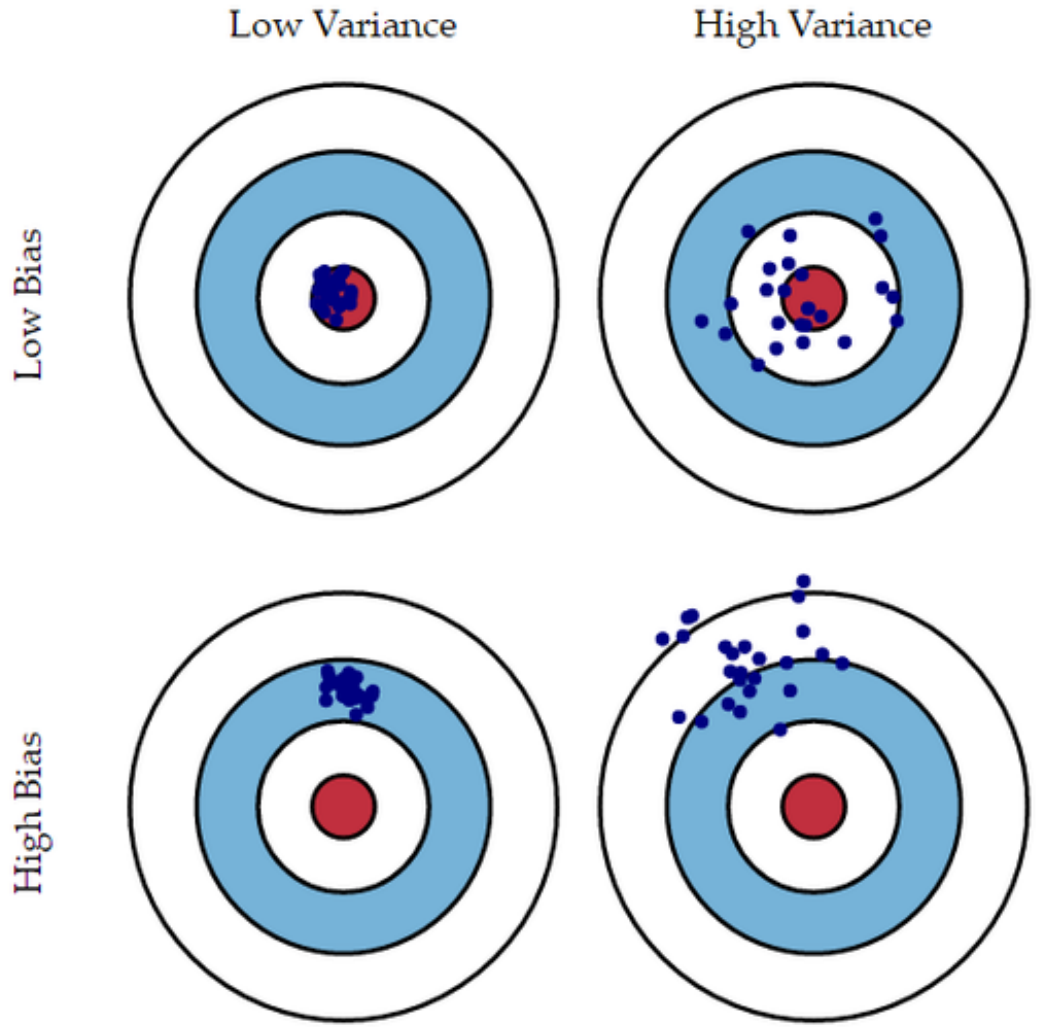
Two sources of sampling error

There are two main reasons why trends observed in the sample data might differ from those in 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")

Sampling error

Four different sampling procedures:



Each "dot" represent an estimate from a different sample

Remarks on sampling error

- Increasing the sample size will *decrease* sampling variability, but it *will not* alleviate sampling bias

Remarks on sampling error

- Increasing the sample size will *decrease* sampling variability, but it *will not* alleviate sampling bias
- Sampling procedures with high variance might seem problematic, but statisticians have developed tools (rooted in probability theory) to facilitate decision making in the face of this uncertainty

Practice

Shown below is the text of the Gettysburg Address, your goal is to accurately estimate the document's average word length:

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.

To obtain an estimate, take your own sample of 5 words (trying to be *representative*).

Practice

Now, answer the following:

1) What is the *population* and the *sample*?
2) What is the *population parameter* and the corresponding *sample estimate*?
3) Which quadrant of the bias/variance matrix ("two sources of error" slide) do you think your sampling procedure belongs to?

Practice (solution)

1) The population is all words in the Gettysburg Address (with the individual words being cases within this population). The sample is the 5 words you selected.
2) The population parameter is the average word length for the full address. The sample estimate is the average of the 5 words you selected.
3) Your procedure was likely biased and also high variance. For reference, the population mean is $\mu = 4.29$.

Common sampling methods

- **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)

Common sampling methods

- **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
  - Cons: can be difficult to execute

Common sampling methods

- **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
  - 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 becomes complicated (sampling weights, etc.)

Other sources of bias

1. **Non-response Bias** - Subjects who are recruited but decline to participate in a study differ in important ways from those who do participate or respond
2. **Non-ignorable Missing Data** - Subjects who are excluded from analysis due to missing data differ in important ways from those with complete data
3. **Social Desirability Bias** - Respondents tend to answer questions in ways that portray themselves in a positive light - Link
4. **Interviewer Bias** - The interviewer causes subjects to the behave differently than they otherwise would

Practice

With your group, discuss whether each of the following describe a **sample** or a **population**. If the data are a sample, describe the target population and whether the sample is biased or representative.

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. The Department of Transportation announces that of the 250 million registered cars in the US, 2.1% are hybrids
3. 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 (solution)

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 DOT has information on all registered cars in the US.
3. This is a sample, the population is all adult workers. It is a biased sample due being an online poll, and the social desirability typically associated with Steve Jobs.

Conclusion

1. Samples and populations
   - a sample is a subset of cases from a population that is used to make inferences
2. Sources of sampling error
   - Sampling bias is the result of a sampling procedure that systematically over (or under) selects certain types of cases
   - Sampling variability decreases for larger samples
3. Sampling methods
   - Convenience sampling is easy, but can be biased (though not necessarily)
   - Simple random sampling is unbiased, but can be difficult to implement