Module A: The Foundations of Statistics

Slide Deck A2:

Sampling Theory

The section in which we start looking at samples. Here, we examine the difference between the sample and two type of population. We also focus on sampling theory and a few methods for properly collecting data.

Today's Objectives

By the end of this slidedeck, you should

- understand some basic statistical terms
- distinguish between population and sample
- distinguish between the two types of populations
- know the five basic sampling schemes
  - Clearly distinguish between cluster and stratified sampling
- understand the interplay between bias and variance
- understand how these terms relate to how a sample represents a population

Basic Statistics Terminology

These terms will be used throughout the course:

- A **target population** is a particular group of interest.
  - A **sampled population** is a group from which the sample is taken.
  - A **sampling frame** is a physical list of all members of the sampled population.
- A **sample** is a subset of the population from which data are collected.
- A **variable** is a value or characteristic that changes among members of the population.
- The **data** are the counts, measurements, or observations gathered about a specific variable in a population in order to study it.
- A population **parameter** is a numerical description of a population characteristic.
- A sample **statistic** is a (numerical) description of sample characteristics.

Sample vs. Population

It is *essential* that you are mindful of the relationship between the populations and the sample.

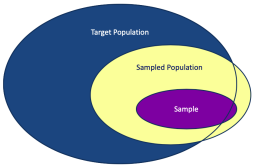
Who Cares?

In statistics, we use the sample to draw conclusions about the target population.

Sample vs. Population

This figure is designed to help you visualize the relationship among the target population (population of interest), the sampled population, and the sample:

- the blue oval represents the target population (a.k.a., the population of interest);
- the yellow oval represents the sampled population (which *may* extend beyond the target population);
- the purple oval represents the actual sample (which *must be* a subset of the sampled population).



Sample vs. Population

There are different features between a population and a sample. This table illustrates many of them:

Population	**Sample**
Group we want to know about	Group we *do* know about
Whole group	Part of the group
Characteristics are called parameters	Characteristics are called statistics
Parameters are generally unknown	Statistics are always known
Parameters are fixed	Statistics change with the sample

Clarifying Examples

Because knowing the difference between the target population, the sampled population, and the sample is so important, please identify the three in the following examples:

- A researcher wanted to determine the popularity of the ice cream flavor of the month (licorice). In a survey taken outside the Gizmo, 200 students at Knox College were asked if they had tried it. Of those surveyed, 192 said yes.
- A researcher wanted to determine if Knox seniors tended to have GPAs greater than 3.50. To determine this, the researcher asked 150 students who walked through the SMC Atrium. Of those 150, only 17 had a GPA greater than 3.50.
- A researcher wanted to know if T-Mobile was the most popular cell phone provider at Knox College. To determine this, she sampled 300 people leaving George Davis Hall (GDH), asking each their provider. Of those 300 people, 184 students indicated they used T-Mobile, 54 used AT&T, eight used US Cellular, and two used Verizon. No student used another provider.

Two related terms in statistics are **bias** and **variance**. They are different and must be treated as such.

Definition (Bias)

The **bias** is the difference between the actual value of the parameter and the expected value of the statistic.

Definition (Variance)

The **variance** is the (statistical) variability of the statistic across multiple samples.

In a perfect world, one wants no bias (**unbiased**) and no variance. In the *real* world, one cannot have both. Also, the tendency is that as one decreases, the other increases. This is referred to as the "**Bias-Variance Tradeoff**."

Typical Sampling Methods

These are the five typical sampling methods covered in introductory statistics courses. We will cover each in turn.

- Simple Random Sampling (SRS)
- Cluster Sampling
- Stratified Sampling
- Systematic Sampling
- Convenience Sampling

Simple Random Sampling

Definition

A sampling method in which each element from the sampled population has an equal chance of being selected.

Advantages

- unbiased
  - i.e., if one performs SRS a gazillion times, the *average* of the samples will be perfectly representative of the population

Disadvantages

- high resource cost
- high variability of estimates
- may not be possible to conduct (requires a **sampling frame**)

Cluster Sampling

Definition

A sampling method in which the sampled population is divided into subgroups (called **clusters**) according to a variable that is *not* related to the variable being estimated. Elements are chosen from one (or more) clusters.

Advantages

- unbiased
  - but only if the grouping variable is uncorrelated with what is being estimated

Disadvantages

- high resource cost, but lower than SRS
- high variability of estimates, but lower than SRS

Stratified Sampling

Definition

A sampling method in which the population is divided into subgroups (called **strata**) according to a variable that *is* related to the variable being estimated. Elements are chosen from each strata.

Advantages

- unbiased
  - but only if the proportion of each strata in the population is known

Disadvantages

- moderate resource cost
- moderate variability of estimates

Systematic Sampling

Definition

A sampling method in which every n^{th} member of the population is selected, after a random starting point.

Advantages

- unbiased
  - as long as the gap (n) is large enough to avoid overlap between groups

Disadvantages

- moderate resource cost
- low variability of estimates
- may not be possibly performed

Convenience Sampling

Definition

Convenience sampling is a method of collecting samples by taking samples that are conveniently located without attempting to ensure that the sample is representative of the population.

Why is this important? According to Edgar and Manz (2017)

> *Convenience sampling is the most common form of nonprobabilistic sampling, mostly because it is misused.*

You may want to read this entry for more information on the effects of convenience sampling

`https://research-methodology.net/sampling-in-primary-data-collection/convenience-sampling/`

Summary of Sampling Schemes

To summarize the above sampling schemes in terms of the important aspects:

	Bias	**Variability**	**Cost**
high	Convenience	Convenience	Simple Random
		Simple Random	Stratified
↕	Stratified	Cluster	Cluster
	Cluster		Systematic
	Systematic	Stratified	
low	Simple Random	Systematic	Convenience

Clarifying Examples

Which of the sampling schemes is described by each of the following research activities? Is the method appropriate in each case?

- A researcher wants to determine the popularity of the ice cream flavor of the month (licorice) at the Gizmo. To do this, he stands outside the Gizmo between 11am and 3pm on Monday asking every 10th person their opinion on the ice cream flavor of the month.

- A researcher wants to determine if Knox seniors tend to have GPAs greater than 3.50. To determine this, the researcher asks the GPA of the first 150 Knox seniors who walk through the SMC Atrium, starting at Monday at 3:00pm (Week 3 in the Fall term).

Clarifying Examples

Which of the sampling schemes is described by each of the following research activities? Is the method appropriate in each case?

- A researcher wants to know if T-Mobile is the most popular cell phone provider at Knox College. To determine this, she sampled the first 300 Knox students leaving George Davis Hall (GDH), starting at Monday at 8:00am (Week 3 in the Fall term), asking each their provider.

- A researcher wants to determine the proportion of students on campus supporting the equality of non-binary students. To do this, they randomly select 500 students from the campus directory.

Clarifying Examples

Which of the sampling schemes is described by each of the following research activities? Is the method appropriate in each case?

- A researcher wants to determine which of the five 'burger joints' on Henderson (Culver's, McDonald's, Steak and Shake, Wendy's, and Dairy Queen) is most popular. To determine this, she mails a random sample of 1000 people in Galesburg using addresses in the white pages. A total of 14 people reply.

- A researcher wants to know if iron, when left to the elements, tends to oxidize as wüstite (FeO), as magnetite (Fe_3O_4), or as something else. To determine this, he collects several ingots of pure iron bars (sent from Parchem, a legitimate supplier of chemicals) and exposes them to Galesburg weather for a year, starting on January 1.

Today's Objectives

Now that we have concluded this lecture, you should be able to

- understand some basic statistical terms
- distinguish between population and sample
- distinguish between the two types of populations
- know the five basic sampling schemes
  - Clearly distinguish between cluster and stratified sampling
- understand the interplay between bias and variance
- understand how these terms relate to how a sample represents a population

Supplemental Activity

To further explore these topics, there is:

- Laboratory Activity A

Note that you can access all Laboratory Activities here:
`https://www.kvasaheim.com/courses/stat200/labs/`

Supplemental Readings

The following are some readings that may be of interest to you in terms of sampling theory:

- Hawkes Learning: Section 1.3
- Intro to Modern Statistics: Section 2.1
- R for Starters: Nothing

- Wikipedia: Bias-Variance Tradeoff

Finally, Laboratory Activity D *will* explore a solution to the Bias-Variance Tradeoff.