Study Design 1

Brief Overview of Sampling Methods

Grinnell College

Review

We have spent time looking at what we can do with data.

- Making graphics + visuals
- Describing graphics + visuals
- Tables
- Probability

Outline

Statistics (largely) involves the following three broad domains:

- Design – how do we obtain our data
- Description – graphics and summaries
- Inference – decision-making

By the end of today you will be able to answer:

- What are some different types of ways to collect samples?
- What are generalizations?
- How do we avoid *biases* in our data collection?

Review (again)

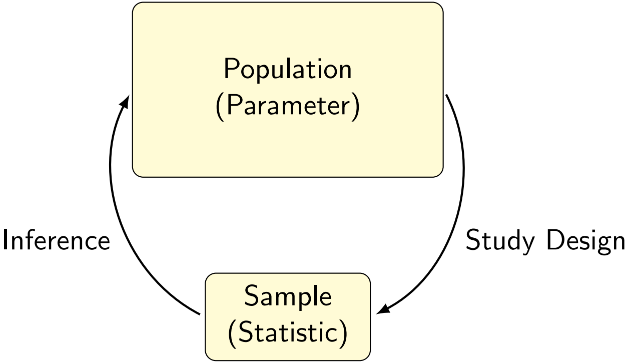
Population is a big group of subjects/events/things about which we wish to learn about

Parameter is a *quantifiable* attribute of a population. Most of the time, the parameter value is unknown

Sample is a much smaller, subgroup of a larger population

Statistic is a numerical summary of the sample that we calculate from our sample data

The Statistical Framework



Anecdotal Evidence

source: IMS Textbook

1) A man on the news got mercury poisoning from eating swordfish, so the average mercury concentration in swordfish must be dangerously high.

2) I met two students who took more than 7 years to graduate from Duke, so it must take longer to graduate at Duke than at many other colleges.

3) My friend's dad had a heart attack and died after they gave him a new heart disease drug, so the drug must not work.

Surveys

Surveys are a type of study where we ask people about their attitudes/opinions/beliefs

Some important things to think about:

- How do we select people we talk to?
- How many people do we talk to?
- How do we obtain information?
  - phone, email, in-person?

Choosing a Sample

How do we select people?

We want our sample to be **representative** of our population.

- this means that our sample is nearly the same as our population, only smaller
  - i.e.: same proportions M/F, same age/ethnic demographics
- a representative sample allows us to generalize our results from the sample to the pop.
- **biased**: a sample that is **not** representative

Not always possible to get a representative sample, but this is what we want to strive for

Generalizations

The process of applying results from a study to a larger group (population) is called *generalization* (making things more general/broad).

Example: The Harvard 1988 heart attack data we looked at previously was conducted on middle-aged male physicians. There is evidence in this study that aspirin reduced heart attack rates. Can this be generalized to everyone?

Example: Most water quality samples from 50 randomly chosen Iowa lakes demonstrated heightened levels of nitrates (a carcinogen). What can we say about nitrates in Iowa's lakes? Can we say anything about the rivers from this data?

How do we select people?

Random Sample

We can choose people at random from our population to reduce the chances of getting a biased sample.

- usually the best way to get a representative sample
- allows us to **generalize** from our sample to the pop.

Sample Size (n = ?)

- number of people we survey is important (more people = more info)
- sample size is very important, but proportion of pop. surveyed is not
- better to have small sample that is representative, than a large sample that is biased

How do we select people?

Census – Would conducting a census of the entire population be better?

Issues

- difficult
- time consuming
- expensive

Sampling Frame

The **sampling Frame** is the actual list of things we have access to for sampling things from the population.

- May not have full access to the people/things we want → issues

Example: Random surveys sent via mail to customers of an electric utility company → may have full access to list of company's customers including address → no issues reaching people

Example: Random surveys sent via email to customers of this same company → missing info on some customers email addresses (older customers) → cannot reach these customers → opinions of older customers can't be obtained

Sampling Methods – SRS

Simple Random Sample (SRS)

- each combination of observations has same chance of being selected
- if we were to select multiple SRSs again, different observations will be chosen
  - variability by sampling → sampling variability

Example: Suppose we are interested in studying bridge degradation across Iowa (a problem in recent years). To take a SRS of Iowa bridges and test them, we need a list of all bridges+locations in Iowa. Practical?

image: https://www.grandforksherald.com/newsmd/aging-rural-bridges-still-a-huge-concern

Sampling Methods – Cluster Sample

A **cluster** is a bunch of *things* grouped together, often based on geographical information.

- clusters are full of *heterogenous* (different) things and look like mini versions of the population
- different clusters should look *homogenous* (similar) when compared to *each other*

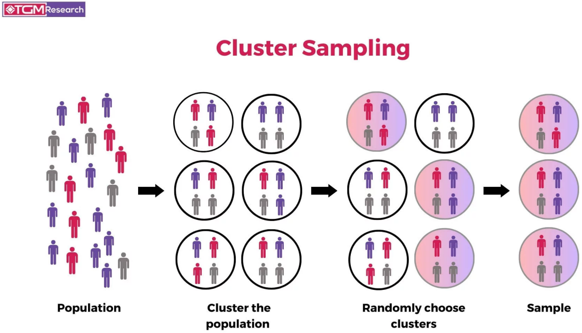


image: https://tgmresearch.com/cluster-sampling.html

Sampling Methods – Cluster Sample

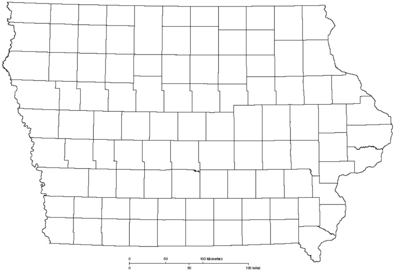
Benefits: Reduced travel time (and cost) for gathering information, don't need sampling frame of all things just the clusters

Drawback: Clusters may not actually be representative

Example: Choosing a random sample of all bridges in Iowa is impractical. Instead maybe we use clusters = counties and randomly choose a handful of counties to go inspect all bridges in.

- Still may be impractical. SRS of bridges within each county?

Iowa Counties



Sampling Methods – Stratified

Stratified Sample

- **strata**: subgroups of pop., within each the individuals are similar
  - individuals within a strata are similar, but strata themselves can be very different
- we take a SRS from each strata to ensure each group has representation

geologylearn.blogspot.com/2015/10/rock-layers.html

Stratified Sample Example

Benefits: Ensures each defined group has representation

Drawbacks: Requires good prior information or thought put into defining strata, strata are not always perfect classifications (think human demographics).

Example: Still thinking about Iowa bridges. Maybe urban vs. rural bridges are an important distinction? Strata = county type (rural vs. urban). Thought process: rural bridges are more similar to each other, urban bridges are more similar to each other. Take a SRS of each type of bridge to make sure both types have representation.

- Issues? Misclassification, required me to know about urban/rural

Sampling Practice 1

Scenario: We want to ask 200 Grinnell students their opinion on if they prefer living in the dorms

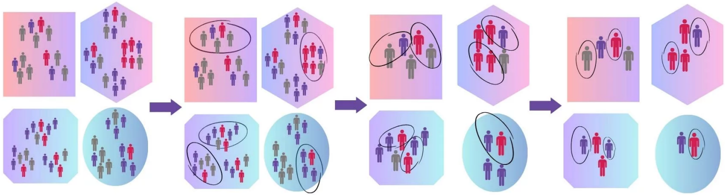
How do we do this with an SRS?

Why might we want to use a stratified sample? What could we stratify by?

Multi-Stage Sampling

These different types of sampling methods can be combined to make working with large populations easier.

Multi-stage Cluster Sampling



Selected countries (clusters) from global population

Within each country, randomly chose regions

From each region, randomly selected cities/towns

Randomly sampled participants from each selected city/town

Could even apply stratified sampling to this last step so that each city/town gives us particular demographics

image: https://tgmresearch.com/cluster-sampling.html

Biased Samples

Biased samples are not representative of the population

Voluntary Response Sample

- people select themselves to participate
- usually people with strong opinions respond to surveys

Convenience Sample

- people are chosen in a non-random way
  - poll at a specific location
- name comes from the fact that this is 'easy' to do

Sampling Practice 2

Scenario: We want to ask a bunch of people their opinion on which football team they like more: ISU cyclones vs Iowa Hawkeyes.

Suppose I put a poll on social media: What types of bias might happen?

Suppose I interview people outside of the ISU football stadium during a gameday. What type of bias will I get?

Biased Samples

Sampling Bias

- **undercoverage**: certain groups may not be represented in samples
- **sampling frame** (list of who we can sample from) may be missing some of the population

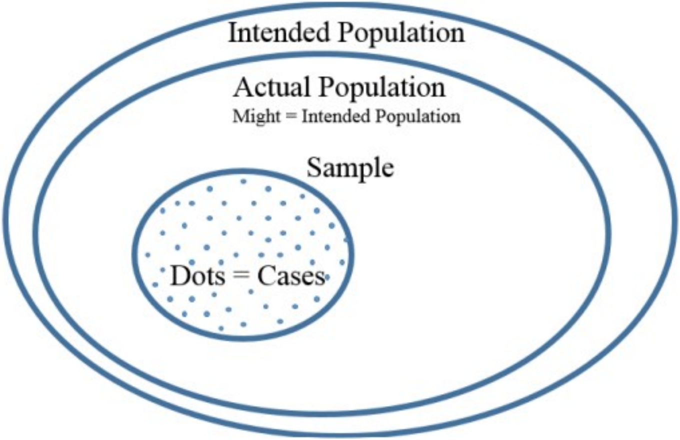
Non-response Bias

- some people can't be surveyed or choose not to participate

Response Bias

- we don't always get accurate info from people
- question wording
- not wanting to provide truthful answers

Intended vs. Actual population



source: Dr. Ziegler's Stat 104 notes (ISU)

Wrapping up – Reflection

What are some different types of ways to collect samples?

What are generalizations? What allows us to make them?

What are some ways we can avoid biases when getting our samples?