SAMPLING AND DATA

MAT 152 – Statistical Methods I

Lecture 2

Instructor: Dustin Roten

Fall 2020

Constructing a Representative Sample

Sometimes it is possible to conduct a **census** (measuring the whole population).

Other times, a **sample** must be constructed.

How can we be sure that a sample has the same characteristics as the population it is representing?

Random sampling – Each member of a population initially has an equal chance of being selected in the sample.

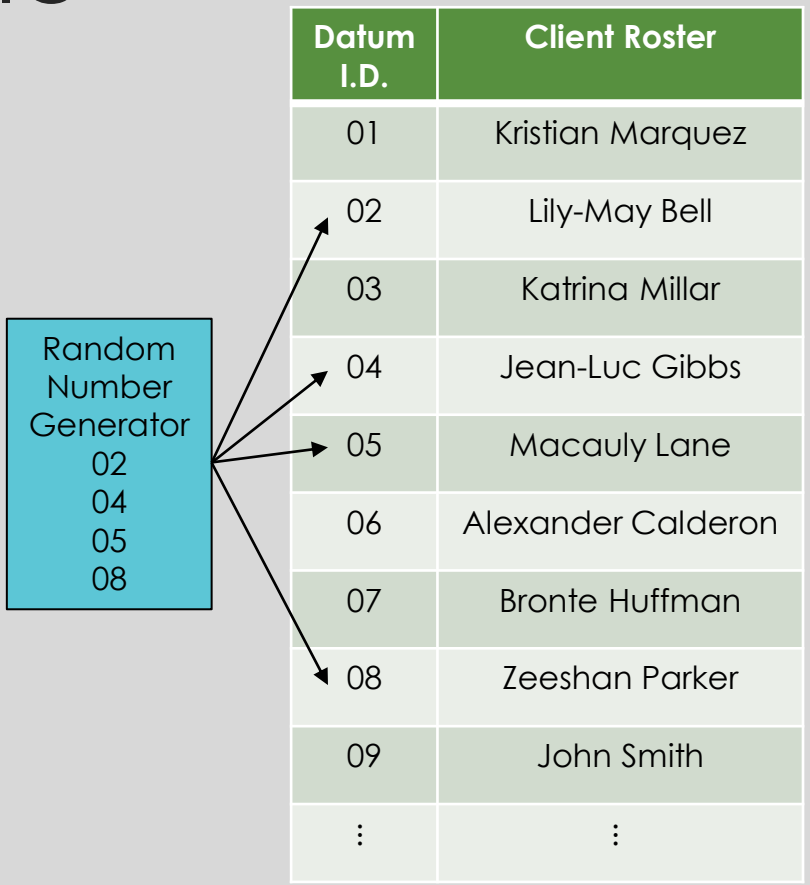
There are several methods of random sampling…

Method #1
Simple Random Sample

- Any group of *n* individuals is equally likely to be chosen as any other group of *n* individuals.

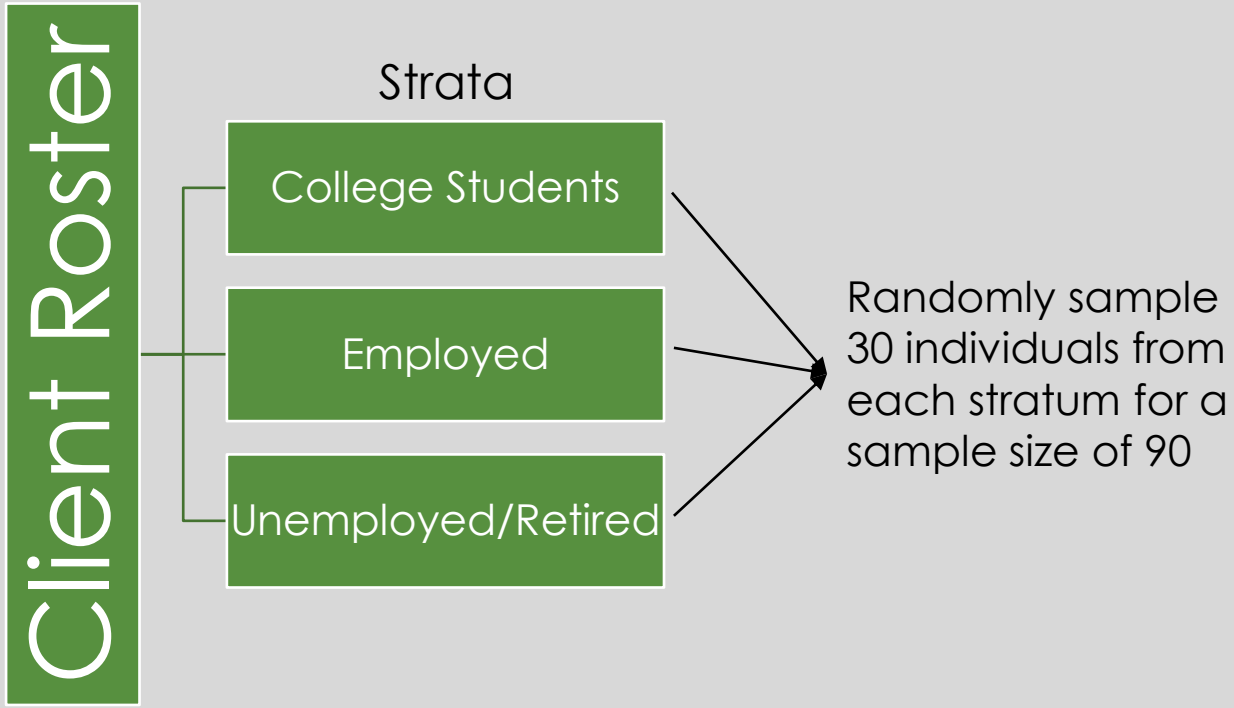
- Revisiting an Example:
  - A fitness center is interested in the average amount of time a client exercises in the center each week.

- A random number generator is used to select individuals from the population



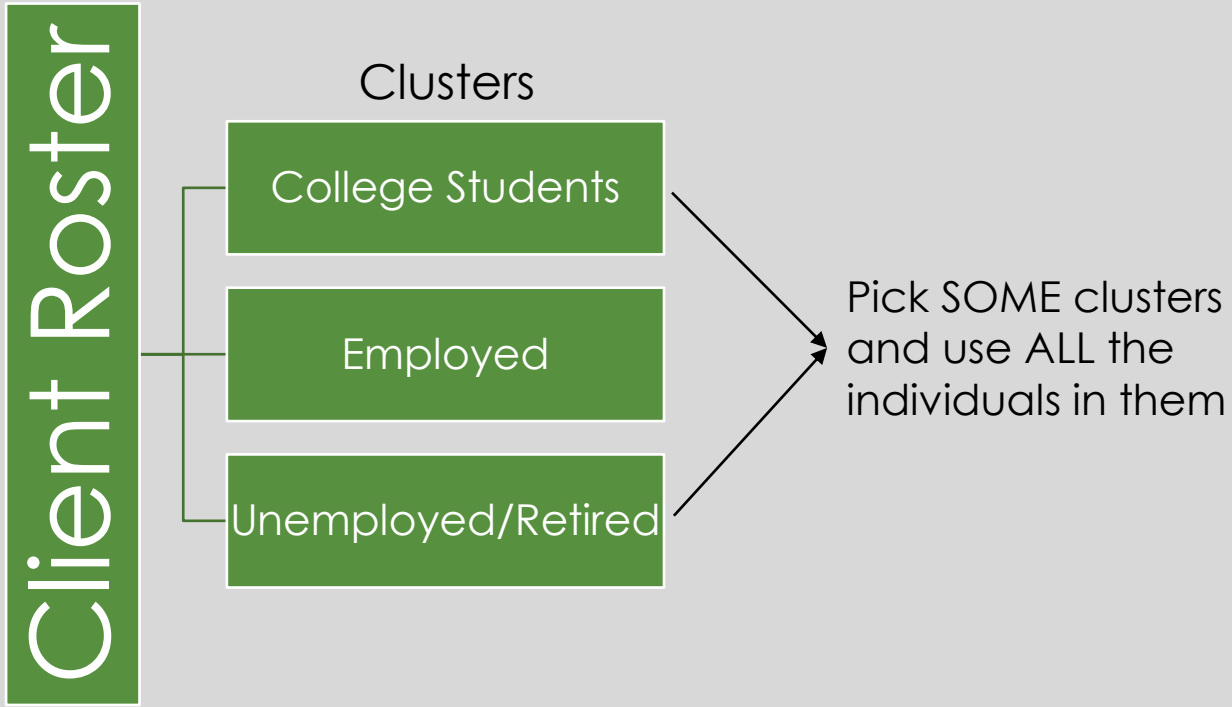
Method #2
Stratified Sample

- A population is divided into groups called **strata** and then a proportionate number of individuals is selected from each **stratum**.



Method #3
Cluster Sample

- The population is divided into **clusters** (like strata) and then *some* of the clusters are selected. All members from the selected clusters make up the sample.



Method #4
Systematic Sample

- A starting point within the data is randomly selected and every n^{th} piece of data is selected from a list of the population.

Datum I.D.	Client Roster
01	Kristian Marquez
02	Lily-May Bell
03	Katrina Millar
04	Jean-Luc Gibbs
05	Macauly Lane
06	Alexander Calderon
07	Bronte Huffman
08	Zeeshan Parker
09	John Smith
⋮	⋮

Method #5
Convenience Sampling

- NON-RANDOM SAMPLING METHOD
- Using data readily available

"The exit poll is the crucial bit of information available to tell you who came out to vote, what candidate they voted for, and why. Exit polls are conducted as voters leave their polling place, thus the name "exit" polls. For the states that hold caucuses, exit polls are called entrance polls, simply because interviewers talk to voters as they're entering their polling place instead of exiting." – CNN Politics

Not likely to reflect the entire population

Methods of Random Sampling

- Simple Random Sample
- Stratified Sample
- Cluster Sample
- Systematic Sample
- Non-random: Convenience Sample

- **Sampling Errors:**
  - The sample isn't big enough
  - The sample doesn't reflect the population
  - Sampling bias – some members are more likely to be chosen than others.
- **Non-sampling Errors:**
  - A broken measurement device
- Because of **Variation** within data, a sample isn't EXACTLY reflective of a population.
- The larger the sample sizes, the smaller the sampling error.

What if we wanted to conduct an experiment?

- Surveys gather data.
- Experiments investigate a relationship between two variables in a *controlled* way.
- Sampling from a population provides a group of **experimental units**. (individual people or objects)
- Suppose variable *X* causes a change in variable *Y*.
  - Example: $X = \text{Breakfast before an exam (yes, no)}$; $Y = \text{Exam Score}$
  - Question: Does eating breakfast before an exam increase your score?
  - Hypothesis: Eating breakfast before an exam will increase your score.

- In this example, *X* is the **explanatory variable**. It is the variable that causes the change in *Y*.
- *Y* is the **response variable**. It is the variable being affected.
- The values of *X* are called **treatments** (breakfast, no breakfast)
- In a **randomized experiment**, the researcher manipulates values of the explanatory variable and measures the resulting changes in the response variable.

Example 1

- Suppose you want to investigate the effectiveness of vitamin E in preventing disease. You **sample** a **population** of interest (teenagers, ages 65+, etc.) and ask them if they regularly take vitamin E. You notice that the subjects who take vitamin E exhibit better health on average than those who do not.
- *Does this REALLY prove that vitamin E is effective in disease prevention?*

- In this scenario, it does NOT prove vitamin E is effective in disease prevention. Why?
  - Typically, people who take vitamin supplements make additional healthy choices: choosing not to smoke, dieting, exercising, etc. Thus, it's possible that the individuals taking vitamin E have MULTIPLE characteristics that also keep them healthy.
  - How can we tease out the effects of vitamin E only?

- Other factors/variables that may cloud a study are called **lurking variables**.

- This can be addressed by **random assignment** where **experimental units** are randomly assigned. **Lurking variables** are then spread equally between treatment groups.

Example 2

(An example of the whole process)

Suppose a statistics instructor wanted to determine how caffeine affects college students' exam scores.

Hypothesis: Ingesting caffeine before an exam will increase an individual's score.

1.) 20 students are randomly selected from each department at Forsyth Tech. (participants give **informed consent** and **IRB** guidelines are followed.)

2.) Each student is randomly assigned to a treatment group.

group #1: 12oz of caffeinated coffee 1hr before a "general knowledge" exam

group #2: 12oz of decaffeinated coffee 1hr before a "general knowledge" exam

3.) Afterwards, the exams are graded. Analyses are performed (specifics given later in the course)

Identify the population of interest.
all college students

Identify the sample.
a randomly sampled group of Forsyth Tech students

Identify the experimental units.
The college students in the study

Identify the explanatory variable.
Caffeine

Identify the treatments
Caffeinated and decaffeinated coffee.

Identify the response variable.
Exam score

Example 2 (cont.)

- When a study prompts a physical response from participants, it is difficult to isolate the effects of the explanatory variable.

- Thus, a **control group** is needed: experimental units given decaffeinated coffee as a treatment (a **placebo**).

- It's best if participants do not know which type of coffee they are given.
  - If a participant KNOWS that they are given caffeinated coffee, he/she may feel pressured to perform better on the exam. Similarly, if a participant KNOWS that the he/she is given decaffeinated coffee, this may adversely affect performs.

- **Blinding** is essential to remove the *power of suggestion.*
  - Participants do not know which treatment they are receiving.
  - A **double-blind** study indicates that participants and administrators of the treatment do not know which is the active treatment or the placebo.

A Quick Review

- Sampling methods:

  Simple Random Sample, Stratified Sample, Cluster Sample, Systematic Sample, Convenience Sample (NR)

- Experiments have the following components:
  - Explanatory variable – A variable that causes change in another variable
  - Response variable – The variable being changed.
  - Treatment – The different values of an explanatory variable
  - Experimental unit – A single object or individual to be measured

- Control groups are assigned a placebo (one of the treatments)

- Blinding and double-blinding removes the "power of suggestion"
- Institutional Review Boards (IRB) ensure the safety of individuals participating in studies