Week 1

Data Basics & Sampling

ST 314
Introduction to Statistics for Engineers

Oregon State University

What is data and why do we need it?

Records

Information

Decision making

Trends

Data, Individuals, and Variables

- An observational unit or individual is an object described by data.
- A characteristic of an individual is a variable.
- A data set describes a set of individuals and their variable values.
- Variable types:
  - Categorical: based on a quality assigned to a category
  - Quantitative: based on a measured quantity
    - Discrete: countable set of values $\{0, 1, 2, 3, \ldots\}$
    - Continuous: takes on all values over an interval $[0, 10]$

Data Example

These data are from applicants to a rigorous engineering course that requires students to take an entrance exam prior to admission into the course.

Individual	ExamScore	GPA	Gender	Year
1	79	3.3	Man	Junior
2	88	2.9	Man	Senior
3	92	3.8	Non-binary/non-conforming	Soph.
4	87	3.7	Man	Junior
5	68	3.4	Woman	Junior

- **Individual -** Identifying number ← not variable
- **ExamScore -** Entrance exam score out of 100 } quantitative
- **GPA -** College GPA at time of application
- **Gender -** Gender with which applicant most closely identifies } categorical
- **Year -** Year is school at time of application

How are data collected?

- Observational studies
- Designed experiments

Populations, Samples, and Inference

- The population is the complete collection of subjects or things in which we are interested.
- A sample is a subset of the population.
- Statistical inference is the process of using known sampled information to form a conclusion about unknown population characteristics.

Parameters & Statistics

Parameters → Population

- A **parameter** is characteristic of the complete collection of interest
- Parameters are often unknown
- Estimated from a statistic

General notation θ

"theta"

Statistics → Sample

- A **statistic** is a characteristic from a subset of the population
- Used to estimate parameter values

General notation $\hat{\theta}$

"theta hat"

Parameters & Statistics Examples

An internet site reports that around 10% of the population is left handed. Since it is important to provide enough left-handed desks suppose the facilities department at OSU is curious if this is true at our university. Let's use our class information to answer the following:

Population:

All OSU students

Sample:

314 students in 8:30 am class in spring 2022

Do you think this sample will be representative of the population of interest?

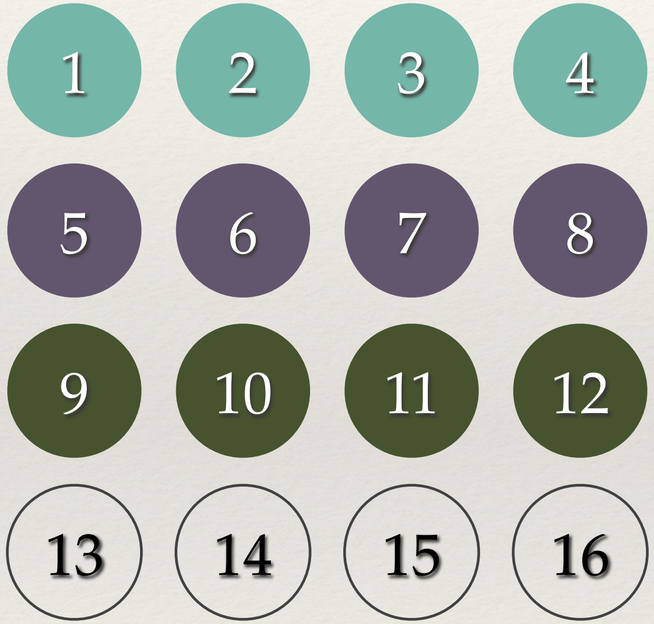
NO probably not

Random Sampling

- Using a random mechanism is the best way to reduce bias in a sample.
- Bias is the tendency to systematically favor certain parts of the population over others.

Sampling Designs: Simple Random Sample

Population



$n = 8$

↑ sample size

6 7 8

9 10 12 13

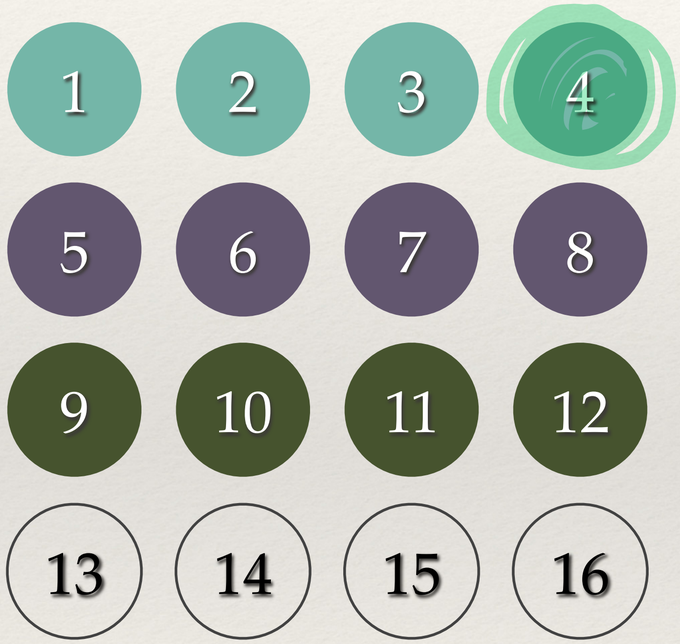
15

Sample

Every combo of n individuals has equal chance of being selected

Sampling Designs: Systematic Random Sample

Population



Randomly select starting value

n = 8

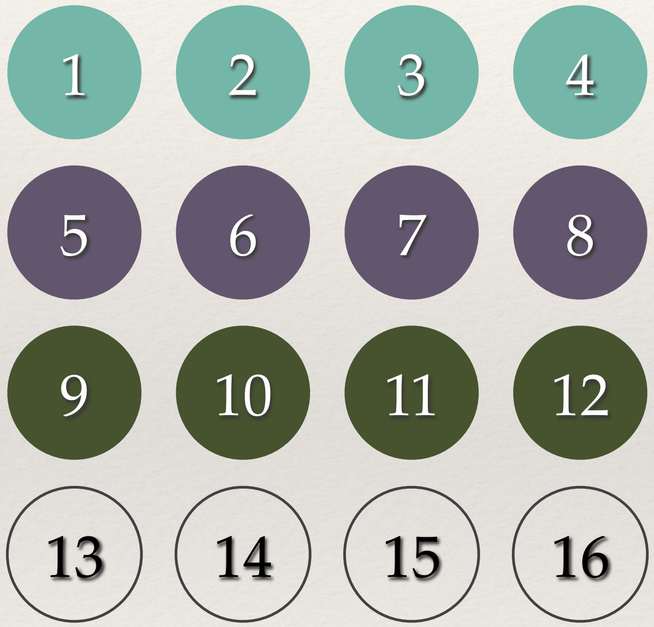
start = 4

Sample

4, 6, 8, 10, 12, 14, 16, 2

Sampling Designs: Stratified Random Sample

Population



Sample

2 3

5 8

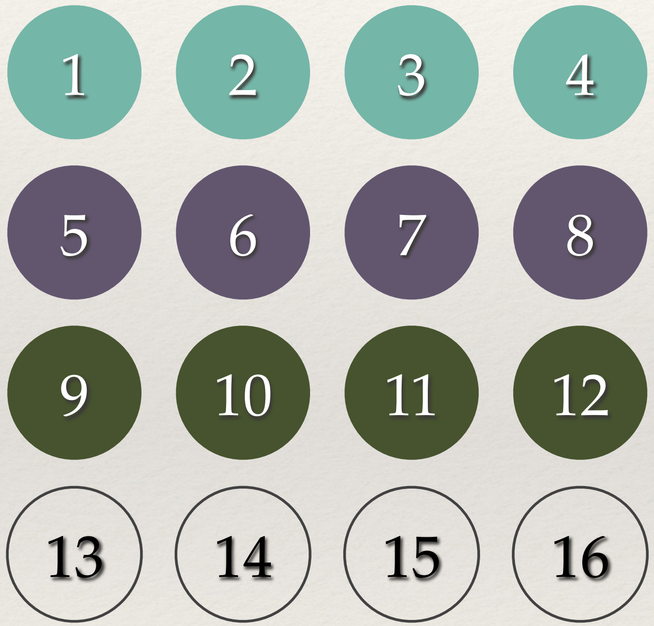
11 12

13 15

Population is broken into strata (groups) then perform a simple random sample within each group

Sampling Designs: Cluster Sample

Population



n = 8

Sample

1 2 3 4

9 10 11 12

Pop. broken into groups

Random Sampling Example

A university would like to assess the experience for the 3000 students living in campus dorms. They come up with three ways to sample 100 students. Match the type of random sampling scheme.

1. They obtain a list of all dorm residents and randomly generate a starting point from 1 to 3000, then select every 30th student on list for a total of 100 students.

systematic

2. They obtain a list of all students living in dorms on campus. They use software to randomly select a sample of n=100 students. Each selected student is surveyed via phone, email or in-person.

simple random sample

3. They can take a simple random sample of 20 students from each of the five dormitories on campus for a total of 100 students.

stratified random sample

Bias & Poor Sampling Design

- A convenience sample is a sample in which sampled observations are the easiest to obtain.
- A voluntary sample is special case of convenience sampling. Participants are not selected, instead they volunteer their response, i.e. an online poll.

Data from poor sampling designs are most likely biased and should not be used to make inference. They may lead to overestimating or underestimating certain characteristics of the population.

Types of Bias

Sampling Bias

specific subpopulations are not adequately represented

Nonresponse Bias

subjects selected cannot be reached or refuse to participate

Response Bias

can occur when wording is misleading or subjective

can occur when the subject matter is sensitive and participants respond untruthfully

Designed Experiments

How does an experiment differ from an observational study?

Individuals/units are assigned to the treatment(s) they receive

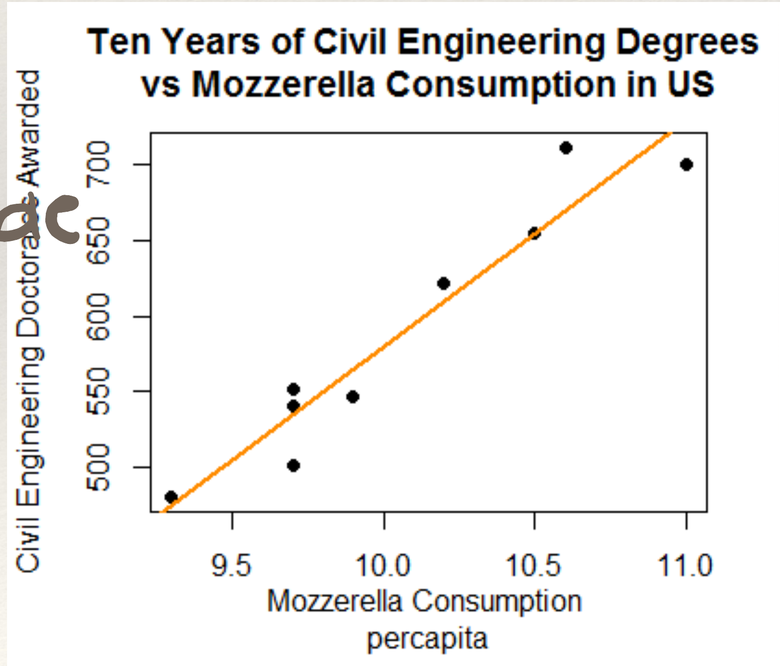
Observational Studies

An **observational study** observes and collects information on units but does not attempt to change or influence the units. Information is collected based on an established factor. Data represents a finite observable population.

An observational study cannot establish a causal relationship.

An unaccounted for confounding variable associated with both factor and response variables, may be the underlying cause of the relationship.

Does cheese cause degrees?



Principles of Experimental Design

- Controlling keep components that are not of interest consistent
- Randomization randomly assign units to the treatments they are receiving to control for factors that cannot be physically changed
- Replication multiple units receive same treatment to get a more accurate understanding of treatment behavior.
- Blocking Break units into groups based on a shared characteristic(s) and then randomly assigning treatments within the blocks

Completely Randomized Block Design