Unit 4: Sampling

2. The magic of randomness: sampling

Stat 140 - 04

Mount Holyoke College

Dr. Shan Shan

Slides posted at *http://sshanshans.github.io/stat140*

Outline

1. This week: Randomness

2. Application 2: sampling
    1. Population vs sample
    2. How to sample badly
    3. Good: Sample Randomly

3. Summary

Outline

- A **population** includes all individuals or objects of interest.
- A **sample** is a part of the population that is observed.

Poll question

Which of the following is most important to you?

a. Athletics
b. Academics
c. Social Life
d. Community Service
e. Other

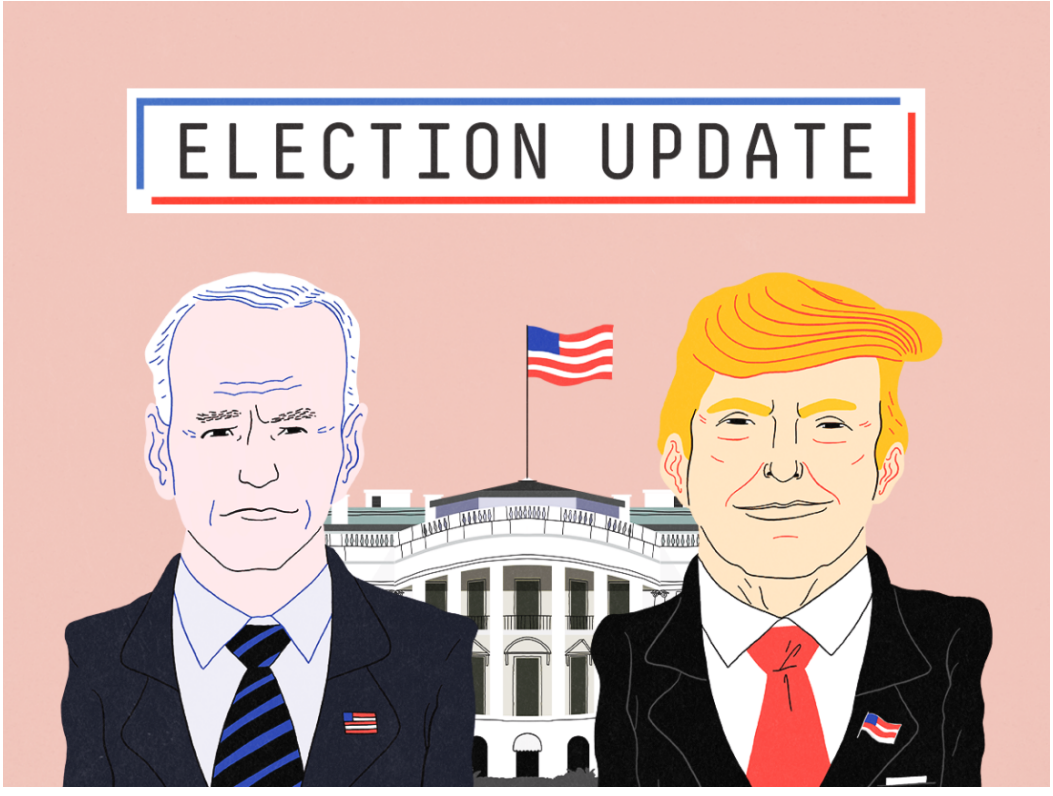
- Suppose researchers studying student life at MHC use the results of our poll question to investigate what MHC students find important about college life

- What is the sample?
- What is the population?
- Can the sample data be generalized to make conclusion about the population? Why or why not?

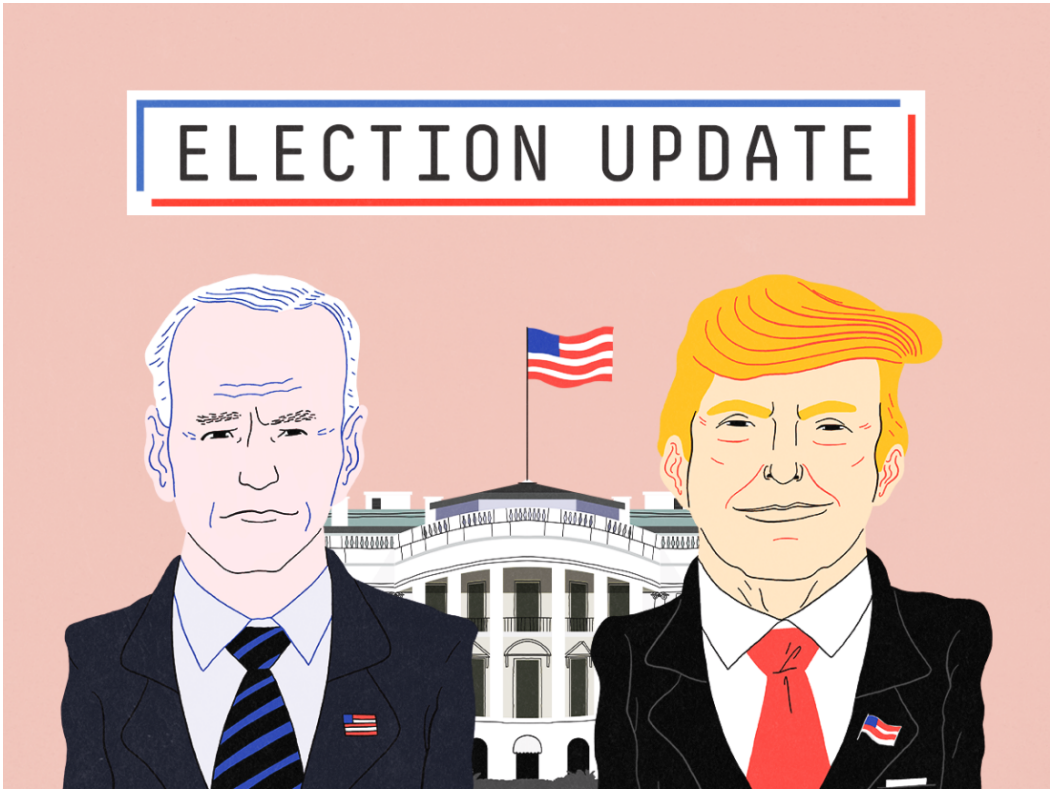
- When you taste a spoonful of soup and decide the spoonful you tasted isn't salty enough, that's *exploratory analysis*
- If you generalize and conclude that your entire soup needs salt, that's an *inference*
- For your inference to be valid, the spoonful you tasted (the sample) needs to be *representative* of the entire pot (the population)

We'd like to know about an entire group of individuals – a **population** – but examining all of them is usually impractical, if not impossible. So we settle for examining a smaller group of individuals – a **sample** – selected from the population.

Election polling



Election polling



Population: Registered voters in U.S.
Sample: Individuals who were polled

Dewey Defeats Truman?

- The paper was published before the conclusion of the 1948 presidential election, and was based on the results of a large telephone poll which showed Dewey sweeping Truman
- However, Harry S. Truman won the election
- What went wrong?

Some terminology and definitions

- For the sample statistic to be a good estimate of the population parameter, the sample needs to be **representative** of the population.
- Definition: Sampling methods that tend to over-emphasize or under-emphasize some characteristics of the population are **biased**.

Below are a few examples of bias in the sampling stage.

1. **Sample volunteers**: In a voluntary response sample, a large group of individuals is invited to respond, and those who choose to respond are counted. E.g., letters written to members of Congress.
2. **Sample Conveniently**: In convenience sampling we simply include the individuals who are convenient for us to sample. E.g, Surveys of World Wide Web users that over-represent frequent users
3. **Undercoverage**: In undercoverage, some portion of the population is not sampled at all or has a smaller representation in the sample than it has in the population.

Below are a few (more) examples of bias in the sampling stage.

4. **Nonresponse bias**: A common and serious potential source of bias for most surveys is nonresponse bias. No survey succeeds in getting responses from everyone. The problem is that those who don't respond may differ from those who do. And they may differ on just the variables we care about
5. **Response bias**: Response bias refers to anything in the survey design that influences the responses.

How to select samples to represent the entire population?

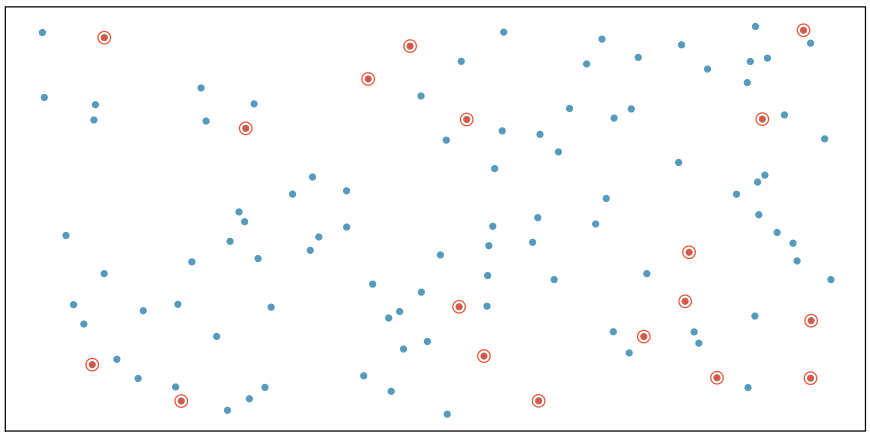
We select them at random. Randomizing protects us from the influences of all the features of our population by making sure that, on average, the sample looks like the rest of the population

- Before the 2008 election, the Gallup Poll took a random sample of 2,847 Americans. 52% of those sampled supported Obama
- In the actual election, 53% voted for Obama

Random sampling is a very powerful tool!!!

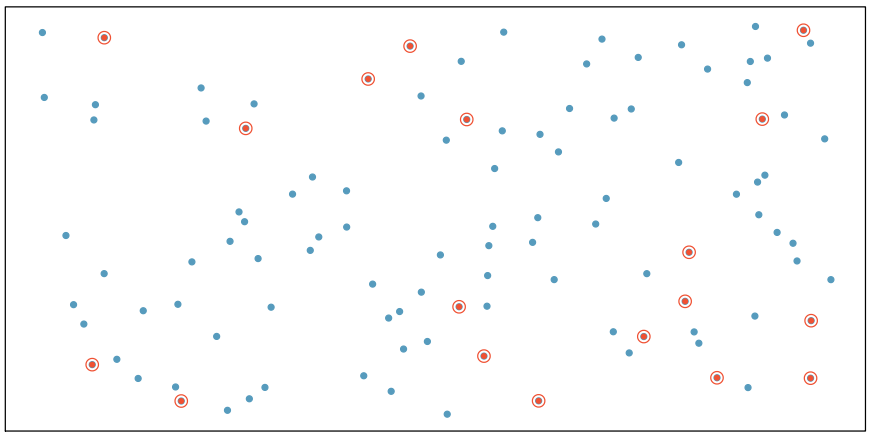
Simple Random Samples (SRS):

Drawing names from a hat



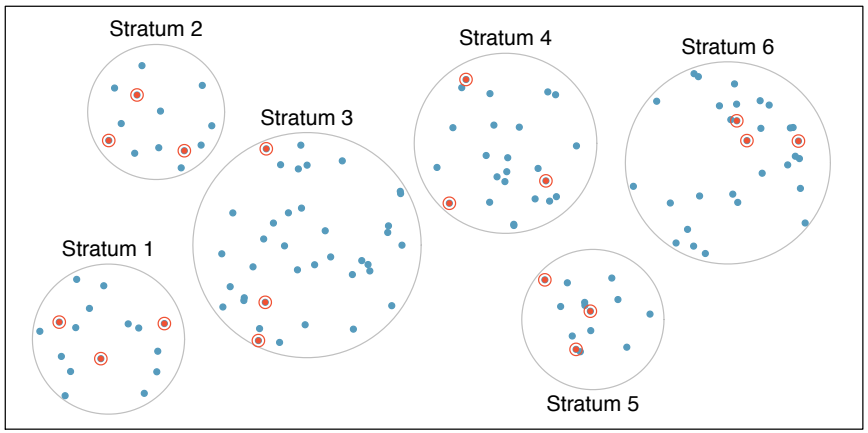
Simple Random Samples (SRS):

Drawing names from a hat



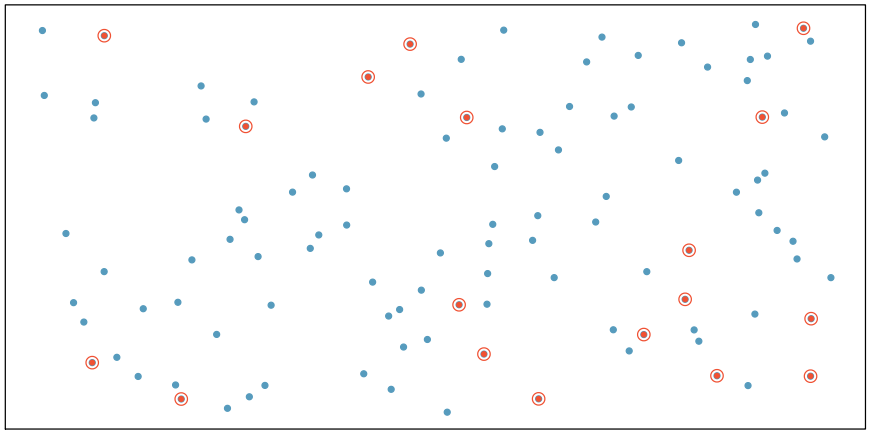
Stratified: homogenous strata

SRS in each stratum



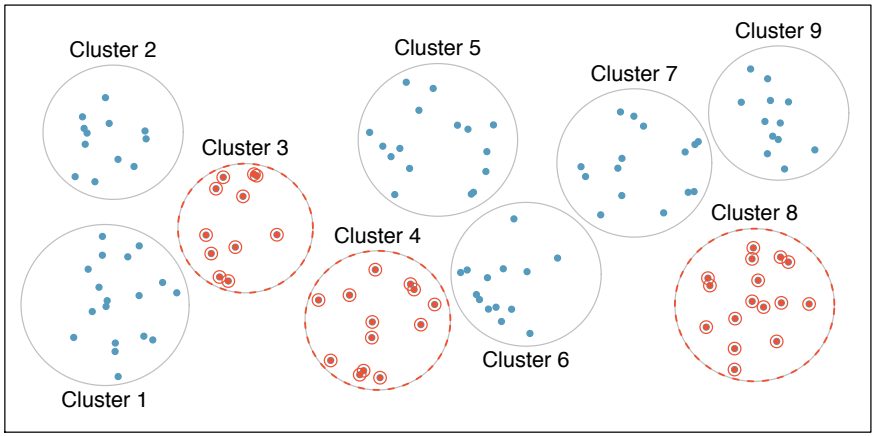
Simple Random Samples (SRS):

Drawing names from a hat



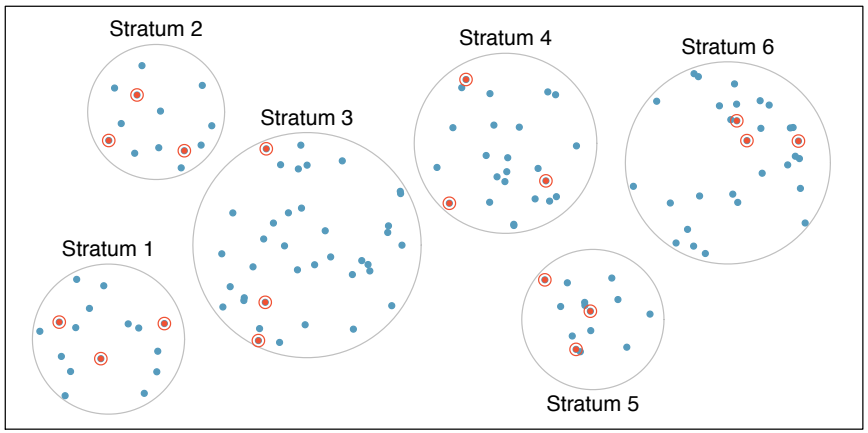
Cluster: sample of a few clusters

Sample all chosen clusters



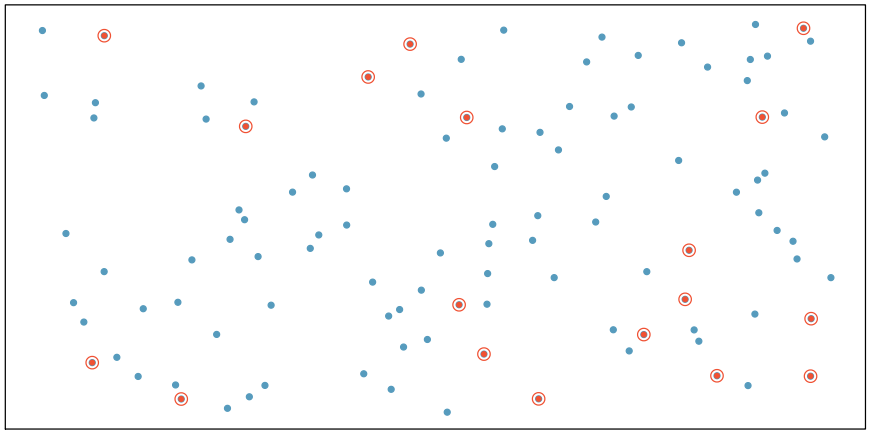
Stratified: homogenous strata

SRS in each stratum



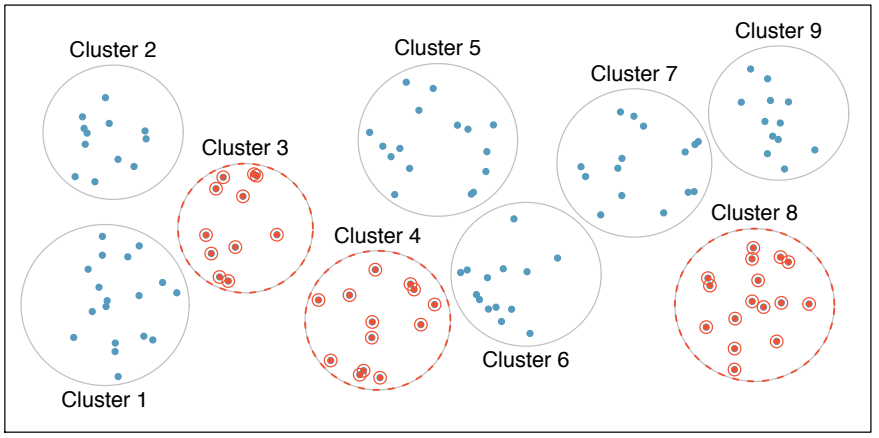
Simple Random Samples (SRS):

Drawing names from a hat



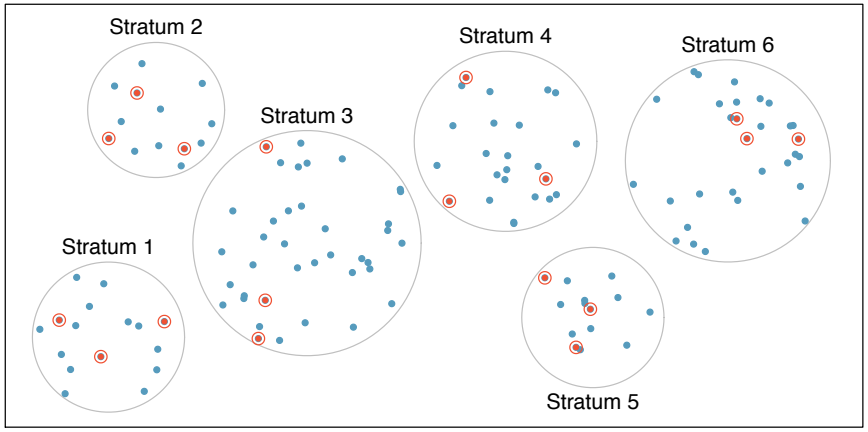
Cluster: sample of a few clusters

Sample all chosen clusters



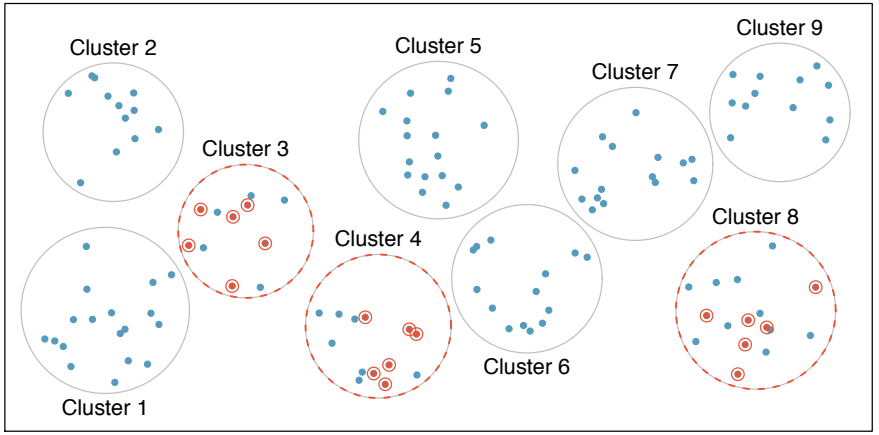
Stratified: homogenous strata

SRS in each stratum



Multistage:

Random sample in chosen clusters



Poop tests stop COVID-19 outbreak at University of Arizona

Outline

1. Population vs sample
2. Good: Sample Randomly
3. Good: Sample Randomly

Important message of the day:

- We use observations in a sample to estimate something about a population.
- Watch out for bias in your sampling. Ouch.