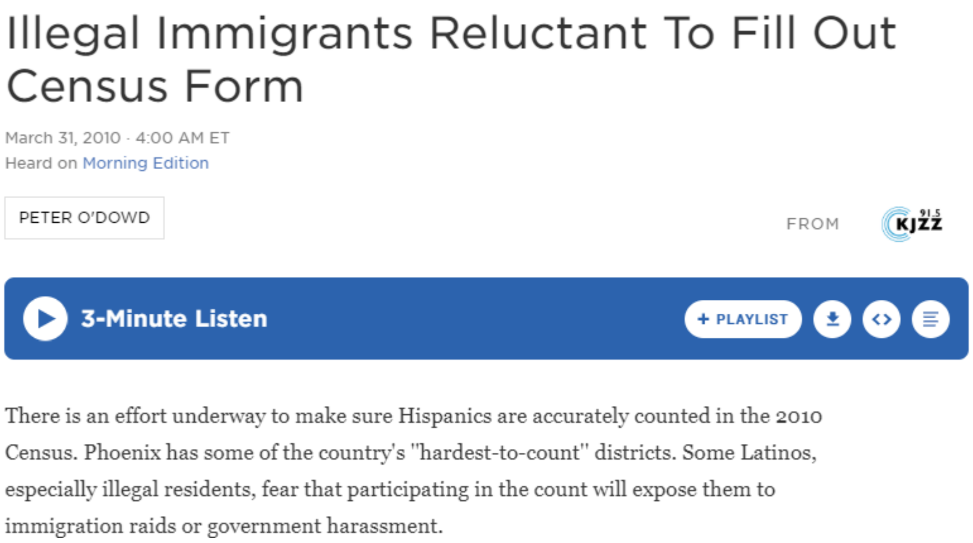
Sampling & sources of bias

- census vs sample
- sources of bias
- sampling methods

Original slides developed by Mine Çetinkaya-Rundel of OpenIntro.

Census



Illegal Immigrants Reluctant To Fill Out Census Form

March 31, 2010 · 4:00 AM ET

Heard on Morning Edition

PETER O'DOWD

FROM KJZZ 91.5

There is an effort underway to make sure Hispanics are accurately counted in the 2010 Census. Phoenix has some of the country's "hardest-to-count" districts. Some Latinos, especially illegal residents, fear that participating in the count will expose them to immigration raids or government harassment.

Wouldn't it be better to just include everyone and "sample" the entire population?

- This is called a *census*.

There are problems with taking a census:

- Conducting a census takes lots of resources.
- Some individuals are hard to locate or hard to measure, and these people may be defferent from the rest of the population.
- Populations rarely stand still

exploratory analysis to inference

exploratory analysis

inference

- Sampling is natural.
- Think about sampling something you are cooking - you taste (examine) a small part of what you're cooking to get an idea about the dish as a whole.
- 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*.

exploratory analysis to inference

exploratory analysis

representative sample

©SpendWithPennies.com

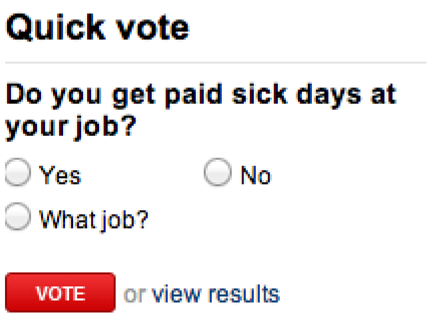
inference

- For your inference to be valid, the spoonful you tasted (the sample) needs to be *representative* of the entire pot (the population).
  - If your spoonful comes only from the surface and the salt is collected at the bottom of the pot, what you tasted is probably not representative of the whole pot.
  - If you first stir the soup thoroughly before you taste, your spoonful will more likely be representative of the whole pot.

a few sources of sampling bias

- Convenience sample bias: Individuals who are easily accessible are more likely to be included in the sample.

- Non-response bias: If only a small fraction of the randomly sampled people choose to respond to a survey, the sample may no longer be representative of the population.
- Voluntary response bias: Occurs when the sample consists of people who volunteer to respond because they have strong opinions on the issue. Such a sample will also not be representative of the population.



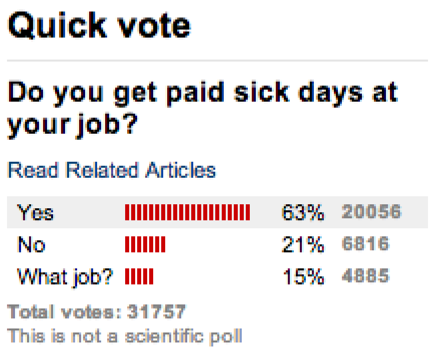
Quick vote

Do you get paid sick days at your job?

○ Yes ○ No

○ What job?

VOTE or view results



Quick vote

Do you get paid sick days at your job?

Read Related Articles

Yes	63%	20056
No	21%	6816
What job?	15%	4885

Total votes: 31757

This is not a scientific poll

Practice

A retail store considering updates to their credit card policies randomly samples 1000 of their credit card holders to survey on the phone. The phone calls are made during business hours, therefore there is a lower rate of responses from members who work during these hours. What type of bias is this indicative of?

a) Convenience sample

b) Voluntary response

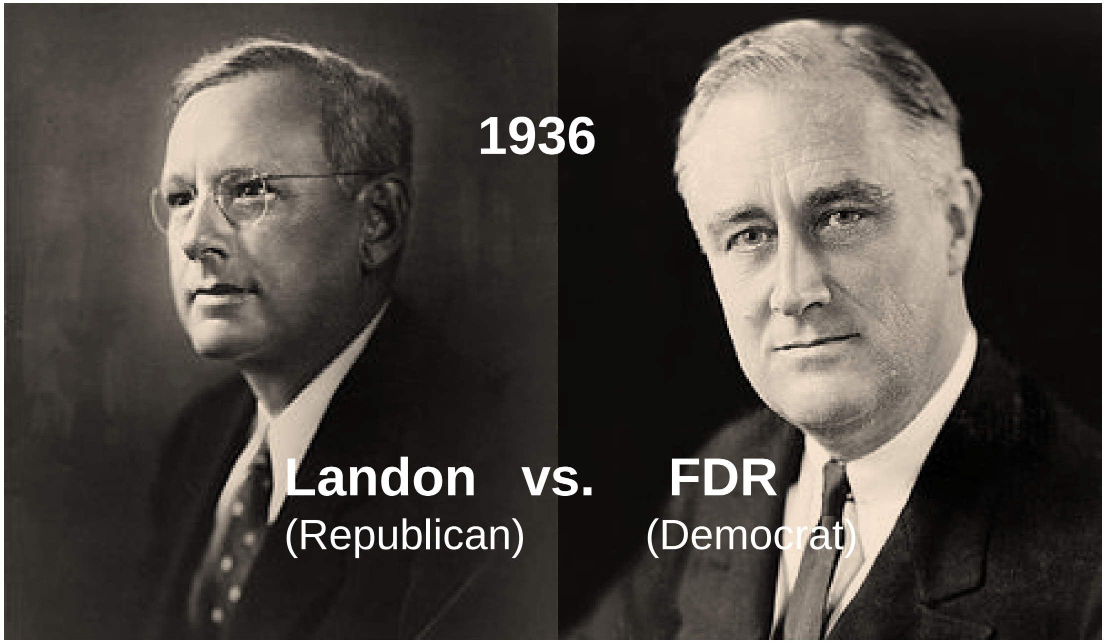
c) Non-response

d) None of the above

There is an initial random sample, but not everyone in this random sample is reached. Therefore the issue is non-response of the sampled individuals

Sampling bias example:

A historical example of a biased sample yielding misleading results

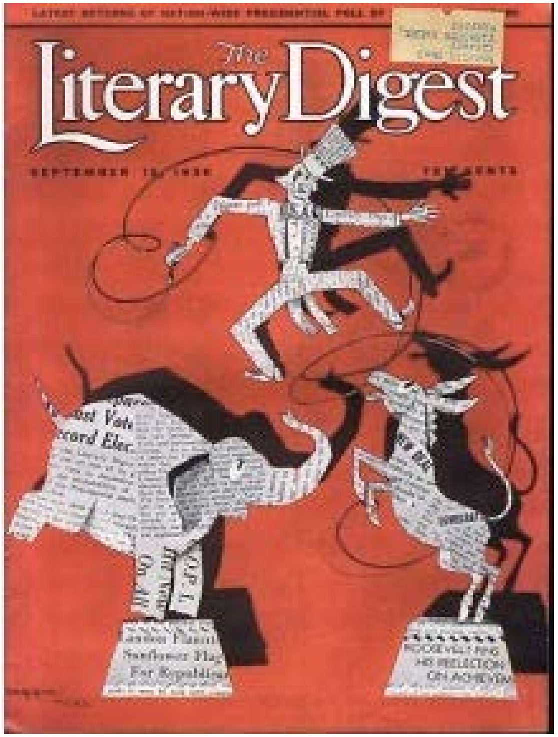


In 1936, Landon sought the Republican presidential nomination opposing the re-election of FDR.

The Literary Digest	Lose with 43% of the votes
Election results	Win with 62% of the votes

- The Literary Digest polled about 10 million Americans, and got responses from about 2.4 million.
- The poll showed that Landon would likely be the overwhelming winner and FDR would get only 43% of the votes.
- Election result: FDR won, with 62% of the votes.
- The magazine was completely discredited because of the poll, and was soon discontinued.

The Literary Digest Poll - what went wrong?



- The magazine had surveyed
  - its own readers,
  - registered automobile owners, and
  - registered telephone users.

- These groups had incomes well above the national average of the day (remember, this is Great Depression era) which resulted in lists of voters far more likely to support Republicans than a truly *typical* voter of the time,

i.e. the sample was not representative of the American population at the time.

Large samples are preferable, but...

The Literary Digest election poll was based on a sample size of 2.4 million, which is huge, but since the sample was *biased*, the sample did not yield an accurate prediction.

Back to the soup analogy:

- If the soup is not well stirred, it doesn't matter how large a spoon you have, it will still not taste right.
- If the soup is well stirred, a small spoon will suffice to test the soup.

Practice

A school district is considering whether it will no longer allow high school students to park at school after two recent accidents where students were severely injured. As a first step, they survey parents by mail, asking them whether or not the parents would object to this policy change. Of 6,000 surveys that go out, 1,200 are returned. Of these 1,200 surveys that were completed, 960 agreed with the policy change and 240 disagreed. Which of the following statements are true?

I. Some of the mailings may have never reached the parents.

II. The school district has strong support from parents to move forward with the policy approval.

III. It is possible that majority of the parents of high school students disagree with the policy change.

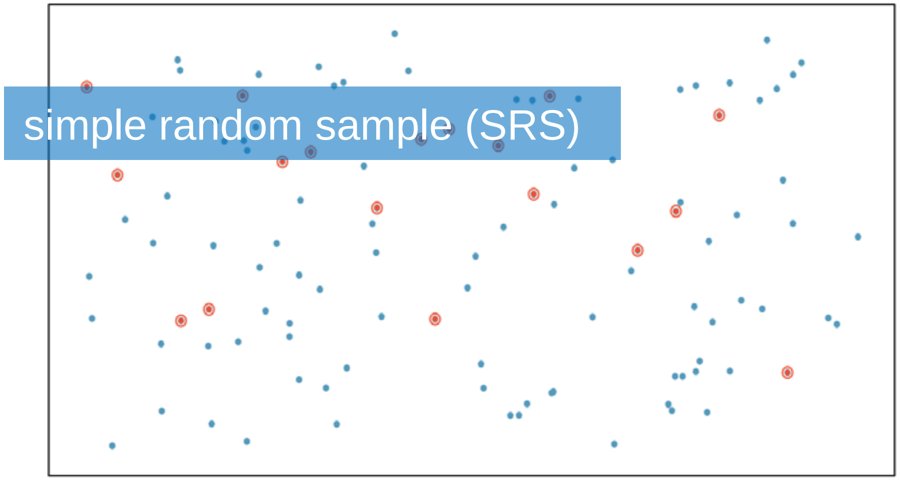
IV. The survey results are unlikely to be biased because all parents were mailed a survey.

(a) Only I (b) I and II [✓ (c) I and III] (d) III and IV (e) Only IV

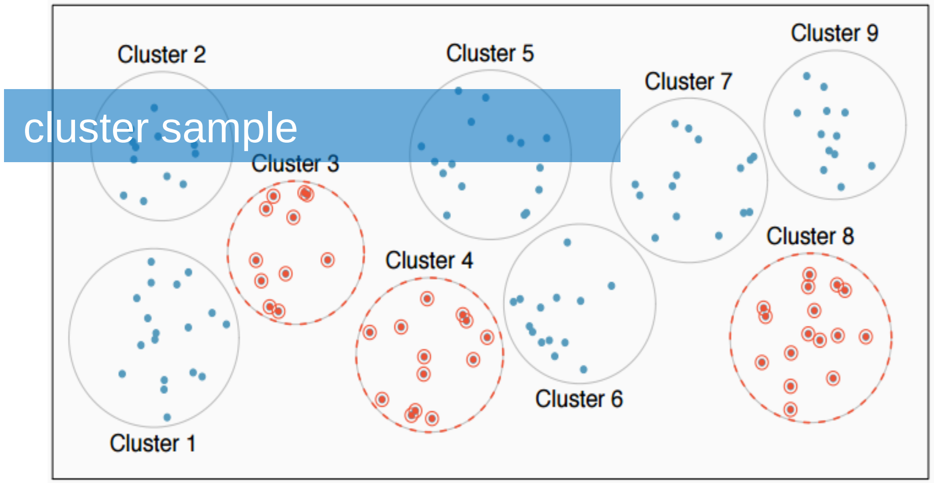
Obtaining good samples

- Almost all statistical methods are based on the notion of implied randomness.
- If observational data are not collected in a random framework from a population, these statistical methods – the estimates and errors associated with the estimates – are not reliable.
- Most commonly used random sampling techniques are *simple*, *stratified*, and *cluster* sampling.

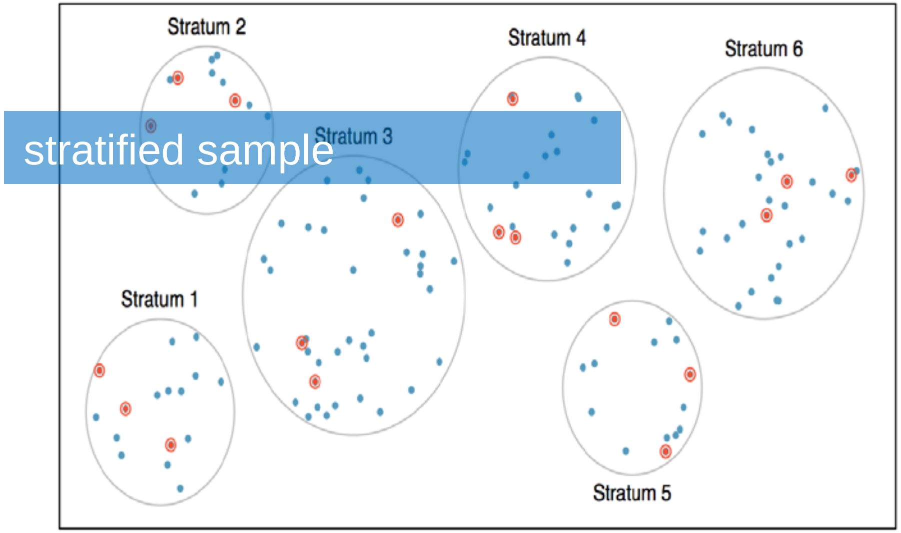
Sampling methods



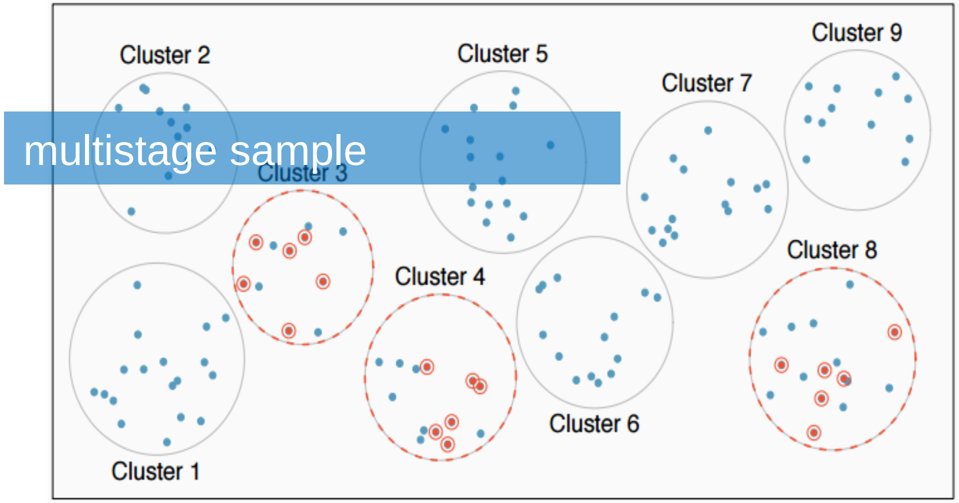
simple random sample (SRS)



cluster sample



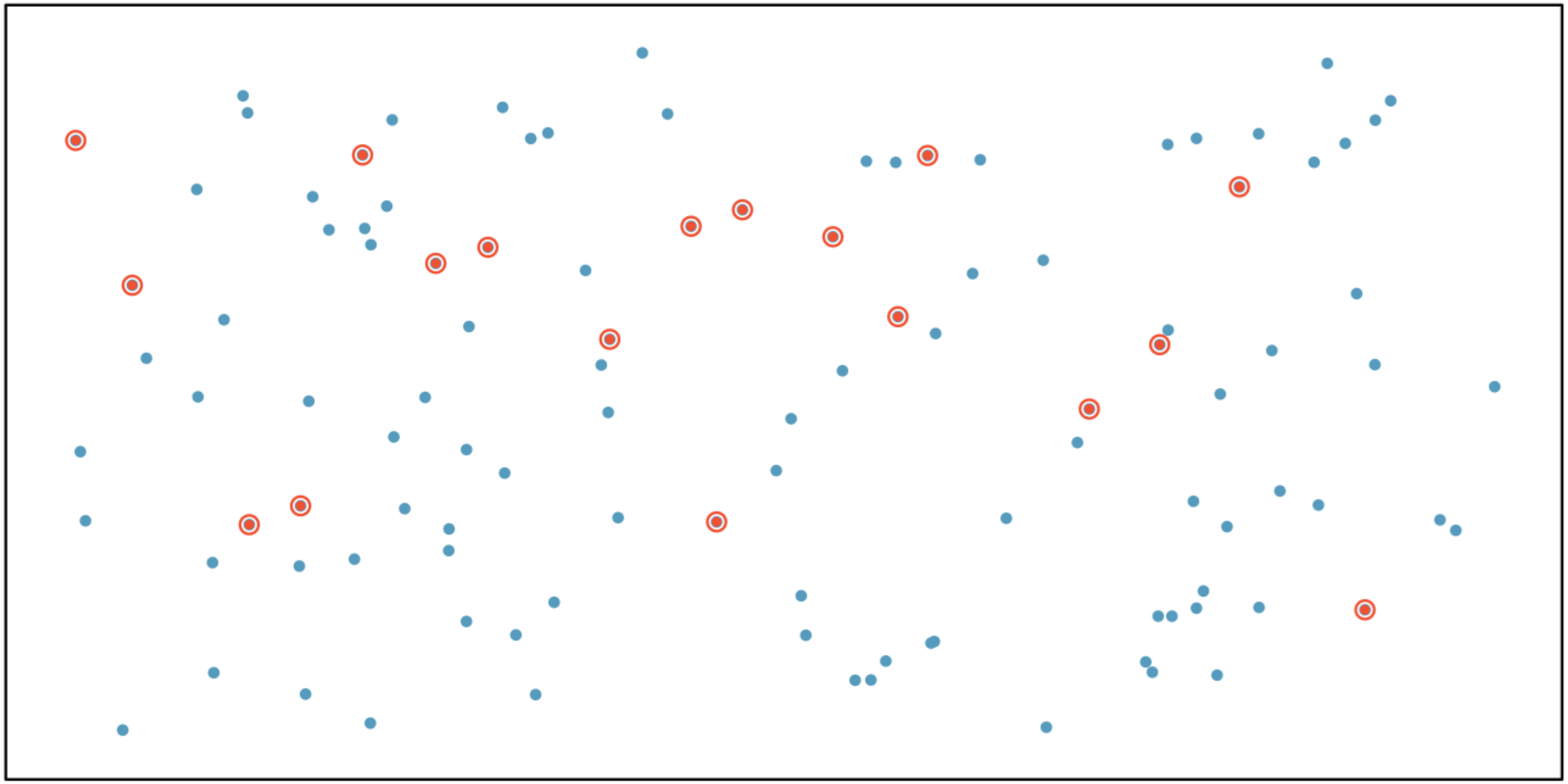
stratified sample



multistage sample

Simple Random Sample

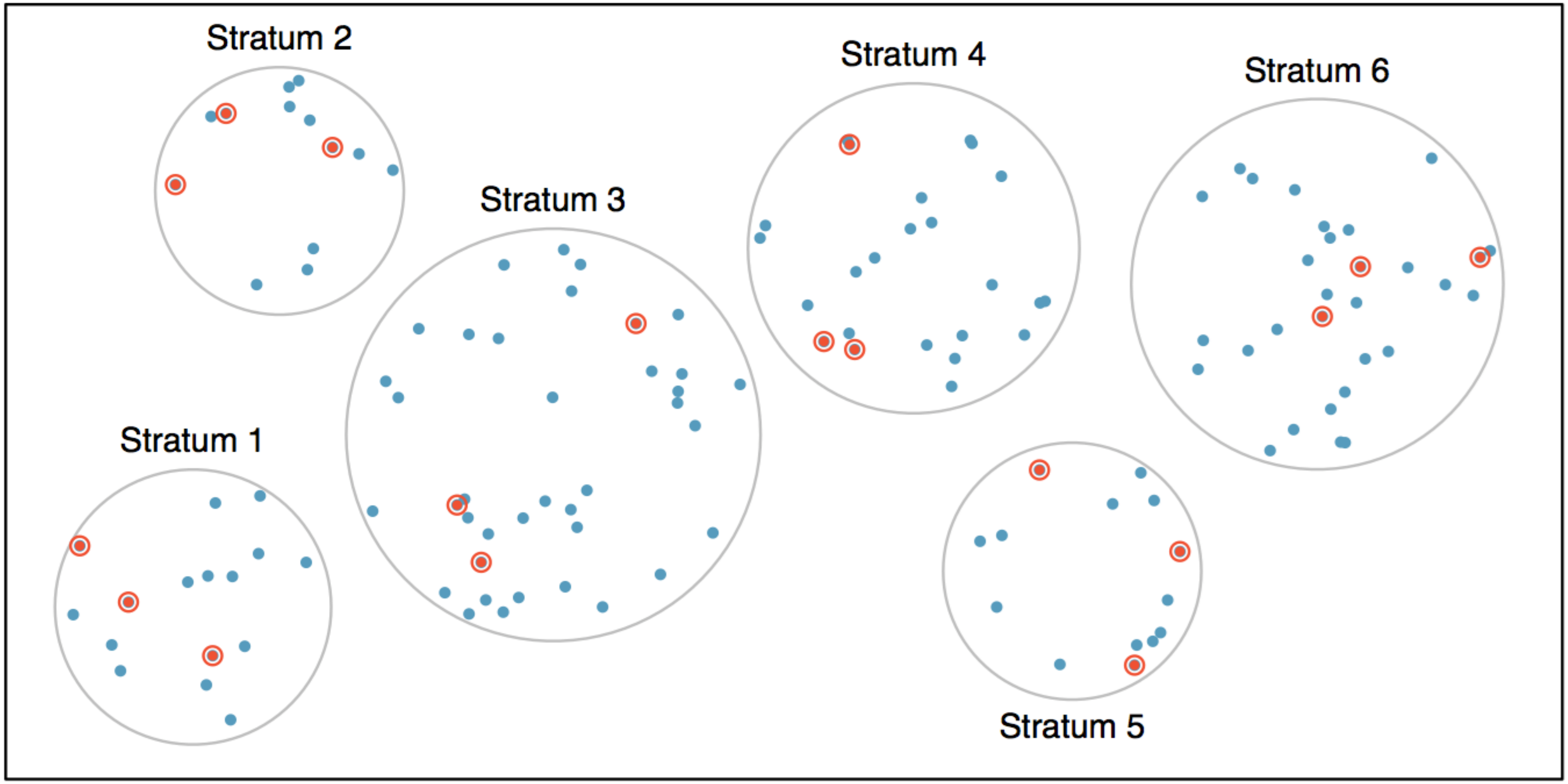
Randomly select cases from the population



equally likely to be selected

Stratified Sample

Strata are made up of similar observations. We take a simple random sample from each stratum.

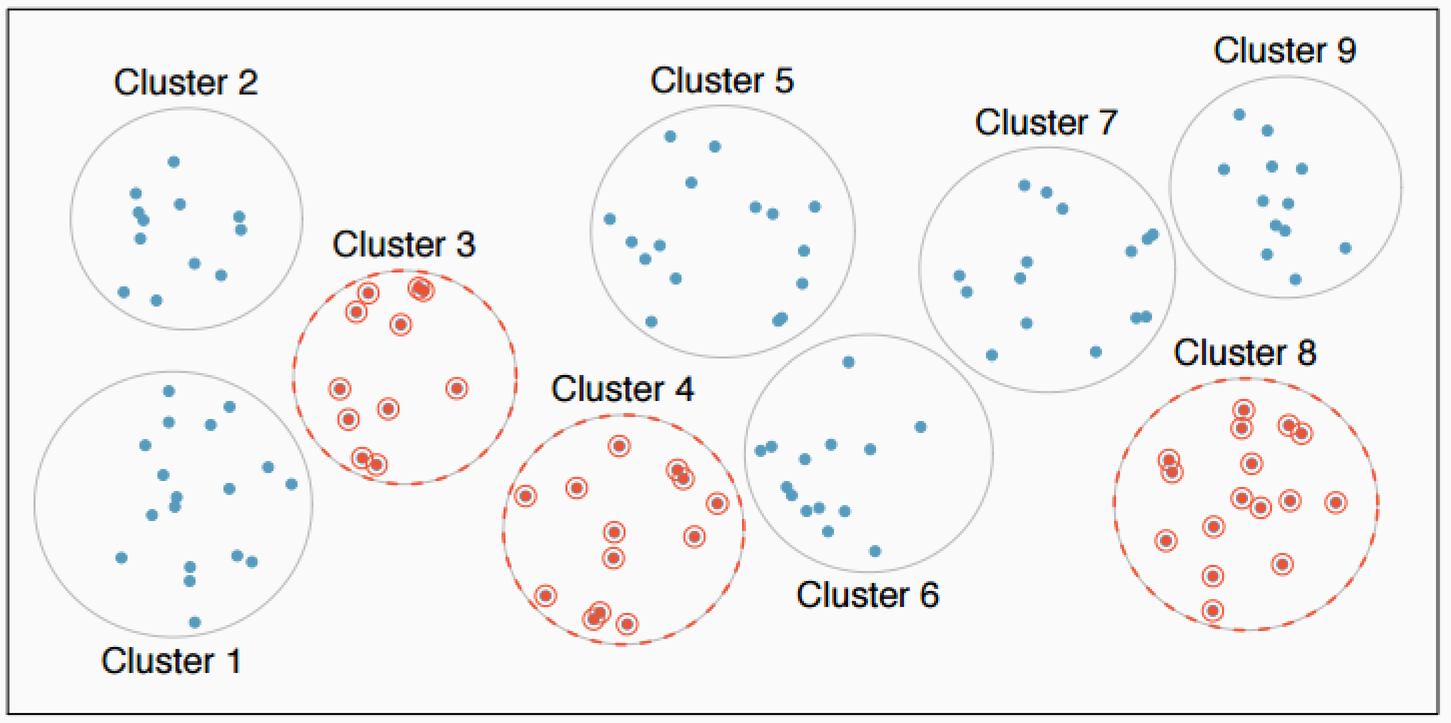


divide the population into homogenous strata,
then randomly sample from with in each stratum

Cluster Sample

Clusters are usually not made up of homogeneous observations.

Usually preferred for economical reasons.



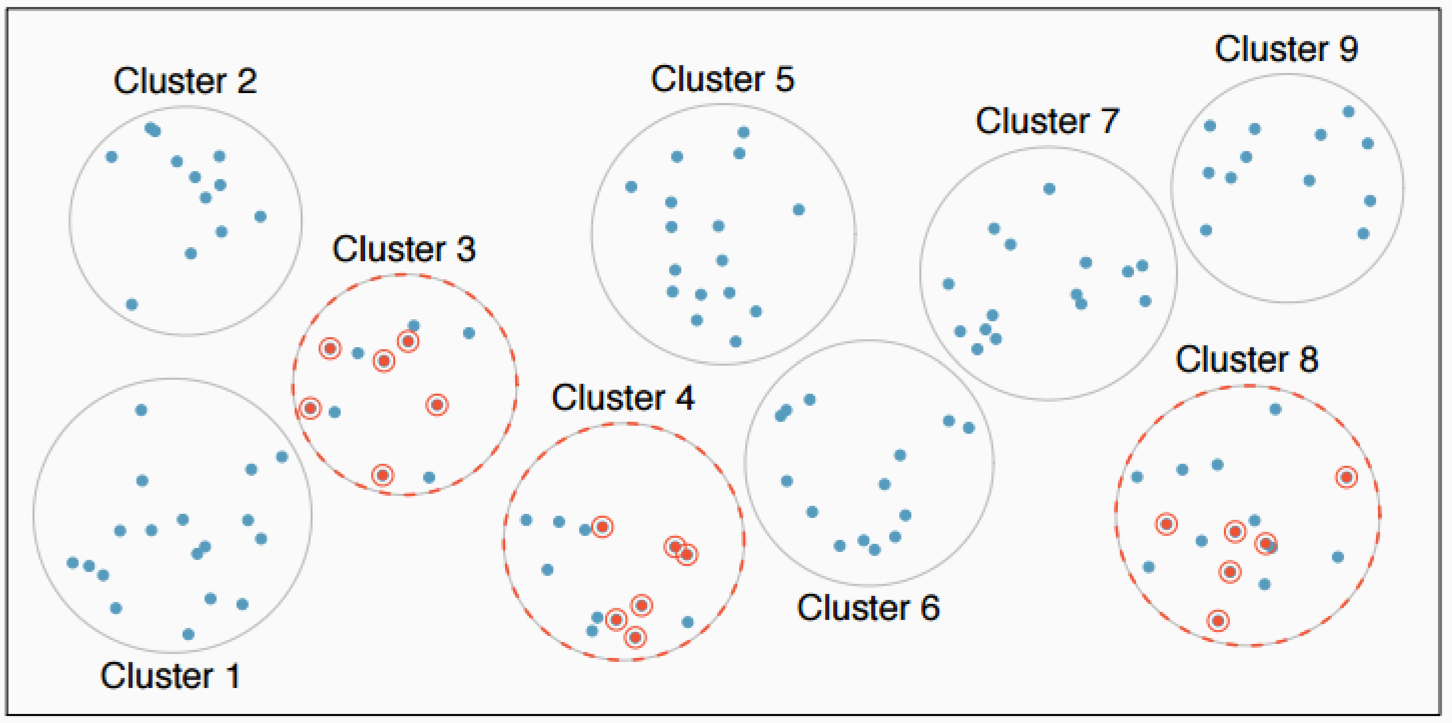
divide the population into clusters,

randomly sample a few clusters,

then sample all observations within these clusters

Multistage Sample

adds another step to cluster sampling.



divide the population clusters,

randomly sample a few clusters,

then randomly sample observations from within these clusters

Practice

A city council has requested a household survey be conducted in a suburban area of their city. The area is broken into many distinct and unique neighborhoods, some including large homes, some with only apartments. Which approach would likely be the *least* effective?

(a) Simple random sampling

[✓ (b) Cluster sampling]

(c) Stratified sampling

(d) Blocked sampling

Homework Week 2 #1:

- Find results from a recent new report about an opinion poll carried out by a news organization such as Gallup, KBS and so on. Briefly describe the sample and how it was chosen. Was the sample chosen in a way that was likely to introduce bias? Explain.