Section 1.3
Sampling

Observational studies

- Researchers collect data in a way that does not directly interfere with how the data arise.
- Results of an observational study can generally be used to establish an association between the explanatory and response variables.

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.

Simple random sample

Randomly select cases from the population, where there is no implied connection between the points that are selected.

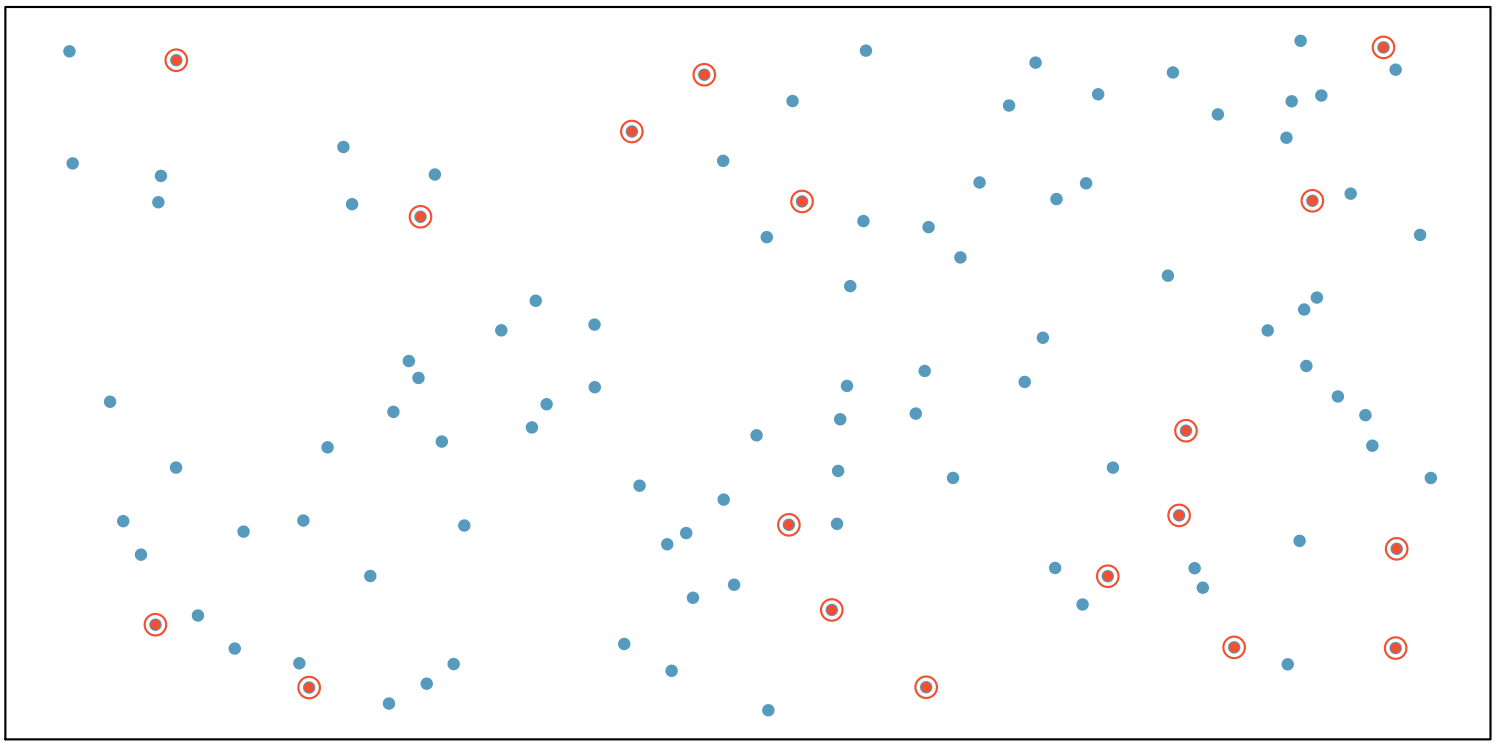


Figure 1.14(a), *OpenIntro Stats*

Stratified sample

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

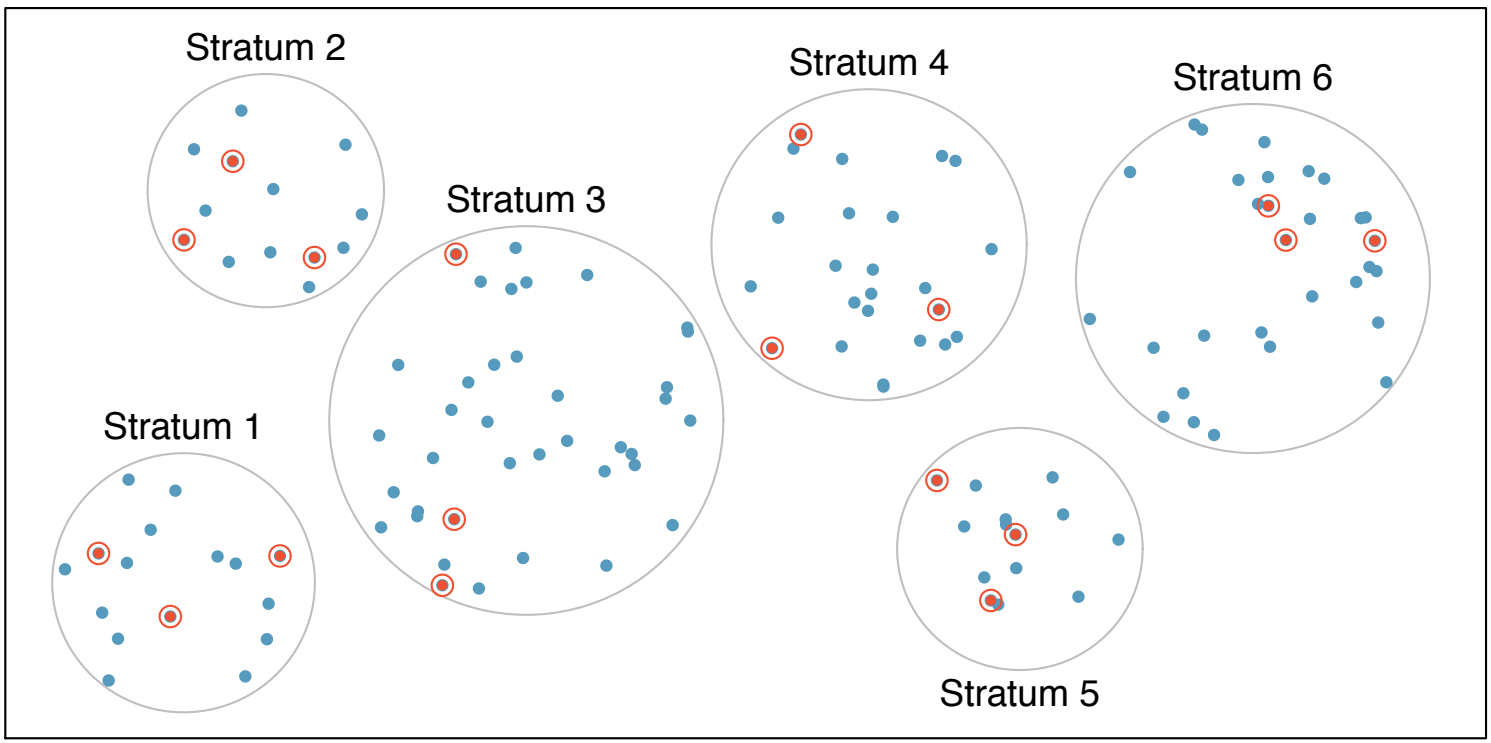


Figure 1.14(b), *OpenIntro Stats*

Cluster sample

Clusters are usually not made up of homogeneous observations. We take a simple random sample of clusters, and then sample all observations in that cluster. Usually preferred for economical reasons.

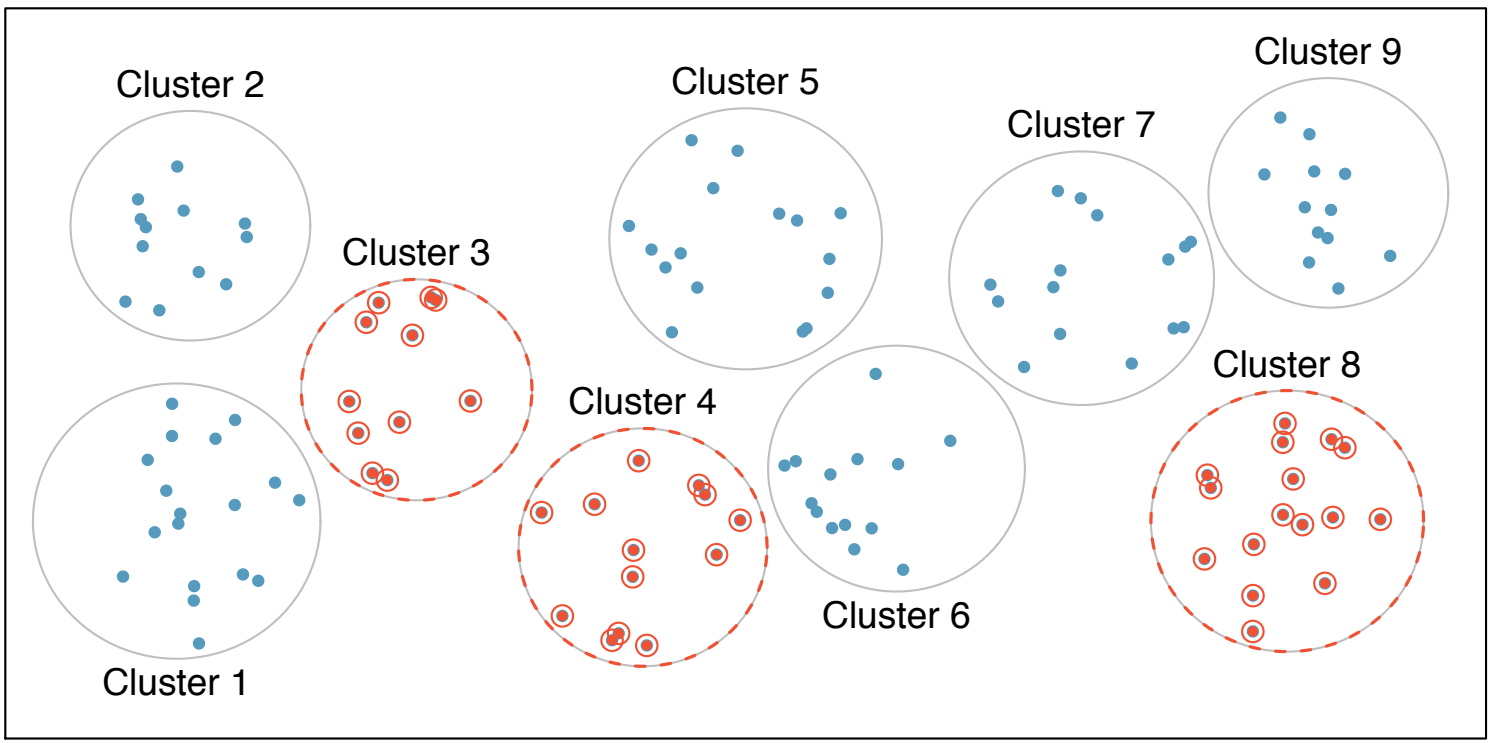


Figure 1.15(a), *OpenIntro Stats*

Multistage sample

Clusters are usually not made up of homogeneous observations. We take a simple random sample of clusters, and then take a simple random sample of observations from the sampled clusters.

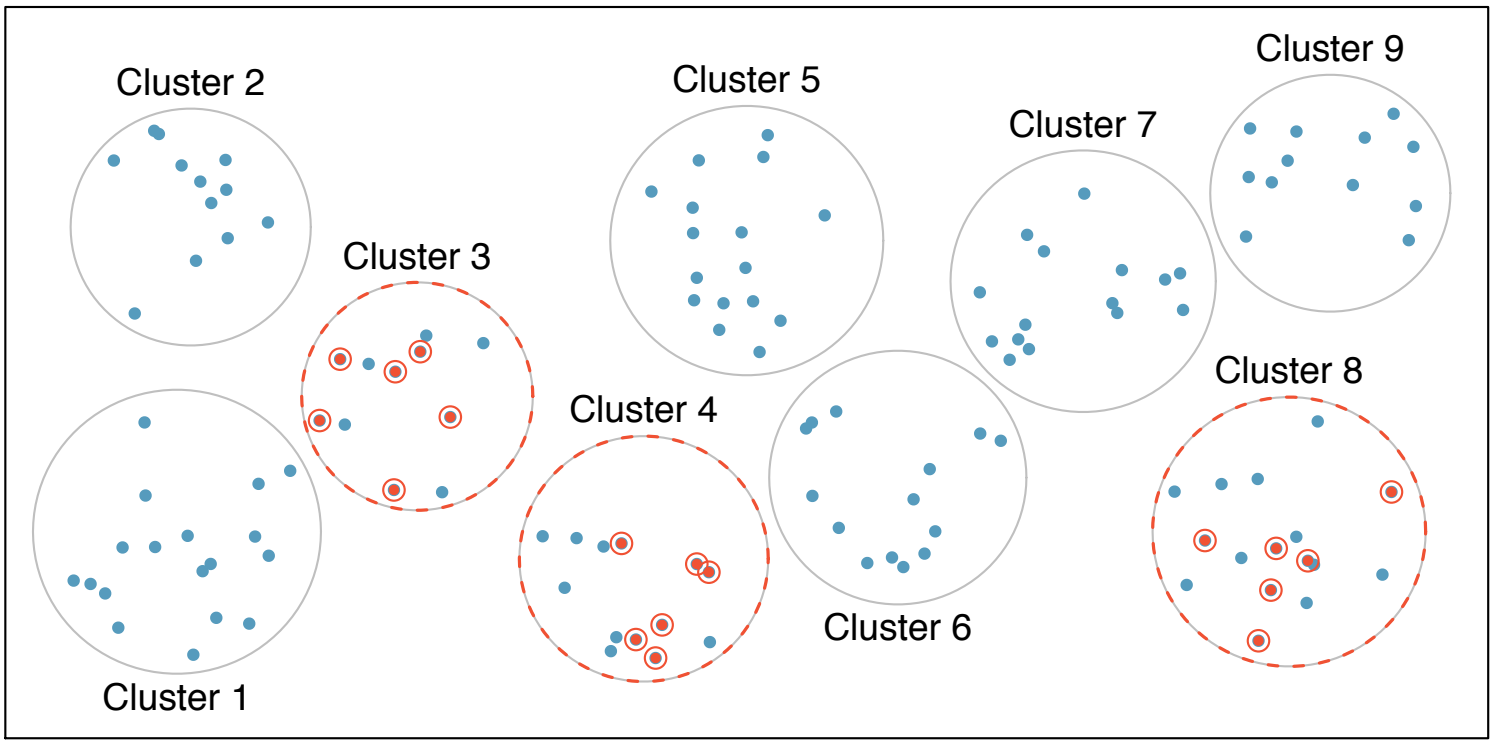


Figure 1.15(b), *OpenIntro Stats*

Example

A city council has requested a household survey be conducted in a suburban area of their city. The area has many distinct and unique neighborhoods, some including large homes, some with only apartments.

Which sampling method would likely be the *least* effective?

(a) Simple random sampling
(b) Cluster sampling
(c) Stratified sampling

Example

A city council has requested a household survey be conducted in a suburban area of their city. The area has many distinct and unique neighborhoods, some including large homes, some with only apartments.

Which sampling method would likely be the *least* effective?

(a) Simple random sampling
(b) *Cluster sampling*
(c) Stratified sampling

Cluster sampling would also likely to be the most convenient for the researchers.

Section 1.4
Experiments

Principles of experimental design

1. *Control:* Control for the (potential) effect of variables other than the ones directly being studied.
2. *Randomize:* Randomly assign subjects to treatments, and randomly sample from the population whenever possible.
3. *Replicate:* Within a study, replicate by collecting a sufficiently large sample. Or replicate the entire study.
4. *Block:* If there are variables that are known or suspected to affect the response variable, first group subjects into *blocks* based on these variables, and then randomize cases within each block to treatment groups.

Example

- We would like to design an experiment to investigate if energy gels makes you run faster:

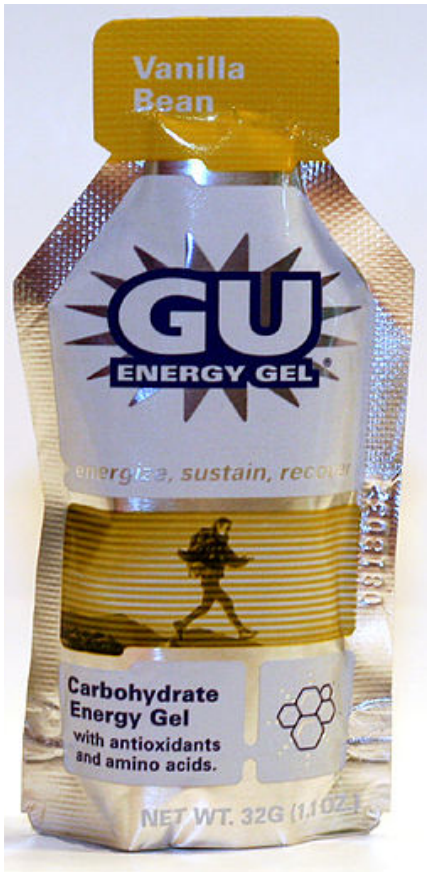
Example

- We would like to design an experiment to investigate if energy gels makes you run faster:
  - Treatment: energy gel
  - Control: no energy gel

Example

- We would like to design an experiment to investigate if energy gels makes you run faster:
  - Treatment: energy gel
  - Control: no energy gel
- It is suspected that energy gels might affect pro and amateur athletes differently, therefore we block for pro status:

Example



- We would like to design an experiment to investigate if energy gels makes you run faster:
  - Treatment: energy gel
  - Control: no energy gel
- It is suspected that energy gels might affect pro and amateur athletes differently, therefore we block for pro status:
  - Divide the sample to pro and amateur
  - Randomly assign pro athletes to treatment and control groups
  - Randomly assign amateur athletes to treatment and control groups
  - Pro/amateur status is equally represented in the resulting treatment and control groups

Example

- We would like to design an experiment to investigate if energy gels makes you run faster:
  - Treatment: energy gel
  - Control: no energy gel
- It is suspected that energy gels might affect pro and amateur athletes differently, therefore we block for pro status:
  - Divide the sample to pro and amateur
  - Randomly assign pro athletes to treatment and control groups
  - Randomly assign amateur athletes to treatment and control groups
  - Pro/amateur status is equally represented in the resulting treatment and control groups

Q: *Why is this important? Can you think of other variables to block for?*

Practice

A study is designed to test the effect of light level and noise level on exam performance of students. The researcher also believes that light and noise levels might have different effects on males and females, so wants to make sure both genders are equally represented in each group. Which of the below is correct?

(a) There are 3 explanatory variables (light, noise, gender) and 1 response variable (exam performance)

(b) There are 2 explanatory variables (light and noise), 1 blocking variable (gender), and 1 response variable (exam performance)

(c) There is 1 explanatory variable (gender) and 3 response variables (light, noise, exam performance)

(d) There are 2 blocking variables (light and noise), 1 explanatory variable (gender), and 1 response variable (exam performance)

Practice

A study is designed to test the effect of light level and noise level on exam performance of students. The researcher also believes that light and noise levels might have different effects on males and females, so wants to make sure both genders are equally represented in each group. Which of the below is correct?

(a) There are 3 explanatory variables (light, noise, gender) and 1 response variable (exam performance)

(b) *There are 2 explanatory variables (light and noise), 1 blocking variable (gender), and 1 response variable (exam performance)*

(c) There is 1 explanatory variable (gender) and 3 response variables (light, noise, exam performance)

(d) There are 2 blocking variables (light and noise), 1 explanatory variable (gender), and 1 response variable (exam performance)

Difference between blocking and explanatory variables

- Factors are conditions we can impose on the experimental units.
- Blocking variables are characteristics that the experimental units come with, that we would like to control for.
- Blocking is like stratifying, except used in experimental settings when randomly assigning, as opposed to when sampling.

More experimental design terminology...

- *Placebo:* fake treatment, often used as the control group for medical studies
- *Placebo effect:* experimental units showing improvement simply because they believe they are receiving a special treatment
- *Blinding:* when experimental units do not know whether they are in the control or treatment group
- *Double-blind:* when both the experimental units and the researchers who interact with the patients do not know who is in the control and who is in the treatment group

Practice

What is the main difference between observational studies and experiments?

(a) Experiments take place in a lab while observational studies do not need to.

(b) In an observational study we only look at what happened in the past.

(c) Most experiments use random assignment while observational studies do not.

(d) Observational studies are completely useless since no causal inference can be made based on their findings.

Practice

What is the main difference between observational studies and experiments?

(a) Experiments take place in a lab while observational studies do not need to.

(b) In an observational study we only look at what happened in the past.

(c) *Most experiments use random assignment while observational studies do not.*

(d) Observational studies are completely useless since no causal inference can be made based on their findings.