Chapter 1: Introduction to Data

Math 140

Hitchman

based on content in OpenIntro Stats, 4th Ed

August 30, 2022

Section 1.1
A Case Study

Case Study: Using stents to prevent strokes

Identify a question

Stents are devices put inside blood vessels that assist in patient recovery after cardiac events and reduce the risk of an additional heart attack or death. Many doctors have hoped that there would be similar benefits for patients at risk of stroke.

Does the use of stents reduce the risk of stroke?

Case Study: Using stents to prevent strokes

Collect Relevant Data on the Subject

Researchers conducted an experiment with 451 at-risk patients. Each volunteer patient was randomly assigned to one of two groups

- **Treatment Group**: Patients in this group received a stent and medical management. The medical management included medications, management of risk factors, and help in lifestyle modification.
- **Control group**: Patients in this group received the same medical management as the treatment group, but they did not receive stents.

Why two groups?
The control group provides a reference point against which we can measure the medical impact of stents in the treatment group.

Case Study: Using stents to prevent strokes

Collecting data in control and treatment groups

- Researchers randomly assigned 224 patients to the treatment group and 227 to the control group.
- They studied the effect of stents at two time points: 30 days after enrollment and 365 days after enrollment.
- Patient outcomes are recorded as "stroke" or "no event", representing whether or not the patient had a stroke at the end of a time period.

Case Study: Using stents to prevent strokes

The Data

Patient	group	0-30 days	0-365 days
1	treatment	no event	no event
2	treatment	stroke	stroke
3	treatment	no event	no event
⋮	⋮	⋮	
450	control	no event	no event
451	control	no event	no event

Case Study: Using stents to prevent strokes

Summarizing the data at 365 days

group	no event	stroke
control	199	28
treatment	179	45

Thoughts?
Proportion of patients in control group with a stroke in a year:

$$28/(199+28) \approx 0.123, \quad \text{or} \quad 12.3\%.$$

Proportion of patients in treatment group with a stroke in a year:

$$45/(179+45) \approx 0.201, \quad \text{or} \quad 20.1\%.$$

Case Study: Using stents to prevent strokes

Form a Conclusion

Does the use of stents reduce the risk of stroke?

- The researchers expected to find that stents helped reduce the rate of stroke.
- Perhaps it's the other way around?

Case Study: Using stents to prevent strokes

Two possible conclusions

1. Stents *do help* reduce the rate of stroke, and we just happened to observe a sample with an unusually high number of strokes
2. Stents *do not help* reduce the rate of stroke.
3. Other possibilities?

Key Statistical Question

Is the observed difference due to chance, or do the data show a "real" difference?

Simulation: A tool with which to address this key question

Intermission: A first look at RStudio as a simulation tool

- Simulation results

What is Statistics?

We may place statistics in the context of an investigative process:

1. Identify a question or problem.
2. Collect relevant data on the topic.
3. Analyze the data.
4. Form a conclusion.
5. Make decisions based on the conclusion.

What is Statistics?

We may place statistics in the context of an investigative process:

1. Identify a question or problem.
2. Collect relevant data on the topic.
3. Analyze the data.
4. Form a conclusion.
5. Make decisions based on the conclusion.

Statistics as a subject focuses on making steps 2-4 objective, rigorous, and efficient.

What is Statistics?

In other words, statistics has three primary components:

- How best can we collect data?
- How should it be analyzed?
- And what can we infer from the analysis?

Section 1.2
Data Basics

Storing Data

Definition

A **data matrix** is a 2-dimensional array, in which each row corresponds to an **observational unit** (individual cases) and each column corresponds to a **variable** (that is being measured).

Storing Data

	Name	height (in)	Age (yrs)	FavNum	TimeinUD (ep)	TeleAbil
1	Dustin	61.3	12	π	0	no
2	Will	61.9	12	4	7.2	no
3	Lucas	63.8	13	8	0	no
4	Eleven	62.1	11	315	nan	yes
5	Mike	64.3	13	11	0	no

- 5 observations (characters in *Stranger Things*)
- 6 variables
- Each variable gives a piece of information about each character (name, height, time spent in the upside down (in units of 'episodes'), etc).
- The data matrix in this example has 5 rows and 6 columns

Data Matrix

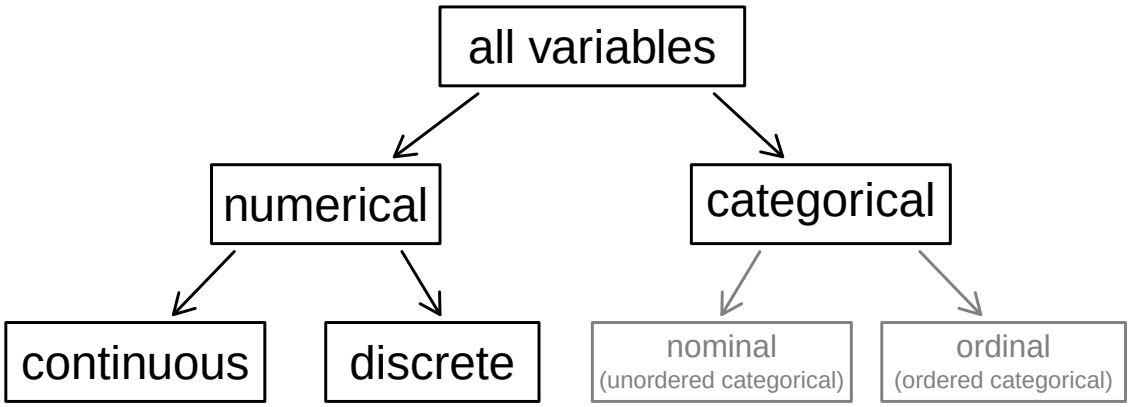
- Convenient and common way to organize data
- Spreadsheets!
- This structure allows new cases to be added as rows and/or new variables as new columns.

Example

We consider data for 3,142 counties in the United States, which includes each county's name, the state where it resides, its population in 2017, how its population changed from 2010 to 2017, poverty rate, and six additional characteristics. How might these data be organized in a data matrix? How many rows, how many columns in this matrix?

ANSWER: Each county is a case, giving a row in the matrix. The characteristics recorded (i.e., the variables) give columns in the matrix. So the matrix will have 3,142 rows and 11 columns, one for each of the variables

Types of Variables



Section 1.2.2 has a nice discussion of these terms. Generally speaking, if it makes sense to "do math" on a variable (like add, subtract, find the average), the variable is better thought of as numerical than categorical.
NOTE: A numerical variable is sometimes called a quantitative variable.

Figure 1.7 in *OpenIntro Stats*

Example (*Stranger Things*)

	Name	height (in)	Age (yrs)	FavNum	TimeinUD (ep)	TeleAbil
1	Dustin	61.3	12	π	0	no
2	Will	61.9	12	4	7.2	no
3	Lucas	63.8	13	8	0	no
4	Eleven	62.1	11	315	nan	yes
5	Mike	64.3	13	11	0	no

- Categorical (nominal): 'Name', 'TeleAbil'
- Numerical (discrete): 'Age'
- Numerical (continuous): 'height', 'FavNum', 'TimeinUD'
- Note: 'nan' commonly used as a place holder for a missing data value.

Relationships between variables

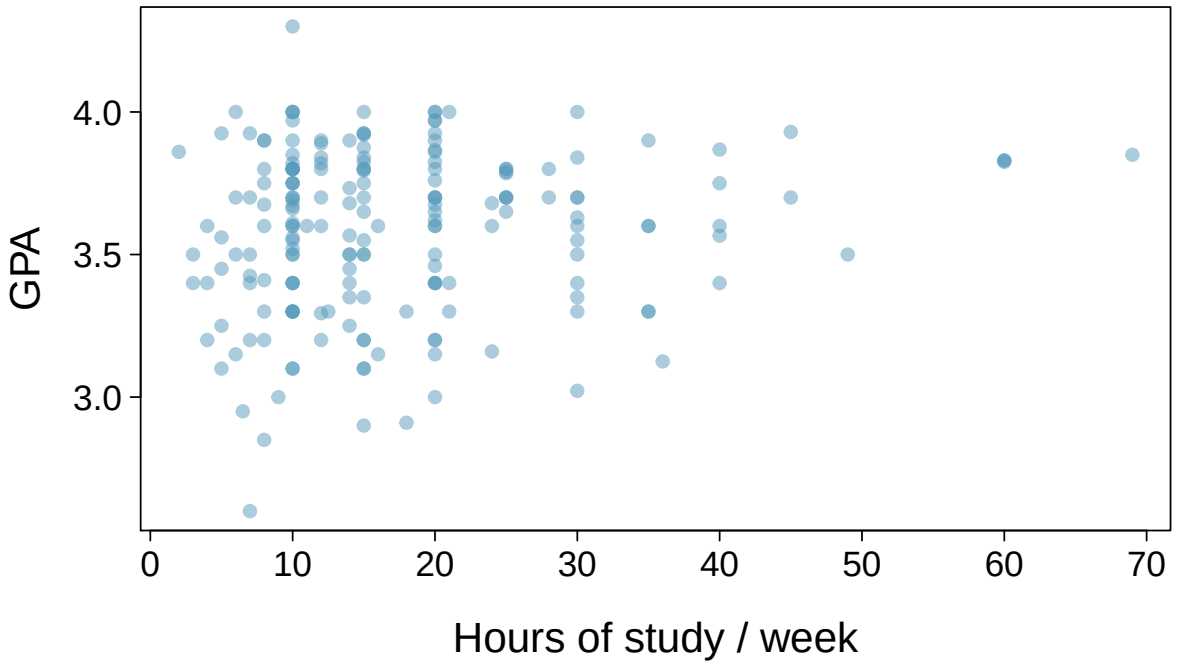
- Many analyses are motivated by a researcher looking for a relationship between two or more variables.
- If two variables are numerical, a **scatterplot** gives a visual description of any association between them.
- Categorical variables can be pictured in scatterplots as well, perhaps with colors or dot sizes!

Check out gapminder

Relationships between variables

- When two variables show some connection with one another, they are called **associated variables**.
- Two numerical variables are **negatively associated** if one tends to decrease as the other increases,
- and are **positively associated** if one tends to increase as the other increases.
- If two variables do not appear to be associated, they are said to be **independent**.

Relationships among variables



Does there appear to be a relationship between GPA and number of hours students study per week?

Can you spot anything unusual about any of the data points?

ANSWER: There is one student with GPA > 4.0, this is likely a data error.

Explanatory and response variables

- When we suspect one variable might causally affect another, we label the first variable the **explanatory variable** and the second the **response variable**.

$$\text{explanatory variable} \xrightarrow{\textit{might affect}} \text{response variable}$$

- Labeling variables as explanatory and response does not guarantee a causal relationship, of course. We use these labels only to keep track of which variable we suspect affects the other.

Explanatory and response variables

Example (Migraines and Acupuncture)

- The patients in the treatment group received acupuncture that was specifically designed to treat migraines.
- The patients in the control group received placebo acupuncture (needle insertion at non-acupoint locations).
  24 hours after patients received acupuncture, they were asked if they were pain free.
- What are the explanatory and response variables in this study?

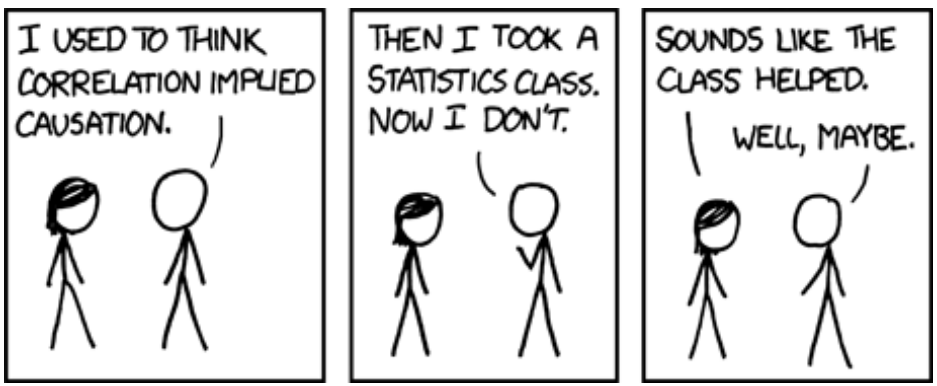
Explanatory: type of acupuncture; Response: pain free (y or n)

Association does not mean Causation

- Is there an association between the percentage of people in a country not using the internet and the life expectancy? Does failure to use the internet decrease life expectancy?
- gapminder

Association vs. causation

- When two variables show some connection with one another, they are called **associated** variables.
- If two variables are not associated, i.e. there is no evident connection between the two, then they are said to be **independent**.
- In general, association does not imply causation, and causation can only be inferred from a randomized experiment...



http://xkcd.com/552/

Observational Studies vs Experiments

Two primary types of data collection:

- Researchers perform an **observational study** when they collect data in a way that does not directly interfere with how the data arise.
- When researchers want to investigate the possibility of a causal connection, they conduct an **experiment**, which generally has explanatory and response variables.

Observational Studies vs Experiments

Example (Does drinking alcohol affect body temperature?)

- Researchers give varying amounts of alcohol to volunteer subjects
- They measure the change in each subject's temperature in the 15 minutes after taking the alcohol.
- Observational study or Experiment?

Experiment. The amount of alcohol consumed is the explanatory variable, change in body temp is the response.

Observational Studies vs Experiments

Example (Do biology majors spend more on textbooks than psychology majors?)

- Find 10 bio students and 10 psych students and compare average costs on books.
- Observational study or Experiment?

Observational study. "Major" is the explanatory variable, "cost of books" the response.

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.

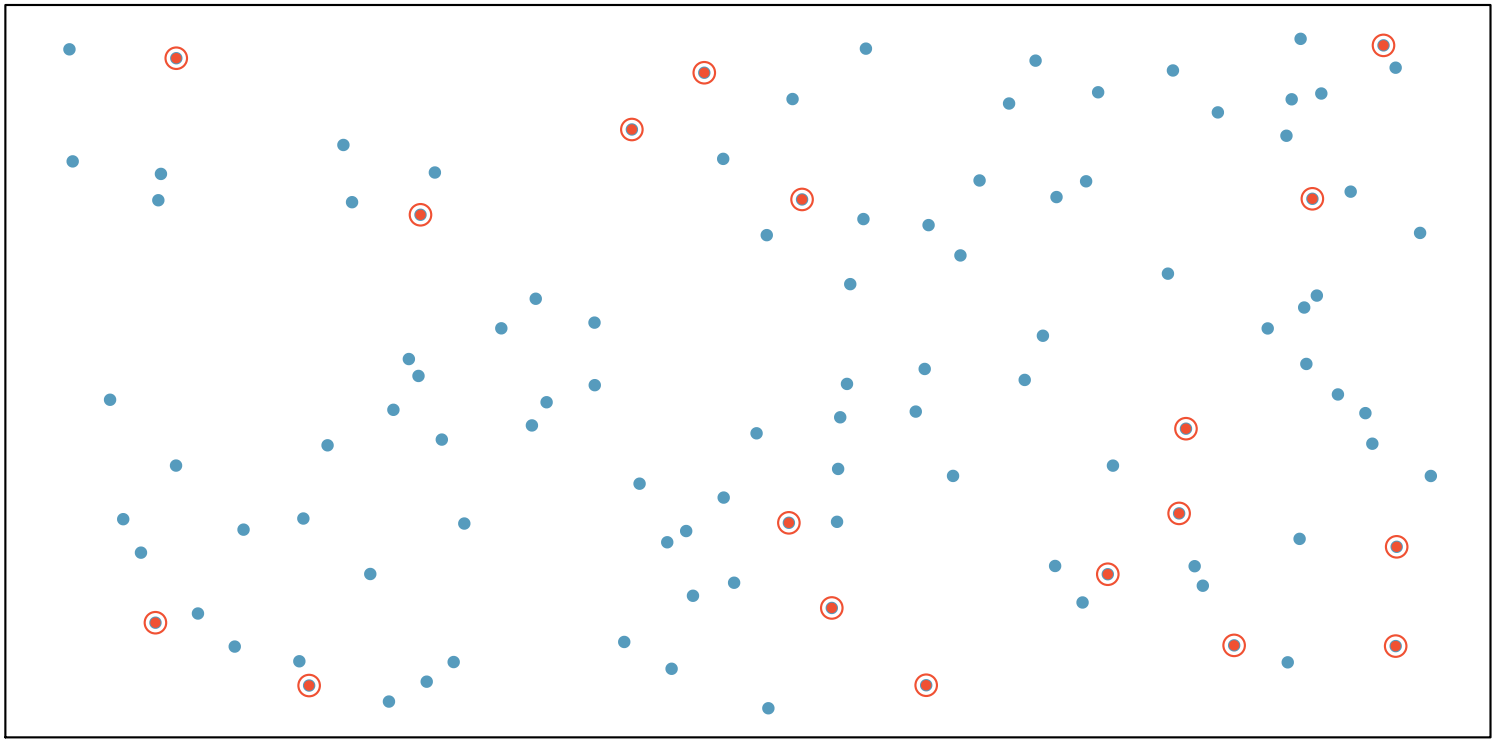


Figure 1.14(a), *OpenIntro Stats*

Stratified sample

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

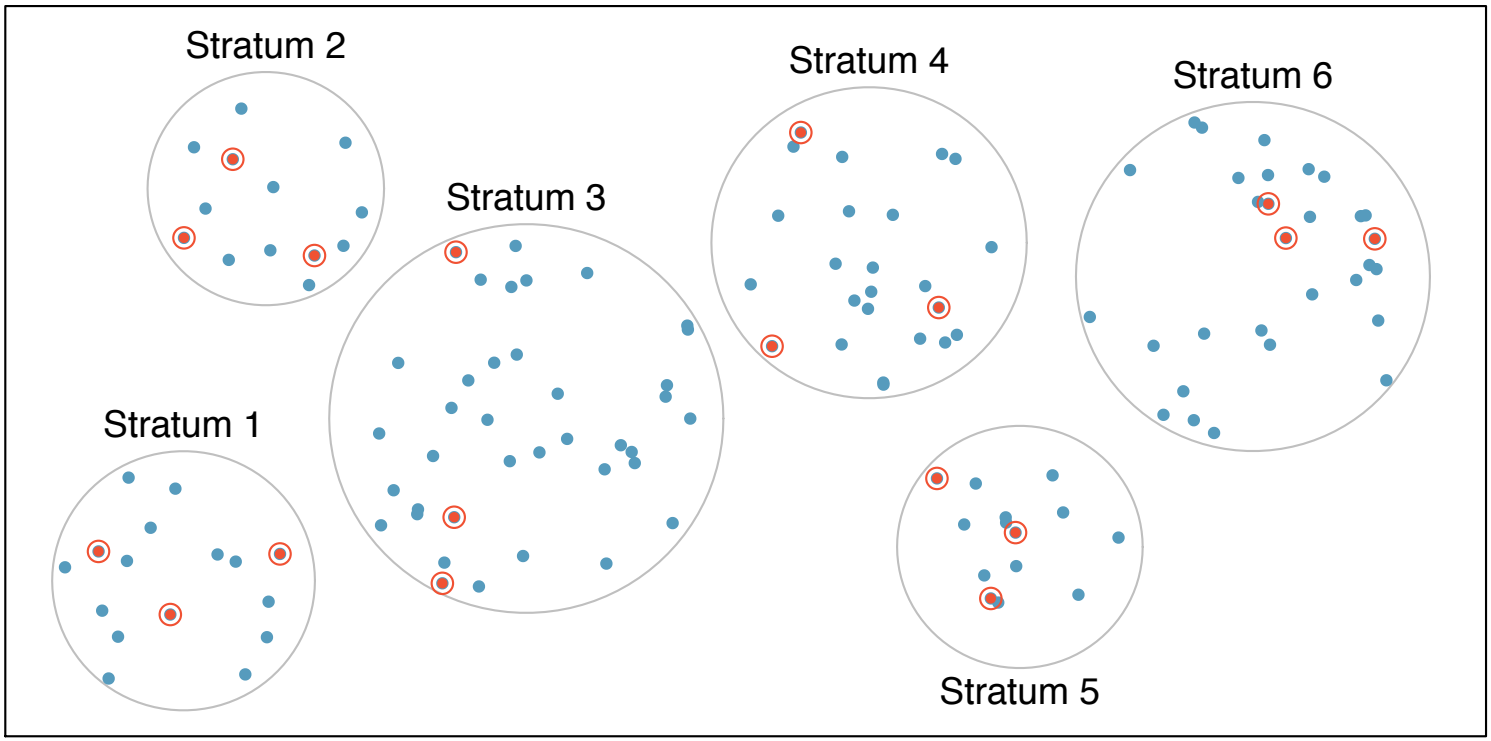


Figure 1.14(b), *OpenIntro Stats*

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.

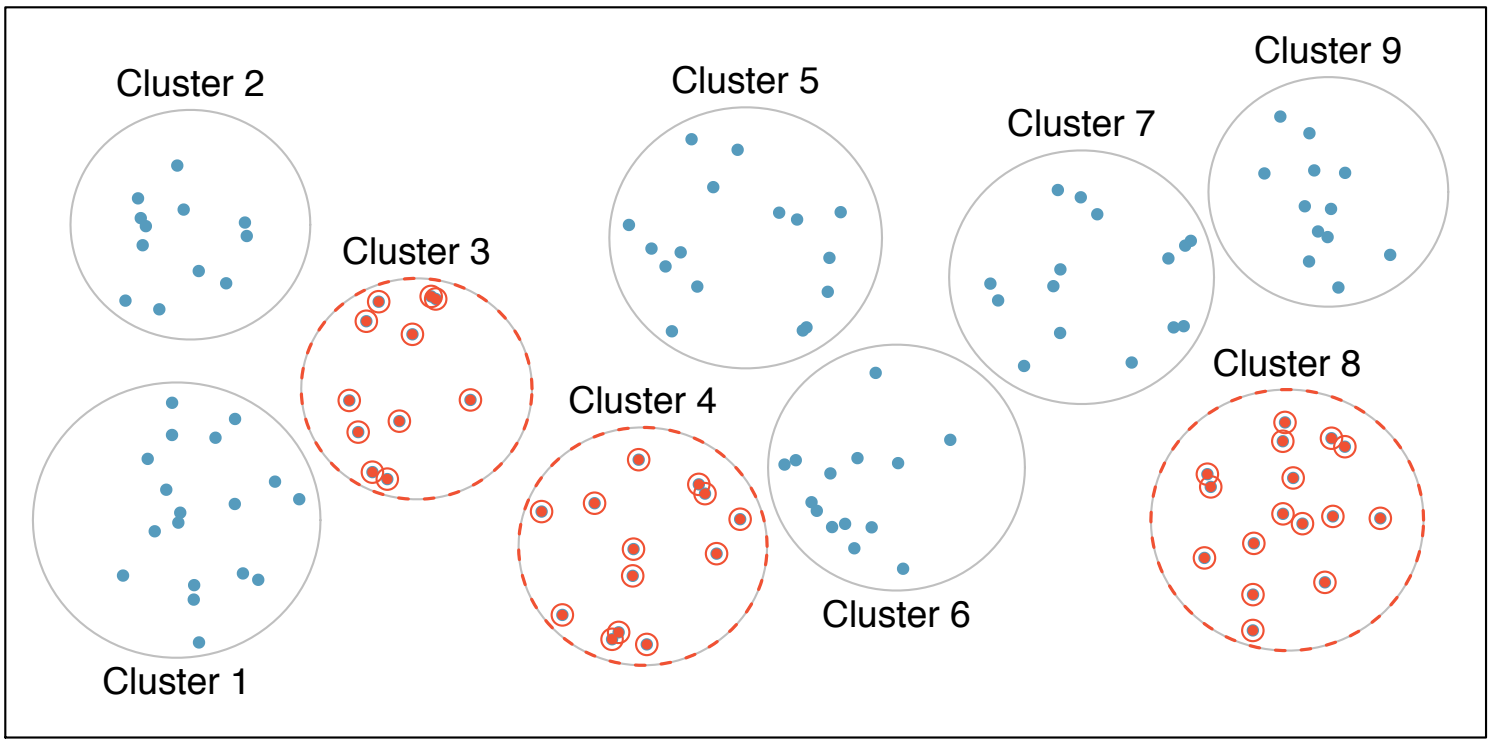


Figure 1.15(a), *OpenIntro Stats*

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.

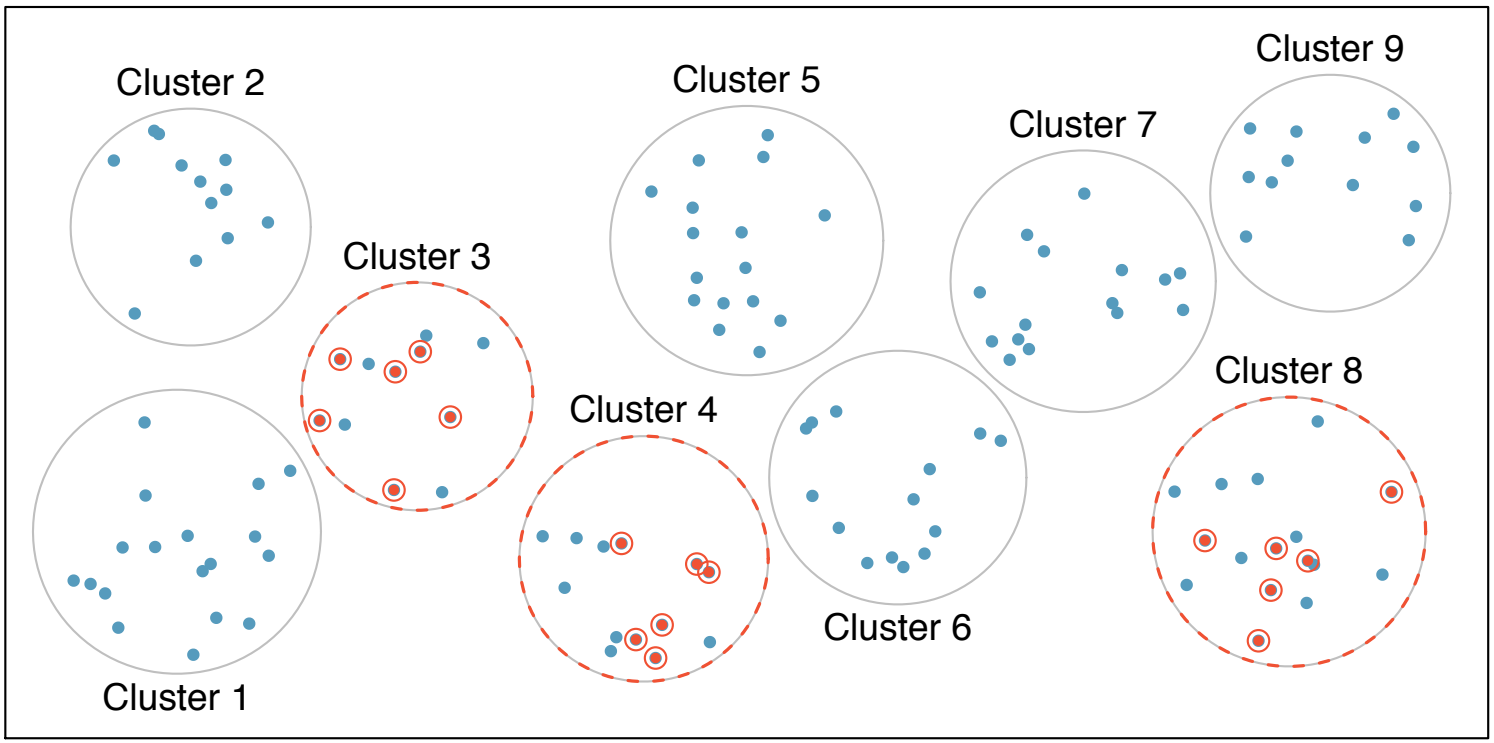


Figure 1.15(b), *OpenIntro Stats*

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) *Cluster sampling*
(d) Stratified sampling

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

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

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

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

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