# Data Analytics with R

Sumit Mishra

Institute for Financial Management and Research, Sri City

**Introduction to Data**

28 October 2020

# Agenda

- Experiments Case Study: using stents to prevent strokes
- Data basics.
- Sampling principles and strategies.

# Case: Stents to prevent strokes

# Using stents to prevent strokes

| 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 |

Figure 1: Results for five patients from the stent study.

# Using stents to prevent strokes

| | 0-30 days | | 0-365 days | |
|---|---|---|---|---|
| | stroke | no event | stroke | no event |
| treatment | 33 | 191 | 45 | 179 |
| control | 13 | 214 | 28 | 199 |
| Total | 46 | 405 | 73 | 378 |

Figure 2: Descriptive statistics for the stent study.

# Using stents to prevent strokes

| | 0-30 days | | 0-365 days | |
|---|---|---|---|---|
| | stroke | no event | stroke | no event |
| treatment | 33 | 191 | 45 | 179 |
| control | 13 | 214 | 28 | 199 |
| Total | 46 | 405 | 73 | 378 |

Figure 2: Descriptive statistics for the stent study.

Exercise: Calculate the proportion of people who had a stroke in the treatment and control groups.

# Data Basics

# Observations, variables, and data matrices

Consider the following table.

| | loan_amount | interest_rate | term | grade | state | total_income | homeownership |
|---|---|---|---|---|---|---|---|
| 1 | 7500 | 7.34 | 36 | A | MD | 70000 | rent |
| 2 | 25000 | 9.43 | 60 | B | OH | 254000 | mortgage |
| 3 | 14500 | 6.08 | 36 | A | MO | 80000 | mortgage |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 50 | 3000 | 7.96 | 36 | A | CA | 34000 | rent |

Figure 3: Four rows from the `loan50` data matrix.

# Observations, variables, and data matrices

Consider the following table.

| | loan_amount | interest_rate | term | grade | state | total_income | homeownership |
|---|---|---|---|---|---|---|---|
| 1 | 7500 | 7.34 | 36 | A | MD | 70000 | rent |
| 2 | 25000 | 9.43 | 60 | B | OH | 254000 | mortgage |
| 3 | 14500 | 6.08 | 36 | A | MO | 80000 | mortgage |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 50 | 3000 | 7.96 | 36 | A | CA | 34000 | rent |

Figure 3: Four rows from the `loan50` data matrix.

- Each row represents a single loan.
  - The formal name: observation

# Observations, variables, and data matrices

Consider the following table.

| | loan_amount | interest_rate | term | grade | state | total_income | homeownership |
|---|---|---|---|---|---|---|---|
| 1 | 7500 | 7.34 | 36 | A | MD | 70000 | rent |
| 2 | 25000 | 9.43 | 60 | B | OH | 254000 | mortgage |
| 3 | 14500 | 6.08 | 36 | A | MO | 80000 | mortgage |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 50 | 3000 | 7.96 | 36 | A | CA | 34000 | rent |

Figure 3: Four rows from the `loan50` data matrix.

- Each row represents a single loan.
  - The formal name: observation
- Each header represents characteristics of each loan.
  - The formal name: variable

# Observations, variables, and data matrices

Consider the following table.

| | loan_amount | interest_rate | term | grade | state | total_income | homeownership |
|---|---|---|---|---|---|---|---|
| 1 | 7500 | 7.34 | 36 | A | MD | 70000 | rent |
| 2 | 25000 | 9.43 | 60 | B | OH | 254000 | mortgage |
| 3 | 14500 | 6.08 | 36 | A | MO | 80000 | mortgage |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 50 | 3000 | 7.96 | 36 | A | CA | 34000 | rent |

Figure 3: Four rows from the `loan50` data matrix.

- Each row represents a single loan.
  - The formal name: observation
- Each header represents characteristics of each loan.
  - The formal name: variable

- What does the first row represent?
  - Loan amount: $ 7500
  - Borrower's location: MD (Maryland)
  - Interest rate on the loan: 7.34%.

# Practice

Construct the gradebook for a course in R assuming that there are five students, two assignments, and one exam.

# Types of variables

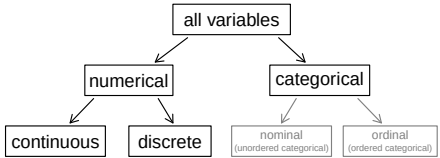
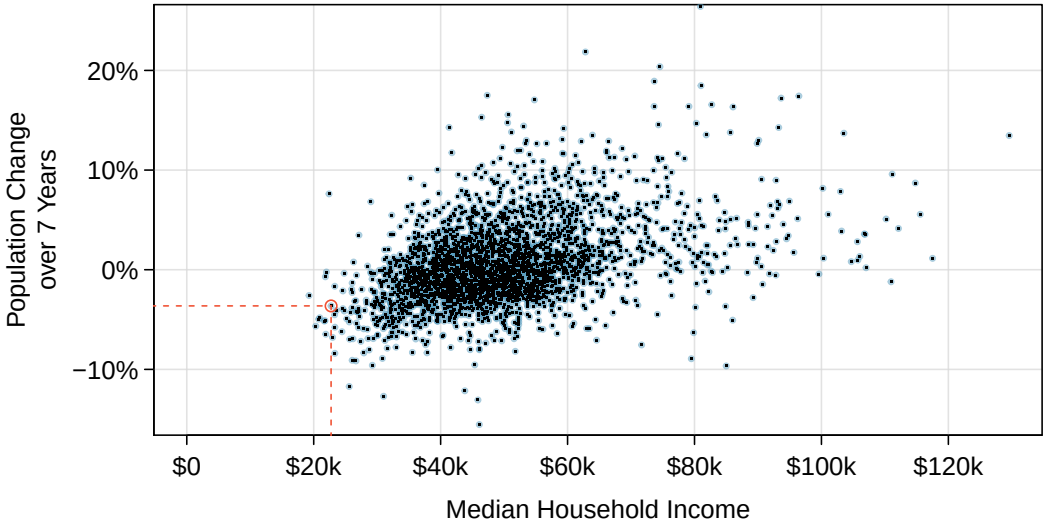



Figure 4: Breakdown of variables into their respective types.

# Relationship between variables



# Practice

Examine the variables in the `loan50` data set. Create two questions about possible relationships between variables in `loan50` that are of interest to you.

# Explanatory and Response Variables

Think about the following question.
What will be the increase in sales if there is a one percent rise in ad expenditure?

# Explanatory and Response Variables

Think about the following question.
What will be the increase in sales if there is a one percent rise in ad expenditure?
The framing of the question is such that we seek a directional relationship between spending on ads and sales.

`Explanatory Variable` $\Rightarrow$ `Response Variable`

# Sampling Principles and Strategies

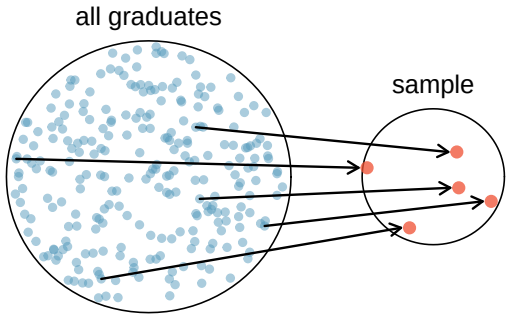
# Introduction

Consider the following possible responses to the three research questions:

- I subscribed to Spotify after I got tired of their annoying ads. Those annoying ads lead to greater number of subscriptions.
- A student attending Random Coaching Classes was ranked first in the Pointless Entrance Exam. Therefore, on average, kids who attend Random end up in Pretentious Institute of Meritocracy.
- An acquantaince who was infected with coronavirus was cured due to Coronil. Coronil is the cure for coronavirus.

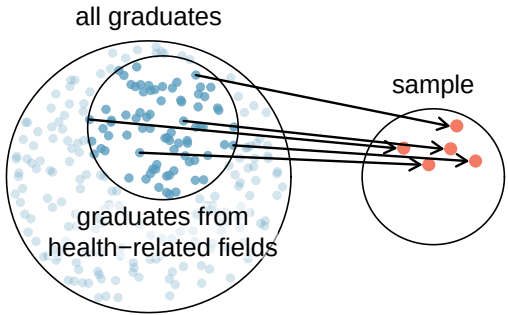
# Sampling from a population

Objective: Estimate the time to graduation for university undergraduates in the last five years by collecting a sample.



# Sampling from a population: Example

Let's suppose that a student from medicine is being asked to select survey respondents and he ends up oversampling other students from his own field. This introduces a **bias**.



# Simple random sample

To avoid the problem that we saw in the previous slide, we will employ a simple random sample.

# Simple random sample

To avoid the problem that we saw in the previous slide, we will employ a simple random sample.

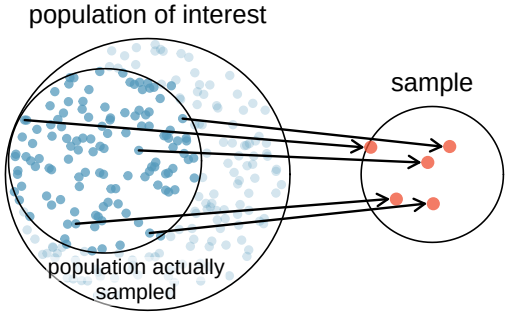
This means that each person in the population has an equal chance of being included.

# Simple random sample

To avoid the problem that we saw in the previous slide, we will employ a simple random sample.

This means that each person in the population has an equal chance of being included.

Even after having chosen the sample carefully, some people may choose not respond. This introduces another form of bias known as the **non-response bias**.



# Sampling Methods

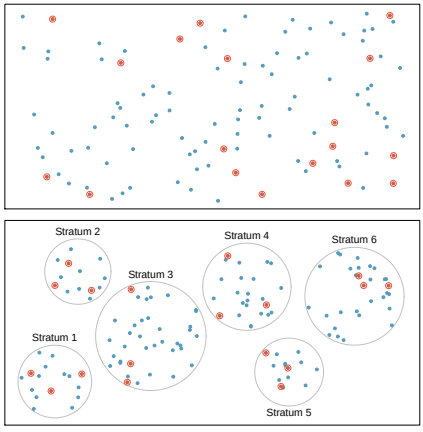



Figure 5: Examples of simple random and stratified sampling. In the top panel, simple random sampling was used to randomly select the 18 cases. In the bottom panel, stratified sampling was used: cases were grouped into strata, then simple random sampling was employed within each stratum.

# Sampling Methods

- **Simple random sampling**
  - Example: You want to estimate salaries of IPL players (there are eight IPL teams).
  - Let's assume that there are 160 players, and you are going to sample 16 of them.
  - You write the names of the players onto slips of paper.
  - Randomly pick 16 players.

# Sampling Methods

- **Simple random sampling**
  - Example: You want to estimate salaries of IPL players (there are eight IPL teams).
  - Let's assume that there are 160 players, and you are going to sample 16 of them.
  - You write the names of the players onto slips of paper.
  - Randomly pick 16 players.
- **Stratified sampling**
  - You will choose two players from each team (randomly of course).
  - This is especially useful when the observations in each *stratum* are very similar.

# Sampling Methods

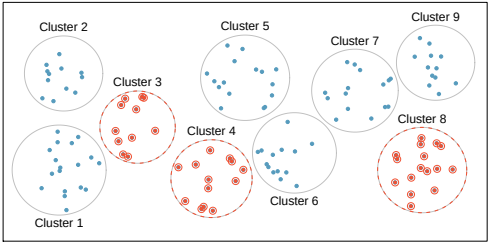
- **Cluster sample**:
  - Break up the population into many groups.
  - Choose a fixed sample of clusters.
  - Survey all observations from each of these clusters.

# Sampling Methods

- **Cluster sample**:
  - Break up the population into many groups.
  - Choose a fixed sample of clusters.
  - Survey all observations from each of these clusters.
- **Multi-stage sample**
  - Random sampling within each cluster.

These approaches are useful when clusters are largely similar, but there's lot of variation within each of these clusters.

# Sampling Methods




- **Cluster sampling (Top Panel)**
  - Data binned into 9 clusters.
  - 3 out of 9 were sampled.
  - Everyone in those 3 clusters were sampled.
- **Multistage cluster sampling (Bottom Panel)**
  - Randomly select three clusters.
  - Randomly draw people from each of these clusters.