STAT 216 Coursepack

Fall 2023
Montana State University

Melinda Yager
Jade Schmidt
Stacey Hancock

Preface

Placeholder

Fall 2023 Calendar of In-Class Activities

Placeholder

WEEK 1

Basics of Data

Placeholder

1.1 Reading Guides

1.2 Week 1 Reading Guide: Basics of Data

Textbook Sections 1.1: Case study and 1.2: Data basics

Vocabulary

Notes

Example: Section 1.1 — Case study: Using stents to prevent strokes

1.3 Activity 1: Intro to Data

1.3.1 Learning outcomes

1.3.2 Terminology review

1.3.3 General information on the Coursepack

1.3.4 Steps of the statistical investigation process

1.3.5 Take-home messages

1.3.6 Additional notes

1.4 Lecture Notes Week 1: Intro to data

Read through Sections 1.2.1 – 1.2.5 in the course textbook prior to coming to class on Friday using the reading guides at the beginning of week 1 material.

Data basics: Sections 1.2.1 – 1.2.2

Data: ____ used to answer research questions

Observational unit or case: the people or things we ____________________________ data from

Variable: what is measured on each ______________________ or ________________________.

Types of variables

- Categorical variable:

  Ordinal: levels of the variable have a ______________________________________ ordering

  Examples: 'Scale' questions, Years of schooling completed

  Nominal:levels of the variable do ______________________ have a natural ordering

  Examples: hair color, eye color, zipcode

- Quantitative variable:

  Continuous variables: value can be any ______________________________ within a range.

  Examples: percentage of students who are nursing majors, average hours of exercise per week; distance or time (measured with enough precision)

  Discrete variables: can only be ____ values, with jumps between

  Examples: years of schooling completed; SAT score, number of car accidents

Example: The Bureau of Transportation Statistics collects data on all forms of public transportation. The data set seen here includes several variables collect on flights departing on a random sample of 150 US airports in December of 2019.

```
airport <- read.csv("data/airport_delay.csv")
glimpse(airport)
#> Rows: 150
#> Columns: 19
#> $ airport            <chr> "ABI", "ABY", "ACV", "ACY", "ADQ", "AEX", "ALB", "~
#> $ city               <chr> "Abilene", "Albany", "Arcata/Eureka", "Atlantic Ci~
#> $ state              <chr> " TX", " GA", " CA", " NJ", " AK", " LA", " NY", "~
#> $ airport_name       <chr> " Abilene Regional", " Southwest Georgia Regional"~
#> $ hub                <chr> "no", "no", "no", "no", "no", "no", "no", "no", "n~
#> $ international      <chr> "no", "no", "no", "yes", "no", "yes", "yes", "yes"~
#> $ elevation_1000     <dbl> 1.7906, 0.1932, 0.2223, 0.0748, 0.0787, 0.0881, 0.~
#> $ latitude           <dbl> 32.4, 31.5, 41.0, 39.5, 57.7, 31.3, 42.7, 35.2, 45~
#> $ longitude          <dbl> -99.7, -81.2, -124.1, -74.6, -152.5, -92.5, -73.8,~
#> $ arr_flights        <int> 195, 81, 215, 293, 54, 282, 943, 410, 53, 32314, 6~
#> $ perc_delay15       <dbl> 16.410256, 13.580247, 23.255814, 15.358362, 12.962~
#> $ perc_cancelled     <dbl> 0.5128205, 0.0000000, 4.1860465, 0.6825939, 14.814~
#> $ perc_diverted      <dbl> 0.00000000, 0.00000000, 2.32558139, 0.68259386, 0.~
#> $ arr_delay          <int> 1563, 1244, 4763, 2905, 329, 1293, 15127, 9705, 25~
#> $ carrier_delay      <int> 459, 890, 1613, 476, 180, 302, 5627, 2253, 439, 10~
#> $ weather_delay      <int> 21, 43, 549, 124, 1, 58, 2346, 168, 1236, 13331, 2~
#> $ nas_delay          <int> 257, 39, 154, 771, 51, 112, 2096, 616, 746, 45674,~
#> $ security_delay     <int> 0, 0, 0, 25, 0, 0, 44, 0, 0, 375, 0, 83, 0, 23, 0,~
#> $ late_aircraft_delay <int> 826, 272, 2447, 1509, 97, 821, 5014, 6668, 108, 10~
```

- What are the observational units?

- Identify which variables are categorical.

- Identify which variables are quantitative.

Exploratory data analysis (EDA)

Summary statistic: a number which ______________________________ an entire data set

- Also called the ______________ ________________

Examples:

proportion of people who had a stroke

mean (or average) age

- Summary statistic and type of plot used depends on the type of variable(s)!

Roles of variables: Sections 1.2.3 – 1.2.5

Explanatory variable: predictor variable

- The variable researchers think *may be* _________________ the other variable.
- In an experiment, what the researchers _________________ or _____________________.
- The groups that we are comparing from the data set.

Response variable:

- The variable researchers think *may be* _________________________ by the other variable.
- Always simply _____________________ or ________________________; never controlled by researchers.

Examples:

Can you predict a criminal's height based on the footprint left at the scene of a crime?

- Identify the explanatory variable:

- Identify the response variable:

Does marking an item on sale (even without changing the price) increase the number of units sold per day, on average?

- Identify the explanatory variable:

- Identify the response variable:

In the Physician's Health Study, male physicians participated in a study to determine whether taking a daily low-dose aspirin reduced the risk of heart attacks. The male physicians were randomly assigned to the treatment groups. After five years, 104 of the 11,037 male physicians taking a daily low-dose aspirin had experienced a heart attack while 189 of the 11,034 male physicians taking a placebo had experienced a heart attack.

- Identify the explanatory variable:

- Identify the response variable:

Relationships between variables

- Association: the ____________________ between variables create a pattern; knowing something about one variable tells us about the other.
  - Positive association: as one variable ____________________, the other tends to _______________________ also.
  - Negative association: as one variable ____________________, the other tends to ____________________.
- Independent: no clear pattern can be seen between the _______________.

WEEK 2

Study Design

Placeholder

2.1 Week 2 Reading Guide: Sampling, Study Design, and Scope of Inference

Textbook Chapter 2: Study Design

Section 2.1: Sampling principles and strategies

Vocabulary

Notes

Notes on types of sampling bias

Section 2.2 & 2.3: Study Design

Reminders from Section 1.2

Vocabulary

Notes

Section 2.4: Scope of inference

2.2 Lecture Notes Week 2: Study Design

Sampling Methods: Section 2.1 in the course textbook

Good vs. bad sampling

Types of Sampling Bias

Examples

Observational studies, experiments, and scope of inference: Sections 2.2 – 2.4 in the course textbook

Study design

Scope of Inference

2.3 Out-of-Class Activity Week 2: American Indian Address

2.3.1 Learning outcomes

2.3.2 Terminology review

2.3.3 American Indian Address

By eye selection

Types of bias

2.3.4 Take-home messages

2.3.5 Additional notes

2.4 Activity 2: American Indian Address (continued)

2.4.1 Learning outcomes

2.4.2 Terminology review

Random selection

Effect of sample size

2.4.3 Take-home messages

WEEK 3

Exploring Categorical and Quantitative Data

Placeholder

3.1 Week 3 Reading Guide: Introduction to R, Categorical Variables, and a Single Quantitative Variable

Textbook Chapter 3 Applications: Data

Notes

Functions

Textbook Chapter 4: Exploring categorical data

Vocabulary

Notes

Review of Simpson's Paradox

Textbook Chapter 5: Exploring quantitative data

Type of Plots

Vocabulary

Notes

Summarizing Chapters 4 and 5

Notes

Data visualization summary

3.2 Lecture Notes Week 3: Exploratory Data Analysis

Summarizing categorical data

Displaying categorical variables

Simpson's paradox

Summarizing quantitative data

Types of plots

Four characteristics of plots for quantitative variables

Robust statistics

3.3 Out-of-Class Activity Week 3: Summarizing Categorical Variables

3.3.1 Learning outcomes

3.3.2 Terminology review

3.3.3 Graphing categorical variables

WEEK 4

Exploring Multivariable Data

Placeholder

4.1 Week 4 Reading Guide: Two Quantitative Variables and Multivariable Concepts

Textbook Chapter 6: Correlation and regression

Section 6.1 (Fitting a line, residuals, and correlation)

Reminders from Section 5.1

Vocabulary

Notes

Example: Brushtail possums

Section 6.2: Least squares regression

Vocabulary

Notes

Example: Elmhurst College

Section 6.3: Outliers in linear regression

Vocabulary

Notes

Section 6.4: Chapter 6 review

Notes

Section 7.1: Gapminder world

Vocabulary

Notes

Section 7.2: Simpson's Paradox, revisited

Reminder from Section 4.4

Notes

Example: SAT scores

4.2 Lecture Notes Week 4: Regression and Correlation

Summary measures and plots for two quantitative variables

Multivariable plots

4.3 Out-of-Class Activity Week 4: Movie Profits — Correlation and Coefficient of Determination

WEEK 5

Group Exam 1 Review

Placeholder