Exploring Data Frames

Nate Wells

Math 141, 1/27/21

Outline

In this lecture, we will. . .

Outline

In this lecture, we will...

- Review **statistical thinking** and Simpson's Paradox
- Investigate Data Frames and the structure of data

Section 1

Statistical Thinking

Stand Your Ground Defense

Below are the 220 cases where the defense was used, along with the defendant's and the victim's race and the trial outcome.

Table 2

	Minority Defendant Convicted	Minority Defendant Acquitted	White Defendant Convicted	White Defendant Acquitted
Minority Victim	19	45	5	19
White Victim	10	15	40	67

1. Overall, which group was convicted at a higher rate?

2. When the victim was white, which group was convicted at a higher rate?

3. When the victim was a minority, which group was convicted at a higher rate?

Stand Your Ground Defense

Below are the 220 cases where the defense was used, along with the defendant's and the victim's race and the trial outcome.

Table 2

	Minority Defendant Convicted	Minority Defendant Acquitted	White Defendant Convicted	White Defendant Acquitted
Minority Victim	19	45	5	19
White Victim	10	15	40	67

1. Overall, which group was convicted at a higher rate?

2. When the victim was white, which group was convicted at a higher rate?

3. When the victim was a minority, which group was convicted at a higher rate?

How is this possible?

Simpson's Paradox

Table 2

	Minority Defendant Convicted	Minority Defendant Acquitted	White Defendant Convicted	White Defendant Acquitted
Minority Victim	19	45	5	19
White Victim	10	15	40	67

1. What was the conviction rate when the victim was white? A minority?

2. When the defendent is white, what tends to be the race of the victim?

Simpson's Paradox

Table 2

	Minority Defendant Convicted	Minority Defendant Acquitted	White Defendant Convicted	White Defendant Acquitted
Minority Victim	19	45	5	19
White Victim	10	15	40	67

1. What was the conviction rate when the victim was white? A minority?

2. When the defendent is white, what tends to be the race of the victim?

Should we conclude that white defendants are convicted at a higher rate than minority defendants?

Statistical Thinking

- Statistical thinking means. . .

Statistical Thinking

- Statistical thinking means. . .
  - Looking holistically at the data

Statistical Thinking

- Statistical thinking means. . .
  - Looking holistically at the data
  - Using domain knowledge to assess appropriateness of measurements

Statistical Thinking

- Statistical thinking means. . .
  - Looking holistically at the data
  - Using domain knowledge to assess appropriateness of measurements
  - Evaluating whether more data would give a different picture, and in what respect

Statistical Thinking

- Statistical thinking means. . .
  - Looking holistically at the data
  - Using domain knowledge to assess appropriateness of measurements
  - Evaluating whether more data would give a different picture, and in what respect
  - Understanding the context of the data

Statistical Thinking

- Statistical thinking means. . .
  - Looking holistically at the data
  - Using domain knowledge to assess appropriateness of measurements
  - Evaluating whether more data would give a different picture, and in what respect
  - Understanding the context of the data
  - Respecting what the data says (and what it does not say!)

Section 2

Structure of Data

What are Data Frames?

- In order to perform *any* meaningful statistical analysis or inference, need real world observations.

What are Data Frames?

- In order to perform *any* meaningful statistical analysis or inference, need real world observations.
- **Data** are characteristics or measurements describing some individuals in a population.

What are Data Frames?

- In order to perform *any* meaningful statistical analysis or inference, need real world observations.
- **Data** are characteristics or measurements describing some individuals in a population.
- It may be tempting to suggest that data is objective.

What are Data Frames?

- In order to perform *any* meaningful statistical analysis or inference, need real world observations.
- **Data** are characteristics or measurements describing some individuals in a population.
- It may be tempting to suggest that data is objective.
  - But all data must be gathered, collated, organized, classified, shaped.

What are Data Frames?

- In order to perform *any* meaningful statistical analysis or inference, need real world observations.
- **Data** are characteristics or measurements describing some individuals in a population.
- It may be tempting to suggest that data is objective.
  - But all data must be gathered, collated, organized, classified, shaped.
  - Each step in the data collection and distribution chain requires decisions.

What are Data Frames?

- In order to perform *any* meaningful statistical analysis or inference, need real world observations.
- **Data** are characteristics or measurements describing some individuals in a population.
- It may be tempting to suggest that data is objective.
  - But all data must be gathered, collated, organized, classified, shaped.
  - Each step in the data collection and distribution chain requires decisions.
  - Data tells a story.

Data in R

- For convenience, data is often stored in a spreadsheet format (like an google sheet)

Data in R

- For convenience, data is often stored in a spreadsheet format (like an google sheet)
- In R, these spreadsheet-type data sets are called **data frames**

Data in R

- For convenience, data is often stored in a spreadsheet format (like an google sheet)
- In R, these spreadsheet-type data sets are called **data frames**
  - A **tibble** is a special kind of data frame that has better display properties (but is otherwise the same as a data frame)

Table 1: The Planets (Sorry Pluto)

name	type	diameter	rotation	rings
Mercury	Terrestrial planet	0.382	58.64	FALSE
Venus	Terrestrial planet	0.949	-243.02	FALSE
Earth	Terrestrial planet	1.000	1.00	FALSE
Mars	Terrestrial planet	0.532	1.03	FALSE
Jupiter	Gas giant	11.209	0.41	TRUE
Saturn	Gas giant	9.449	0.43	TRUE
Uranus	Gas giant	4.007	-0.72	TRUE
Neptune	Gas giant	3.883	0.67	TRUE

Exploring Data Frames

- A data frame is **tidy** when each row corresponds to an **observation** and each column corresponds to a **variable**

Table 2: The Planets (Sorry Pluto)

name	type	diameter	rotation	rings
Mercury	Terrestrial planet	0.382	58.64	FALSE
Venus	Terrestrial planet	0.949	-243.02	FALSE
Earth	Terrestrial planet	1.000	1.00	FALSE
Mars	Terrestrial planet	0.532	1.03	FALSE
Jupiter	Gas giant	11.209	0.41	TRUE
Saturn	Gas giant	9.449	0.43	TRUE
Uranus	Gas giant	4.007	-0.72	TRUE
Neptune	Gas giant	3.883	0.67	TRUE

Exploring Data Frames

- A data frame is **tidy** when each row corresponds to an **observation** and each column corresponds to a **variable**

Table 2: The Planets (Sorry Pluto)

name	type	diameter	rotation	rings
Mercury	Terrestrial planet	0.382	58.64	FALSE
Venus	Terrestrial planet	0.949	-243.02	FALSE
Earth	Terrestrial planet	1.000	1.00	FALSE
Mars	Terrestrial planet	0.532	1.03	FALSE
Jupiter	Gas giant	11.209	0.41	TRUE
Saturn	Gas giant	9.449	0.43	TRUE
Uranus	Gas giant	4.007	-0.72	TRUE
Neptune	Gas giant	3.883	0.67	TRUE

- The first step in digesting data is identifying the *observational unit* (i.e. the individual things whose properties we are measuring)

Observational Unit Example

Consider the following data frame. What might be the observational units? How do we know?

UserID	Common_Name	Tree_Height	Park	Carbon_Storage_lb
505	Norway Maple	39	Berkeley Park	408.8
551	Douglas-Fir	137	Berkeley Park	4722.2
7884	Pacific Dogwood	24	Kenilworth Park	461.5
563	Douglas-Fir	138	Berkeley Park	3580.3
7843	Hinoki Falsecypress	18	Kenilworth Park	230.2
464	Bigleaf Maple	110	Berkeley Park	11640.1
7901	Japanese Flowering Cherry	28	Kenilworth Park	2041.4
7929	Douglas-Fir	112	Kenilworth Park	1852.2
528	Flowering Plum	44	Berkeley Park	3014.6
510	American Elm	81	Berkeley Park	5790.6
7906	Hinoki Falsecypress	18	Kenilworth Park	79.7
7810	Norway Maple	32	Kenilworth Park	206.4

Observational Unit Example

Consider the following data frame. What might be the observational units? How do we know?

UserID	Common_Name	Tree_Height	Park	Carbon_Storage_lb
505	Norway Maple	39	Berkeley Park	408.8
551	Douglas-Fir	137	Berkeley Park	4722.2
7884	Pacific Dogwood	24	Kenilworth Park	461.5
563	Douglas-Fir	138	Berkeley Park	3580.3
7843	Hinoki Falsecypress	18	Kenilworth Park	230.2
464	Bigleaf Maple	110	Berkeley Park	11640.1
7901	Japanese Flowering Cherry	28	Kenilworth Park	2041.4
7929	Douglas-Fir	112	Kenilworth Park	1852.2
528	Flowering Plum	44	Berkeley Park	3014.6
510	American Elm	81	Berkeley Park	5790.6
7906	Hinoki Falsecypress	18	Kenilworth Park	79.7
7810	Norway Maple	32	Kenilworth Park	206.4

- Data was collected by the Portland Parks and Recreation's Urban Forestry Tree Inventory Project, gathered by 1300 volunteers adn city employees between 2010 and 2016

Exploring Data Frames

- A data frame is **tidy** when each row corresponds to an **observation** and each column corresponds to a quality or **variable**

Table 4: The Planets (Sorry Pluto)

name	type	diameter	rotation	rings
Mercury	Terrestrial planet	0.382	58.64	FALSE
Venus	Terrestrial planet	0.949	-243.02	FALSE
Earth	Terrestrial planet	1.000	1.00	FALSE
Mars	Terrestrial planet	0.532	1.03	FALSE
Jupiter	Gas giant	11.209	0.41	TRUE
Saturn	Gas giant	9.449	0.43	TRUE
Uranus	Gas giant	4.007	-0.72	TRUE
Neptune	Gas giant	3.883	0.67	TRUE

Exploring Data Frames

- A data frame is **tidy** when each row corresponds to an **observation** and each column corresponds to a quality or **variable**

Table 4: The Planets (Sorry Pluto)

name	type	diameter	rotation	rings
Mercury	Terrestrial planet	0.382	58.64	FALSE
Venus	Terrestrial planet	0.949	-243.02	FALSE
Earth	Terrestrial planet	1.000	1.00	FALSE
Mars	Terrestrial planet	0.532	1.03	FALSE
Jupiter	Gas giant	11.209	0.41	TRUE
Saturn	Gas giant	9.449	0.43	TRUE
Uranus	Gas giant	4.007	-0.72	TRUE
Neptune	Gas giant	3.883	0.67	TRUE

- The next step in data analysis is to identify the *variables of interest* (i.e. qualities of the observational units)

Types of Variables

- Variables come in a variety of types:

Types of Variables

- Variables come in a variety of types:
- Variables taking numeric values are called **quantitative**

Types of Variables

- Variables come in a variety of types:
- Variables taking numeric values are called **quantitative**
  - Quantitative variables can be ordered and averaged.

Types of Variables

- Variables come in a variety of types:
- Variables taking numeric values are called **quantitative**
  - Quantitative variables can be ordered and averaged.
  - Not every variable involving numbers in quantitative!

Types of Variables

- Variables come in a variety of types:
- Variables taking numeric values are called **quantitative**
  - Quantitative variables can be ordered and averaged.
  - Not every variable involving numbers in quantitative!
- Variables taking non-numeric values are called **categorical**

Types of Variables

- Variables come in a variety of types:
- Variables taking numeric values are called **quantitative**
  - Quantitative variables can be ordered and averaged.
  - Not every variable involving numbers in quantitative!
- Variables taking non-numeric values are called **categorical**
  - The values that a categorical variable can take are called its **levels**

Types of Variables

- Variables come in a variety of types:
- Variables taking numeric values are called **quantitative**
  - Quantitative variables can be ordered and averaged.
  - Not every variable involving numbers in quantitative!
- Variables taking non-numeric values are called **categorical**
  - The values that a categorical variable can take are called its **levels**
  - Some categorical variables can be ordered (but not averaged). These are called **ordinal** variables

Types of Variables

- Variables come in a variety of types:
- Variables taking numeric values are called **quantitative**
  - Quantitative variables can be ordered and averaged.
  - Not every variable involving numbers in quantitative!
- Variables taking non-numeric values are called **categorical**
  - The values that a categorical variable can take are called its **levels**
  - Some categorical variables can be ordered (but not averaged). These are called **ordinal** variables
- In some data frames, certain variables (called **identification variables**) uniquely specify each observation

Classifying Variables

Which variables in the following data frame are quantitative? Categorical? Identification? What might be one ordinal variable that could be added?

Classifying Variables

Which variables in the following data frame are quantitative? Categorical? Identification? What might be one ordinal variable that could be added?

Table 5: The Planets (Sorry Pluto)

name	type	diameter	rotation	rings
Mercury	Terrestrial planet	0.382	58.64	FALSE
Venus	Terrestrial planet	0.949	-243.02	FALSE
Earth	Terrestrial planet	1.000	1.00	FALSE
Mars	Terrestrial planet	0.532	1.03	FALSE
Jupiter	Gas giant	11.209	0.41	TRUE
Saturn	Gas giant	9.449	0.43	TRUE
Uranus	Gas giant	4.007	-0.72	TRUE
Neptune	Gas giant	3.883	0.67	TRUE