Unit 1: Introduction to data
3. More exploratory data analysis

Sta 101 – Spring 2019

Duke University, Department of Statistical Science

Class Year (*Independent Variable*)	In a Relationship? (*Dependent Variable*)
Freshman	Yes
Senior	Yes
Freshman	It's complicated
Sophomore	No
Junior	No
Junior	Yes
Freshman	Yes
…	…

Dr. Ellison

Slides posted at
https://www2.stat.duke.edu/courses/Spring19/sta101.001/

News/Coming up…

1. Office hours have opened up!
2. Making assigned groups tonight *(email with group assignments sent out tonight).*
3. Find your assigned group members tomorrow in lab *(ask TA if you can't find them.)*
4. Start Lab Assignment 1 in labs tomorrow with your assigned group. **Due next Thursday 1/24** before your next lab.
5. Problem Set 1 due next **Wednesday 1/25 11:55pm** on Sakai!

Outline

1. Housekeeping
2. Main ideas
   - A. Two Categorical Variables
     - A. Use segmented bar plots or mosaic plots for visualizing relationships between two categorical variables
   - B. One Numerical and One Categorical Variable
     - A. Use side-by-side box plots to visualize relationships between a numerical and categorical variable
   - C. Building Intuition For Making Inferences
     - A. Not all observed differences are statistically significant
     - B. Be aware of Simpson's paradox
3. Application Exercise

Outline

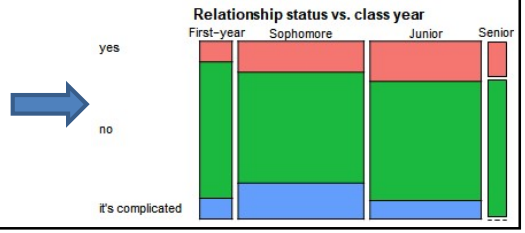
What types of plots can we use to visualize two categorical variables?

Class Year (*Independent Variable*)	In a Relationship? (*Dependent Variable*)
Freshman	Yes
Senior	Yes
Freshman	It's complicated
Sophomore	No
Junior	No
Junior	Yes
Freshman	Yes
…	…

Outline

Why is a mosaic plot a better plot for visualizing the *relationship* between two categorical variables than a segmented bar plot?

Class Year (*Independent Variable*)	In a Relationship? (*Dependent Variable*)
Freshman	Yes
Senior	Yes
Freshman	It's complicated
Sophomore	No
Junior	No
Junior	Yes
Freshman	Yes
…	…

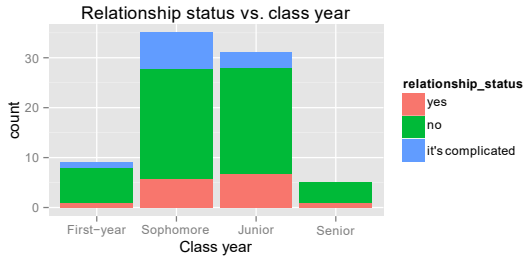


Outline

1. Housekeeping
2. Main ideas
   A. Two Categorical Variables
      A. 🔍 Use segmented bar plots or mosaic plots for visualizing relationships between two categorical variables
   B. One Numerical and One Categorical Variable
      A. 🔍 Use side-by-side box plots to visualize relationships between a numerical and categorical variable
   C. Building Intuition For Making Inferences
      A. 🔍 ✋ Not all observed differences are statistically significant
      B. 🔍 ✋ Be aware of Simpson's paradox
3. Application Exercise

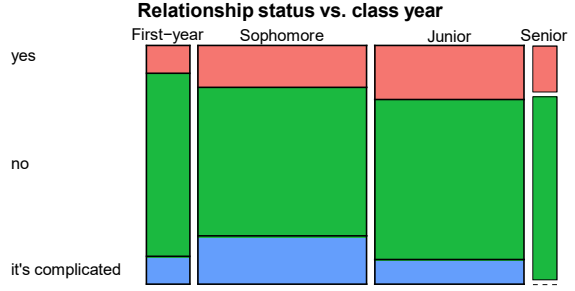
🔍 1. Use segmented bar plots for visualizing relationships bet. 2 categorical variables

What do the heights of the segments represent? Is there a relationship between class year and relationship status? What descriptive statistics can we use to summarize these data? Do the widths of the bars represent anything?



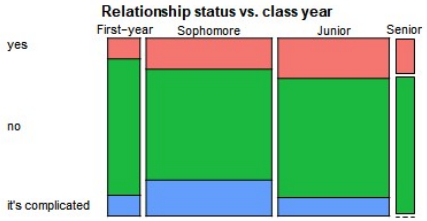
🔍 … or use mosaicplots

What do the widths of the bars represent? What about the heights of the boxes? Is there a relationship between class year and relationship status? What descriptive statistics can we use to summarize these data?



Outline

A mosaic plot looks at <u>percentages</u> of each dependent variable level, given each independent variable level. A frequency bar plot looks at counts. It's easier to see the relationship with a mosaic plot!



Outline

What is a good plot to use for visualizing the relationship between a categorical and numerical variable?

Are You a **Vegetarian**? (Independent Variable)	How many **nights** do you spend **drinking** a week? (Dependent Variable)
yes	0
yes	1
no	7
no	3
no	4
yes	2
no	1
...	...

Outline

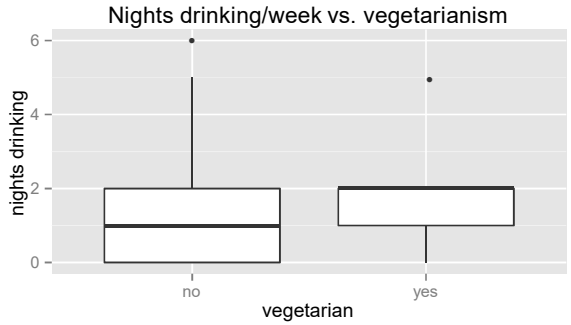
1. Housekeeping

2. Main ideas
   - A. Two Categorical Variables
     - A. Use segmented bar plots or mosaic plots for visualizing relationships between two categorical variables
   - B. One Numerical and One Categorical Variable
     - A. Use side-by-side box plots to visualize relationships between a numerical and categorical variable
   - C. Building Intuition For Making Inferences
     - A. Not all observed differences are statistically significant
     - B. Be aware of Simpson's paradox

3. Application Exercise

2. Use side-by-side box plots to visualize relationships between a numerical and categorical variable

How do drinking habits of vegetarian vs. non-vegetarian students compare?



4

Outline

Side-by-Side Boxplots are useful for visualizing the relationship between a categorical and numerical variable.

Are You a **Vegetarian**? (Independent Variable)	How many **nights** do you spend **drinking** a week? (Dependent Variable)
yes	0
yes	1
no	7
no	3
no	4
yes	2
no	1
...	...

Outline

Building Intuition For Making Inferences

Are all observed differences between two groups actually meaningful?

Outline

1. Housekeeping

2. Main ideas
   - A. Two Categorical Variables
     - A. Use segmented bar plots or mosaic plots for visualizing relationships between two categorical variables
   - B. One Numerical and One Categorical Variable
     - A. Use side-by-side box plots to visualize relationships between a numerical and categorical variable
   - C. Building Intuition For Making Inferences
     - A. Not all observed differences are statistically significant
     - B. Be aware of Simpson's paradox

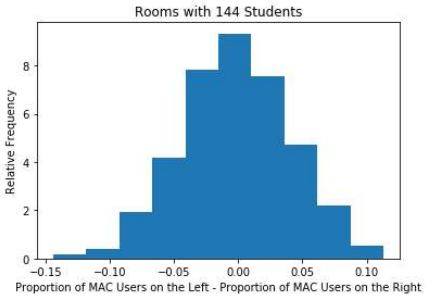
3. Application Exercise

3. Not all observed differences are statistically significant

What percent of the students sitting in the **left side** of the classroom have **Mac computers**? What about on the **right**? Are these numbers exactly the same? If not, do you think the difference is real, or due to random chance?

3. Not all observed differences are statistically significant

What percent of the students sitting in the **left side** of the classroom have **Mac computers**? What about on the **right**? Are these numbers exactly the same? If not, do you think the difference is real, or due to random chance?



5

Outline

Building Intuition For Making Inferences

Some studies can suggest one relationship, but upon a "closer look," can reveal the opposite of this relationship. Beware of Simpson's Paradox!

Outline

1. Housekeeping

2. Main ideas
   - A. Two Categorical Variables
     - A. Use segmented bar plots or mosaic plots for visualizing relationships between two categorical variables
   - B. One Numerical and One Categorical Variable
     - A. Use side-by-side box plots to visualize relationships between a numerical and categorical variable
   - C. Building Intuition For Making Inferences
     - A. Not all observed differences are statistically significant
     - B. Be aware of Simpson's paradox

3. Application Exercise

Race and death-penalty sentences in Florida murder cases

A 1991 study by Radelet and Pierce on race and death-penalty (DP) sentences gives the following table:

Defendant's race	DP	No DP	Total	% DP
Caucasian	53	430	483	
African American	15	176	191	
Total	68	606	674	

Adapted from Subsection 2.3.2 of A. Agresti (2002), Categorical Data Analysis, 2nd ed., and

http://math.stackexchange.com/questions/83756/examples-of-simpsons-paradox.

6

Race and death-penalty sentences in Florida murder cases

A 1991 study by Radelet and Pierce on race and death-penalty (DP) sentences gives the following table:

Defendant's race	DP	No DP	Total	% DP
Caucasian	53	430	483	*11%*
African American	15	176	191	
Total	68	606	674	

Adapted from Subsection 2.3.2 of A. Agresti (2002), Categorical Data Analysis, 2nd ed., and

http://math.stackexchange.com/questions/83756/examples-of-simpsons-paradox.

6

Race and death-penalty sentences in Florida murder cases

A 1991 study by Radelet and Pierce on race and death-penalty (DP) sentences gives the following table:

Defendant's race	DP	No DP	Total	% DP
Caucasian	53	430	483	*11%*
African American	15	176	191	*7.9%*
Total	68	606	674	

Adapted from Subsection 2.3.2 of A. Agresti (2002), Categorical Data Analysis, 2nd ed., and

http://math.stackexchange.com/questions/83756/examples-of-simpsons-paradox.

6

Race and death-penalty sentences in Florida murder cases

A 1991 study by Radelet and Pierce on race and death-penalty (DP) sentences gives the following table:

Defendant's race	DP	No DP	Total	% DP
Caucasian	53	430	483	*11%*
African American	15	176	191	*7.9%*
Total	68	606	674	

Who is more likely to get the death penalty?

Adapted from Subsection 2.3.2 of A. Agresti (2002), Categorical Data Analysis, 2nd ed., and

http://math.stackexchange.com/questions/83756/examples-of-simpsons-paradox.

6

Another look

Same data, taking into consideration victim's race:

Victim's race	Defendant's race	DP	No DP	Total	% DP
Caucasian	Caucasian	53	414	467	
Caucasian	African American	11	37	48	
African American	Caucasian	0	16	16	
African American	African American	4	139	143	
Total		68	606	674	

7

Another look

Same data, taking into consideration victim's race:

Victim's race	Defendant's race	DP	No DP	Total	% DP
Caucasian	Caucasian	53	414	467	11.3%
Caucasian	African American	11	37	48	
African American	Caucasian	0	16	16	
African American	African American	4	139	143	
Total		68	606	674	

7

Another look

Same data, taking into consideration victim's race:

Victim's race	Defendant's race	DP	No DP	Total	% DP
Caucasian	Caucasian	53	414	467	11.3%
Caucasian	African American	11	37	48	22.9%
African American	Caucasian	0	16	16	
African American	African American	4	139	143	
Total		68	606	674	

7

Another look

Same data, taking into consideration victim's race:

Victim's race	Defendant's race	DP	No DP	Total	% DP
Caucasian	Caucasian	53	414	467	11.3%
Caucasian	African American	11	37	48	22.9%
African American	Caucasian	0	16	16	0%
African American	African American	4	139	143	
Total		68	606	674	

7

Another look

Same data, taking into consideration victim's race:

Victim's race	Defendant's race	DP	No DP	Total	% DP
Caucasian	Caucasian	53	414	467	11.3%
Caucasian	African American	11	37	48	22.9%
African American	Caucasian	0	16	16	0%
African American	African American	4	139	143	2.8%
Total		68	606	674	

7

Another look

Same data, taking into consideration victim's race:

Victim's race	Defendant's race	DP	No DP	Total	% DP
Caucasian	Caucasian	53	414	467	*11.3%*
Caucasian	African American	11	37	48	*22.9%*
African American	Caucasian	0	16	16	*0%*
African American	African American	4	139	143	*2.8%*
Total		68	606	674	

Who is more likely to get the death penalty?

Contradiction?

- People of one race are more likely to murder others of the same race, murdering a Caucasian is more likely to result in the death penalty, and there are more Caucasian defendants than African American defendants in the sample.

Victim's race	Defendant's race	DP	No DP	Total	% DP
Caucasian	Caucasian	53	414	467	*11.3%*
Caucasian	African American	11	37	48	*22.9%*
African American	Caucasian	0	16	16	*0%*
African American	African American	4	139	143	*2.8%*
Total		68	606	674	

$90\% \left(= \frac{467 + 143}{674}\right)$ vs $10\% \left(= \frac{48 + 16}{674}\right)$

Contradiction?

- People of one race are more likely to murder others of the same race, murdering a Caucasian is more likely to result in the death penalty, and there are more Caucasian defendants than African American defendants in the sample.

Victim's race	Defendant's race	DP	No DP	Total	% DP
Caucasian	Caucasian	53	414	467	*11.3%*
Caucasian	African American	11	37	48	*22.9%*
African American	Caucasian	0	16	16	*0%*
African American	African American	4	139	143	*2.8%*
Total		68	606	674	

$12\% \left(= \frac{53 + 11}{467 + 48}\right)$ vs $2.5\% \left(= \frac{0 + 4}{16 + 143}\right)$

Contradiction?

- People of one race are more likely to murder others of the same race, murdering a Caucasian is more likely to result in the death penalty, and there are more Caucasian defendants than African American defendants in the sample.

Victim's race	Defendant's race	DP	No DP	Total	% DP
Caucasian	Caucasian	53	414	467	*11.3%*
Caucasian	African American	11	37	48	*22.9%*
African American	Caucasian	0	16	16	*0%*
African American	African American	4	139	143	*2.8%*
Total		68	606	674	

Contradiction?

- People of one race are more likely to murder others of the same race, murdering a Caucasian is more likely to result in the death penalty, and there are more Caucasian defendants than African American defendants in the sample.
- Controlling for the victim's race reveals more insights into the data, and changes the direction of the relationship between race and death penalty.

Defendant Races (Independent Variable)	**Death Penalty? (Dependent Variable)**
White Defendant	yes
Black Defendant	yes
White Defendant	no
White Defendant	no
White Defendant	no
White Defendant	yes
White Defendant	no
Black Defendant	no
...	...

Defendant and Victim Races *(Independent Variable)*	**Death Penalty?** *(Dependent Variable)*
Black Victim - White Defendant	yes
Black Victim - Black Defendant	yes
Black Victim - White Defendant	no
White Victim - White Defendant	no
Black Victim - White Defendant	no
White Victim - White Defendant	yes
White Victim - White Defendant	no
Black Victim - Black Defendant	no
...	...

8

Contradiction?

- People of one race are more likely to murder others of the same race, murdering a Caucasian is more likely to result in the death penalty, and there are more Caucasian defendants than African American defendants in the sample.
- Controlling for the victim's race reveals more insights into the data, and changes the direction of the relationship between race and death penalty.
- This phenomenon is called *Simpson's Paradox*: An association, or a comparison, that holds when we compare two groups can disappear or even be reversed when the original groups are broken down into smaller groups according to some other feature (a confounding/lurking variable).

8

Outline

1. Housekeeping
2. Main ideas
   - A. Two Categorical Variables
     - A. Use segmented bar plots or mosaic plots for visualizing relationships between two categorical variables
   - B. One Numerical and One Categorical Variable
     - A. Use side-by-side box plots to visualize relationships between a numerical and categorical variable
   - C. Building Intuition For Making Inferences
     - A. Not all observed differences are statistically significant
     - B. Be aware of Simpson's paradox
3. Application Exercise

Application exercise: 1.2 Histogram to boxplot

See the course website for instructions.

9

Summary of main ideas

1. Use segmented bar plots or mosaic plots for visualizing relationships between two categorical variables
2. Use side-by-side box plots to visualize relationships between a numerical and categorical variable
3. Not all observed differences are statistically significant
4. Be aware of Simpson's paradox

10

Outline

1. Main ideas
   - A. Two Categorical Variables
     - A. 🔍 Use segmented bar plots or mosaic plots for visualizing relationships between two categorical variables
   - B. One Numerical and One Categorical Variable
     - A. 🔍 Use side-by-side box plots to visualize relationships between a numerical and categorical variable
   - C. Building Intuition For Making Inferences
     - A. 🔍 ✋ Not all observed differences are statistically significant
     - B. 🔍 ✋ Be aware of Simpson's paradox