Unit 1: Introduction to data

3. More exploratory data analysis

Sta 101 - Spring 2016

Duke University, Department of Statistical Science

Dr. Çetinkaya-Rundel

Slides posted at *http://bit.ly/sta101_s16*

Outline

1. Housekeeping

2. 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

3. Application Exercise

4. Summary

Announcements

-

Outline

1. Housekeeping

2. 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

3. Application Exercise

4. Summary

Outline

1. Housekeeping

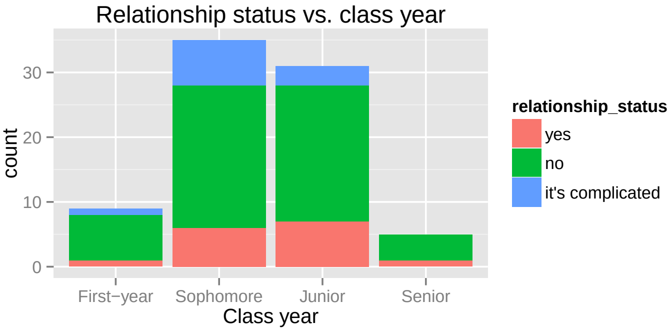
2. Main ideas
    1. Use segmented bar plots or mosaic plots for visualizing relationships between two categorical variables
    2. Use side-by-side box plots to visualize relationships between a numerical and categorical variable
    3. Not all observed differences are statistically significant
   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

3. Application Exercise

4. Summary

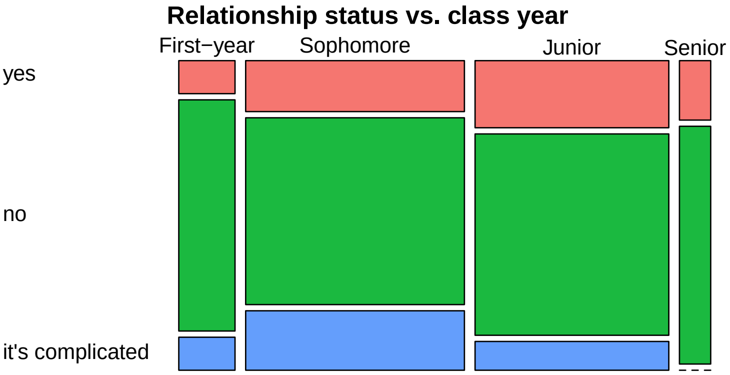
1. Use segmented bar plots for visualizing relationships bet. 2 categorical variables

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 other tools could we use to summarize these data?



Outline

1. Housekeeping

2. 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

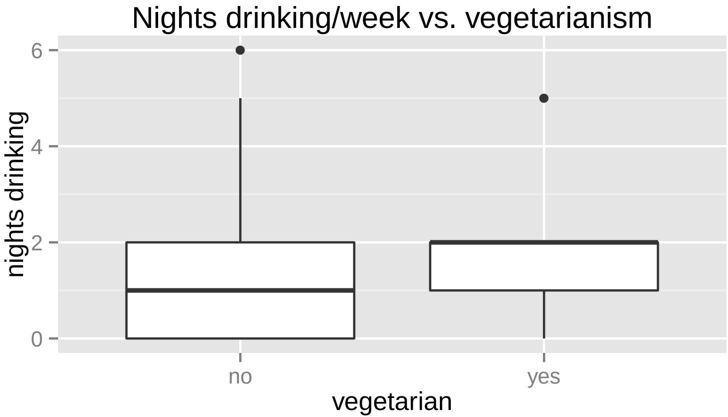
    4. Be aware of Simpson's paradox

3. Application Exercise

4. Summary

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?



Outline

1. Housekeeping

2. 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

3. Application Exercise

4. Summary

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?

1. Housekeeping

2. 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

3. Application Exercise

4. Summary

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.

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.

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.

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.

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	

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	

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	

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	

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	

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.

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.

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).

1. Housekeeping

2. 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

3. Application Exercise

4. Summary

If you finish one, move on to the next.

Application exercise: 1.3 Histogram to boxplot

See the course website for instructions.

Application exercise: 1.3 Scientific studies in the press

See the course website for instructions.

Outline

1. Housekeeping

2. 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

3. Application Exercise

4. Summary

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