Unit 1: Introduction to data

2. Exploratory data analysis

Sta 101 - Spring 2015

Duke University, Department of Statistical Science

January 14, 2015

Dr. Çetinkaya-Rundel

Slides posted at http://bitly.com/sta101sp15

Announcements

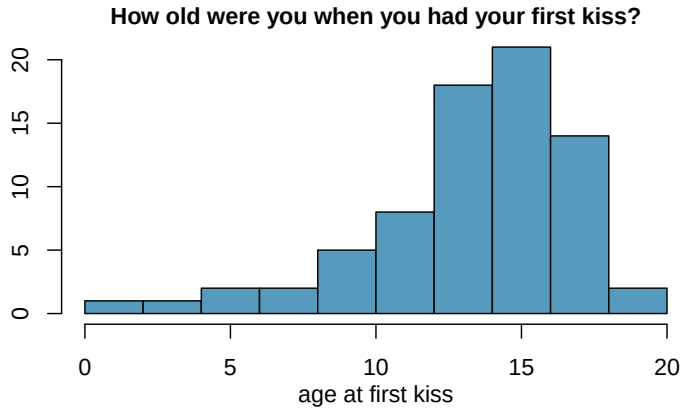
- Sit in teams in class and lab going forward – are you missing a team member?
- If you haven't yet done so, take the class survey
- Lab 1 due by your lab session next Monday – one submission per team sufficient
  - Questions about labs?
- TA office hours:
  - SEC open 4pm - 9pm Sunday - Thursday
  - Sta 101 TA hours:
    - *Christine - Sun, 4-6pm*
    - David - Mon, 4-6pm
    - *Anthony - Mon, 7-9pm*
    - Chris (Xinyi) - Tues, 5-7pm
    - Radhika - Tues, 7-9pm
    - *Mao - Wed, 5-7pm*
    - Tori - Wed, 7-9pm
    - *Fiamma - Thur, 6-7pm*
    - *Phillip - TBA*
- No class and no OH on Monday, review randomization test video before Wednesday's class
- PS 2 due Wednesday

From last time - App Ex 1.1: Haters gonna hate

1. Cases: 200 men and women
2. Response: Attitude towards the microwave oven
3. Explanatory: Whether the participant is a hater or not
4. Random sampling / assignment: Via Amazon's MTurk - self selected sample, no random assignment.
5. Type: Observational, doesn't use random assignment.
6. Causality: No
7. Generalizability: Only if we could assume sample from Amazon's MTurk's sample is representative

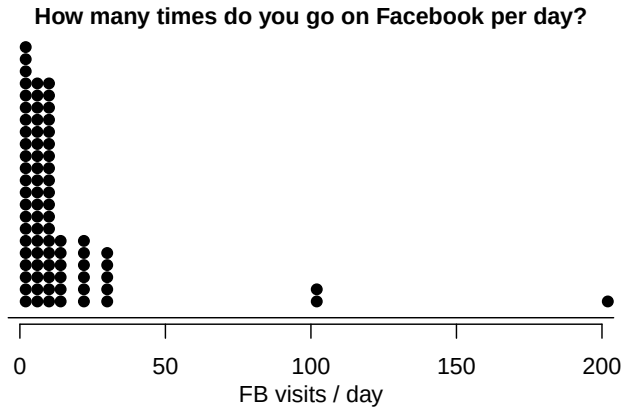
From your survey...

Do you see anything out of the ordinary?

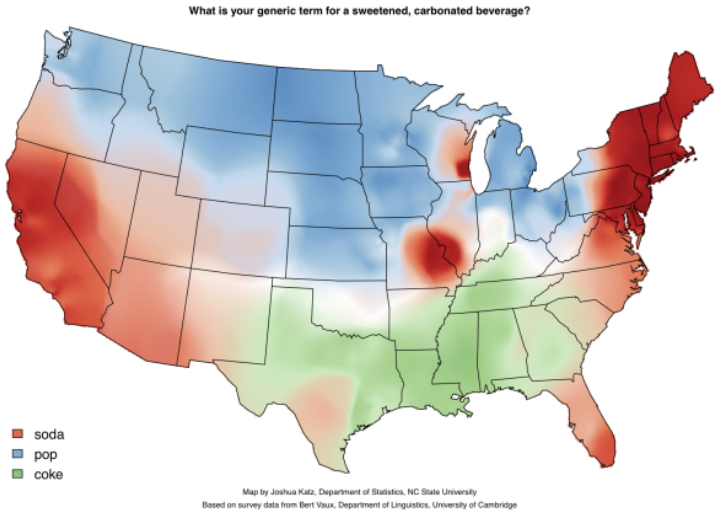


From your survey...

How are people reporting lower vs. higher values of FB visits?



Describe the spatial distribution of preferred sweetened carbonated beverage drink.



http://spark.rstudio.com/jkatz/SurveyMaps

What is missing in this visualization?

http://spark.rstudio.com/jkatz/SurveyMaps

Describing distributions of numerical variables

- *Shape*: skewness, modality
- *Center*: an estimate of a *typical* observation in the distribution (mean, median, mode, etc.)
  - Notation: μ: population mean, $\bar{x}$: sample mean
- *Spread*: measure of variability in the distribution (standard deviation, IQR, range, etc.)
- *Unusual observations*: observations that stand out from the rest of the data that may be suspected outliers

Clicker question

Which of these is most likely to have a roughly symmetric distribution?

(a) salaries of a random sample of people from North Carolina
(b) weights of adult females
(c) scores on an well-designed exam
(d) last digits of phone numbers

Mean vs. median

Clicker question

How do the mean and median of the following two datasets compare?

Dataset 1: 30, 50, 70, 90
Dataset 2: 30, 50, 70, 1000

(a) $\bar{x}_1 = \bar{x}_2$, $\text{median}_1 = \text{median}_2$
(b) $\bar{x}_1 < \bar{x}_2$, $\text{median}_1 = \text{median}_2$
(c) $\bar{x}_1 < \bar{x}_2$, $\text{median}_1 < \text{median}_2$
(d) $\bar{x}_1 > \bar{x}_2$, $\text{median}_1 < \text{median}_2$
(e) $\bar{x}_1 > \bar{x}_2$, $\text{median}_1 = \text{median}_2$

Standard deviation and variance

- Most commonly used measure of variability is the *standard deviation*, which roughly measures the average deviation from the mean
  - Notation: σ: population standard deviation, s: sample standard deviation
- Calculating the standard deviation, for a population (rarely, if ever) and for a sample:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{n}} \qquad s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

- Square of the standard deviation is called the *variance*.

More on SD

Why divide by $n - 1$ instead of n when calculating the sample standard deviation?

Lose a "degree of freedom" for using an estimate (the sample mean, $\bar{x}$), in estimating the sample variance/standard deviation.

Why do we use the squared deviation in the calculation of variance?

- To get rid of negatives so that observations equally distant from the mean are weighed equally.
- To weigh larger deviations more heavily.

Clicker question

True / False: The range is always larger than the IQR for a given dataset.

(a) Yes
(b) No

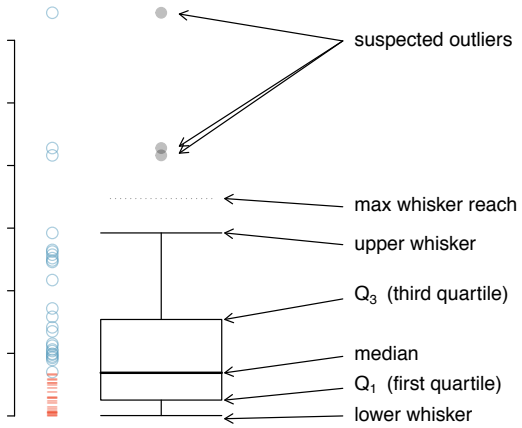
Is the range or the IQR more robust to outliers?

- Mean and standard deviation are easily affected by extreme observations since the value of each data point contributes to their calculation.
- Median and IQR are more robust.
- Therefore we choose median&IQR (over mean&SD) when describing skewed distributions.

Application exercise: 1.2 Distributions of numerical variables

See the course website for instructions.

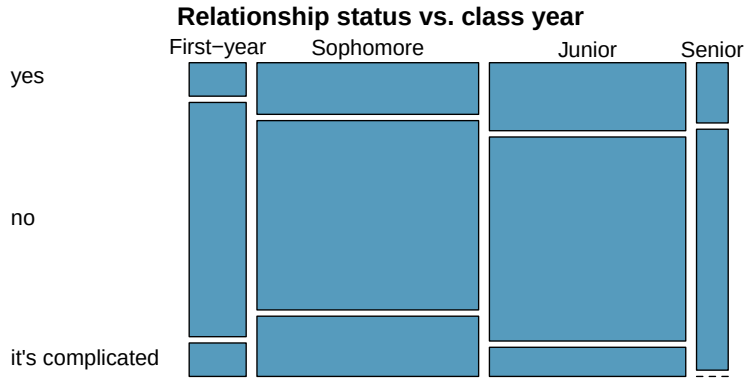
A *box plot* visualizes the median, the quartiles, and suspected outliers. An *outlier* is defined as an observation more than 1.5×IQR away from the quartiles.



Application exercise: 1.3 Boxplots

See the course website for instructions.

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?



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	

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	

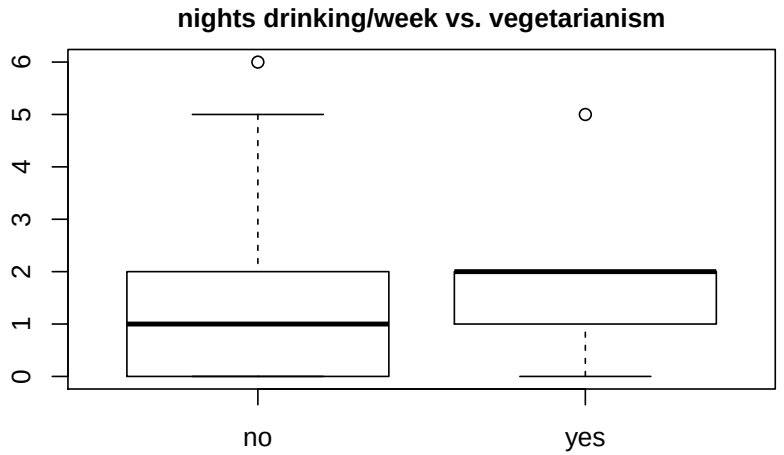
Who is more likely to get the death penalty?

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.
- 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).

Side-by-side box plot

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



Summary of main ideas

1. Always start your exploration with a visualization
2. When describing numerical distributions discuss shape, center, spread, and unusual observations
3. Robust statistics are not easily affected by outliers and extreme skew
4. Use box plots to display quartiles, median, and outliers
5. Use mosaic plots for visualizing relationship between two categorical variables
6. Be aware of Simpson's paradox
7. Use side-by-side box plots to visualize relationships between numerical and categorical variables