Unit 1: Introduction to data

2. 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. Readiness assessment

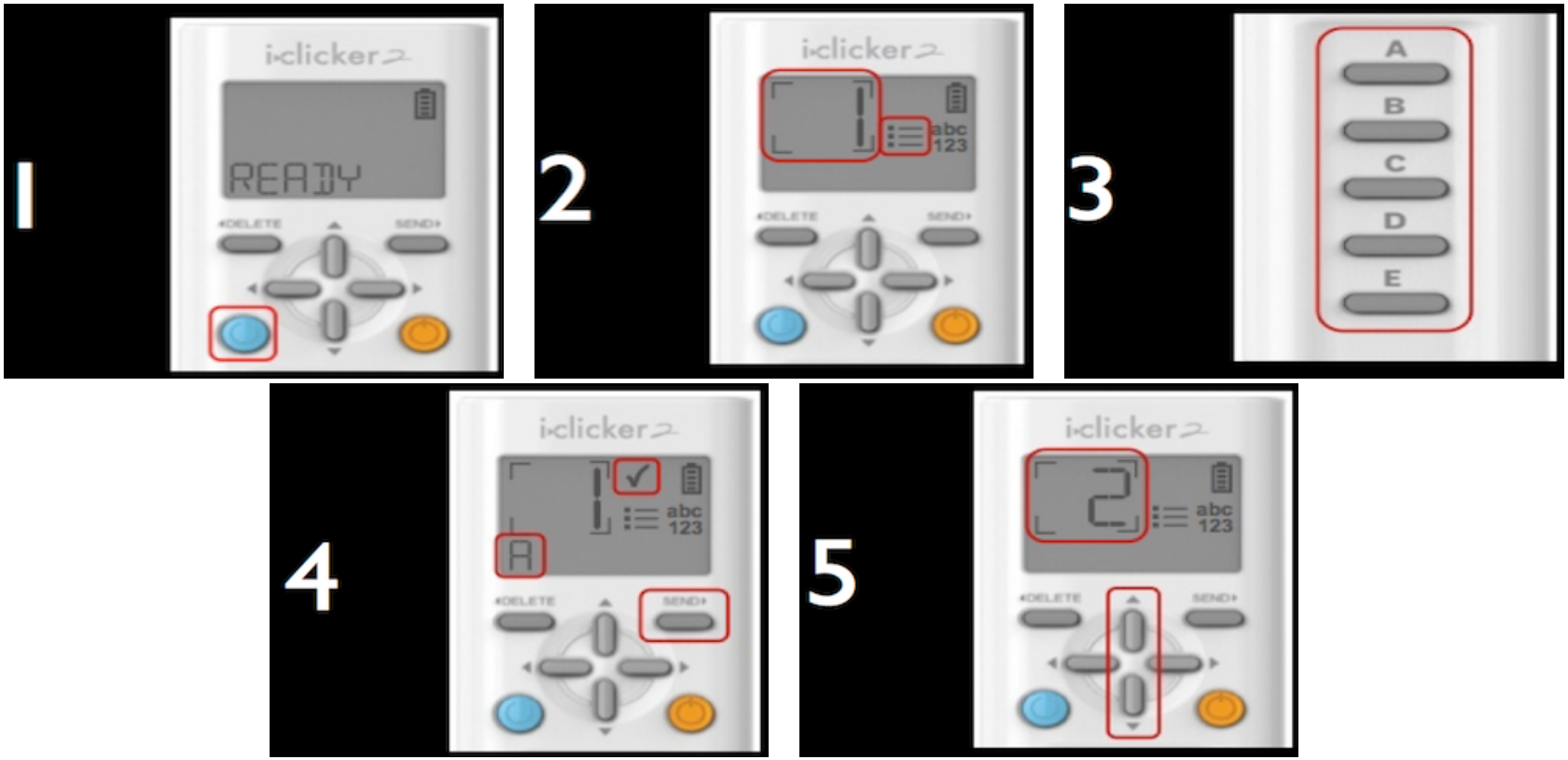
2. Housekeeping

3. 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

4. Application exercises

5. Summary

- *Individual:* 15 minutes, using clickers



- *Team:* 10 minutes, using scratch off sheets (1 per team)

1. Readiness assessment

2. Housekeeping

3. 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

4. Application exercises

5. Summary

- PS 1 is assigned on the course website, start working on it

Outline

1. Readiness assessment

2. Housekeeping

3. 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

4. Application exercises

5. Summary

Outline

1. Readiness assessment

2. Housekeeping

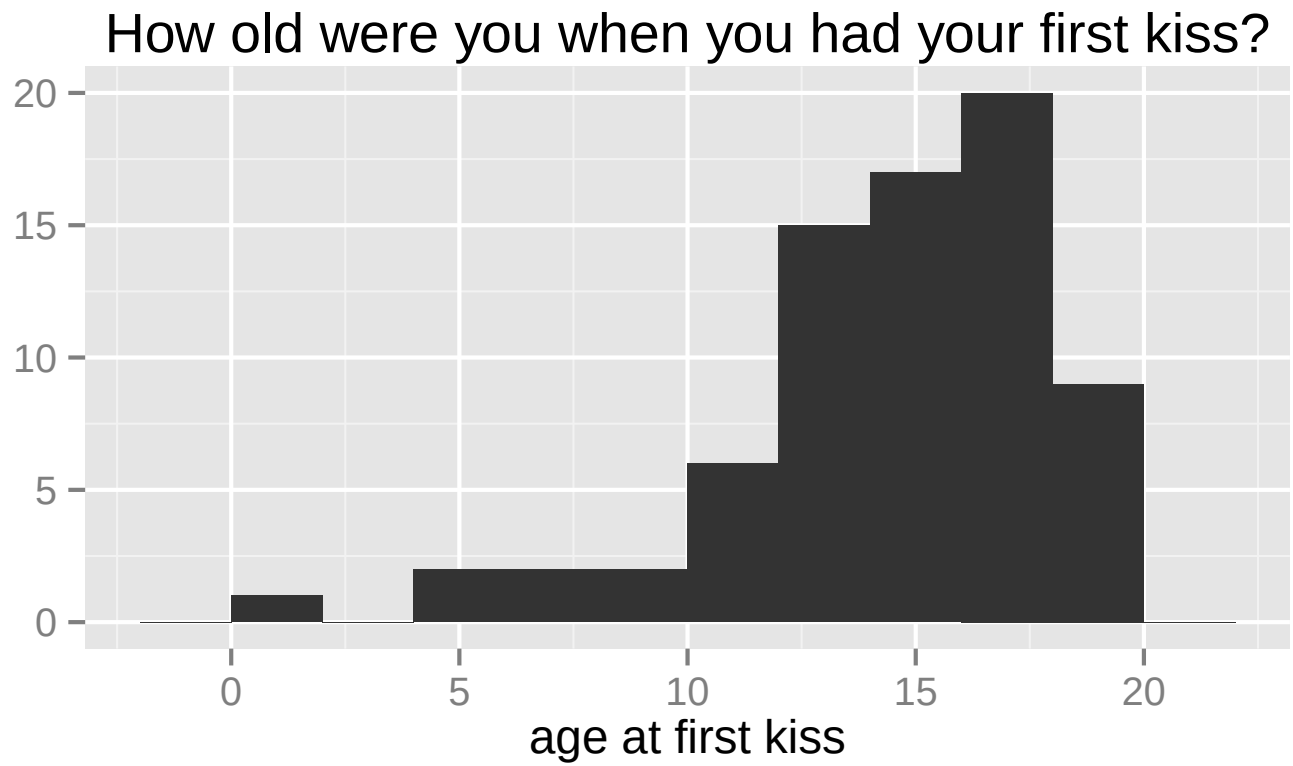
3. Main ideas
    1. Always start your exploration with a visualization
    2. When describing numerical distributions discuss shape, center, spread, and unusual observations
    3. Robust statistics are not easily affected by outliers and extreme skew
    4. Use box plots to display quartiles, median, and outliers

4. Application exercises

5. Summary

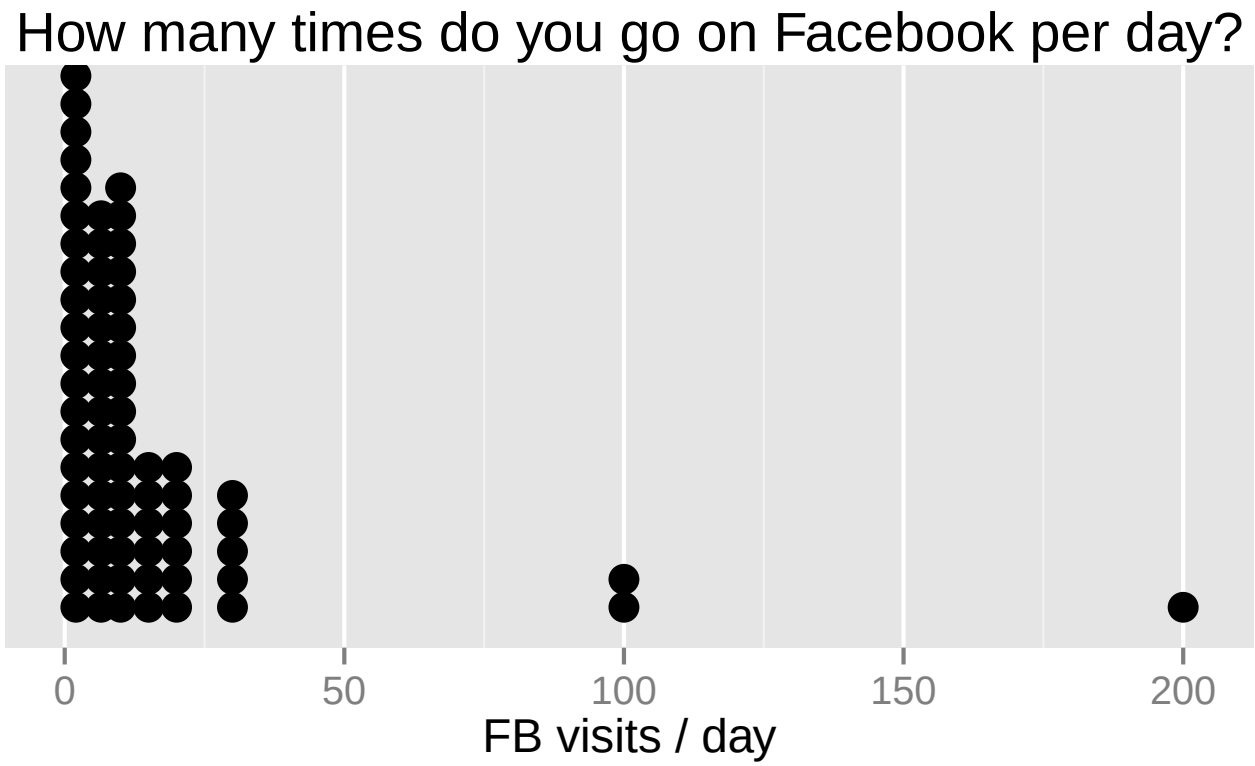
Do you see anything out of the ordinary?

Do you see anything out of the ordinary?

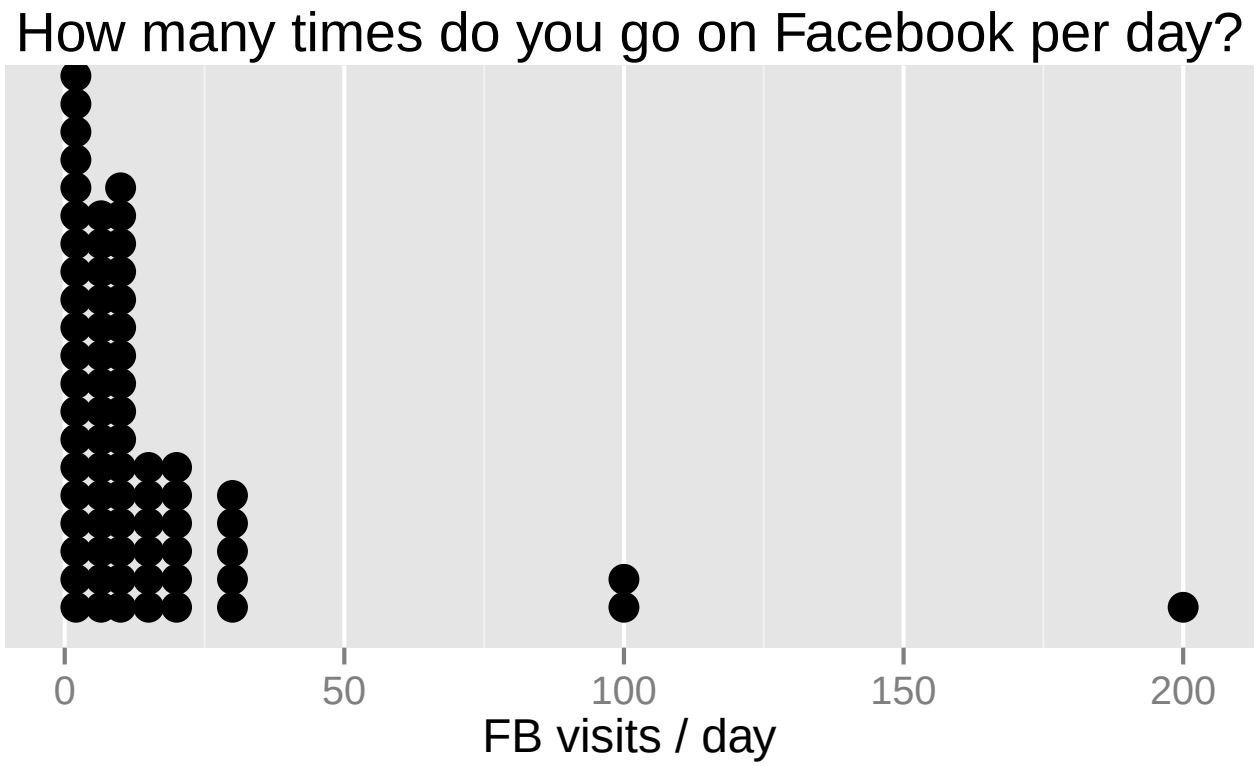


Some people reported very low ages, which might suggest the survey question wasn't clear: romantic kiss or any kiss?

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

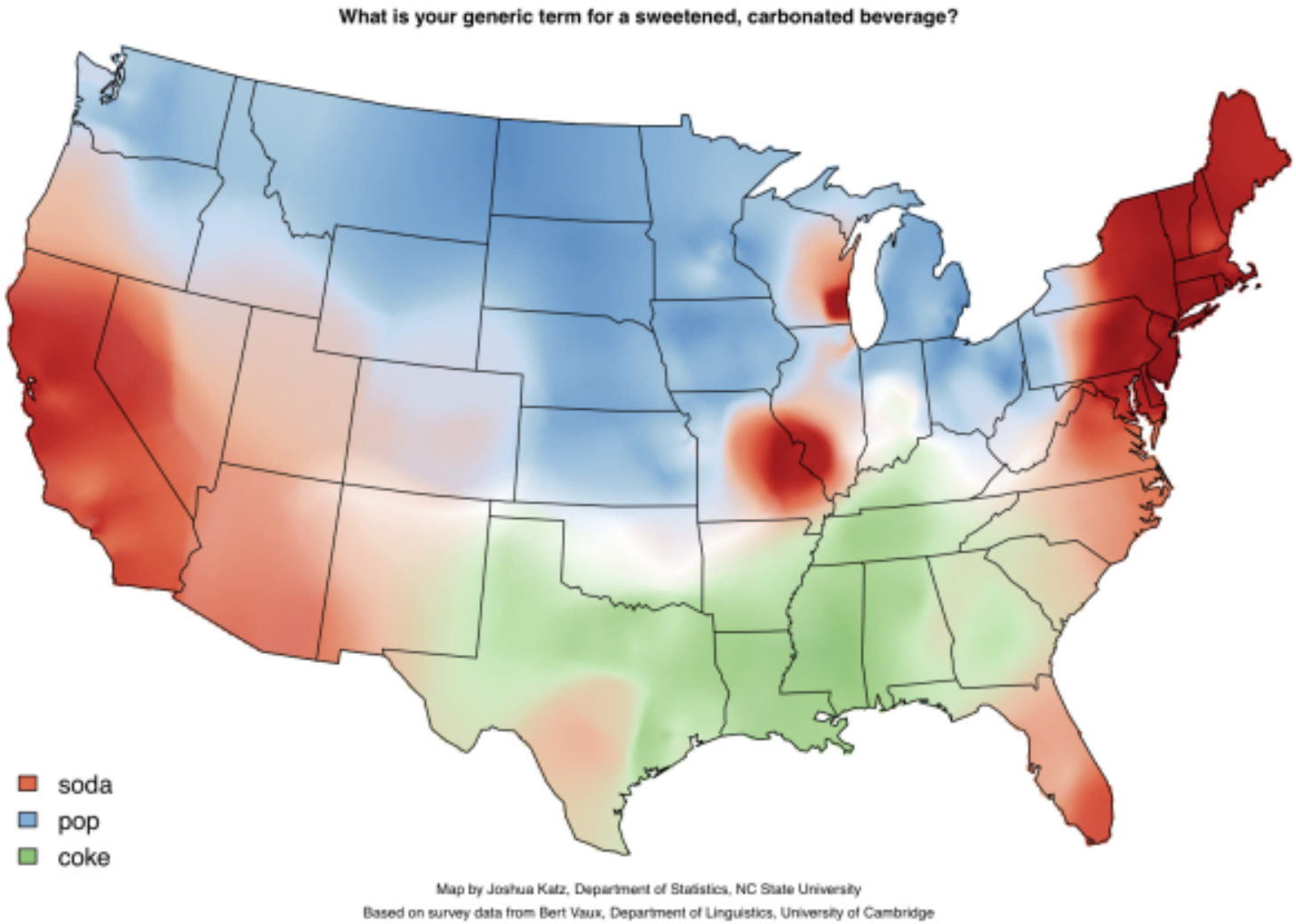


How are people reporting lower vs. higher values of FB visits?



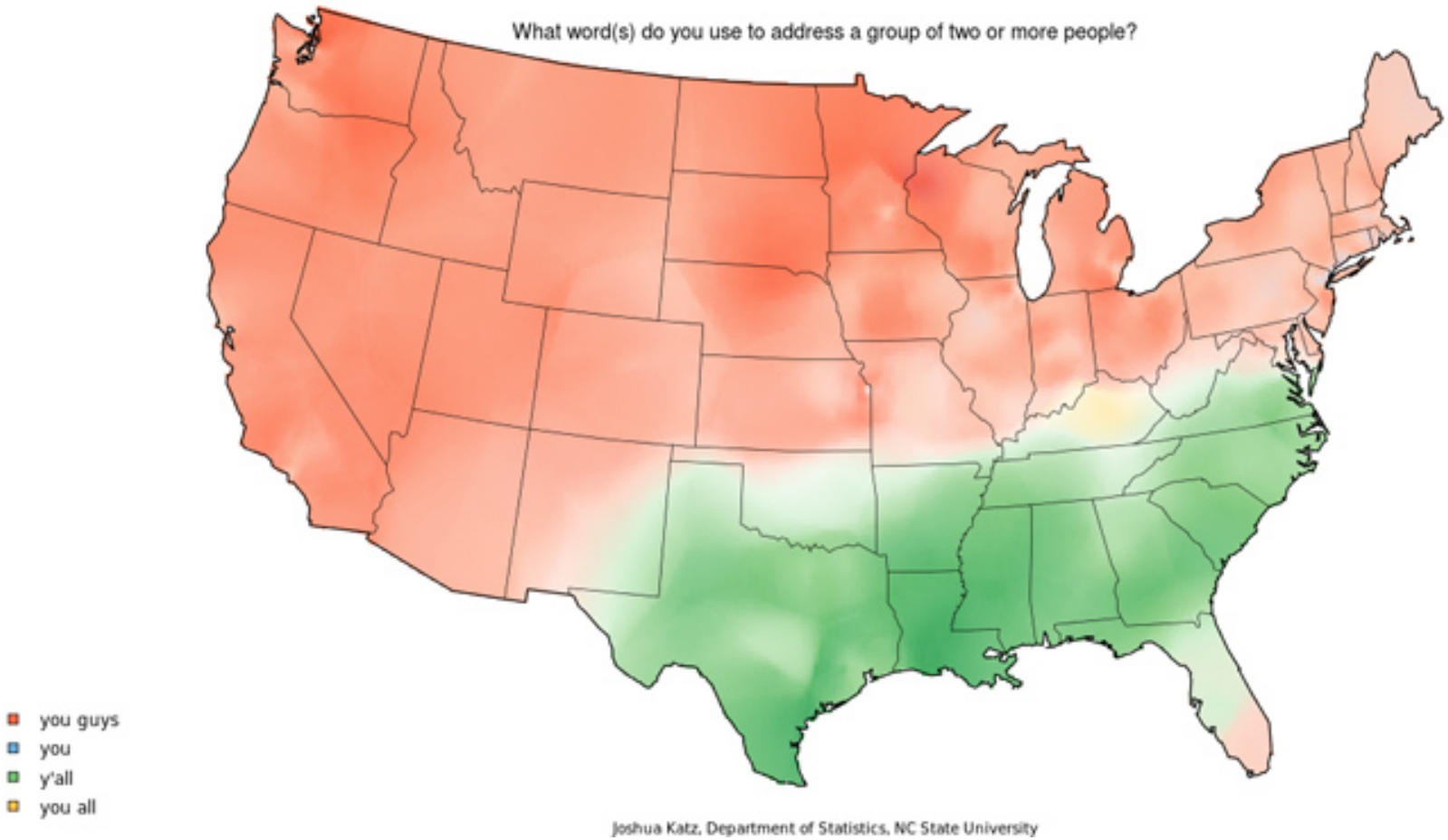
Finer scale for lower numbers.

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

Outline

1. Readiness assessment

2. Housekeeping

3. 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

4. Application exercises

5. Summary

- *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

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

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) $\bar{x}_1 = \bar{x}_2,\ median_1 = median_2$
(b) $\bar{x}_1 < \bar{x}_2,\ median_1 = median_2$
(c) $\bar{x}_1 < \bar{x}_2,\ median_1 < median_2$
(d) $\bar{x}_1 > \bar{x}_2,\ median_1 < median_2$
(e) $\bar{x}_1 > \bar{x}_2,\ median_1 = median_2$

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) $\bar{x}_1 = \bar{x}_2$, $median_1 = median_2$

(b) $\bar{x}_1 < \bar{x}_2$, $median_1 = median_2$

(c) $\bar{x}_1 < \bar{x}_2$, $median_1 < median_2$

(d) $\bar{x}_1 > \bar{x}_2$, $median_1 < median_2$

(e) $\bar{x}_1 > \bar{x}_2$, $median_1 = median_2$

- 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*.

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

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 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?

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 at least as large as the IQR for a given dataset.

(a) Yes

(b) No

Clicker question

True / False: The range is always at least as large as the IQR for a given dataset.

(a) *Yes*

(b) No

Range = max - min, IQR = Q3 - Q1

Clicker question

True / False: The range is always at least as large as the IQR for a given dataset.

(a) *Yes*

(b) No

Range = max - min, IQR = Q3 - Q1

Is the range or the IQR more robust to outliers?

Clicker question

True / False: The range is always at least as large as the IQR for a given dataset.

(a) *Yes*

(b) No

Range = max - min, IQR = Q3 - Q1

Is the range or the IQR more robust to outliers?

IQR

Outline

1. Readiness assessment

2. Housekeeping

3. 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

4. Application exercises

5. Summary

- 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.

Outline

1. Readiness assessment

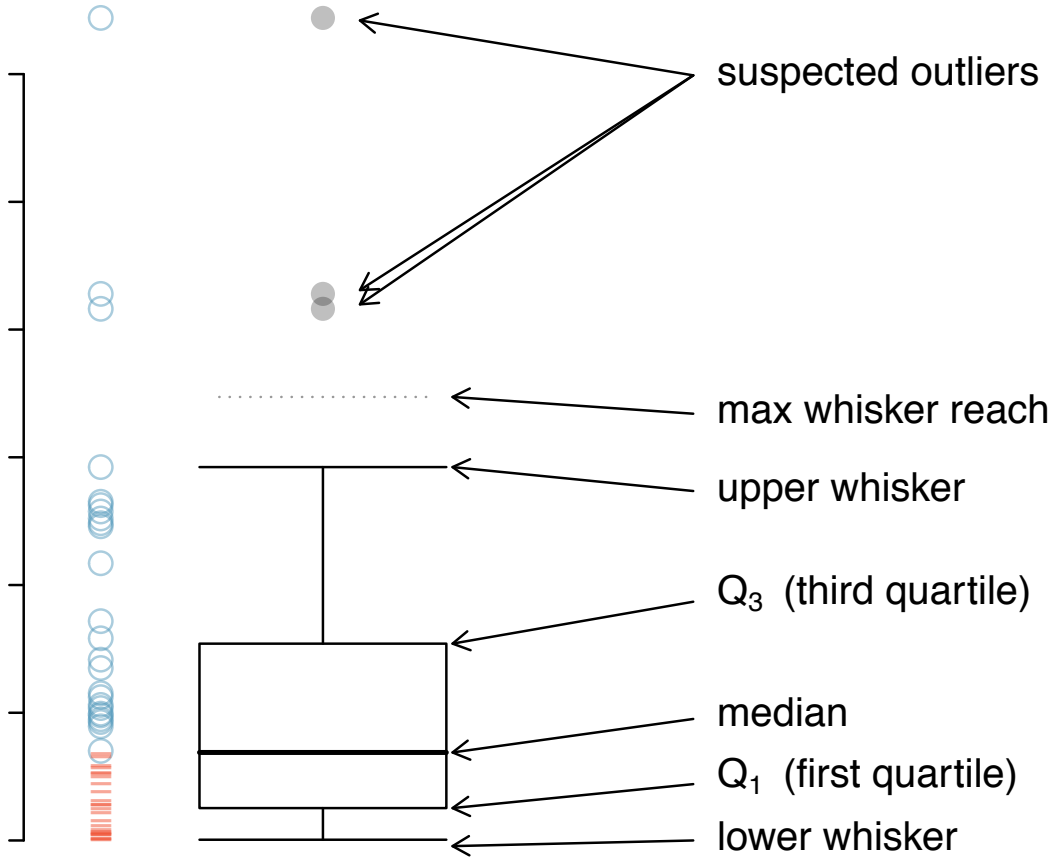
2. Housekeeping

3. 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

4. Application exercises

5. Summary

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



Outline

1. Readiness assessment

2. Housekeeping

3. 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

4. Application exercises

5. Summary

Application exercise: 1.1 Distributions of numerical variables

See the course website for instructions.

Outline

1. Readiness assessment

2. Housekeeping

3. 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

4. Application exercises

5. Summary

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