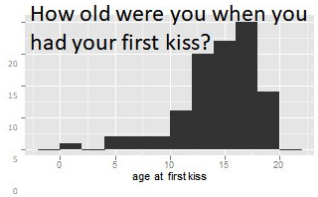
Unit 1: Introduction to data
2. Exploratory data analysis

Sta 101 – Spring 2019

Duke University, Department of Statistical Science



Dr. Ellison

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

Register your Clicker!

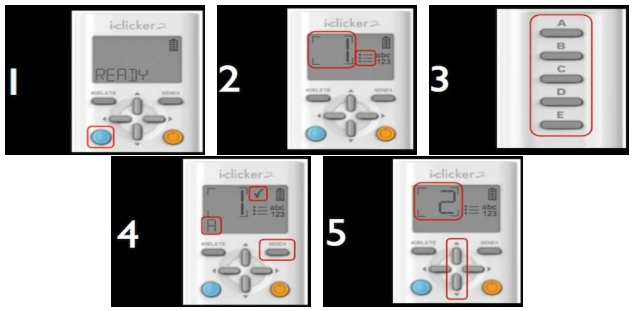
To do now:

1. Turn on your clicker (orange button)
2. Wait about 6 seconds.
3. IF your clicker says "READY,":
   a) Look for when your name appears on the slides. *(If you don't see your name and you are officially enrolled in the course let me know!)*
   b) When you see your name, type the 4 letters you see under it (you have 15 seconds).
   c) If your name box turned another color, your clicker should now be registered to the class!
4. IF your clicker does not say "READY":
   1. Hold down on the orange button until the clicker screen changes.
   2. Quickly press AA.
   3. Wait ~6 seconds.
   4. Your screen should now say "READY". *(If not, ask a TA for help!)*

We will be registering clickers 1/14, 1/16, 1/23… clicker grading begins on 1/28!

Readiness assessment

- *Individual:* 15 minutes, using clickers



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

Summary of main ideas

To Do:

1. Getting to Know you Survey + Pretest **due tomorrow 1/15**
2. Start working on Problem Set 1 **(Due Friday 1/26)**

Outline

I. Readiness assessment

II. Housekeeping

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

IV. Application exercises

V. Summary

Outline

I. Readiness assessment

II. Housekeeping

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

IV. Application exercises

V. Summary

Outline

I. Readiness assessment

II. Housekeeping

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

IV. Application exercises

V. Summary

Outline

I. Readiness assessment

II. Housekeeping

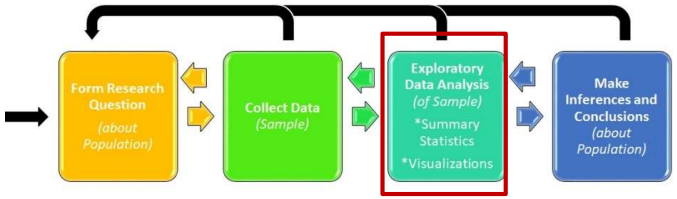
III. Main ideas
   1. Analysis work flow:
      1. Always start your exploration with a visualization
   2. Single Numerical Variable
      1. When describing numerical distributions discuss shape, center, spread, and unusual observations
      2. Robust statistics are not easily affected by outliers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

IV. Application exercises

V. Summary

Outline

Should we calculate a summary statistic or make a data visualization first?



From a past Sta 101 survey...

Do you see anything out of the ordinary?

3

From a past Sta 101 survey...

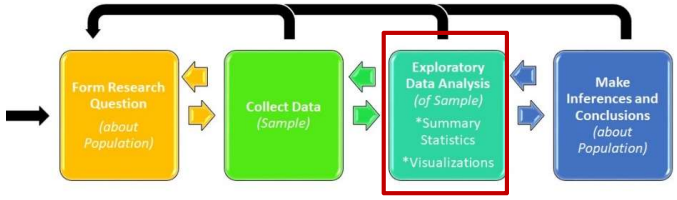
Do you see anything out of the ordinary?

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

3

Outline

We should start our exploratory data analysis with a visualization first!



Outline

Lab Hint: When asked to describe a visualization of a single numerical variable, there are four things we should always discuss. What are they?

Outline

I. Readiness assessment

II. Housekeeping

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

IV. Application exercises

V. Summary

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

Outline

What are some things to think about when guessing the distribution of any variable?

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

8

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

8

Outline

Think about natural boundaries!

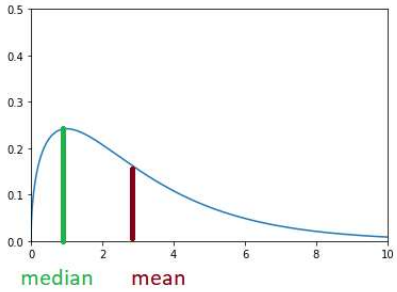
Application exercise: 1.1 Distributions of numerical variables

See the course website for instructions.

15

Outline

How do mean, median, and skewness relate?



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

Mean vs. median

Clicker question

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

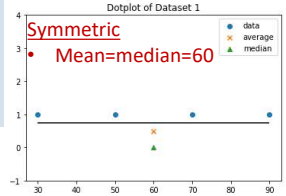
Dataset 1: 30, 50, 70, 90
Dataset 2: 30, 50, 70, 1000

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

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

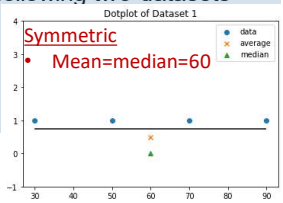
Mean vs. median

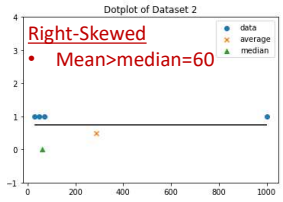
Clicker question

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

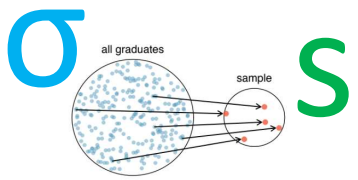




9

Outline

Why are the ways we calculate *population* standard deviation and *sample* standard deviation different?



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*.

10

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*.

10

More on SD

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

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.

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.

- More uncertainty introduced by using $\bar{x}$ instead of μ.
- All else held equal (ie. $\bar{x}$ = μ) :
  - n → n-1 … gets smaller!
  -

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

…gets larger…. incorporates more variation/uncertainty.

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?

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.)

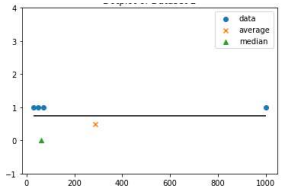
Why do we use the squared deviation in the calculation of variance?

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

11

Outline

Center:
When should we use mean vs. median?



Spread:
When should we use standard deviation vs. IQR vs. range?

Outline

Range and IQR

Clicker question

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

(a) Yes
(b) No

12

Range and IQR

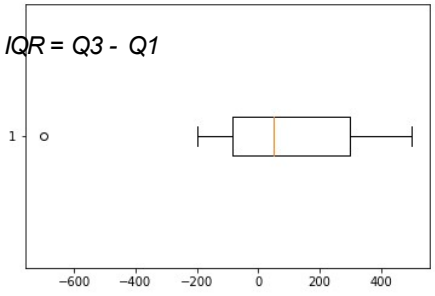
Clicker question

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



Range and IQR

Clicker question

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

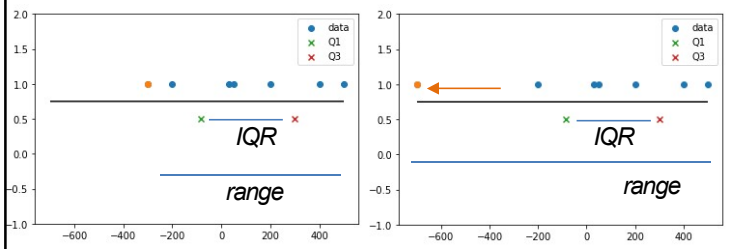
(a) Yes

(b) No

Range = max - min, IQR = Q3 - Q1

Is the range or the IQR more robust to outliers?

Range and IQR



Is the range or the IQR more robust to outliers?

IQR

Robust statistics

- 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 to outliers.**
- Therefore we choose **median & IQR** (over mean&SD) when describing skewed distributions.
- We choose **mean & SD** when describing symmetric distributions, as they are more useful in using mathematical theory to make inferences.

Outline

How do we determine if a data point is an outlier in a numerical distribution?

How do we construct a boxplot?

https://www.kdnuggets.com/2017/01/3-methods-deal-outliers.htm

Outline

I. Readiness assessment

II. Housekeeping

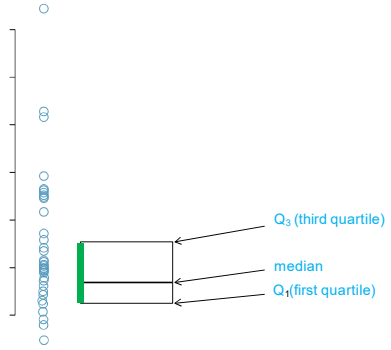
III. Main ideas
   1. Analysis work flow:
      1. NEW Always start your exploration with a visualization
   2. Single Numerical Variable
      1. When describing numerical distributions discuss shape, center, spread, and unusual observations
      2. Robust statistics are not easily affected by outliers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

IV. Application exercises

V. Summary

Box plot

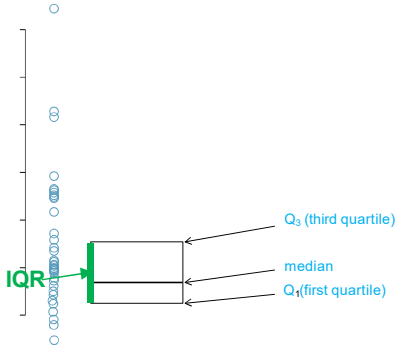
A *box plot* visualizes the median, the quartiles, and suspected outliers. **An *outlier* is defined as an observation more than 1.5×IQR away from the quartiles (Q1 and Q3).**



14

Box plot

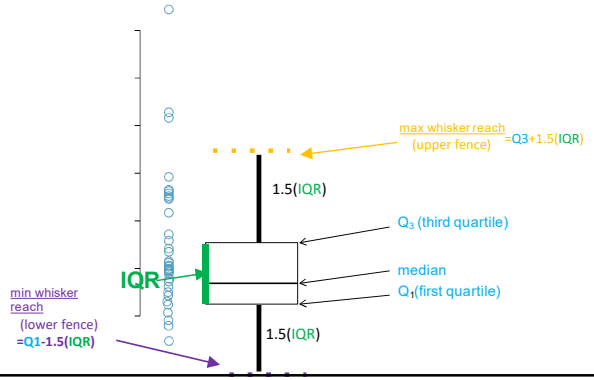
A *box plot* visualizes the median, the quartiles, and suspected outliers. **An *outlier* is defined as an observation more than 1.5×IQR away from the quartiles (Q1 and Q3).**



14

Box plot

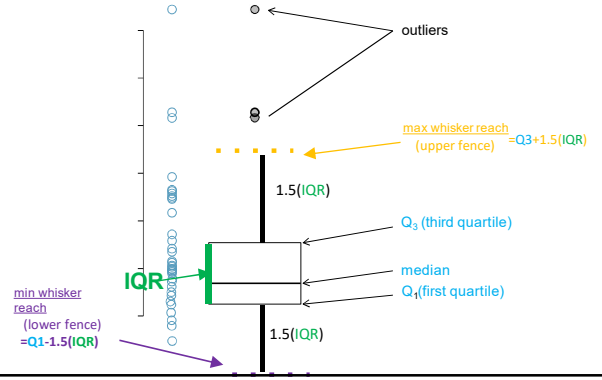
A *box plot* visualizes the median, the quartiles, and suspected outliers. **An *outlier* is defined as an observation more than 1.5×IQR away from the quartiles (Q1 and Q3).**



14

Box plot

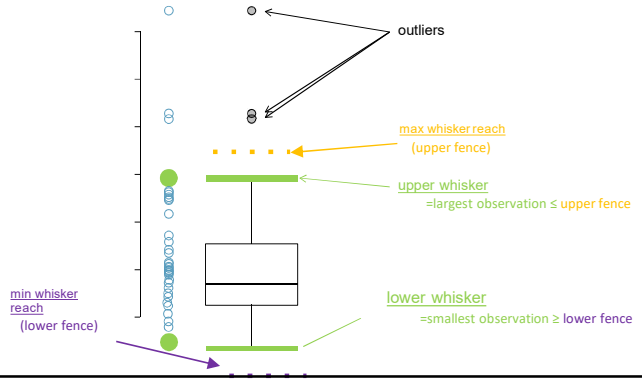
A *box plot* visualizes the median, the quartiles, and suspected outliers. **An *outlier* is defined as an observation more than 1.5×IQR away from the quartiles (Q1 and Q3).**



14

Box plot

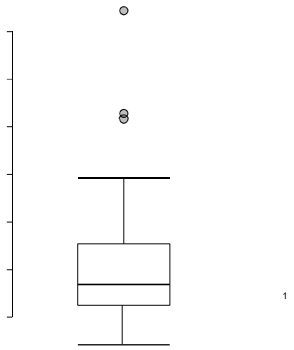
A *box plot* visualizes the median, the quartiles, and suspected outliers. **An *outlier* is defined as an observation more than 1.5×IQR away from the quartiles (Q1 and Q3).**



14

Box plot

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 (Q1 and Q3).**



14

Outline

I. Readiness assessment

II. Housekeeping

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

IV. Application exercises

V. Summary