Unit 1: Introduction to Data

3. More Exploratory Data Analysis

10/4/2017

Descriptive statistics

What's the difference between .mp3 and .FLAC?
.jpeg and .png?

.mp3 and .jpeg are **lossy compression** -- they make data keeping only the most important or representative parts of it.

Descriptive statistics are kind of lossy compression: **What one/few number(s) that best represent my data**?

Key ideas

1. Good visualizations help you understand your data
2. Descriptive statistics compress data so you can communicate about it
3. The “right” statistics depend on the shape of the data

Center and Variability

A common measure of central tendency is the mean, denoted as $\bar{x}$:

$$\bar{x} = \frac{x_1 + x_2 + \cdots x_n}{n}$$

A common measure of central tendency is the standard deviation, denoted as s:

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

Why do we need to talk about both center and spread?

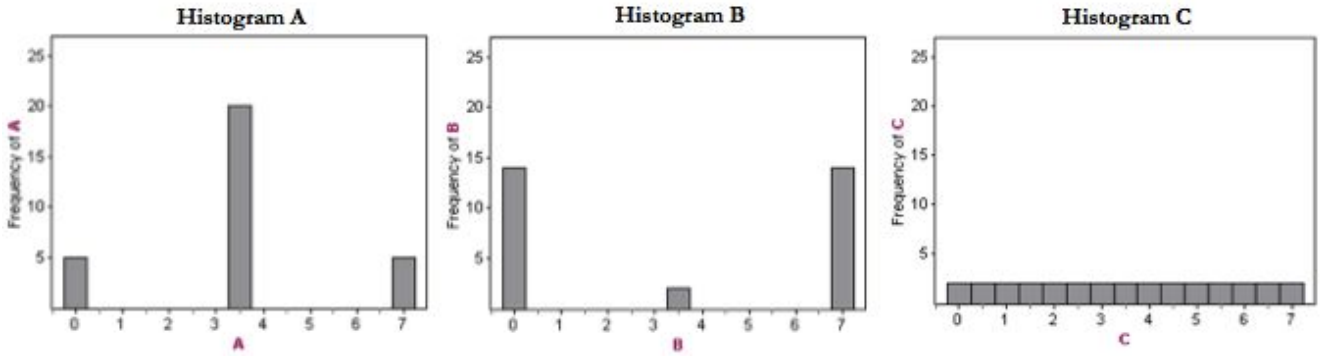
Spread tells you how well your central tendency represents your data

	# people at Sally's book club	# people at Maria's book club
Week 1	8	1
Week 2	10	18
Week 3	11	10
Week 4	9	2
Week 5	12	19
Mean	$= \frac{8+10+11+9+12}{5} = 10$	$= \frac{1+18+10+2+19}{5} = 10$
Standard Deviation	$= \sqrt{\frac{(8-10)^2+(10-10)^2+(11-10)^2+(9-10)^2+(12-10)^2}{4}} \approx 1.6$	$= \sqrt{\frac{(1-10)^2+(18-10)^2+(10-10)^2+(2-10)^2+(19-10)^2}{4}} \approx 8.5$

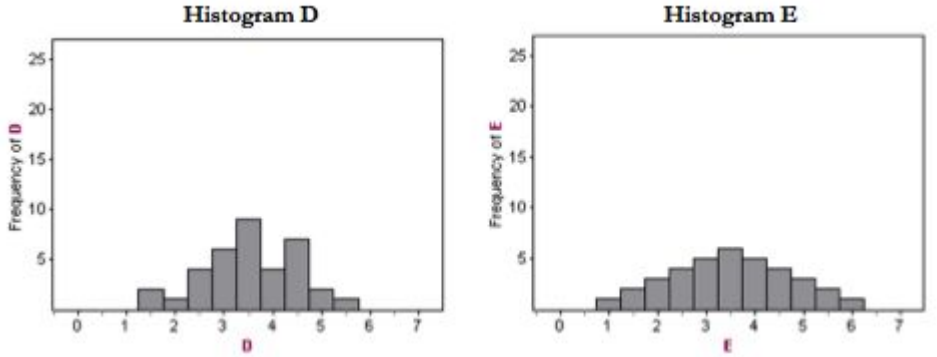
Thanks Deena Bernett

Practice Question 1

Which of these is most variable?



Which of these is more variable?



Let’s look at our data from monday

Should you always use the mean to measure central tendency?

Can you think of something where almost everyone in the population is above the mean?

When distributions are not symmetric, the mean can sometimes be misleading.

Median

The median is the value that splits the data in half when ordered in ascending order.

$$0, 1, 2, 3, 4$$

If there are an even number of observations, then the median is the average of the two values in the middle.

$$0, 1, \underline{2, 3}, 4, 5 \rightarrow \frac{2+3}{2} = 2.5$$

Since the median is the midpoint of the data, 50% of the values are below it. Hence, it is also the 50th percentile.

Practice Question 2

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) $\bar{x}_1 > \bar{x}_2$, $\text{median}_1 = \text{median}_2$

Practice Question 2

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) $\bar{x}_1 > \bar{x}_2$, $\text{median}_1 = \text{median}_2$

Interquartile range to measure spread

The 25th percentile is called the first quartile (Q1)

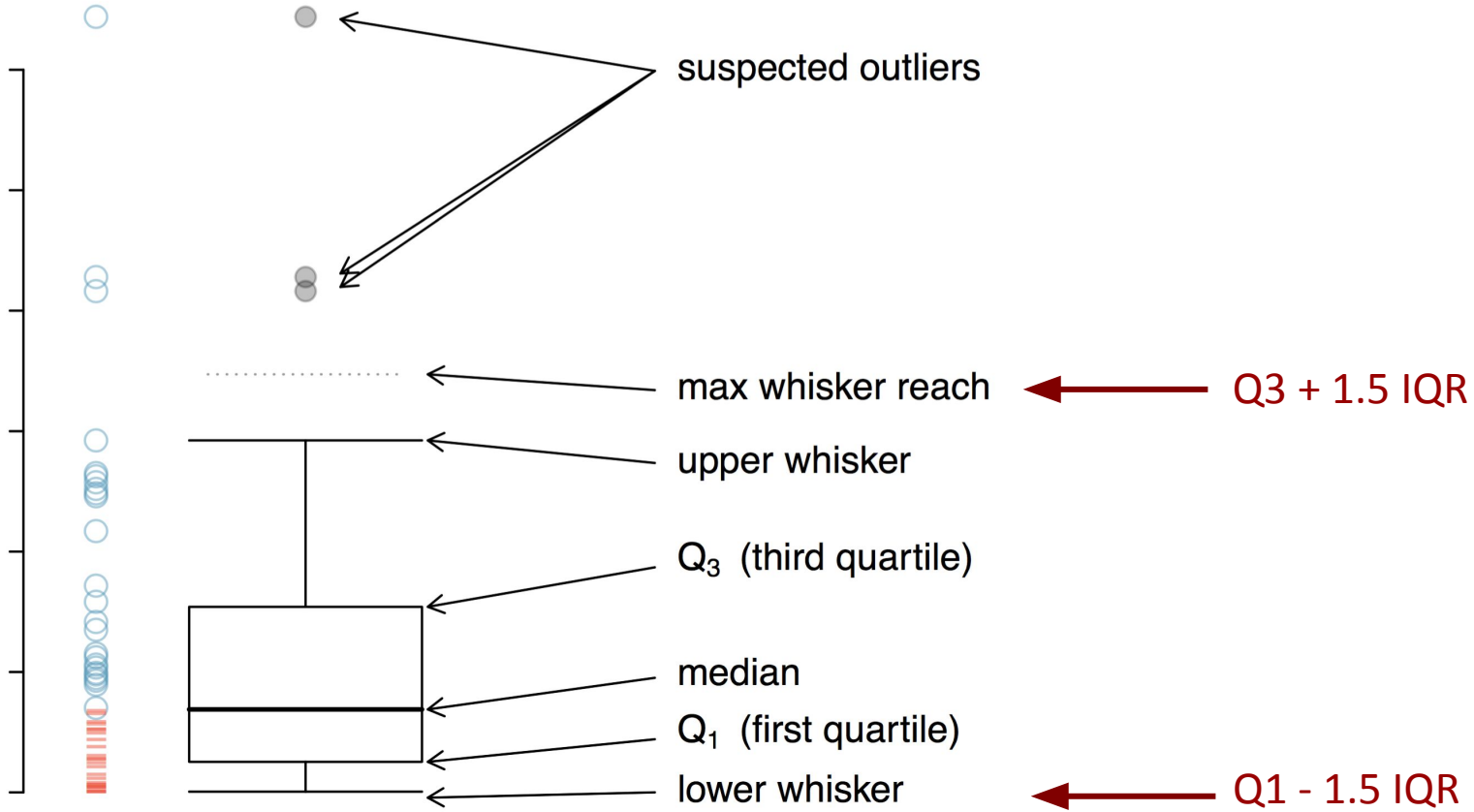
The 50th percentile is called the median

The 75th percentile is called the third quartile (Q3)

Between the first and third quartile are 50% of the data. The range between Q1 and Q3 is called Interquartile Range or IQR

$$IQR = Q3 - Q1$$

Boxplots are visualizations that use percentiles to compress the data



Outliers

A potential outlier is defined as an observation beyond the maximum reach of the whiskers. It is appears extreme relative to the rest of the data.

Why is it important to look for outliers?

- Identify extreme skew in the distribution.
- Identify data collection and entry errors.
- Provide insight into interesting features of the data.

Robust statistics

Median and IQR are more robust to skewness and outliers than mean and SD.

- for skewed distributions it is often more helpful to use median and IQR to describe the center and spread
- for symmetric distributions it is often more helpful to use the mean and SD to describe the center and spread

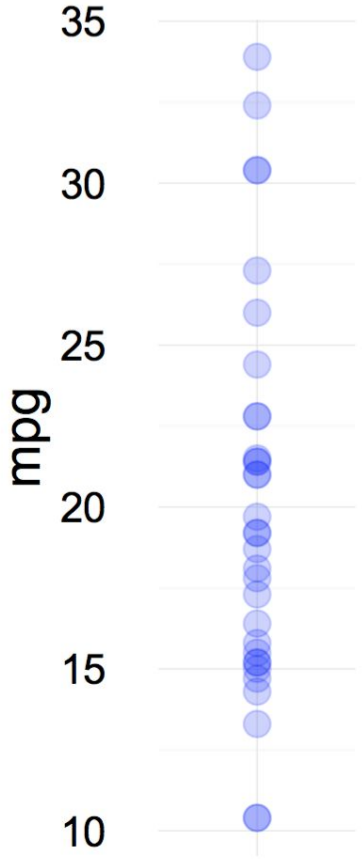
If you would like to estimate the typical household income in the US, would you be more interested in the mean or median income?

Median

Good visualizations

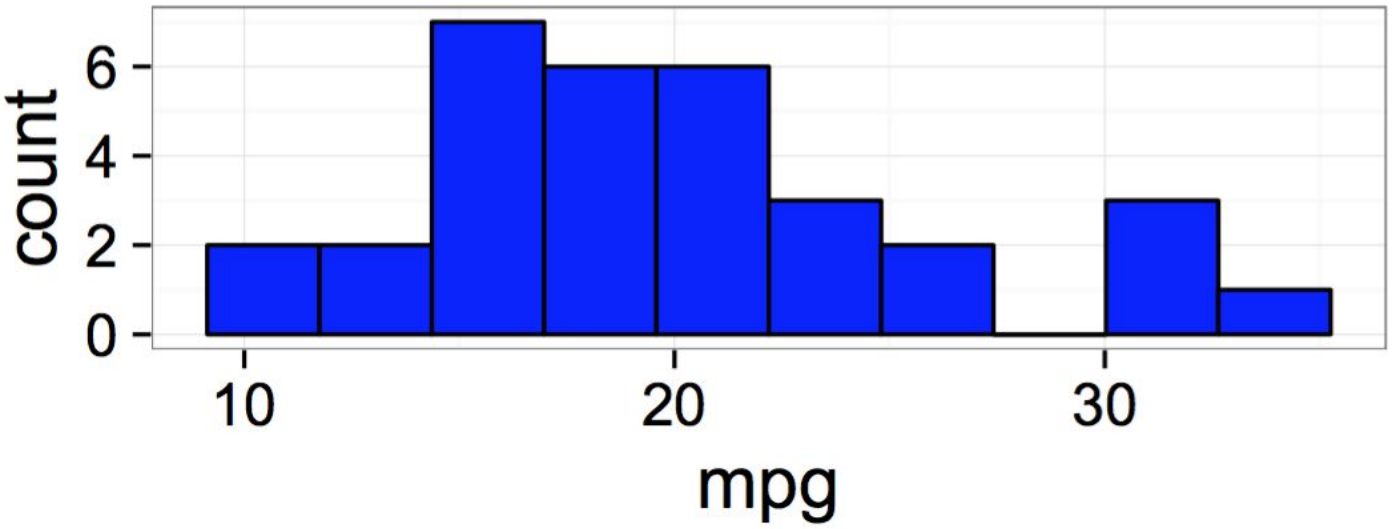
A good visualization makes your intuitions when seeing the data match the results of your statistical analyses

Dot plots make it easy to see where most of the data is.



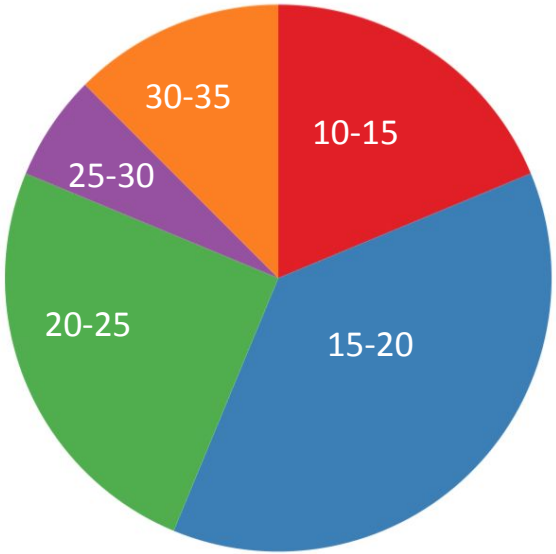
Good visualizations

Histograms make it easy to see where most of the data is.



Bad visualizations

Pie charts make it difficult to see where most of the data is



Why?

We are not good at integrating dimensions

Children under ~7 will fail at this conservation task

But so will you if I don’t pour the water in front of you!

Piaget (1965)

Key ideas

1. Good visualizations help you understand your data
2. Descriptive statistics compress data so you can communicate about it
3. The “right” statistics depend on the shape of the data