

Unit 1: Introduction to Data

2. More Exploratory Data Analysis (Chapter 1.6)

1/27/2020

Central tendency

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

.mp3 and .jpeg are **lossy compression** -- they make data smaller by throwing some of it away.

Central tendency is a kind of lossy compression: **What one number is the most representative of 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

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

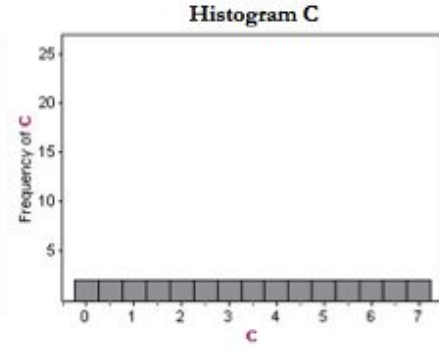
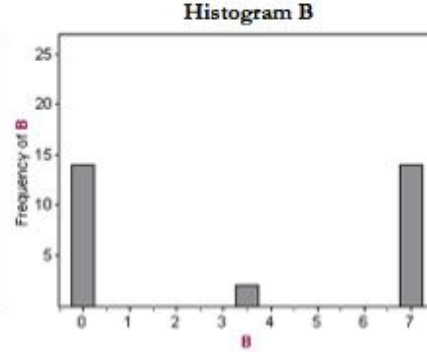
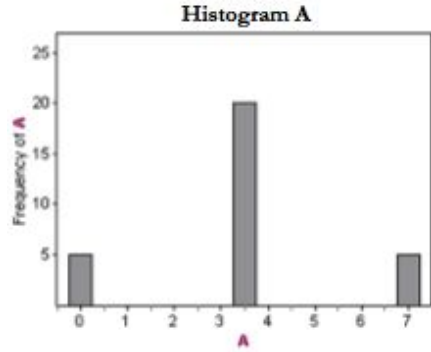
Why do we care about both center and spread?

Spread tells you how well your central tendency represents your data

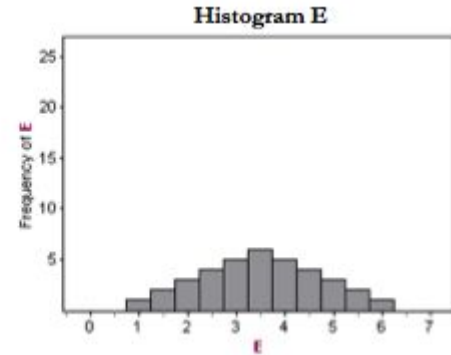
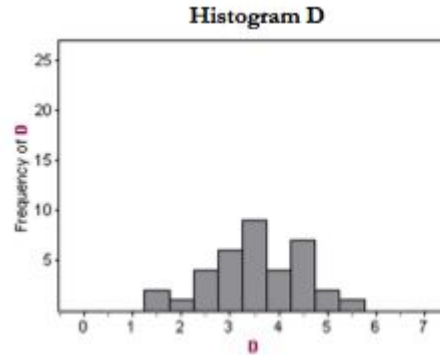
	# people at Sally's book club	# people at Maria's book club
Week 1	8	1
Week 2	10	18
Week 3	11	10
Week 4	9	2
Week 5	12	19
Mean	$= \frac{8 + 10 + 11 + 9 + 12}{5} = 10$	$= \frac{1 + 18 + 10 + 2 + 19}{5} = 10$
Standard Deviation	$= \sqrt{\frac{(8 - 10)^2 + (10 - 10)^2 + (11 - 10)^2 + (9 - 10)^2 + (12 - 10)^2}{4}}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

Practice Question 1

Which of these is most variable?



Which of these is more variable?



Let's look at our data from wednesday

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.

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

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

Can both the mean and median be misleading?

unimodal



bimodal



multimodal



uniform



Interquartile range to measure spread

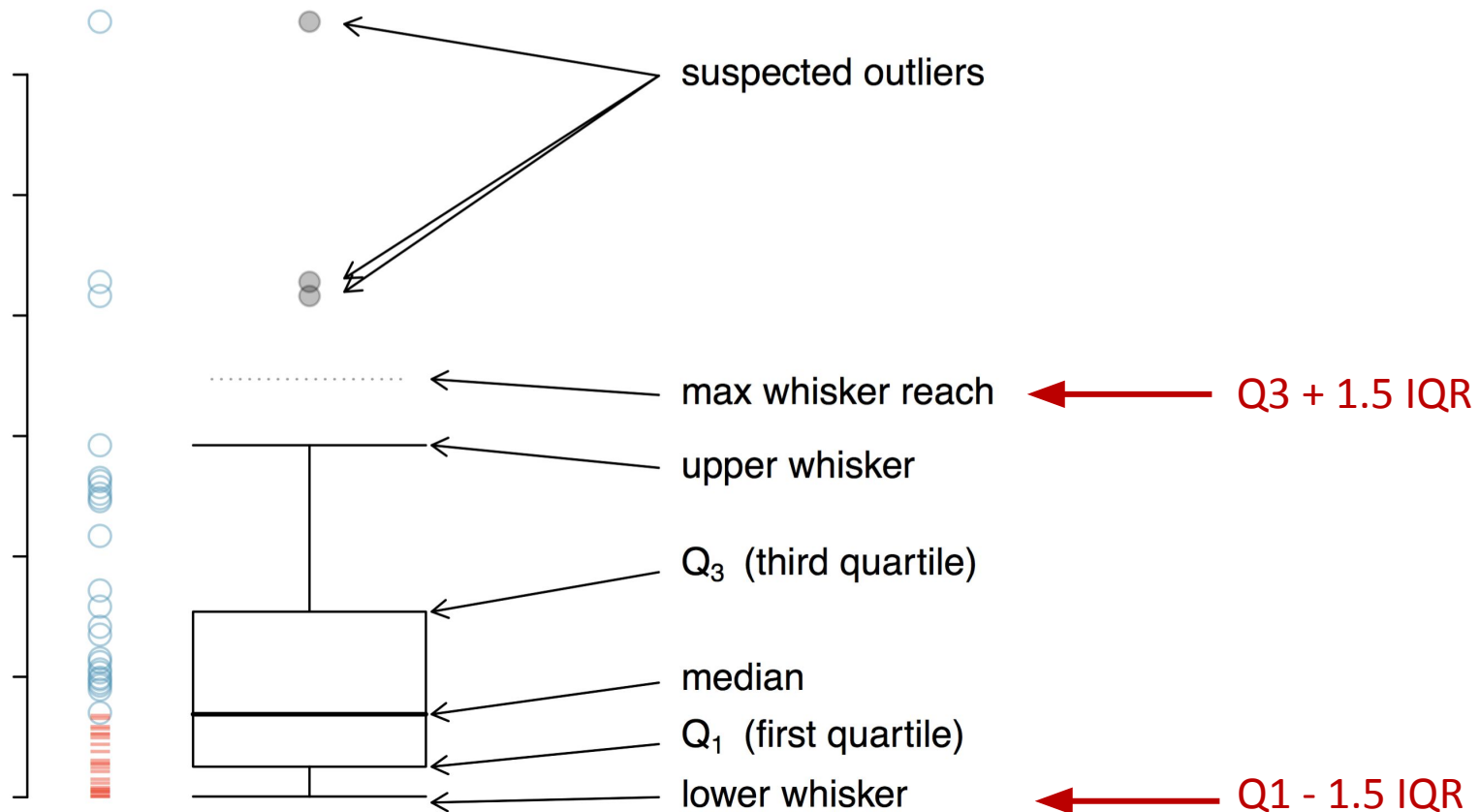
The 25th percentile is called the first quartile (Q1)

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**

Boxplots are visualizations that use percentiles to compress data



Outliers

A potential **outlier** is defined as an observation beyond the maximum reach of the whiskers.

It 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

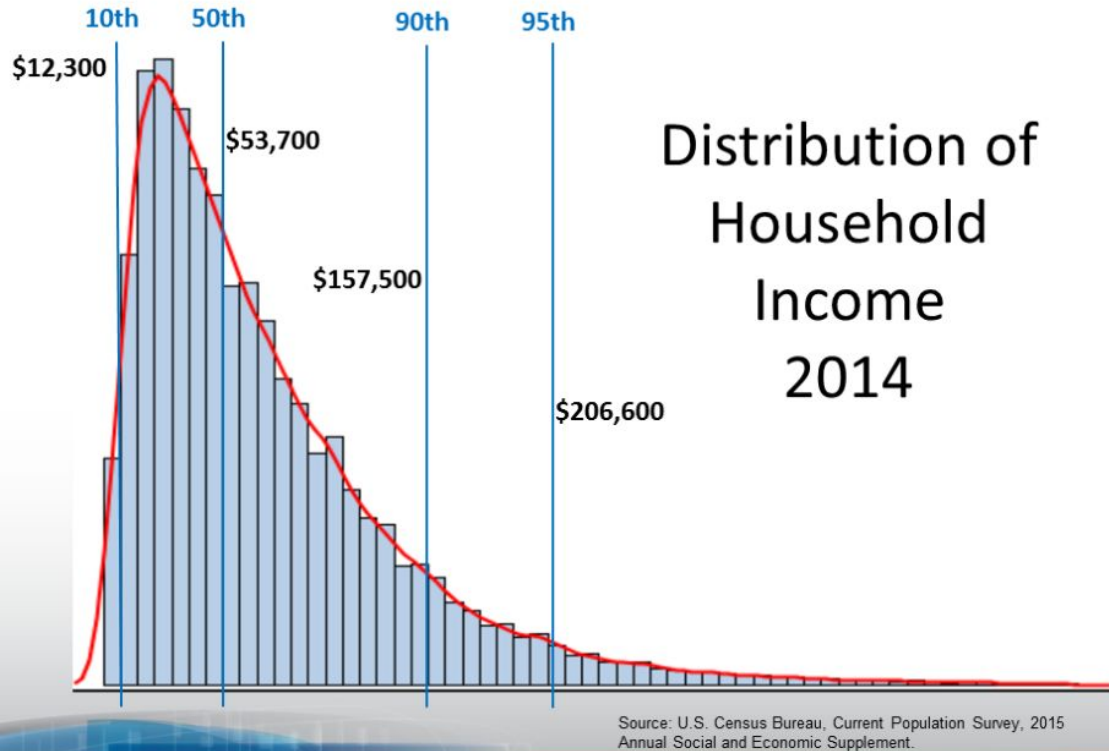
Median and **IQR** are more robust to skewness and outliers than **mean** and **SD**.

- 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 mean or median income?

Median

US Household Income



Median: \$53,700

Mean: \$75,738

What makes a good visualization?

```
> mtcars
      mpg cyl  disp  hp drat   wt  qsec vs am gear carb
Mazda RX4    21.0   6 160.0 110 3.90 2.620 16.46 0 1 4 4
Mazda RX4 Wag 21.0   6 160.0 110 3.90 2.875 17.02 0 1 4 4
Datsun 710    22.8   4 108.0  93 3.85 2.320 18.61 1 1 4 1
Hornet 4 Drive 21.4   6 258.0 110 3.08 3.215 19.44 1 0 3 1
Hornet Sportabout 18.7  8 360.0 175 3.15 3.440 17.02 0 0 3 2
Valiant      18.1   6 225.0 105 2.76 3.460 20.22 1 0 3 1
Duster 360   14.3   8 360.0 245 3.21 3.570 15.84 0 0 3 4
Merc 240D    24.4   4 146.7  62 3.69 3.190 20.00 1 0 4 2
Merc 230     22.8   4 140.8  95 3.92 3.150 22.90 1 0 4 2
Merc 280     19.2   6 167.6 123 3.92 3.440 18.30 1 0 4 4
Merc 280C    17.8   6 167.6 123 3.92 3.440 18.90 1 0 4 4
Merc 450SE   16.4   8 275.8 180 3.07 4.070 17.40 0 0 3 3
Merc 450SL   17.3   8 275.8 180 3.07 3.730 17.60 0 0 3 3
Merc 450SLC  15.2   8 275.8 180 3.07 3.780 18.00 0 0 3 3
Cadillac Fleetwood 10.4  8 472.0 205 2.93 5.250 17.98 0 0 3 4
Lincoln Continental 10.4  8 460.0 215 3.00 5.424 17.82 0 0 3 4
Chrysler Imperial 14.7  8 440.0 230 3.23 5.345 17.42 0 0 3 4
Fiat 128     32.4   4  78.7  66 4.08 2.200 19.47 1 1 4 1
Honda Civic  30.4   4  75.7  52 4.93 1.615 18.52 1 1 4 2
Toyota Corolla 33.9   4  71.1  65 4.22 1.835 19.90 1 1 4 1
Toyota Corona 21.5   4 120.1  97 3.70 2.465 20.01 1 0 3 1
Dodge Challenger 15.5  8 318.0 150 2.76 3.520 16.87 0 0 3 2
AMC Javelin  15.2   8 304.0 150 3.15 3.435 17.30 0 0 3 2
Camaro Z28   13.3   8 350.0 245 3.73 3.840 15.41 0 0 3 4
Pontiac Firebird 19.2  8 400.0 175 3.08 3.845 17.05 0 0 3 2
Fiat X1-9    27.3   4  79.0  66 4.08 1.935 18.90 1 1 4 1
Porsche 914-2 26.0   4 120.3  91 4.43 2.140 16.70 0 1 5 2
Lotus Europa 30.4   4  95.1 113 3.77 1.513 16.90 1 1 5 2
Ford Pantera L 15.8   8 351.0 264 4.22 3.170 14.50 0 1 5 4
Ferrari Dino  19.7   6 145.0 175 3.62 2.770 15.50 0 1 5 6
Maserati Bora 15.0   8 301.0 335 3.54 3.570 14.60 0 1 5 8
Volvo 142E   21.4   4 121.0 109 4.11 2.780 18.60 1 1 4 2
```

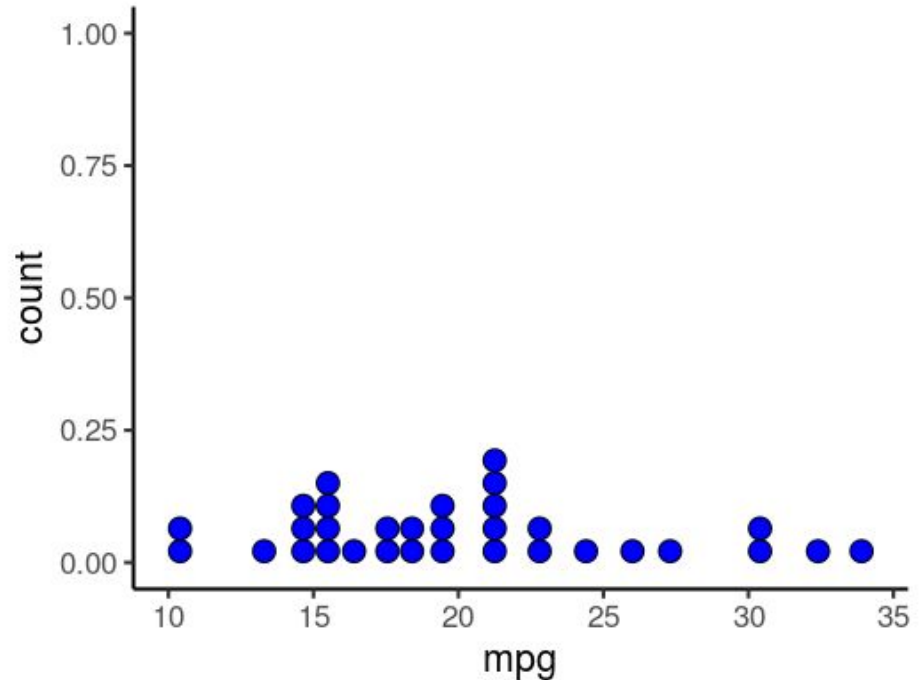
Let's look at the miles per gallon of these cars

Good visualizations

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

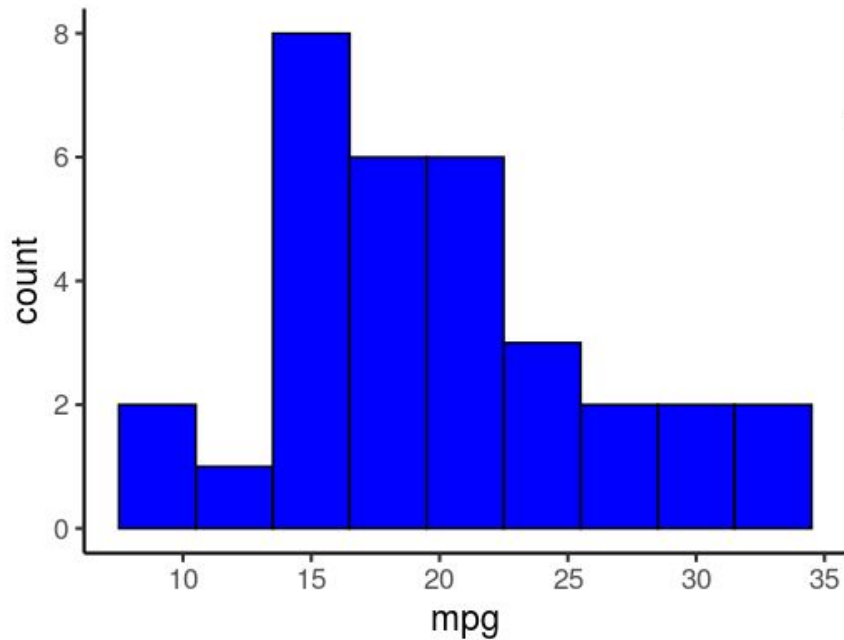
Dot plots make it easy to see where most of the data is.

```
mtcars %>%  
  ggplot(aes(x = mpg)) +  
  geom_dotplot(fill = "blue", color = "black")
```



Good visualizations

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

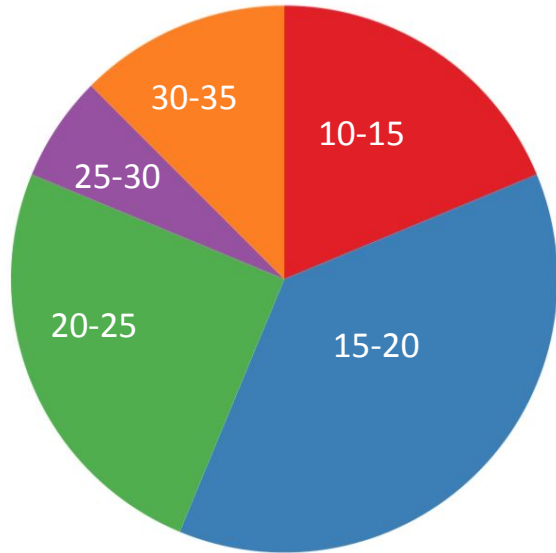


```
mtcars %>%  
  ggplot(aes(x = mpg)) +  
  geom_histogram(fill = "blue", color = "black",  
                 binwidth = 3)
```

Can they be bad sometimes?

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