Measures of spread

Overview

Review of shapes distributions and central tendency

The standard deviation

z-scores

Percentiles

Announcement: homework 1

Homework 1 is due at 11pm on Sunday January 28th

Use Ed Discussions for any questions that come up, and/or attend office hours

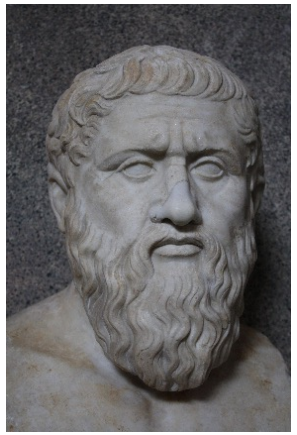
Upload pdfs with your answers to Gradescope

1. Hand in R Markdown pdf under the assignment called Homework 1
2. Make sure to Mark your pages on Gradescope!


ANY
QUESTIONS

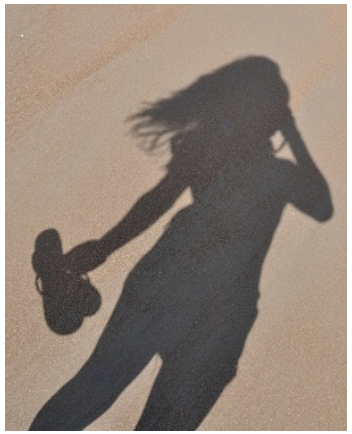
Review and continuation of…

Underlying concepts: the P's and the S's



P-Truth

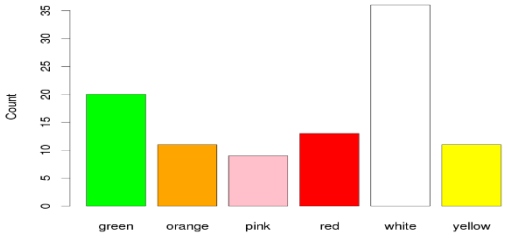
- population or process
- parameter
- Plato (Greek symbols)



S-shadows

- sample
- statistic
- shadow (Latin symbols)

Review: Categorical data and proportions

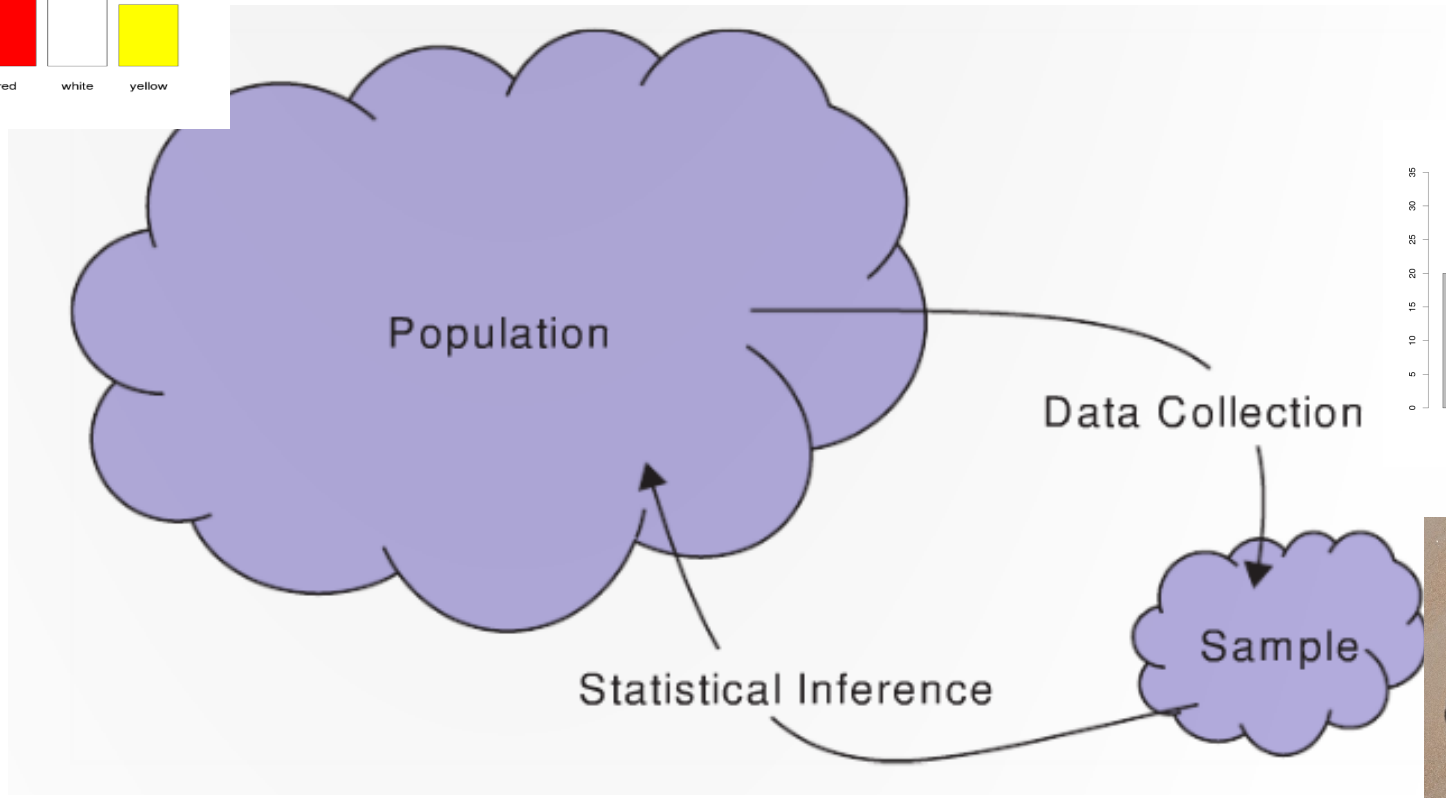


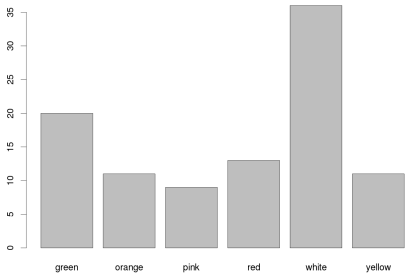
parameter

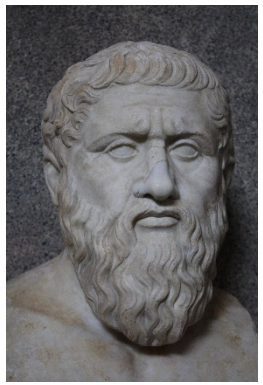
π

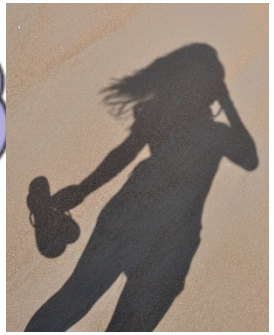
statistic

$\hat{p}$





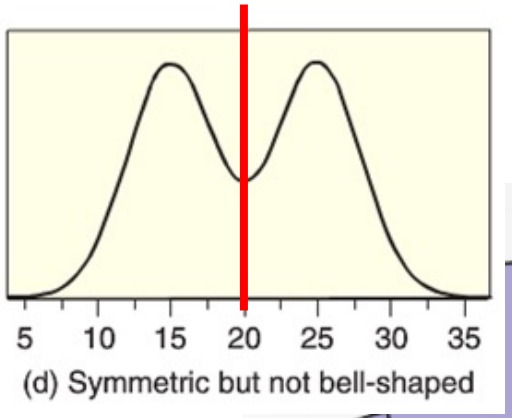




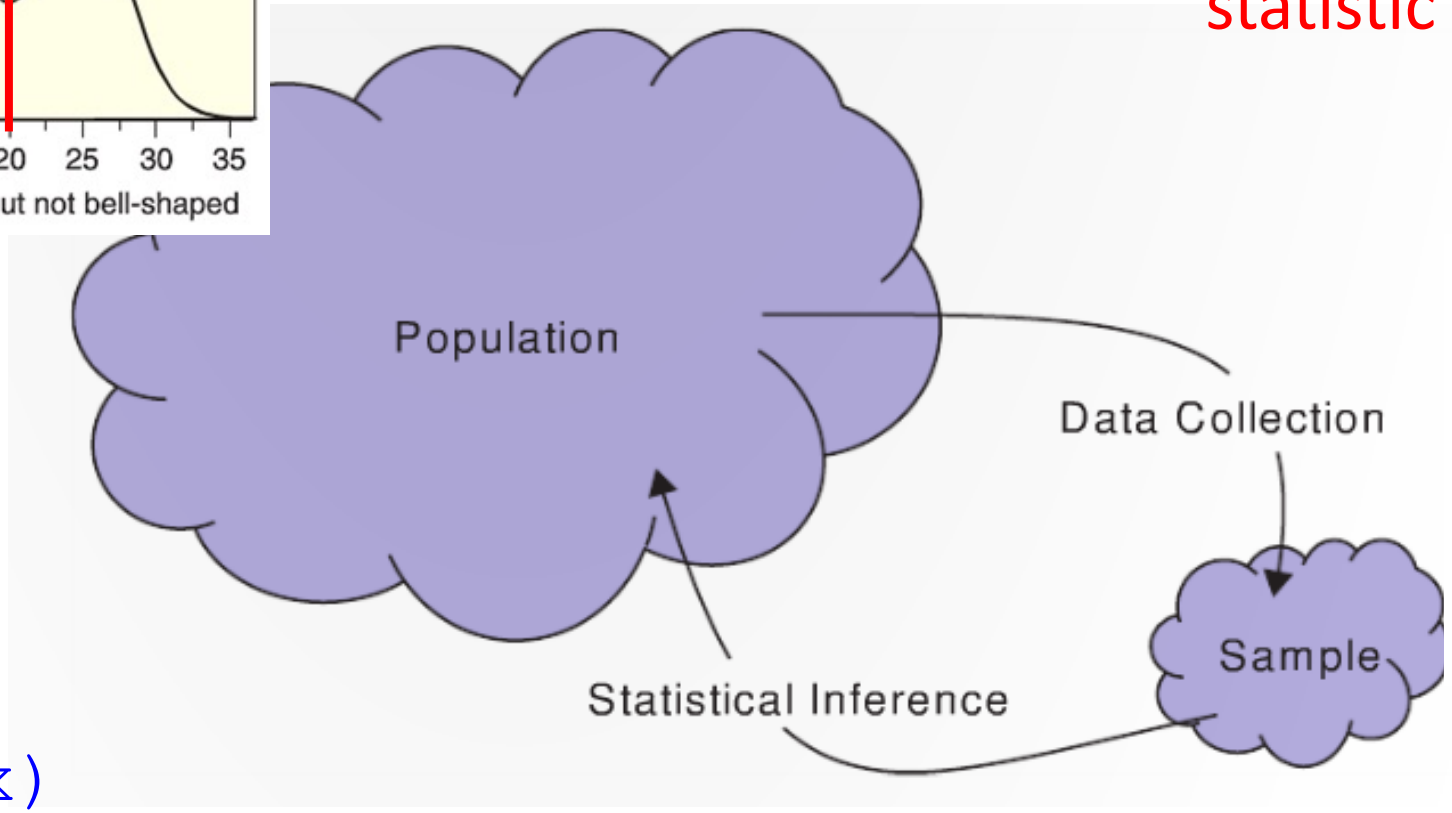
Review: Quantitative data and the mean

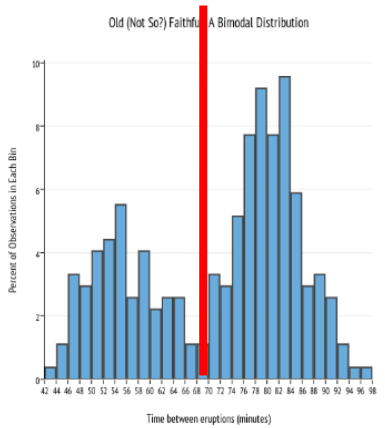
μ ← parameter

$$\bar{x} = \frac{\sum_i^n x_i}{n}$$



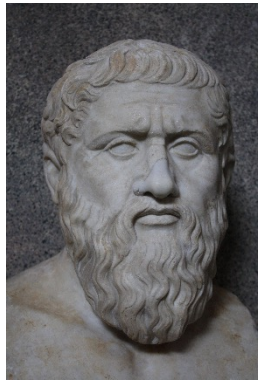
statistic → $\bar{x}$

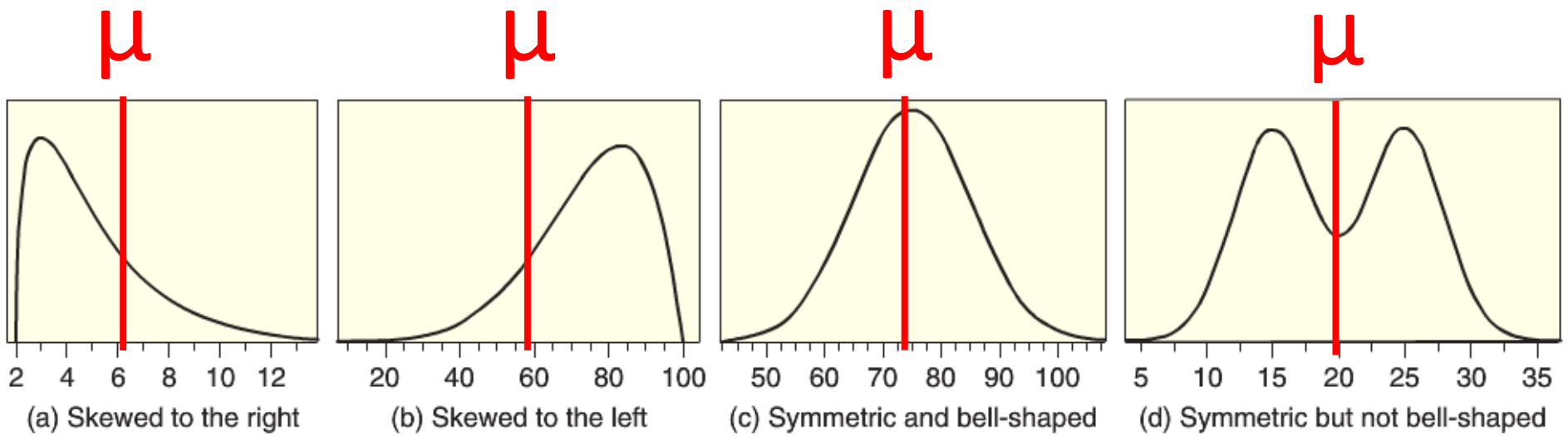


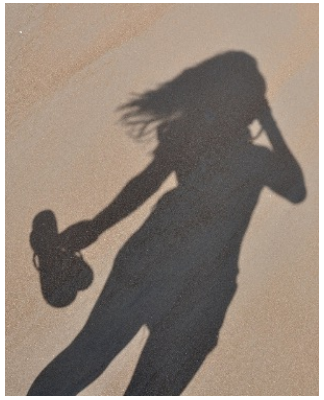


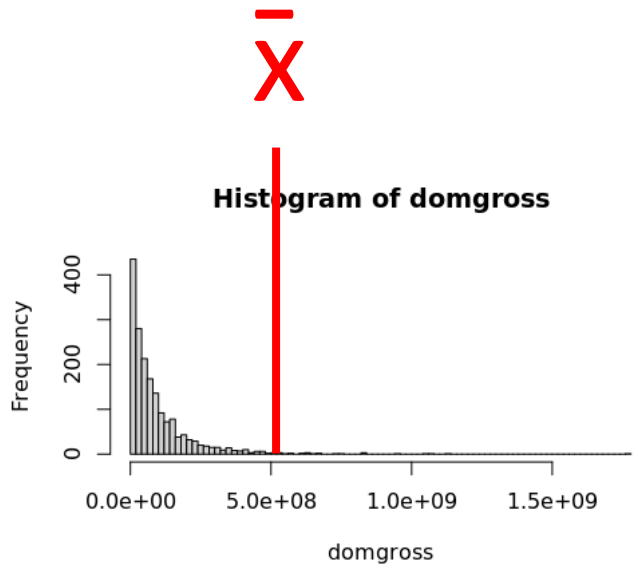
R: `mean(x)`

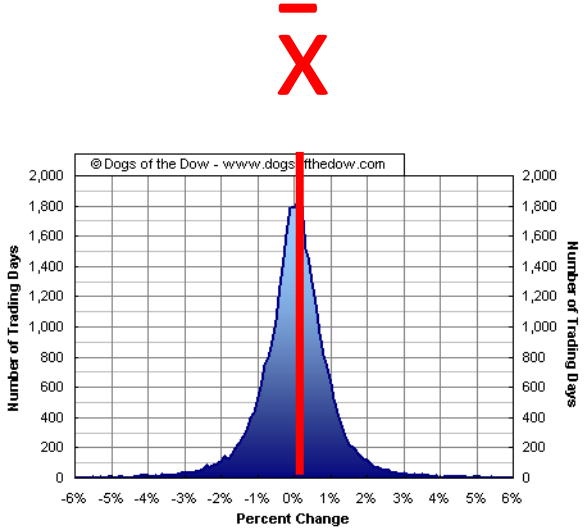
Plato and shadows: distributions and histograms

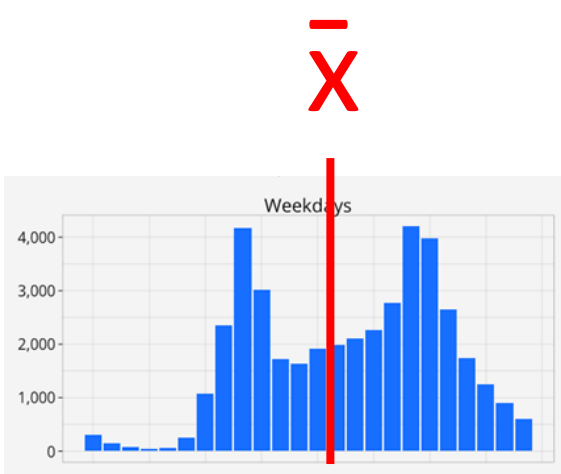






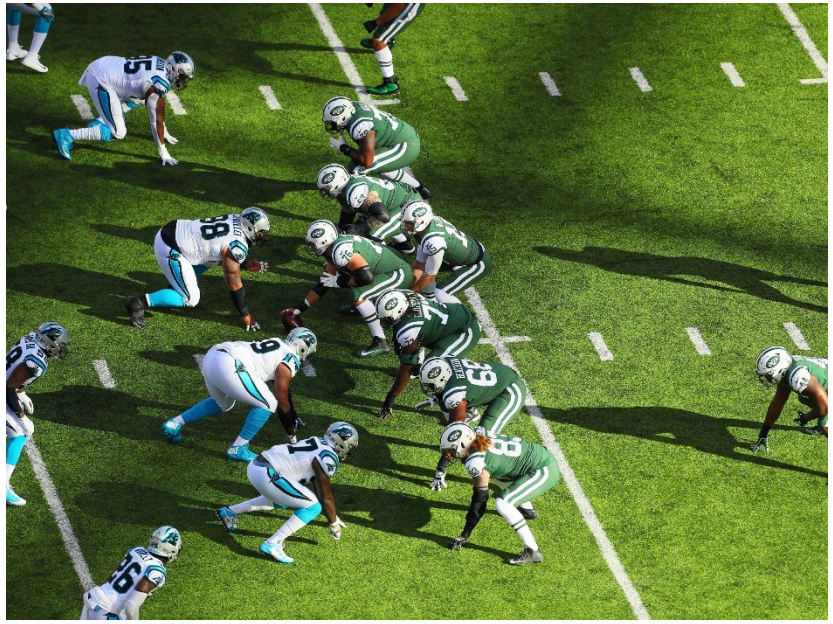


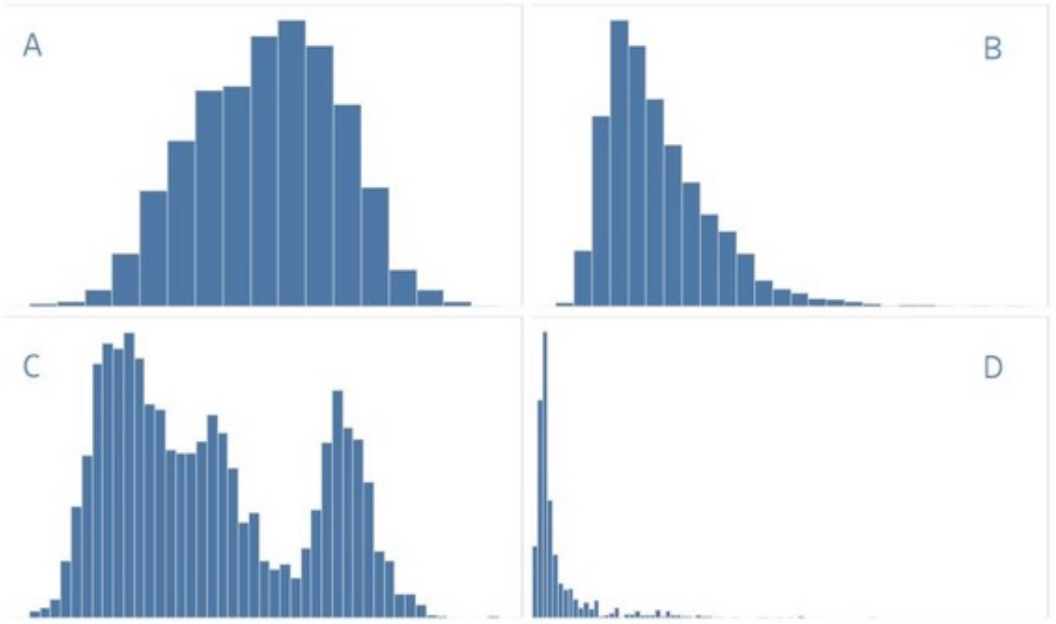




Neat facts – the average NFL player is:

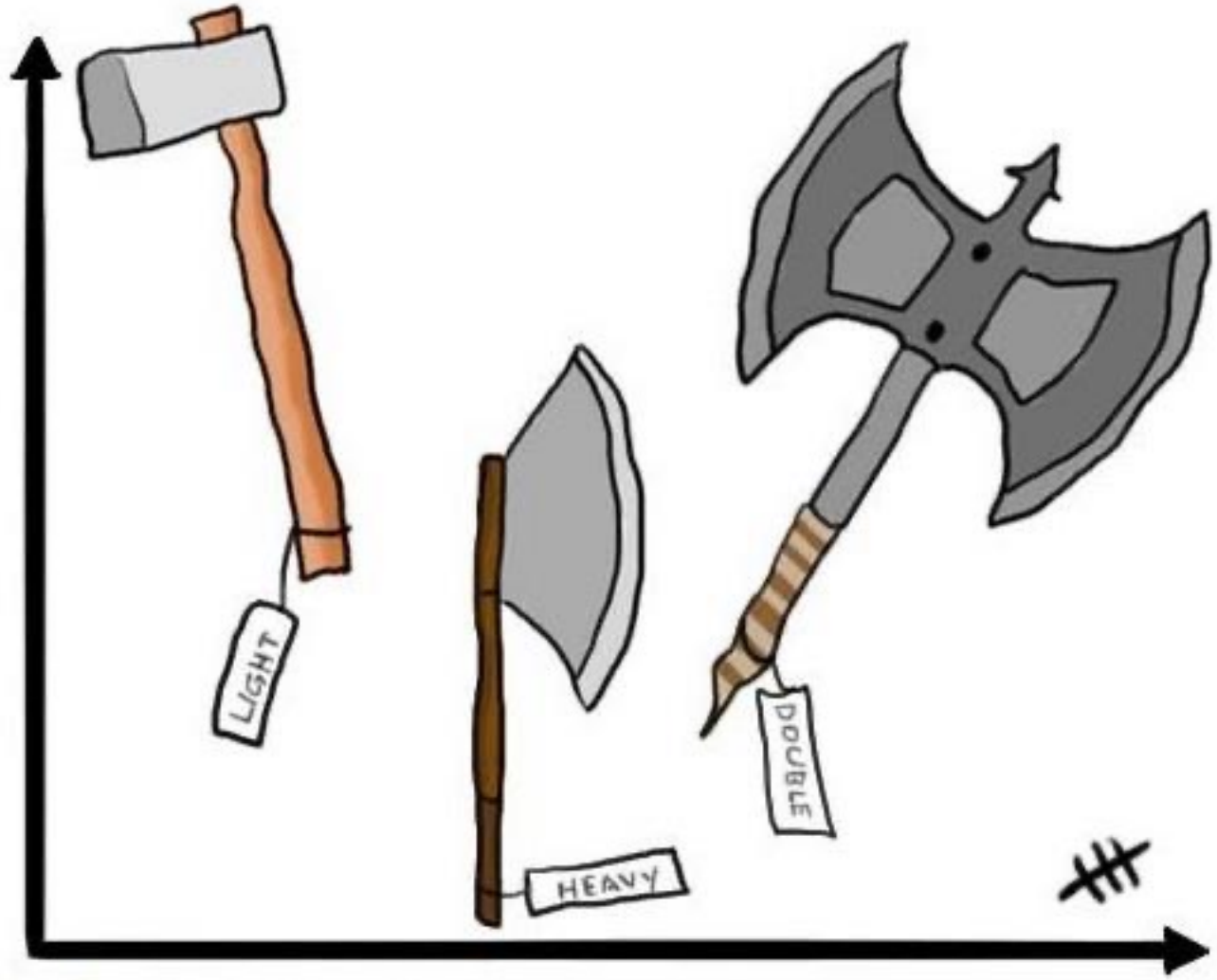
- 1. **Age**: Is about 25 years old
- 2. **Height**: Is just over 6'2" in height
- 3. **Weight**: Weighs a little more than 244lbs
- 4. **Salary**: Makes slightly less than $1.5M in salary per year





Question: Can you tell which histogram goes with which trait?

Always label your axes



Back to the Bechdel data…

```
# get the Bechdel data
> library("fivethirtyeight")
```

Try it in R!

```
SDS100::download_class_code(4)
```

year	title	binary	budget_2013	domgross_2013	intgross_2013
2013	21 & Over	FAIL	13000000	25682380	42195766
2012	Dredd 3D	PASS	45658735	13611086	41467257
2013	12 Years a Slave	FAIL	20000000	53107035	158607035
2013	2 Guns	FAIL	61000000	75612460	132493015
2013	42	FAIL	40000000	95020213	95020213
2013	47 Ronin	FAIL	225000000	38362475	145803842
2013	A Good Day to Die Hard	FAIL	92000000	67349198	304249198

Can you plot a histogram of the movie profits (domgross_2013) with 30 bins?

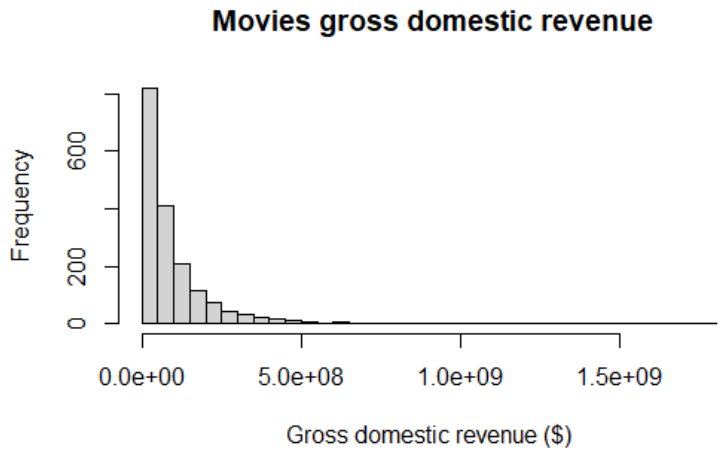
```
> domgross <-  bechdel$domgross_2013  # first create a vector with the population of each country
> hist(domgross , breaks = 30)      # then create the histogram
```

Labeling axes

Question: Can you figure out how to label the axes?

- > ? hist
- Answer: xlab and ylab!

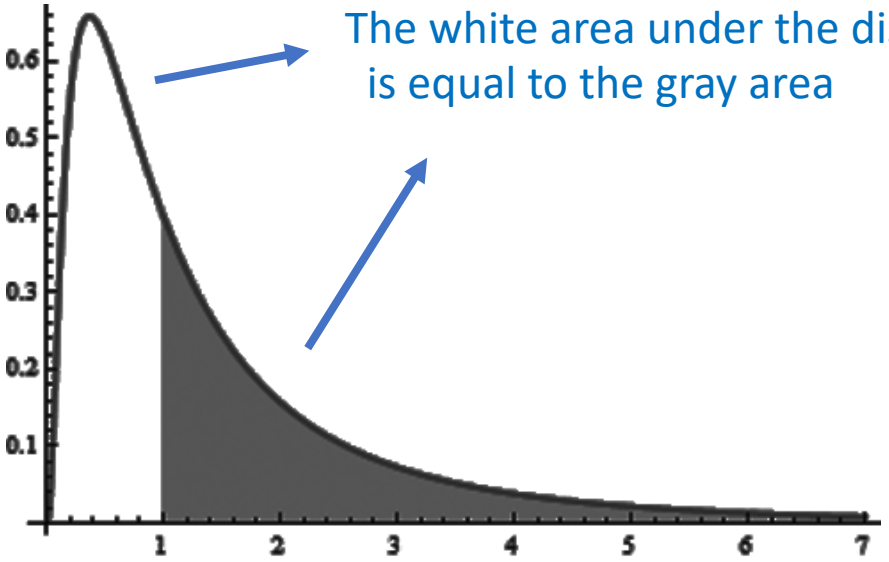
```
> hist(domgross, breaks = 30,
        ylab = "Frequency",
        xlab = "Gross domestic revenue ($)",
        main = " Movies gross domestic revenue")
```

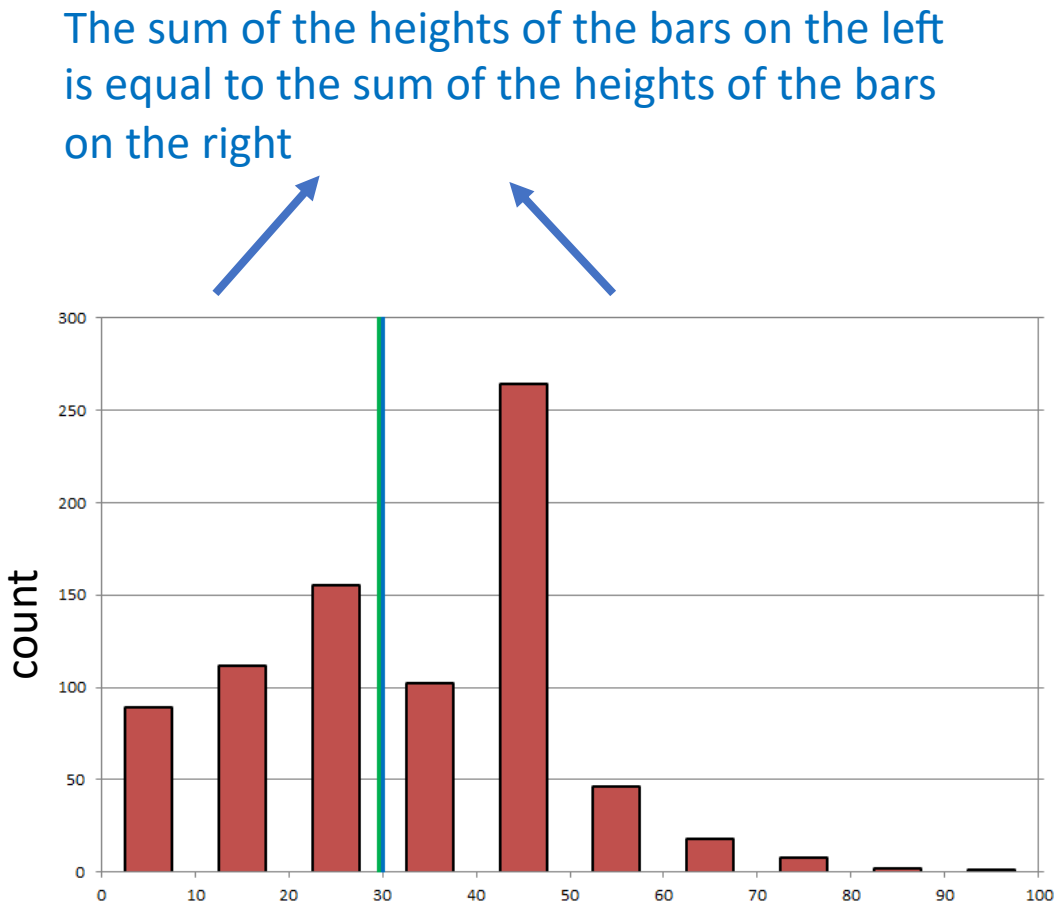


The median

The **median** is a value that splits the data in half

- i.e., half the values in the data are smaller than the median and half are larger






```
R: median(v)
   median(v, na.rm = TRUE)
```

Example of calculating the mean and median

Question: What is the mean and median for the gross domestic profit for the movies in the Bechdel data set?

- Ignore missing data

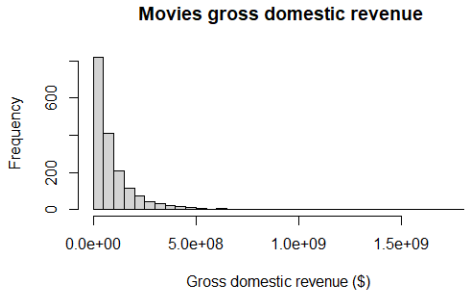

```
mean(domgross, na.rm = TRUE)
median(domgross, na.rm = TRUE)
```

$$\text{Mean} = \frac{\sum_i^n x_i}{n}$$

A: mean = 95,174,784

median = 55,993,641

Does it make sense the mean is larger than the median?



Review: outliers

Q: What is an **outlier?**

- A: An observed value that is notably distinct from the other values in a dataset

Q: Why are they problematic?

- A: can potentially have a large influence on the statistics you calculate

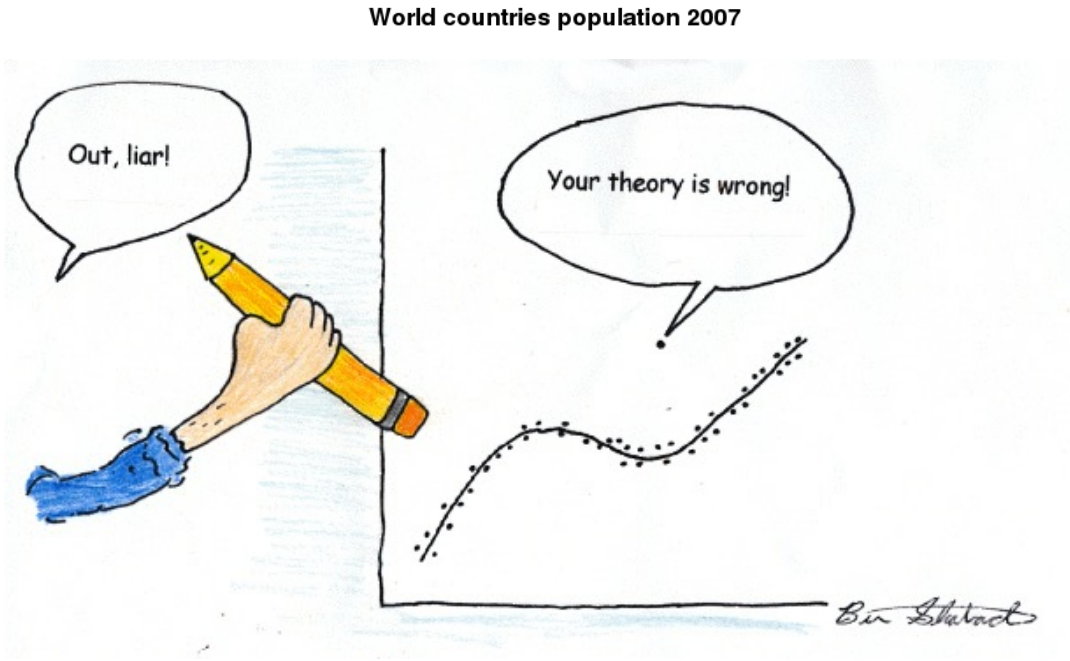
Q: What should you do if you have an outlier in your data?

A: See if you can understand what is causing it!

- If it's an error, delete the point
- If it's a real value, make sure it is not having a big effect on your conclusions, and/or use resistant statistics

Q: Is the mean and/or median resistant?

- A: The median is resistant while the mean is not



Bechdel outliers

Did it pass the Bechdel test?

What movie is this?

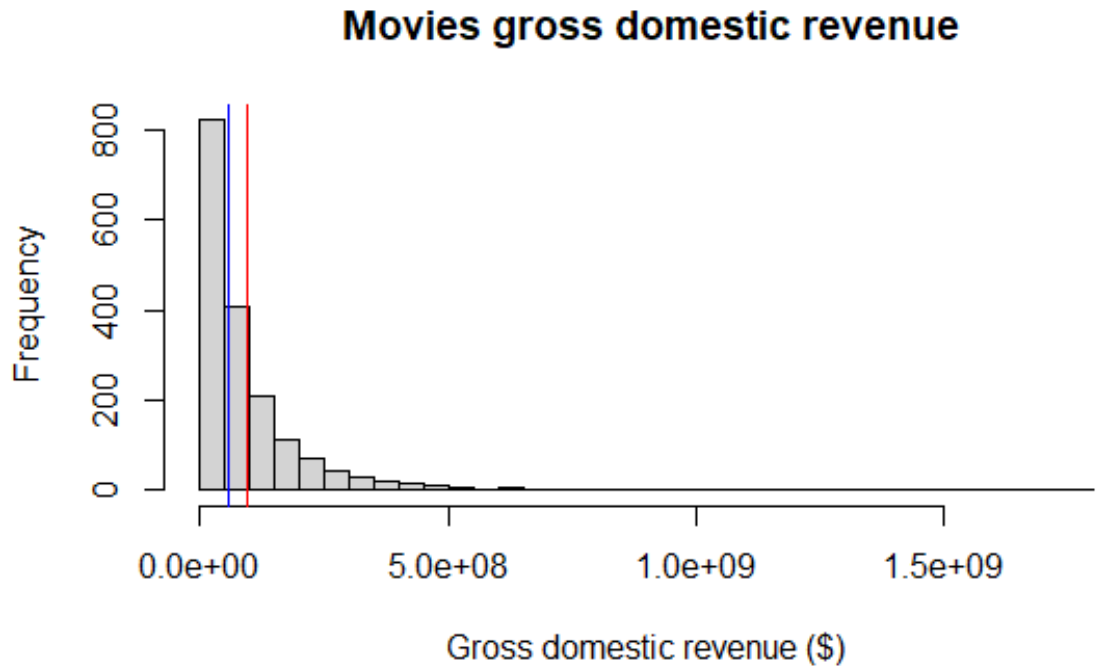
$1,771,682,790

Should we quickly try a few analyses in R?

Measures of spread

Characterizing the spread

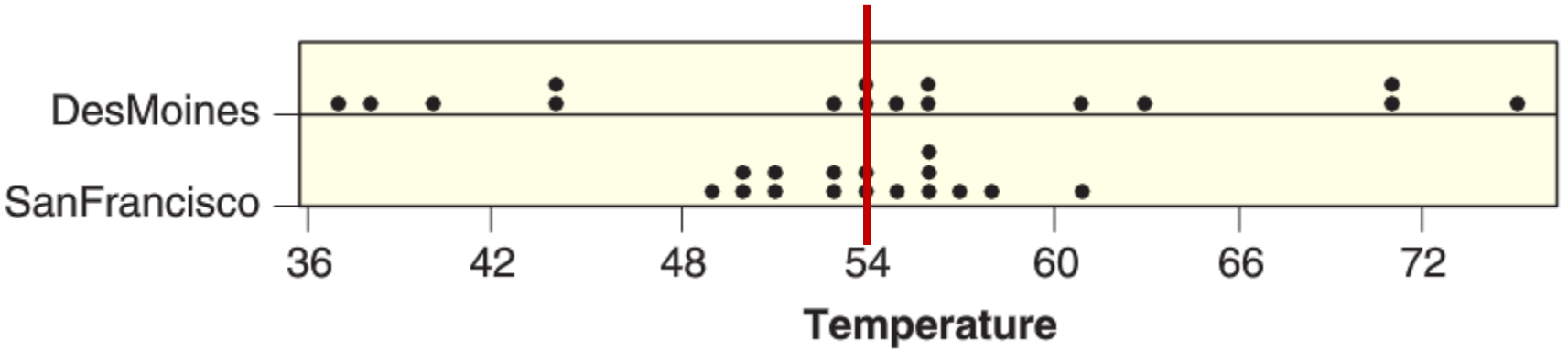
The mean and median are numbers that tell us about the center of a distribution



We can also use numbers to characterize how data is spread

Average monthly temperature: Des Moines vs. San Francisco

Data measured on April 14th from 1997 to 2010:

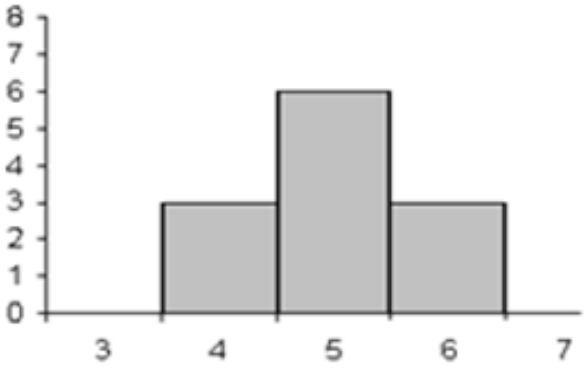


Mean temperature (°F): Des Moines = 54.49 San Fran = 54.01

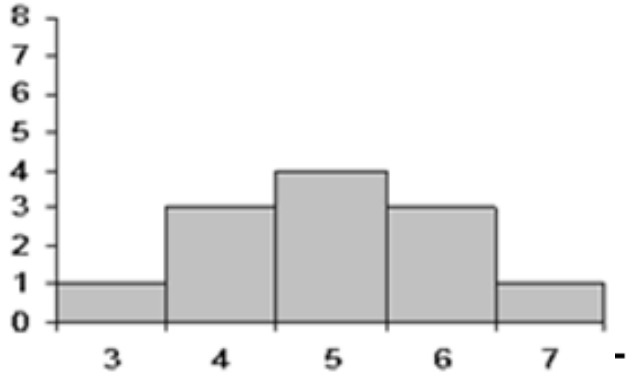
The standard deviation

The **standard deviation** (for a quantitative variable) is a measure of the spread of the data

Smaller standard deviation



Larger standard deviation



It gives a rough estimate for a typical distance a point is from the center

Notation

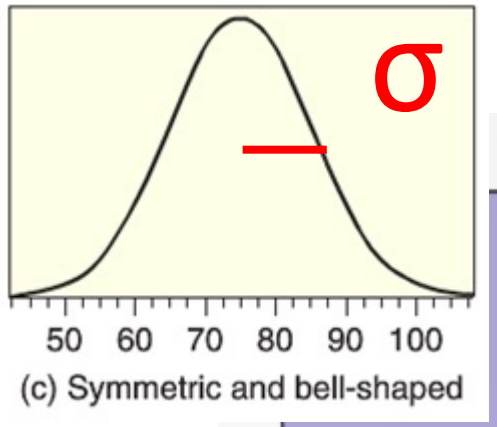
The standard deviation of the ***population*** is denoted **σ**

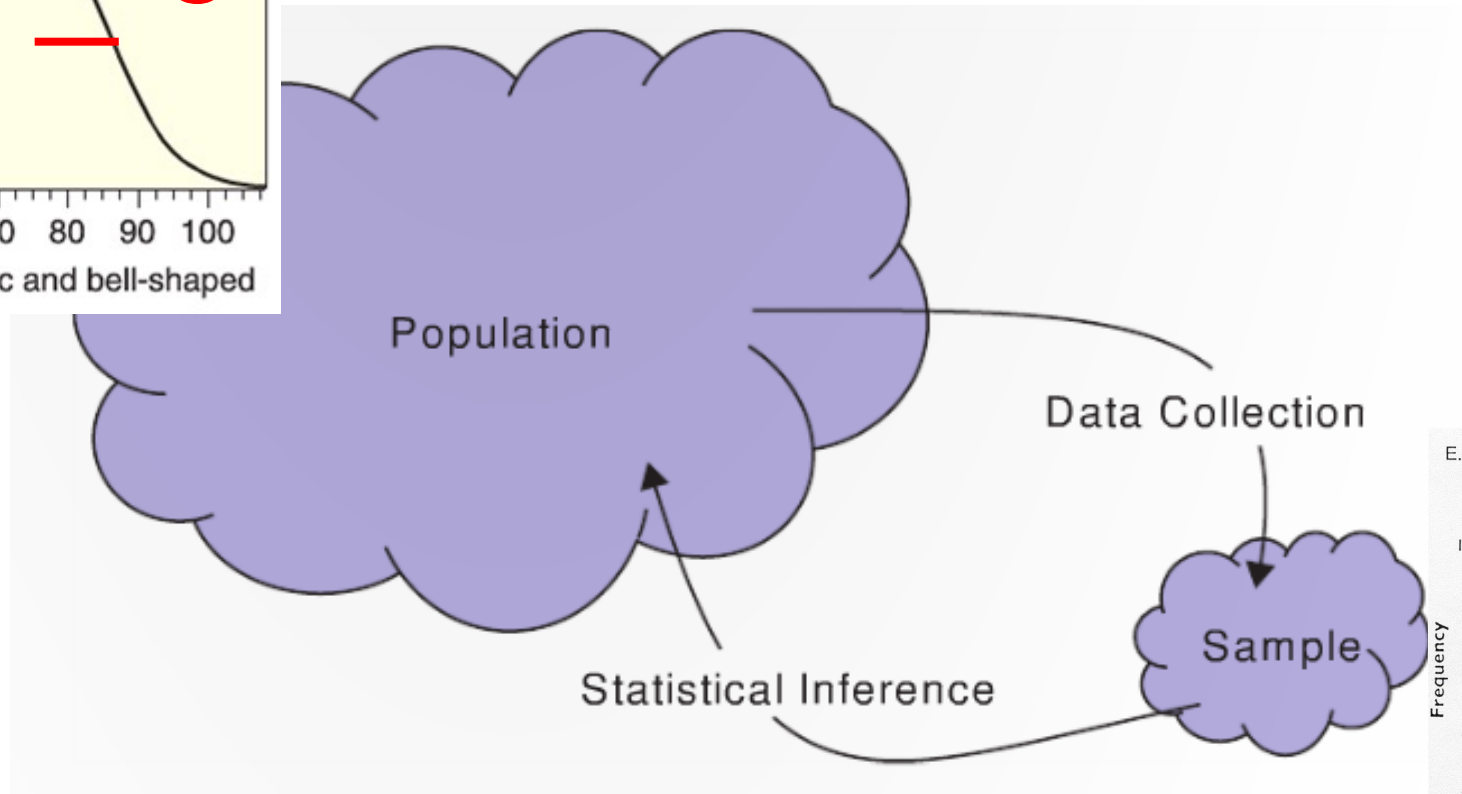
- It measure the spread of the data from the population mean μ

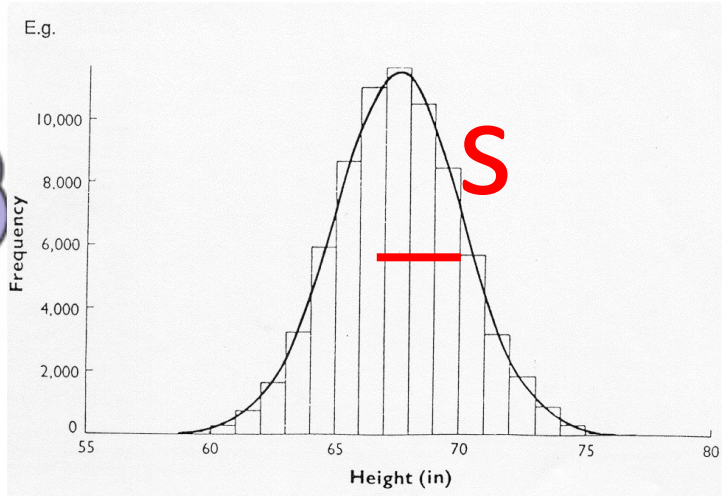
The standard deviation of a ***sample*** is denoted **s**

- It measure the spread of the data from the sample mean $\bar{x}$

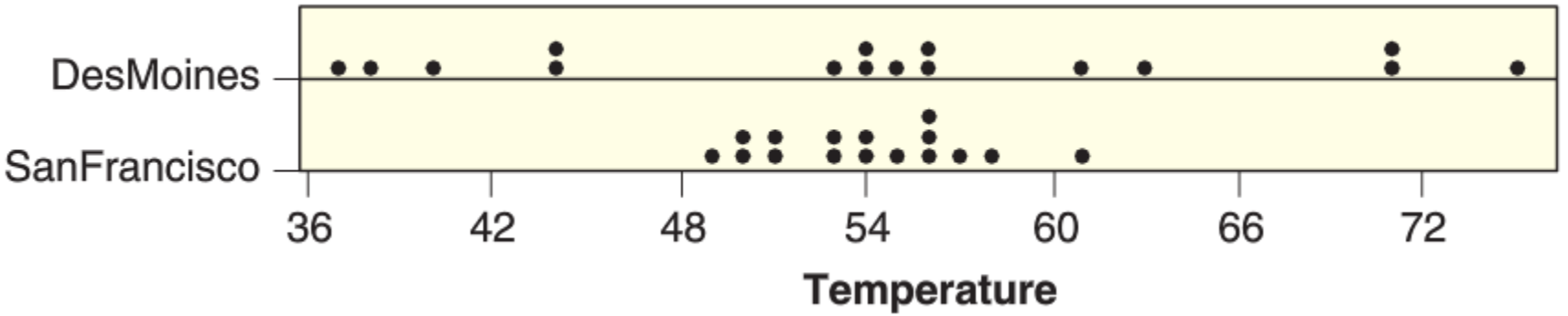
Population and sample standard deviation







Which has the larger standard deviation?



s_{DM} = 11.73 °F s_{SF} = 3.38 °F

The standard deviation

The standard deviation can be computed using the following formula:

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

Example: computing the standard deviation

Suppose we had a sample with n = 4 points:

$x_1 = 8, \quad x_2 = 2, \quad x_3 = 6, \quad x_4 = 4,$

We can compute the mean using the formula:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad = \tfrac{1}{4} \cdot (x_1 + x_2 + x_3 + x_4) \quad = \tfrac{1}{4} \cdot (8 + 2 + 6 + 4)$$

The standard deviation can be computed using the formula:

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

(remember order of operations!)

Hot dogs!

Every 4th of July, Nathan's Famous in NYC holds a hot dog eating contest where contestants try to eat as many hot dogs as they can in 10 minutes

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

Worksheet: Calculate the mean and standard deviation for the number of hot dogs eaten by the winners. Upload the filled out worksheet to Canvas.

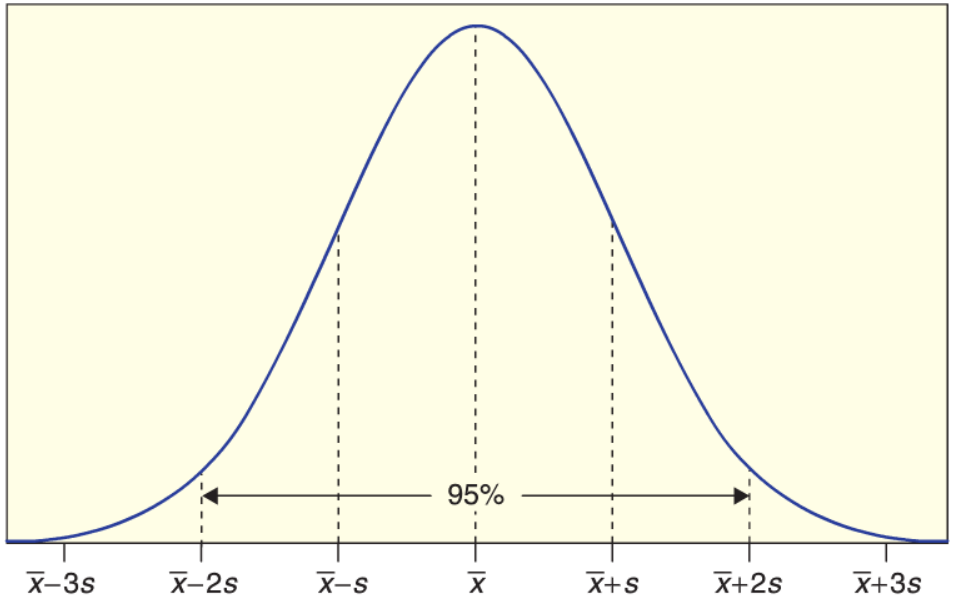
Normal distributions and z-scores

The 95% rule for *normal distributions*

A **normal distribution** is a common distribution that is symmetric and bell shaped

If a distribution of data is approximately normally distributed, about 95% of the data should fall within two standard deviations of the mean

i.e., 95% of the data is in the interval: $\bar{x}$ -2s to $\bar{x}$ +2s



The 95% rule for *normal distributions*

Example: The Dow Jones Industrial Average from 1980 to present has daily percent changes that are approximately normally distributed with a mean of 0.04% and a standard deviation of 1.12%

- SDS100::download_data("DowPrices.csv")

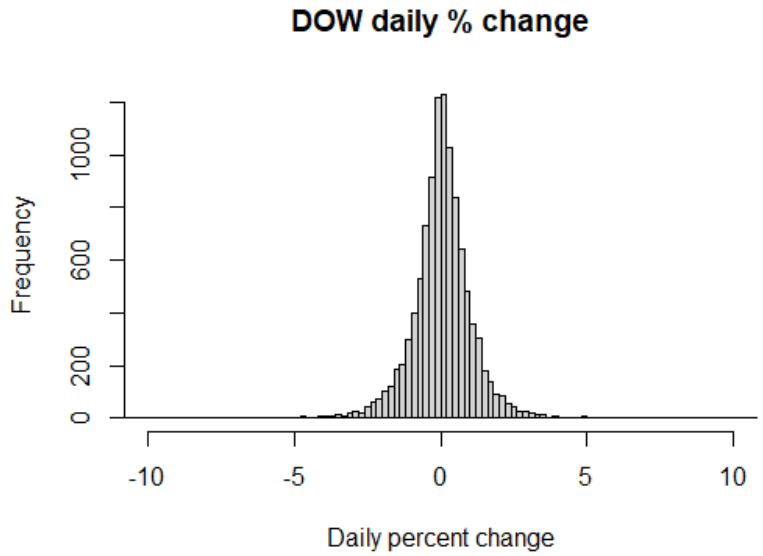
Question: what is the range of values that the middle 95% of Dow percent daily changes fall in?

Answer: (0.04 – 2.24) to (0.4 + 2.24)

95% of daily DOW % changes are in the range:

-2.2% to 2.28%

actual 2.5 to 97.5 quantiles: -2.14 to 2.12

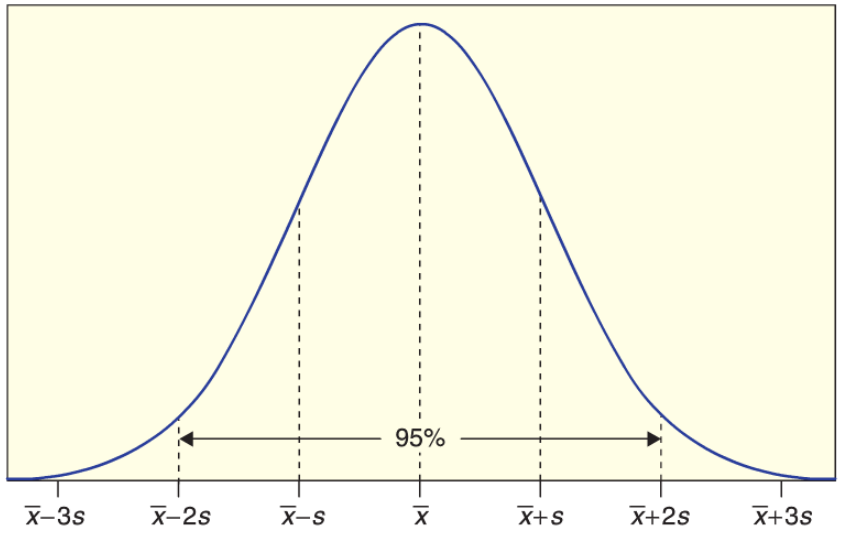


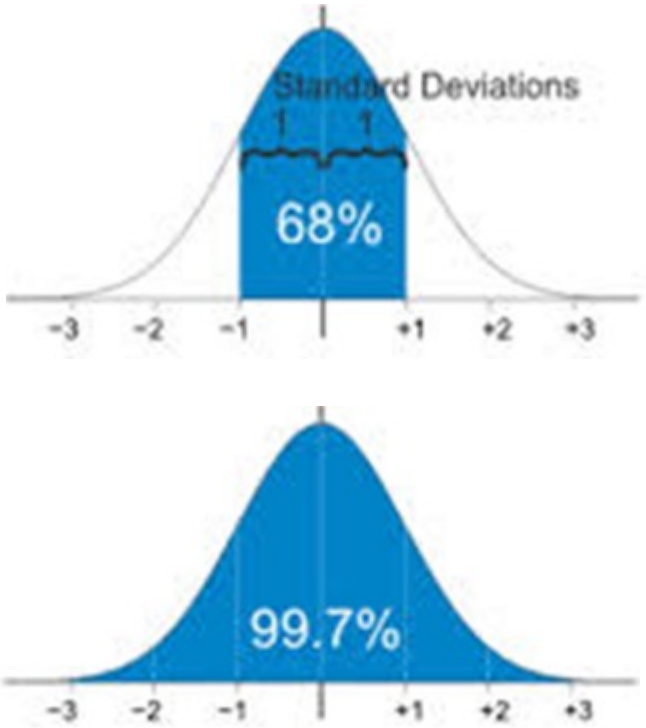
(past performance is not indicative of future results)

The 68%, 95% and 99.7% rules for *normal distributions*

Other properties of normal distributions are:

- 68% of the data falls within **one** standard deviations of the mean
- 95% of the data falls within **two** standard deviations of the mean
- 99.7% of the data falls within **three** standard deviations of the mean





z-scores

The z-scores tells how many standard deviations a value is from the mean

- i.e., how far away a point x_i is from $\bar{x}$ in a way that is independent of the units of measurement

$$\text{z-score}(x_i) = \frac{x_i - \bar{x}}{s}$$

Which accomplishment is most impressive?

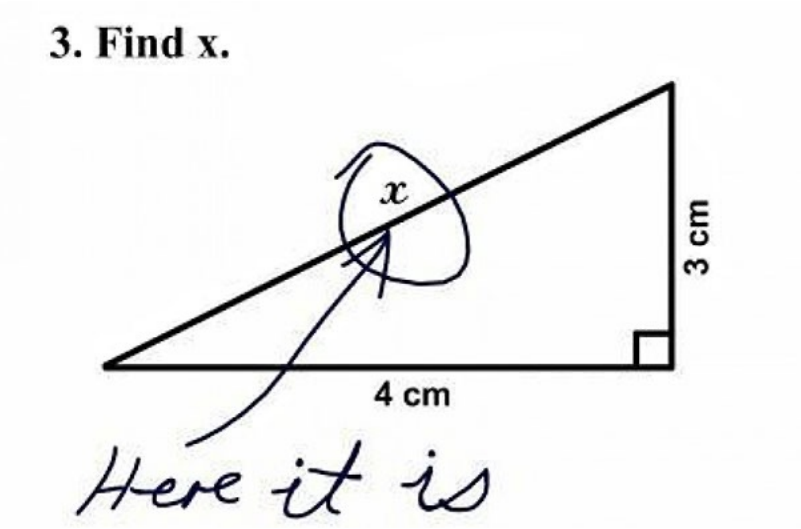
- SAT-Math scores are normally distributed with a mean of 500 and standard deviation of 100.
- ACT-Math scores are normally distributed with a mean of 18 and standard deviation of 6.

- A student has taken both tests. They scored 600 on the SAT-Math and 22 on the ACT-Math.
- Which score is more impressive?

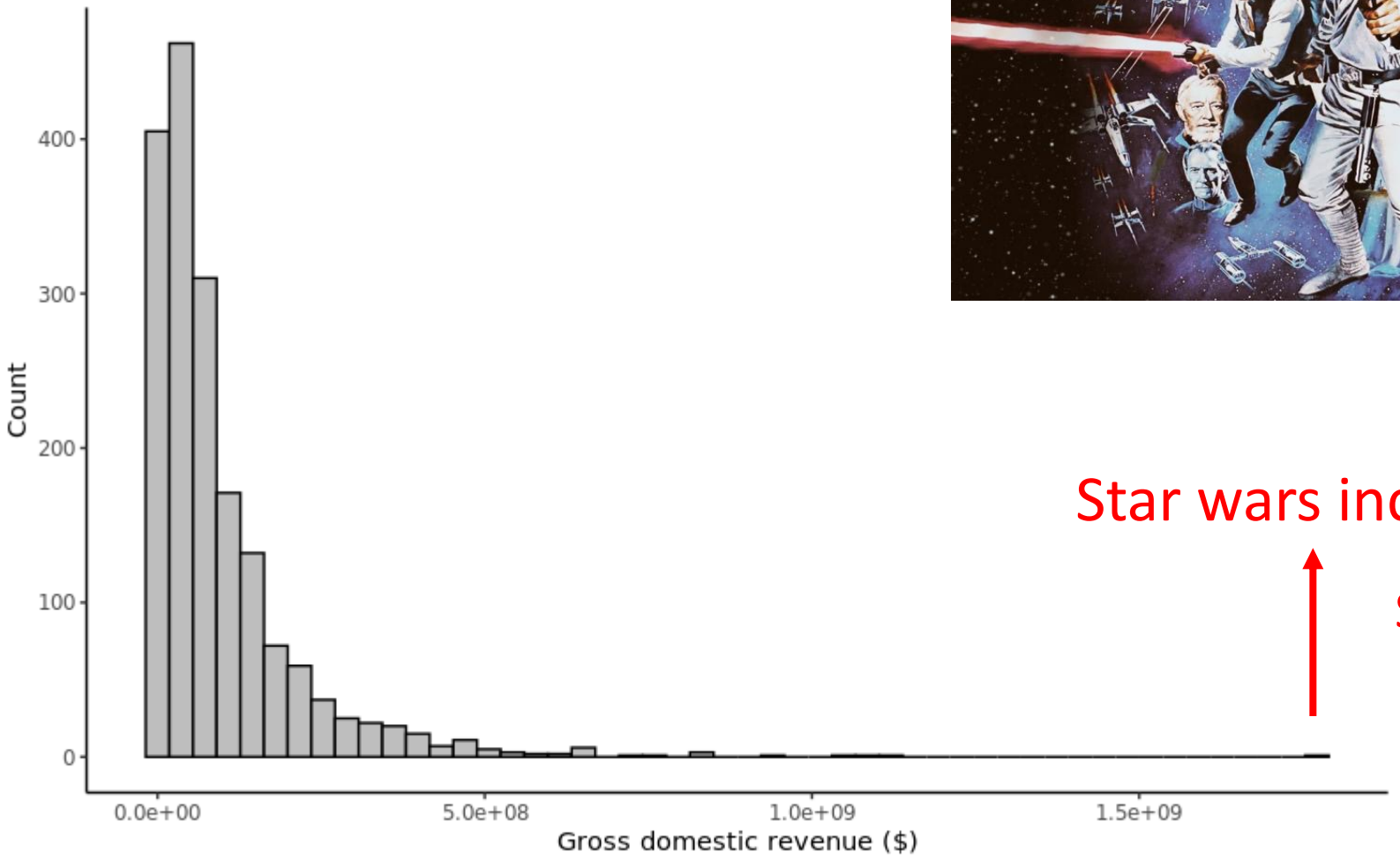
$$\text{z-score}(x_i) = \frac{x_i - \bar{x}}{s}$$

z-score SAT = (600 – 500)/100 = 1

z-score ACT = (22 – 18)/6 = 2/3



What is star wars' z-score?



Star wars income

$1,771,682,790

Should we try it in R?

Percentiles

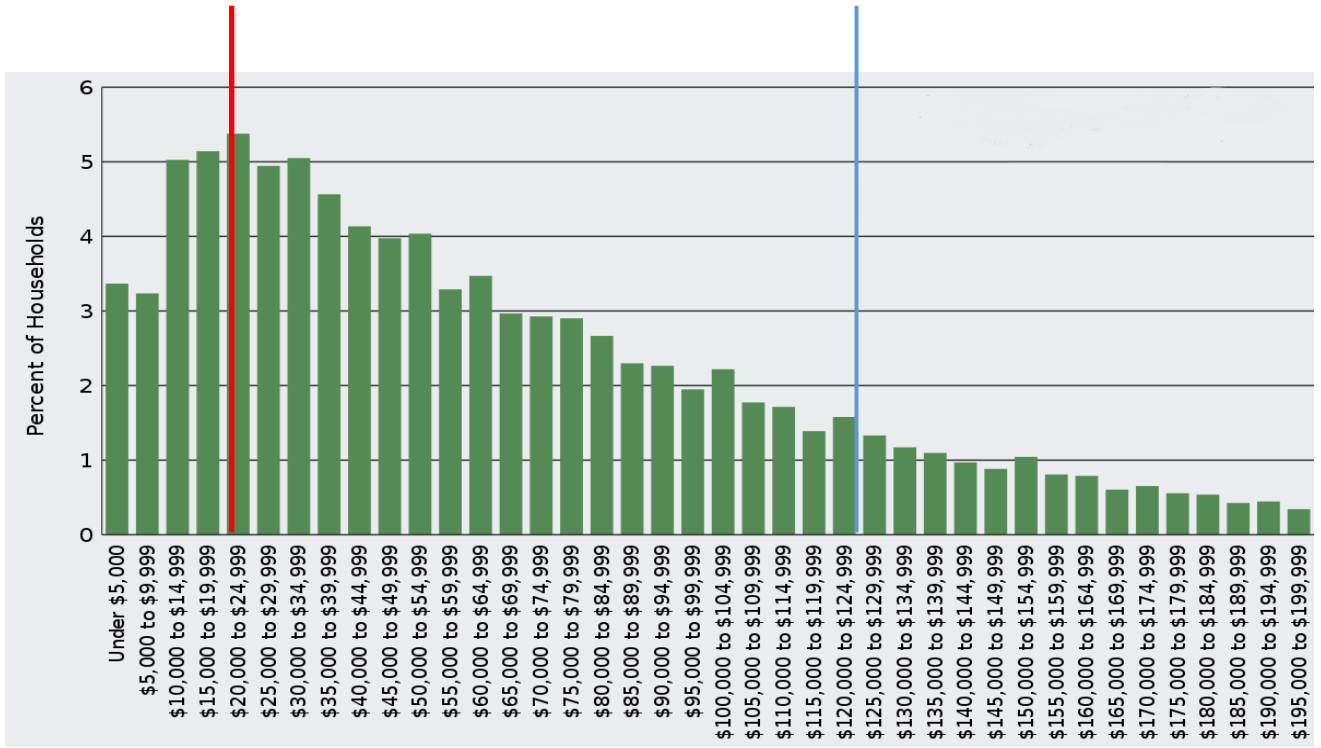
Percentiles

The **Pth percentile** is the value of a quantitative variable which is greater than P percent of the data

For the US income distribution what are the 20th and 80th percentiles?

20th percentile = $21,430

80th percentile = $112, 254



R: `quantile(v, .95)`

Bechdel quantiles

What are the 25th and 75th quantiles for gross domestic revenue for the Bechdel data?

```
> quantile(domgross,  c(.25, .75),
        na.rm = TRUE)
```

25%	75%
$20,546,594	$121,678,352

How do we interpret these numbers?

- i.e., how would we describe these numbers in English?

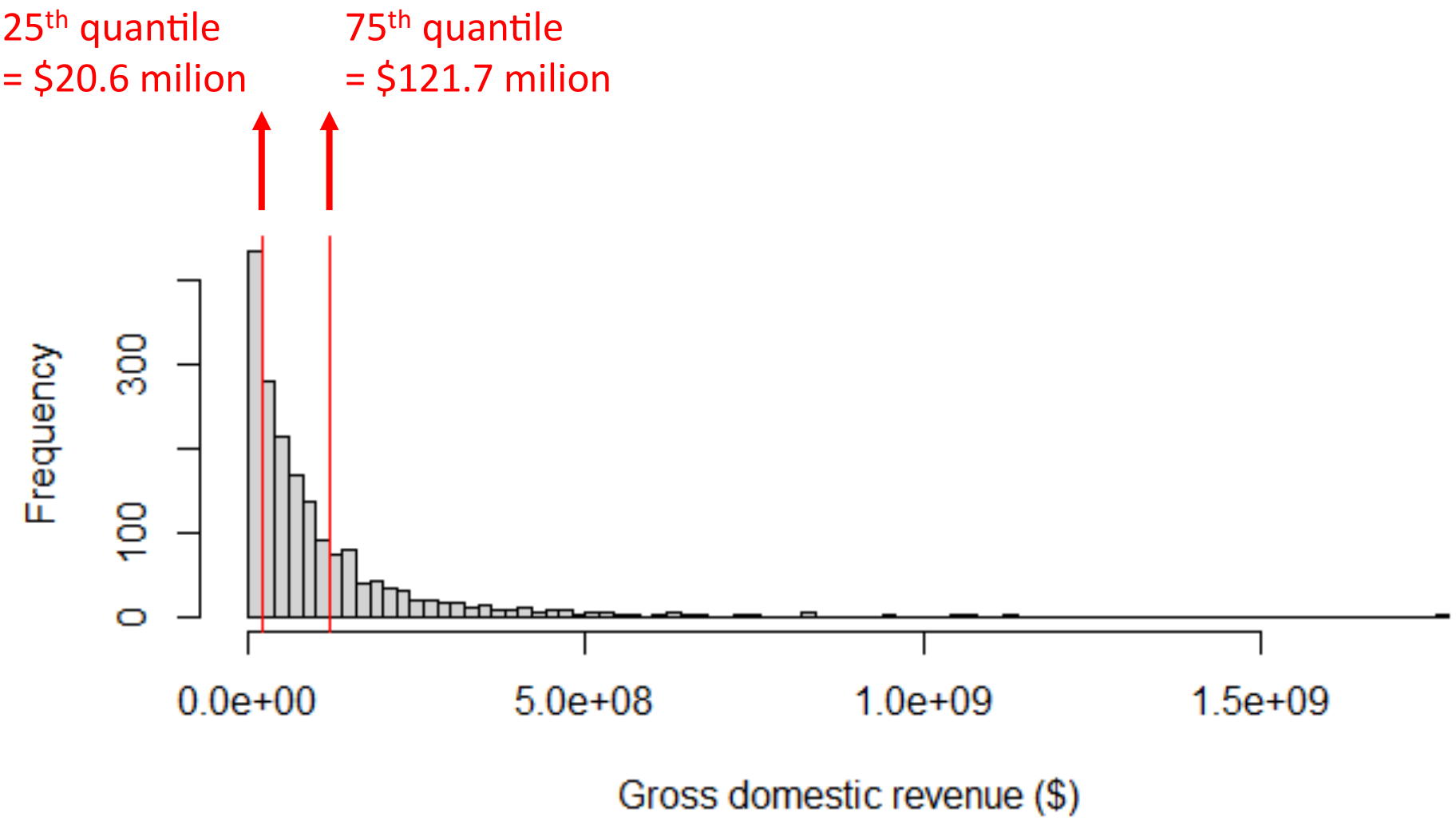
Answer:

25% of movies made less than $20.6 million

75% of movies made less than $121.7 million

Or, the middle 50% of movies made between $20.6 to $121.7 million.

Age of marijuana arrests in Toronto



Five Number Summary

Five Number Summary = (minimum, Q_1, median, Q_3, maximum)

Q_1 = 25^{th} percentile (also called 1^{st} quartile)

Q_3 = 75^{th} percentile (also called 3^{rd} quartile)

Roughly divides the data into fourths

Range and Interquartile Range

Range = maximum – minimum

Interquartile range (IQR) = $Q_3 - Q_1$

Hot dog example – try this at home!

Try this at home: for the hot dog data calculate "by hand"

- The 5 number summary
- The range
- Interquartile range

Cheat sheet:

Five Number Summary = (minimum, Q_1, median, Q_3, maximum)

Range = maximum – minimum

Interquartile range (IQR) = $Q_3 - Q_1$

Q_1 = 25th percentile, Q_3 = 75th percentile

Answer in R: `fivenum(v)`

Should we try it in R on the Bechdel or DOW data?

Year	Hot Dogs
2013	69
2012	68
2011	62
2010	54
2009	68
2008	59
2007	66
2006	54
2005	49
2004	54

Detecting of outliers

As a rule of thumb, we call a data value an **outlier** if it is:

Smaller than: Q_1 - 1.5 * IQR

Larger than: Q_3 + 1.5 * IQR

What is the range that a value would be called an outlier in the hot dog data?

Are there any outliers in the hot dog data?

Homework 1

Homework 1 is due at 11pm on Sunday January 28th

Use Ed Discussions for any questions that come up, and/or attend office hours

Upload pdfs with your answers to Gradescope

1. Hand in R Markdown pdf under the assignment called Homework 1
2. Make sure to Mark your pages on Gradescope!