

Measures of spread



Overview

Review of shapes distributions and central tendency

The standard deviation

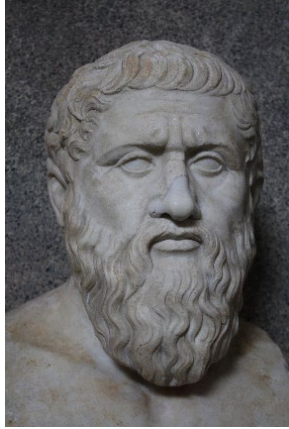
Z-scores

Percentiles

Review and continuation of...

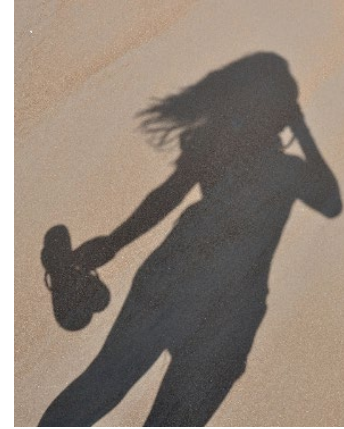
Quantitative variables

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



P-Truth

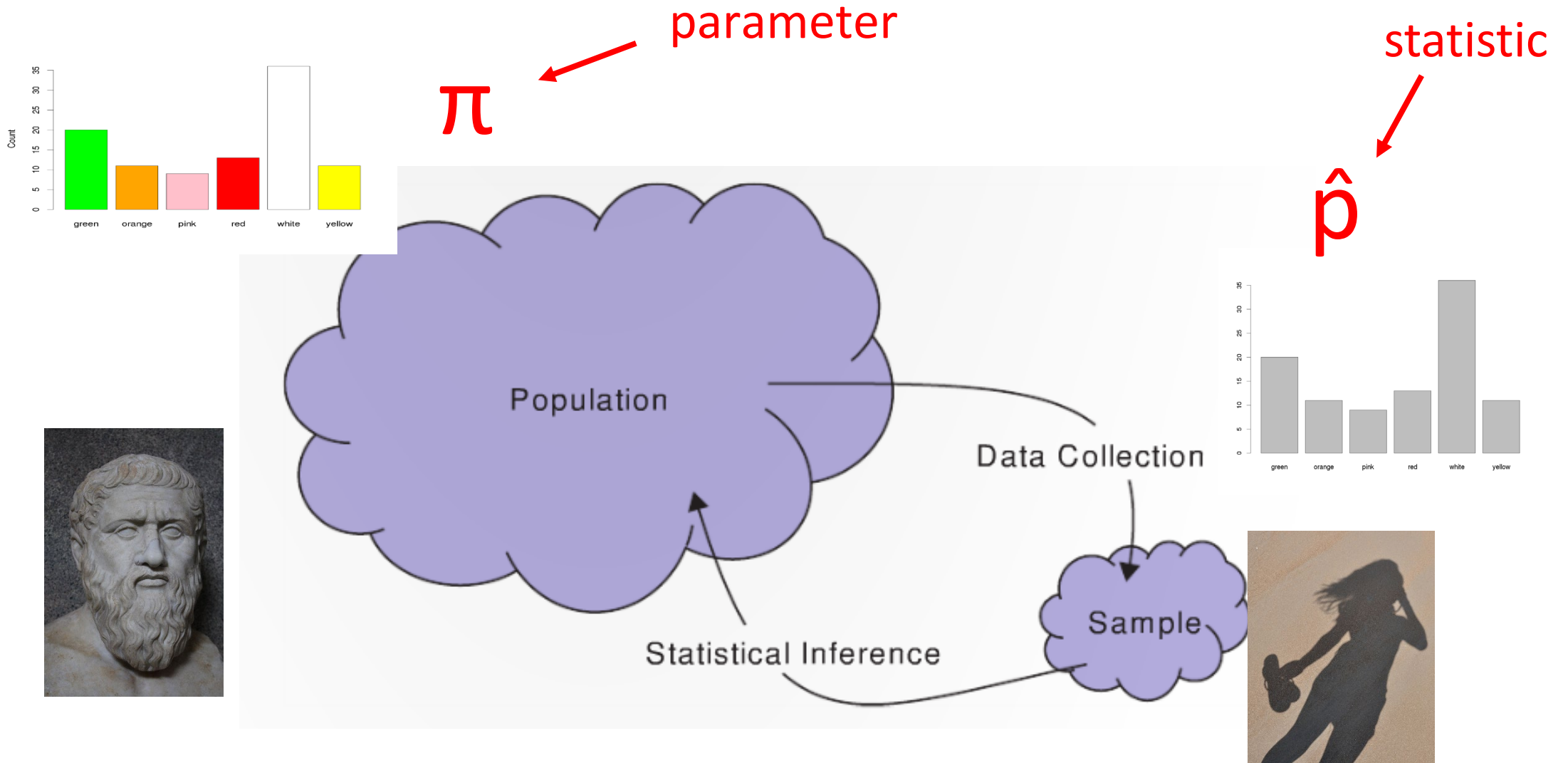
- population or process
- parameter
- Plato (Greek symbols)



S-shadows

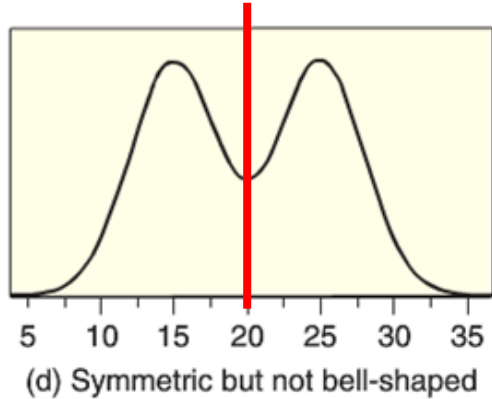
- sample
- statistic
- shadow (Latin symbols)

Review: Categorical data and proportions



Review: Quantitative data and the mean

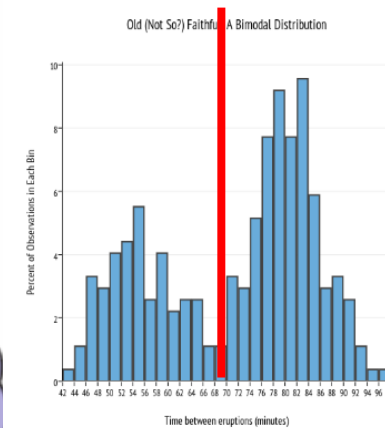
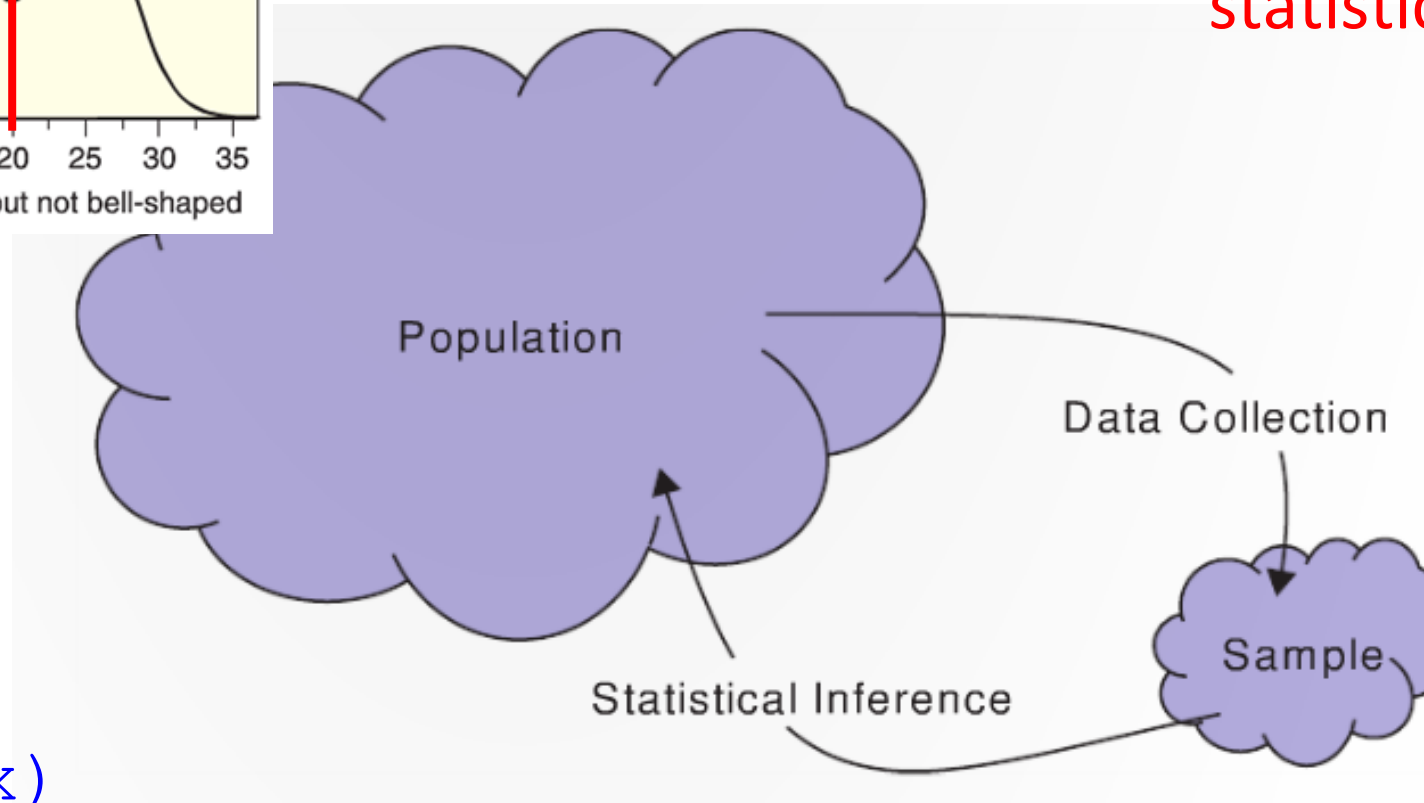
μ ← parameter



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

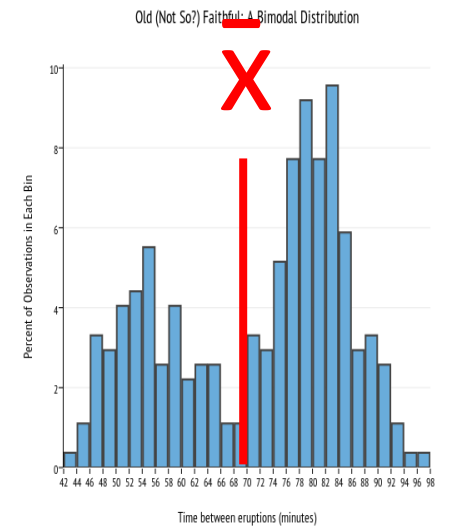
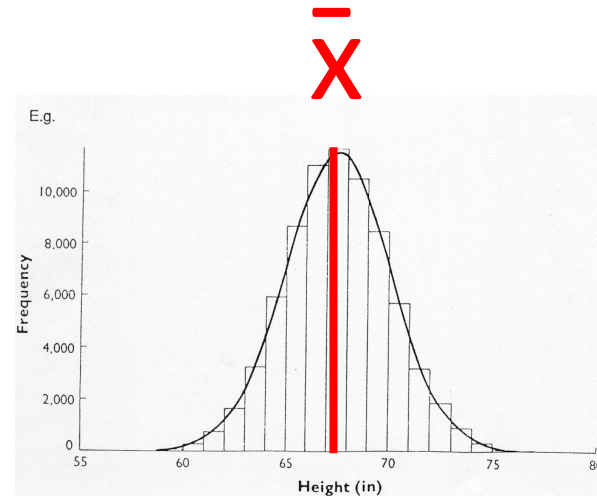
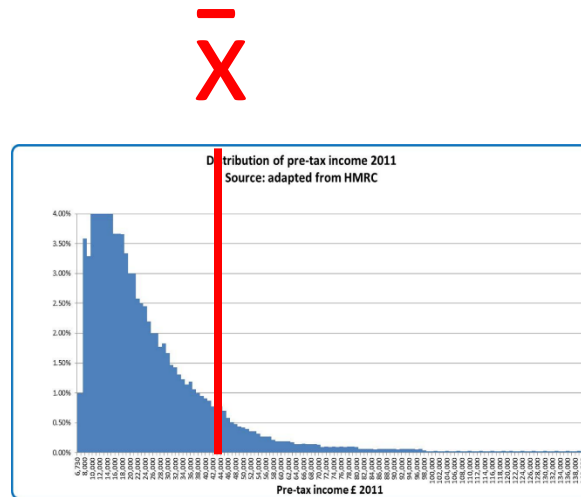
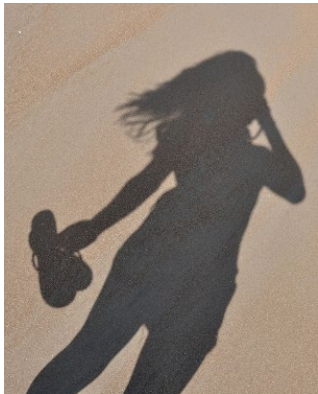
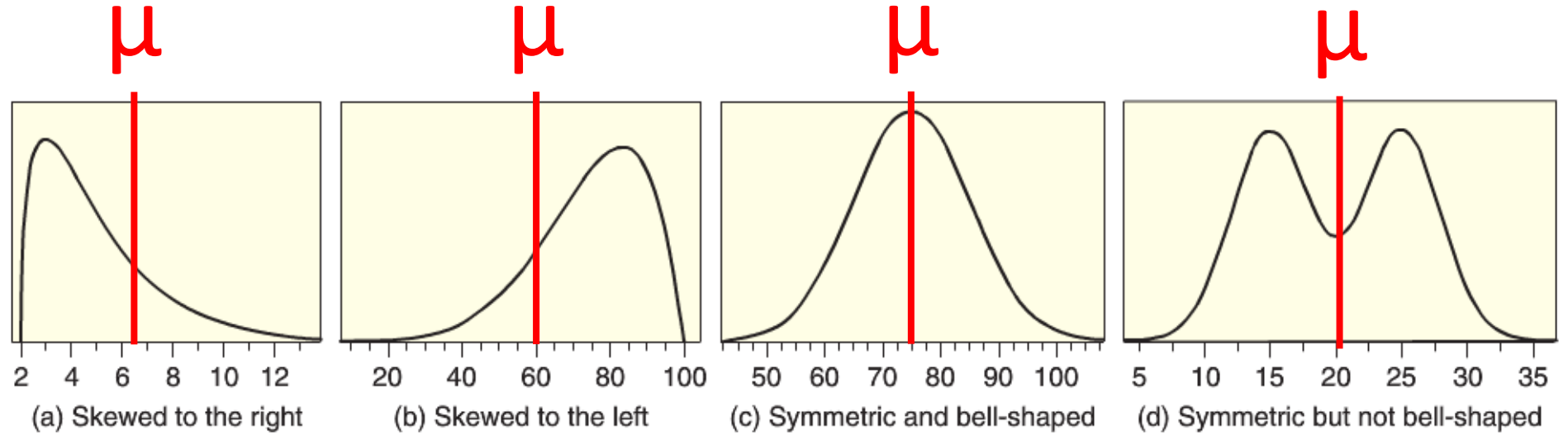
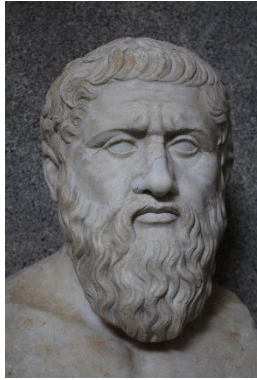
statistic

\bar{x}



R: `mean(x)`

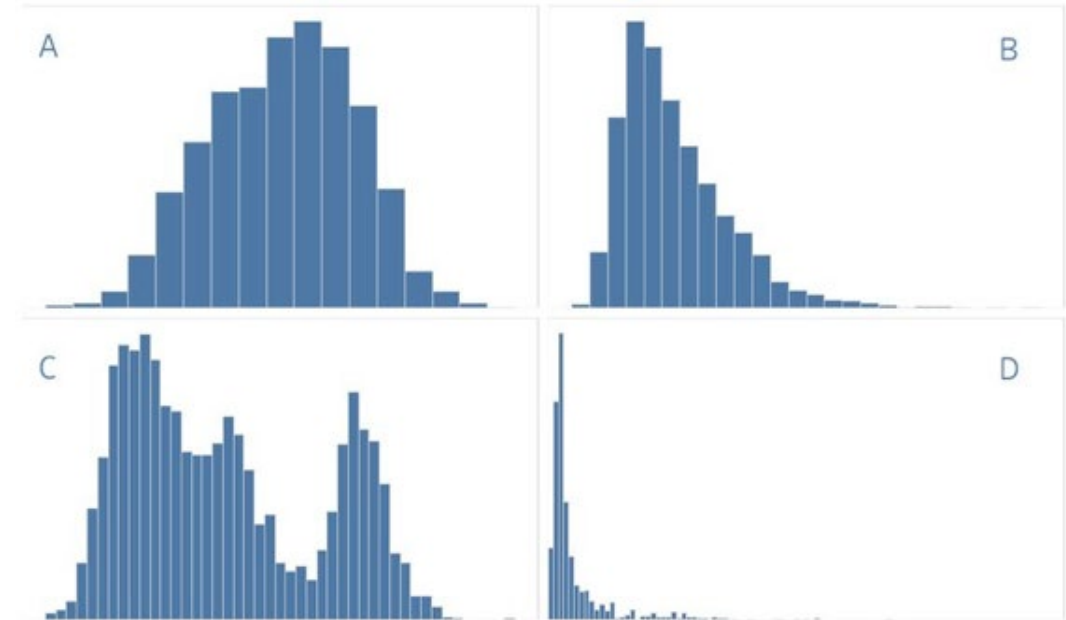
Means for differently shaped distributions





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



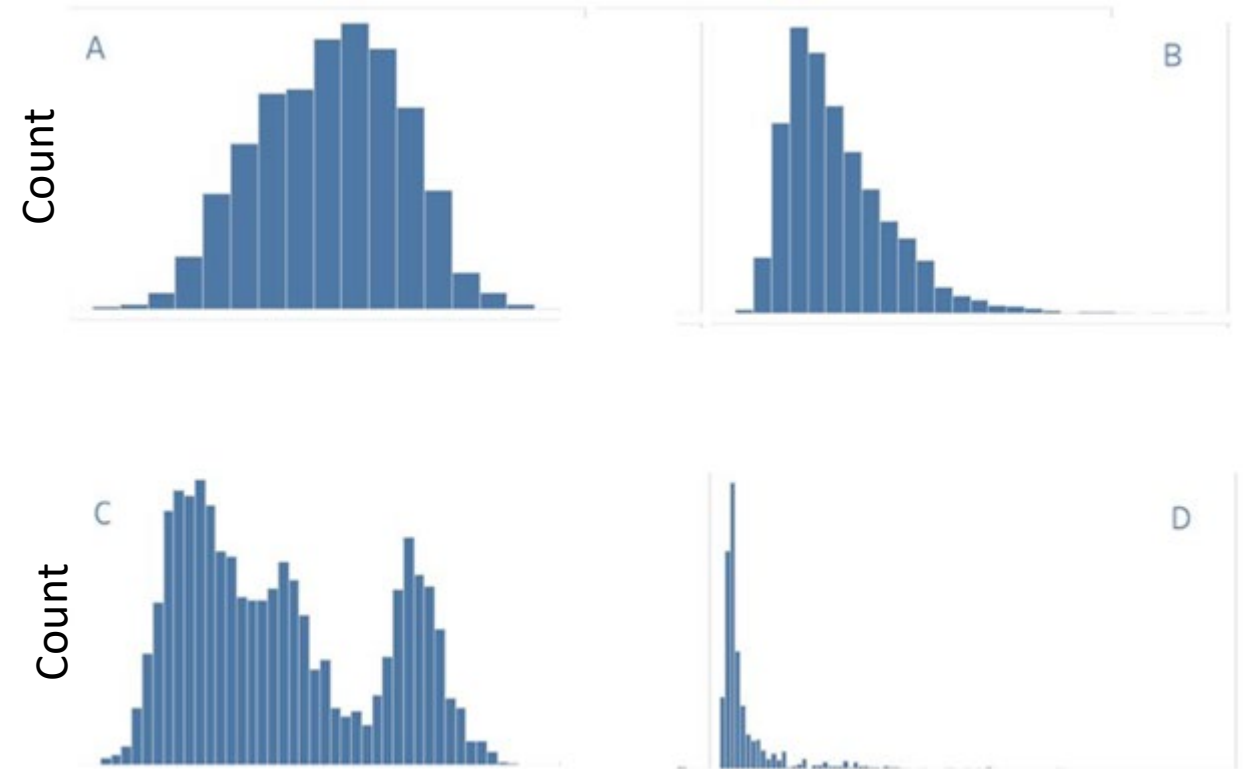
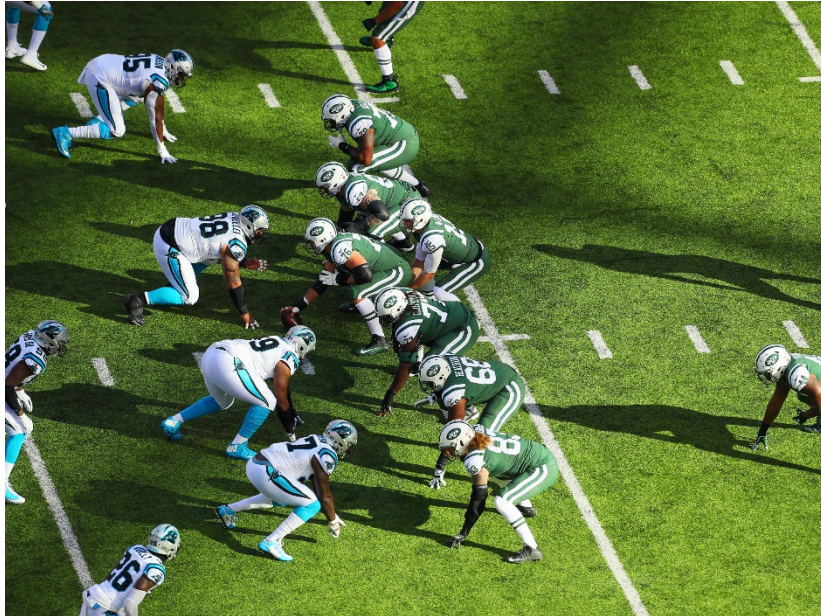
Question: Can you tell which histogram goes with which trait?

Task is to add the labels: **Age, Height, Weight, and Salary**

- Hint: There are a wide range of positions in football that have very different roles
 - E.g., placekickers only play for small factions of the game, while quarterbacks are essentially to a team's success

First: what is the label for the y-axis?

- A: Frequency or count

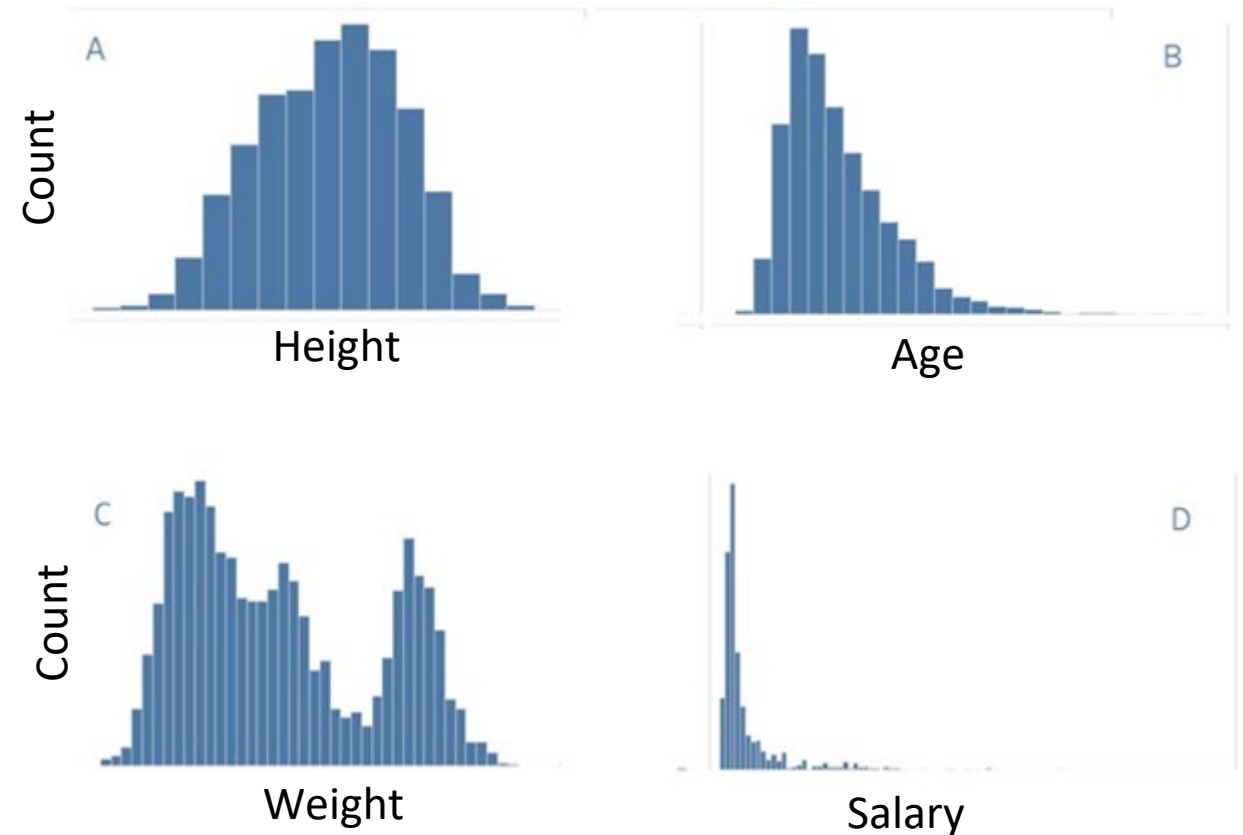


Task is to add the labels: **Age, Height, Weight, and Salary**

- Hint: There are a wide range of positions in football that have very different roles
 - E.g., placekickers only play for small factions of the game, while quarterbacks are essentially to a team's success

First: what is the label for the y-axis?

- A: Frequency or count





If you don't want exes, label you axes!

Back to the Gapminder data...

get a data frame with information about the countries in the world

> download_data("gapminder_2007.Rda") # SDS100 function - only need to run this once

> load("gapminder_2007.Rda")

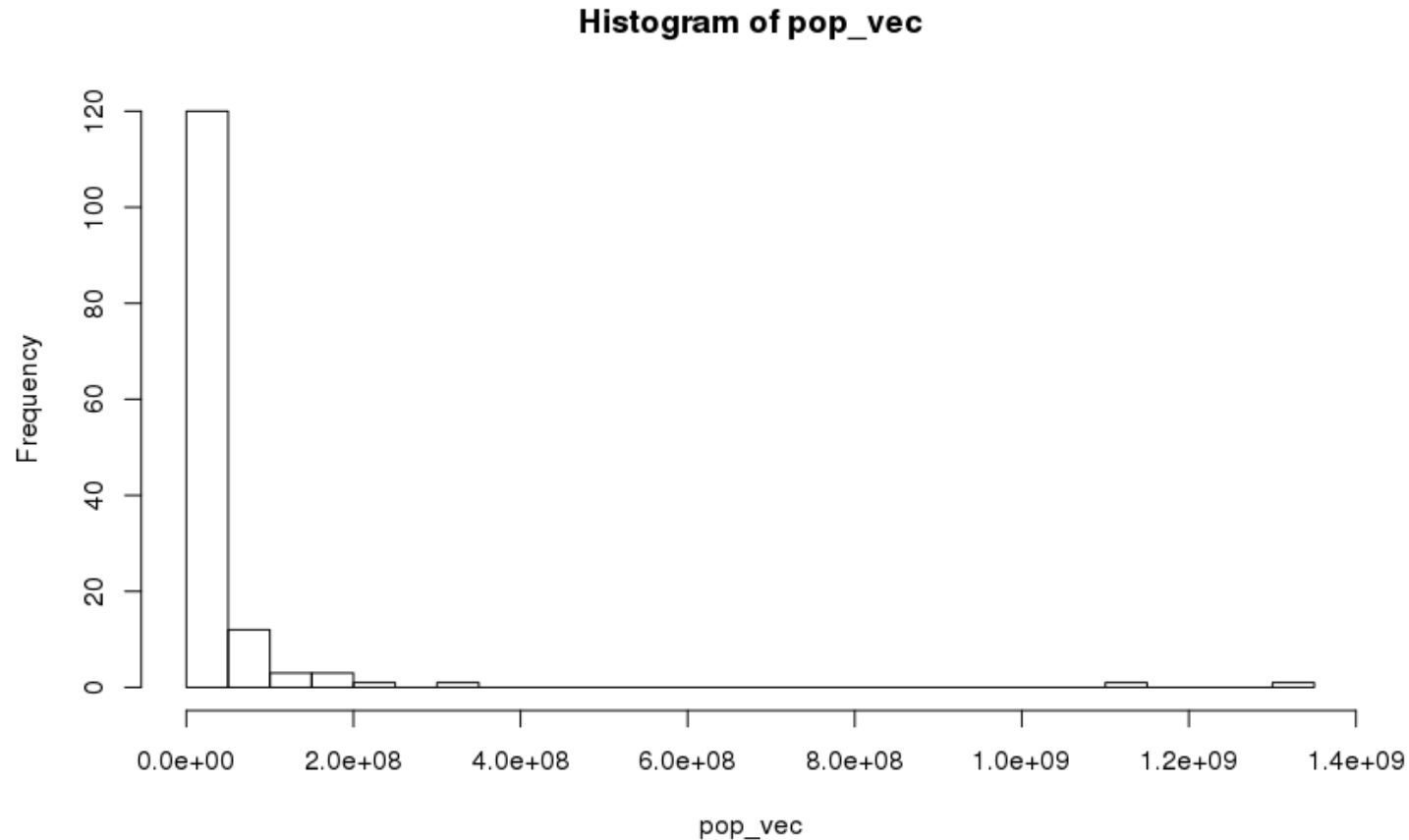
	country	continent	year	lifeExp	pop	gdpPercap
1	Afghanistan	Asia	2007	43.828	31889923	974.5803
2	Albania	Europe	2007	76.423	3600523	5937.0295
3	Algeria	Africa	2007	72.301	33333216	6223.3675
4	Angola	Africa	2007	42.731	12420476	4797.2313
5	Argentina	Americas	2007	75.320	40301927	12779.3796

Can you plot a histogram of the population of each country with 20 bins?

> pop_vec <- gapminder_2007\$pop # first create a vector with the population of each country

> hist(pop_vec, breaks = 20) # then create the histogram

What is missing from this histogram?



Axes labels could be more informative!

Labeling axes

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

- > ? hist
- Answer: xlab and ylab!

```
> hist(pop_vec, breaks = 20,  
       ylab = "Frequency",  
       xlab = "Population",  
       main = "World countries population in 2007")
```

The median

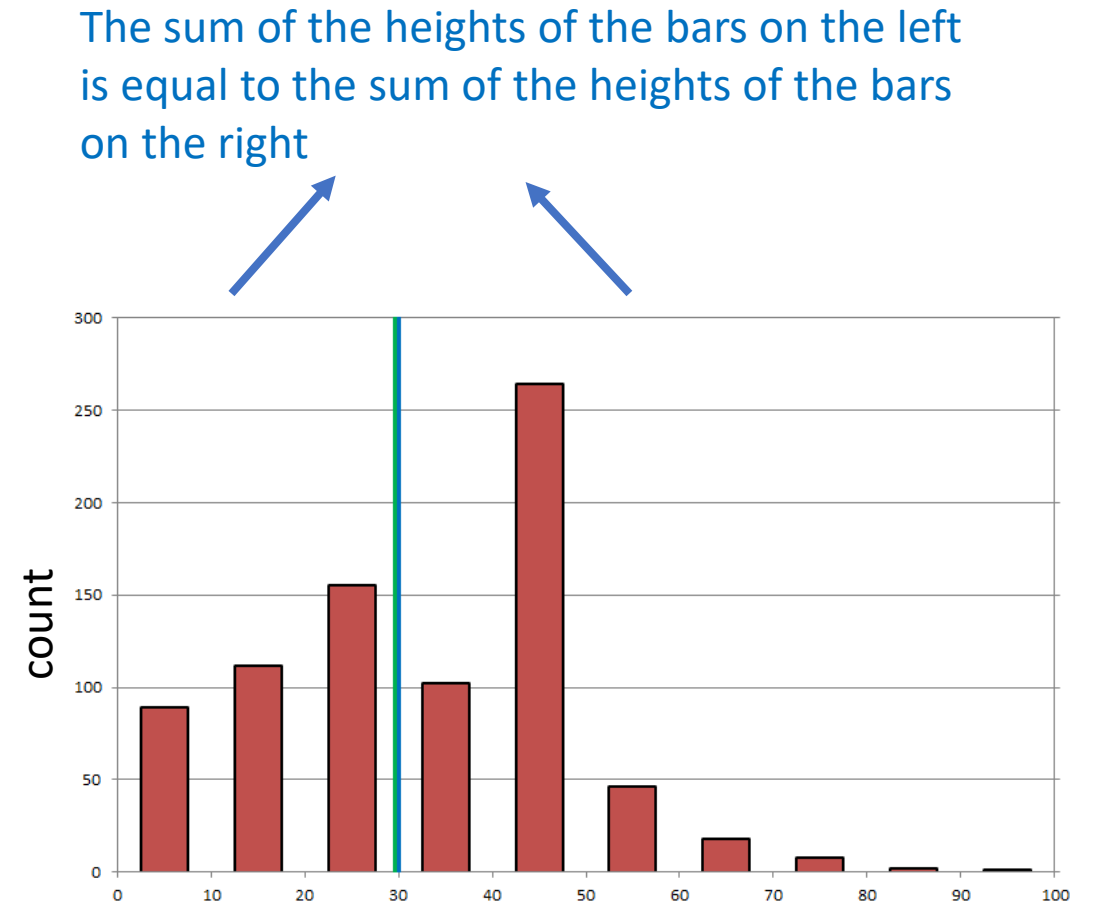
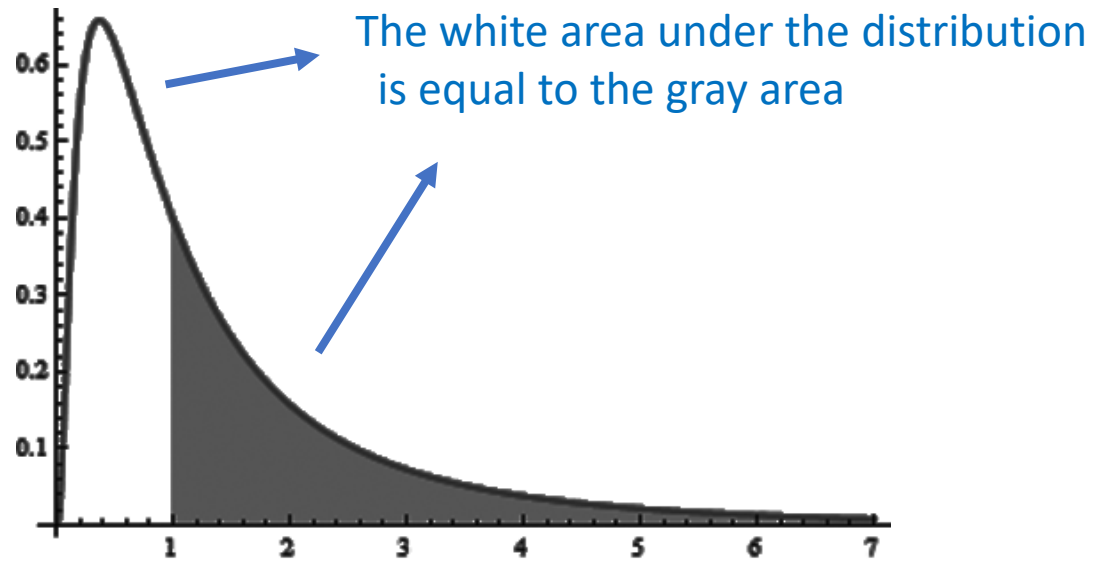
The **median** is a value that splits the data in half

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

To calculate the median for a data sample of size n , sort the data and then:

- If n is odd: The middle value of the sorted data
- If n is even: The average of the middle two values of the sorted data

The median



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

Example of calculating the mean and median

When an individual visits a webpage a 'ping' is generated

Below is a random sample of ping counts from 7 people who pinged a website at least once:

12, 45, 6, 4, 158, 10, 59

Question: What is the mean and median ping count in this sample?

A: mean = 42
median = 12

$$\text{Mean} = \frac{\sum_{i=1}^n x_i}{n}$$



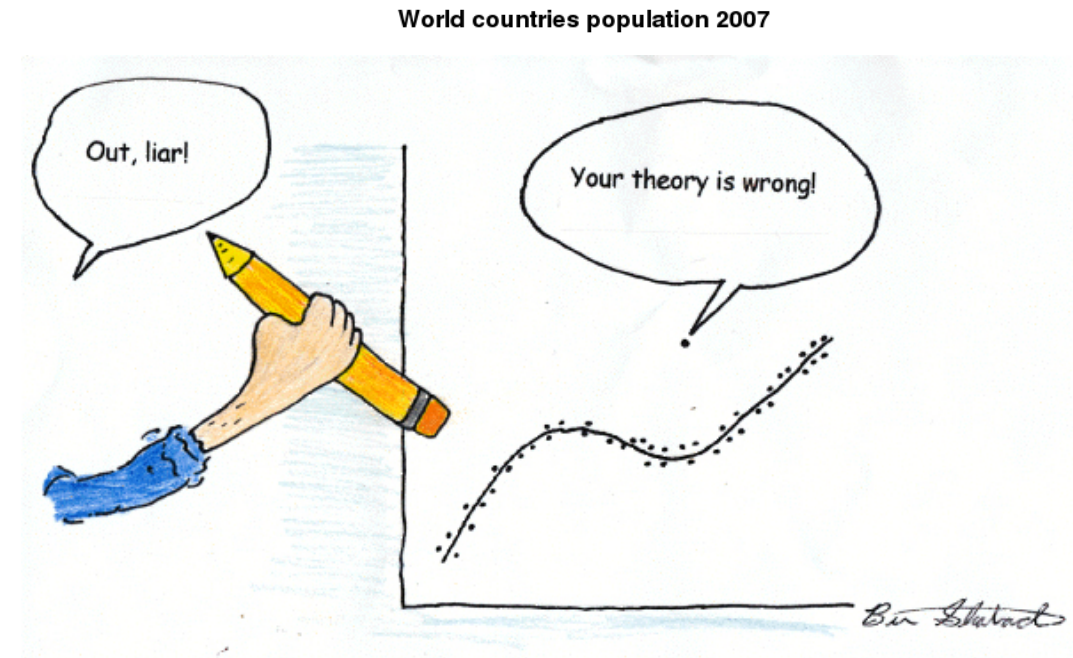
Review: outliers

Q: What is an **outlier**?

Q: Why are they problematic?

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

Q: Is the mean and/or median resistant?



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?

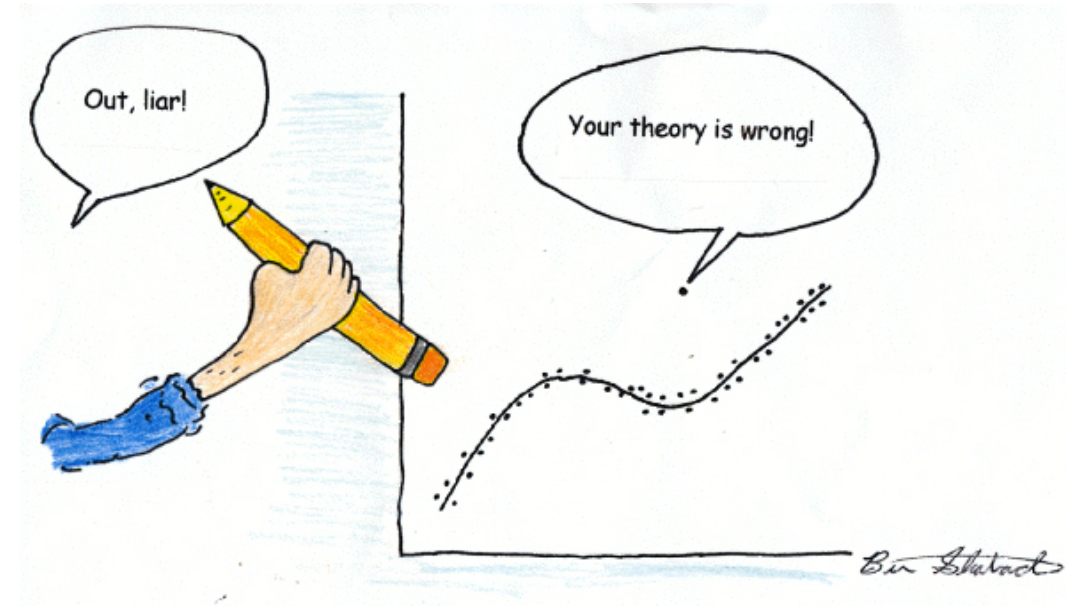
A: See if you can understand what is causing it!

- 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

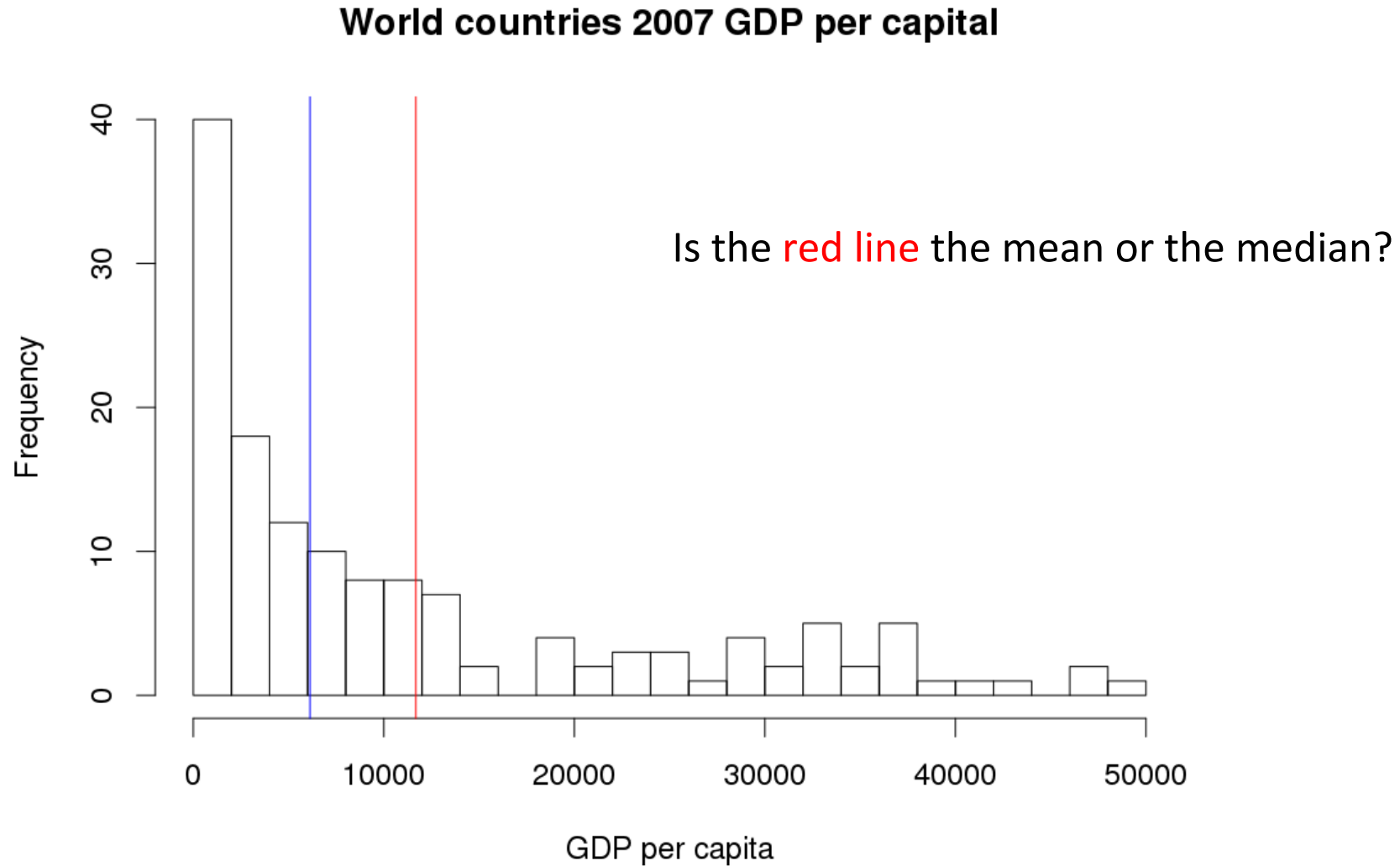
Q: Is the mean and/or median resistant?

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

World countries population 2007



Measure of central tendency: mean and median

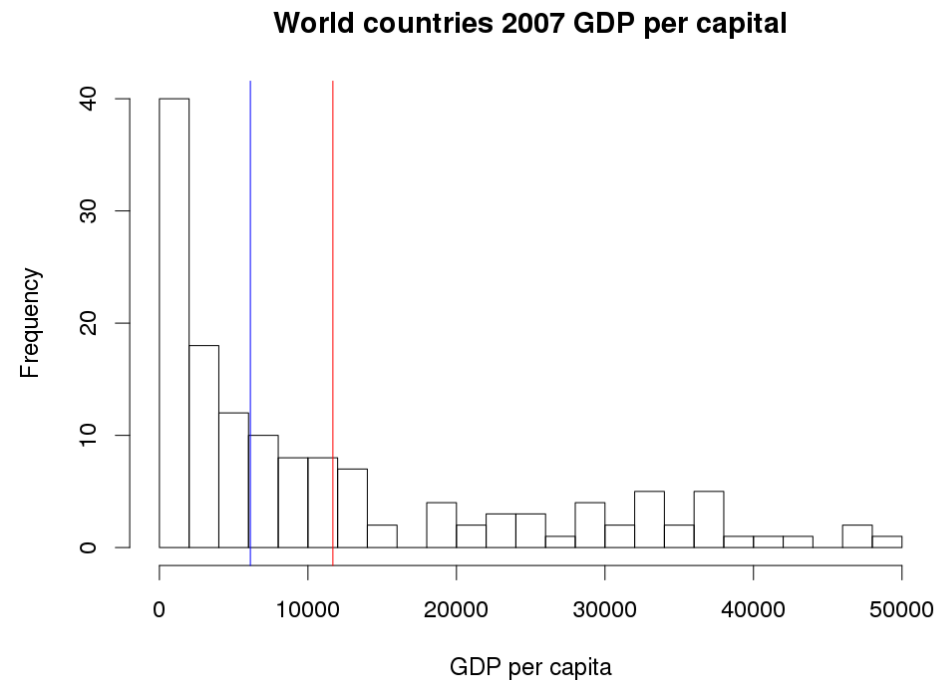


Measures of spread



Characterizing the spread

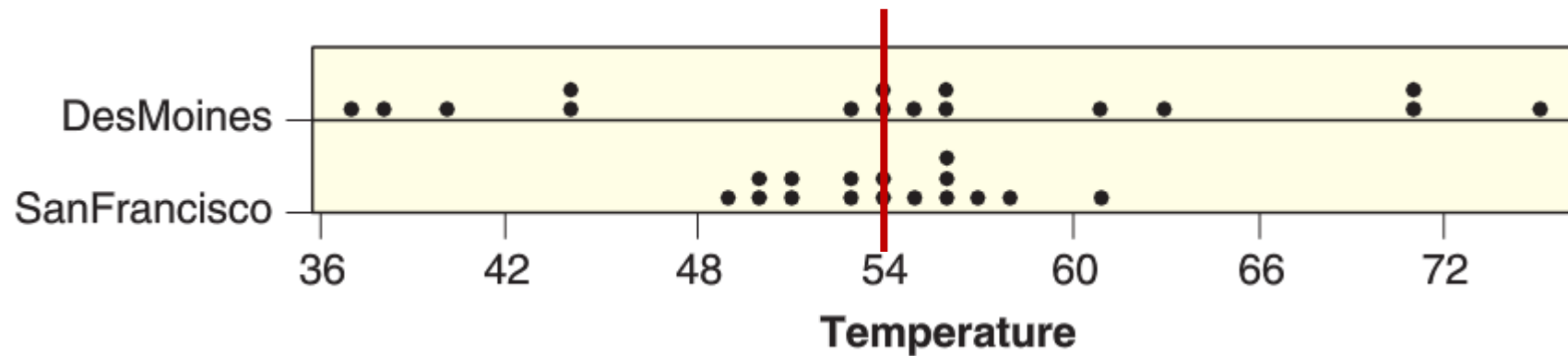
The mean and median are numbers that tell us about the center of a distribution



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:

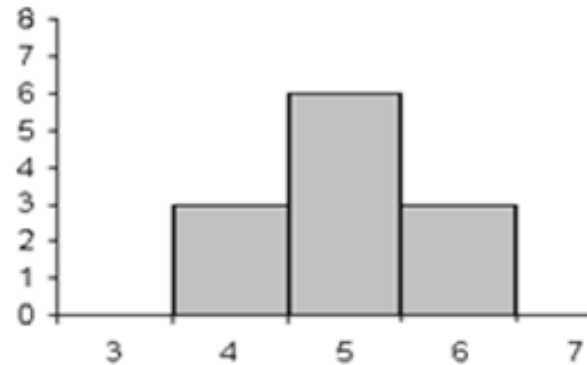


Mean temperature (°F): Des Moines = 54.49 San Fran = 54.01

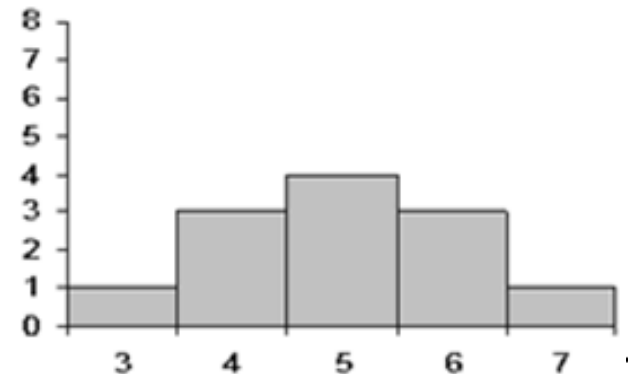
The standard deviation

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

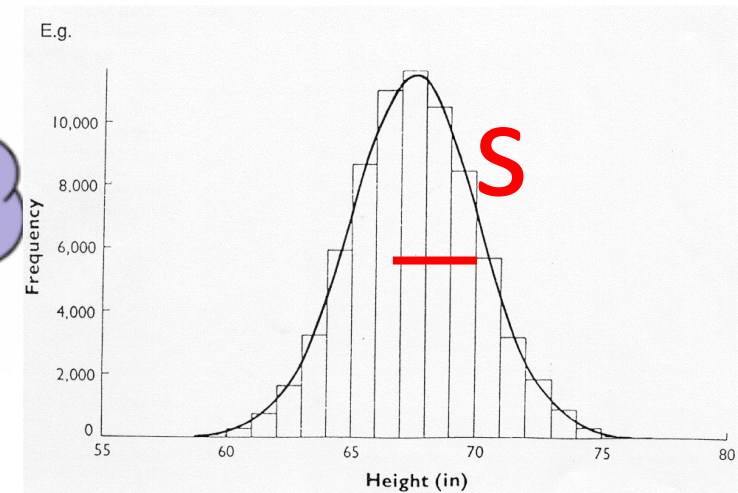
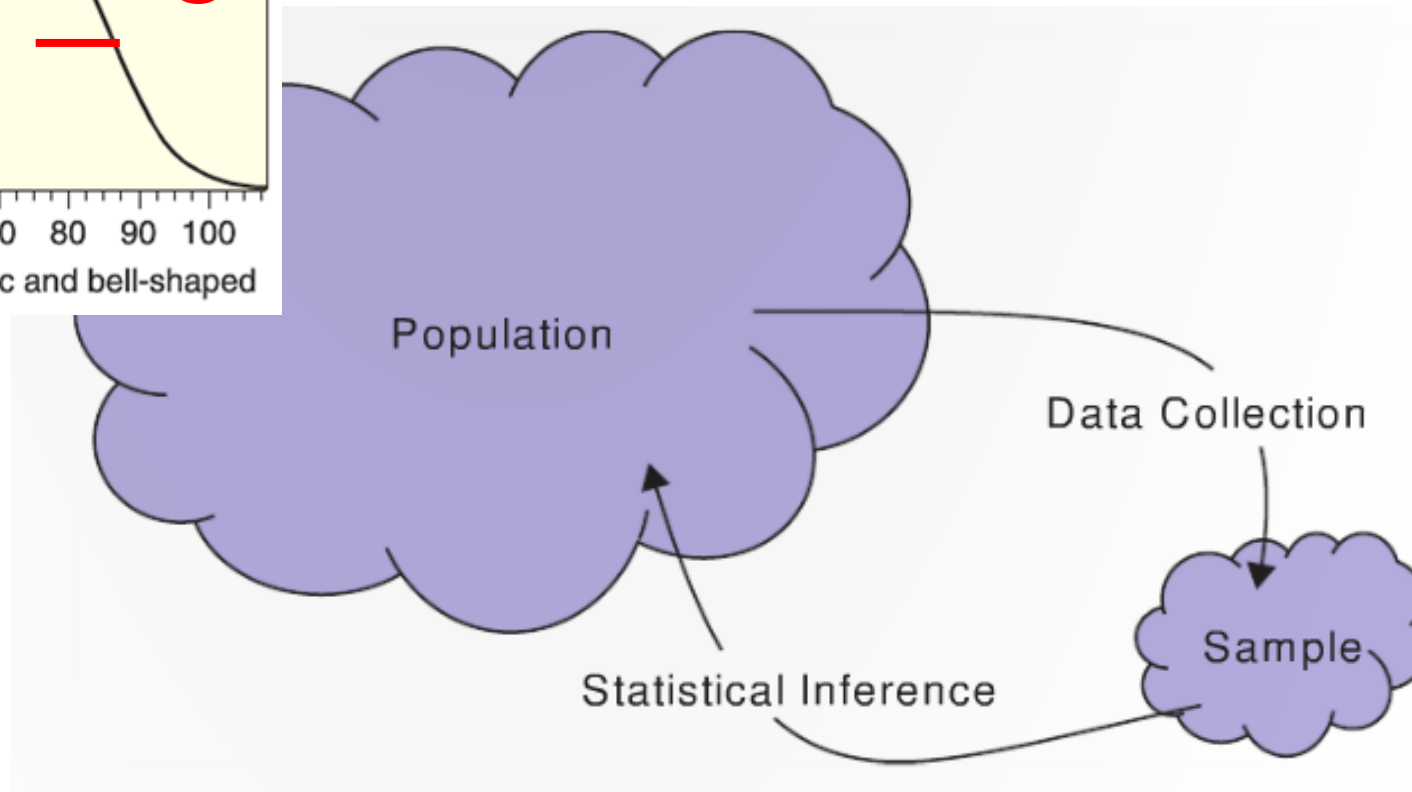
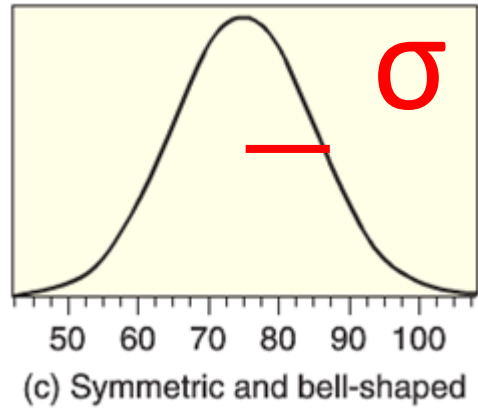
The standard deviation of the ***population*** is denoted σ

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

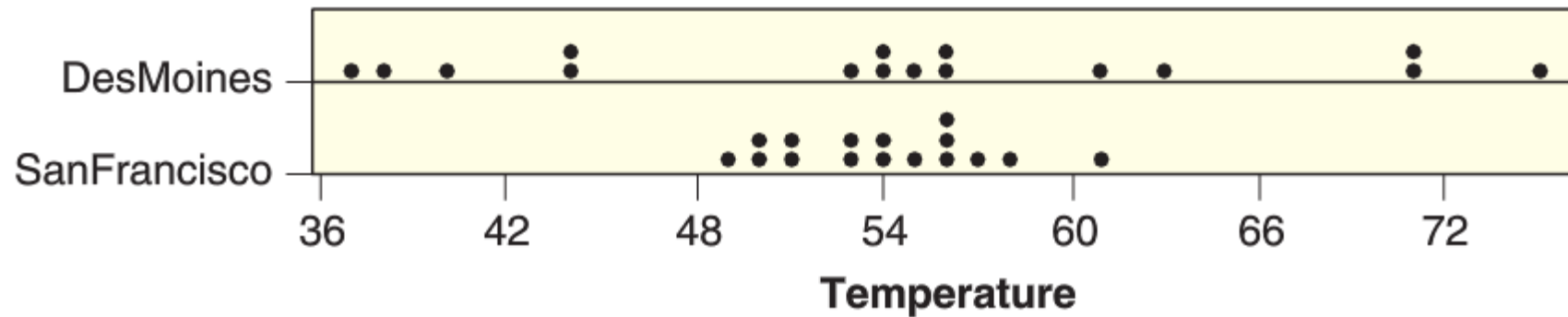
The standard deviation of a ***sample*** is denoted s

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

Population and sample standard deviation



Which has the larger standard deviation?



$$s_{DM} = 11.73 \text{ }^{\circ}\text{F}$$

$$s_{SF} = 3.38 \text{ }^{\circ}\text{F}$$

The standard deviation

The standard deviation can be computed using the following formula:

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

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:

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

The standard deviation can be computed using the formula:

$$s = \sqrt{\frac{1}{(n-1)} \sum_{i=1}^n (x_i - \bar{x})^2} \quad (\text{remember order of operations!})$$

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



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

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

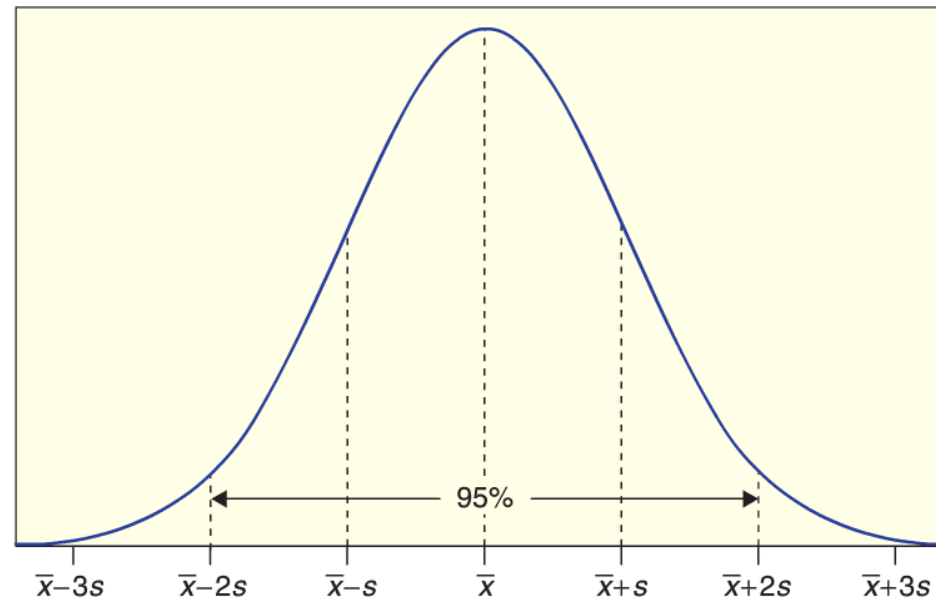
Normal distributions and z-scores

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*

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 $\bar{x} + 2s$

Example: IQ scores are normally distributed with a mean of 100 and a standard deviation of 15.

Question: what is the range of values that the middle 95% of IQ scores fall in?

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 $\bar{x} + 2s$

Example: IQ scores are normally distributed with a mean of 100 and a standard deviation of 15.

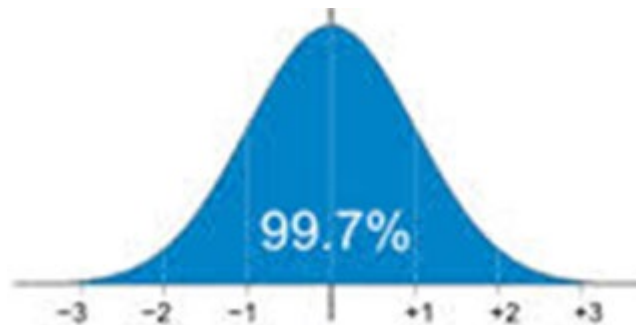
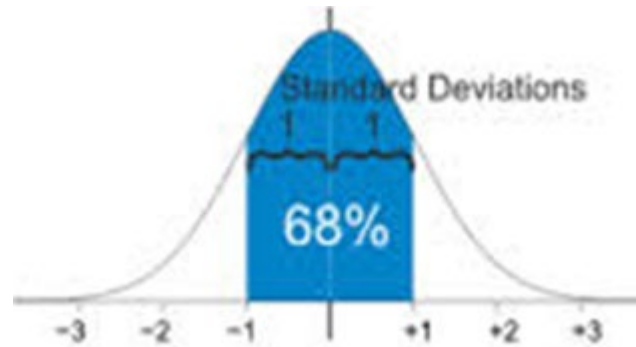
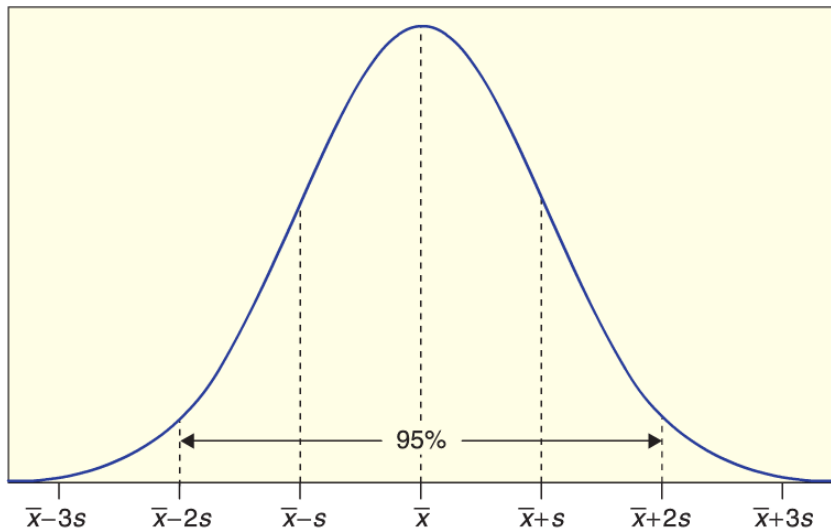
Question: what is the range of values that the middle 95% of IQ scores fall in?

Answer: $(100 - 30)$ to $(100 + 30)$, 95% of IQ scores are in the range 70 to 130

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?

LeBron James is a basketball player who had the following statistics in 2011:

- Field goal percentage (FGPct) = 0.510
- Points scored = 2111
- Assists = 554
- Steals = 124



The summary statistics of the NBA in 2011 are given below

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

	Mean	Standard Deviation
FGPct	0.464	0.053
Points	994	414
Assists	220	170
Steals	68.2	31.5

Question: Relative to his peers, which statistic is most and least impressive?

Which Accomplishment is most impressive?

LeBron James is a basketball player who had the following statistics in 2011:

- Field goal percentage (FGPct) = 0.510
- Points scored = 2111
- Assists = 554
- Steals = 124

The summary statistics of the NBA in 2011 are given below

<u>z</u>	=	<u>$(x - \bar{x}) / s$</u>		
Z-score FGPct	=	$(0.510 - 0.464)/0.053$	=	0.868
Z- score Points	=	$(2111 - 994)/414$	=	2.698
Z-score Assists	=	$(554 - 220)/170$	=	1.965
Z-score Steals	=	$(124 - 68.2)/31.5$	=	1.771

Percentiles

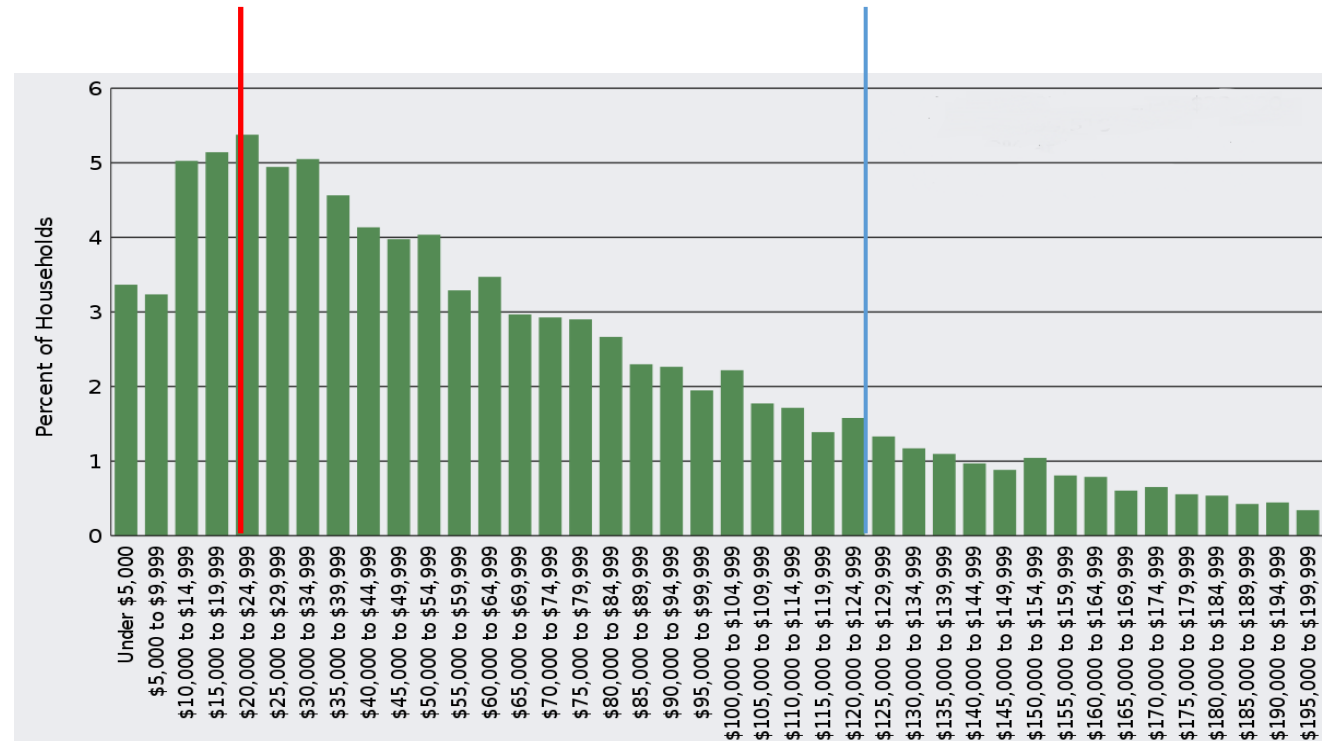
Percentiles

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)`

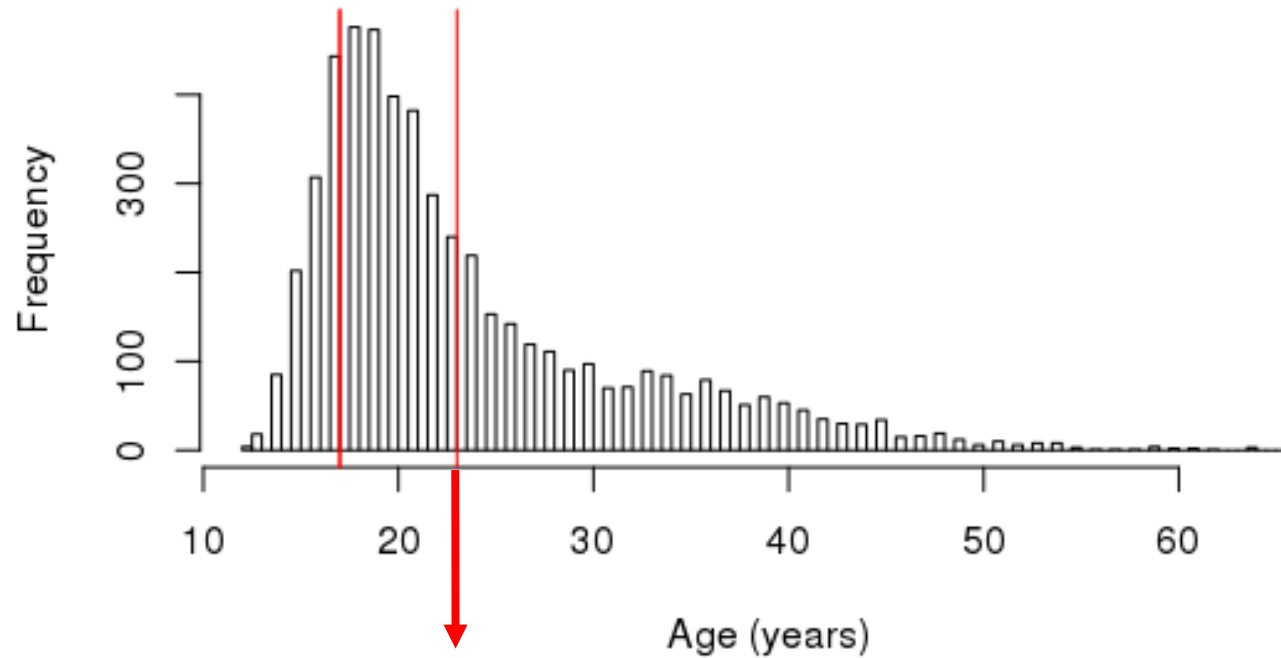
Age of marijuana arrests in Toronto



- > `install.packages('carData')`
 - > `library(carData)` # load the data
 - > `quantile(Arrests$age, .2)` # get the 20th percentile value from a vector of ages of arrests
- 20th percentile value is 17
i.e., 20% of the arrests were of ages 17 or less

Age of marijuana arrests in Toronto

Histogram of Ages of people arrested for marijuana use



60th percentile value is 23

i.e., 60% of the arrests were of ages 23 or less

```
> quantile(Arrests$age, c(.2, .6)) # get the 20th and 60th percentile values from a vector of ages of arrests
```


Five Number Summary

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

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

Q_3 = 75th percentile (also called 3rd 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

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

Answer in R: `fivenum(v)`

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 11:30pm on Sunday February 14th

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