Measures of spread

Overview

Quick review of distributions

Outliers

The mean and median

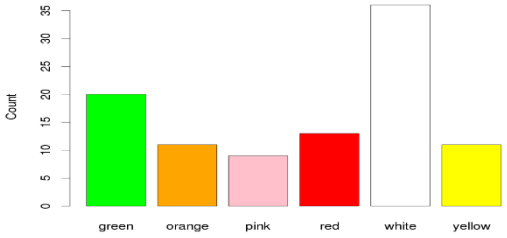
The standard deviation

Z-scores

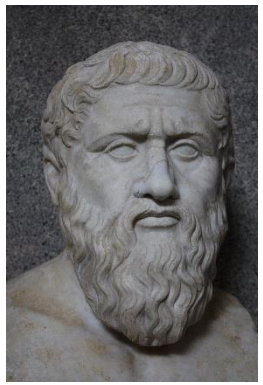
Percentiles

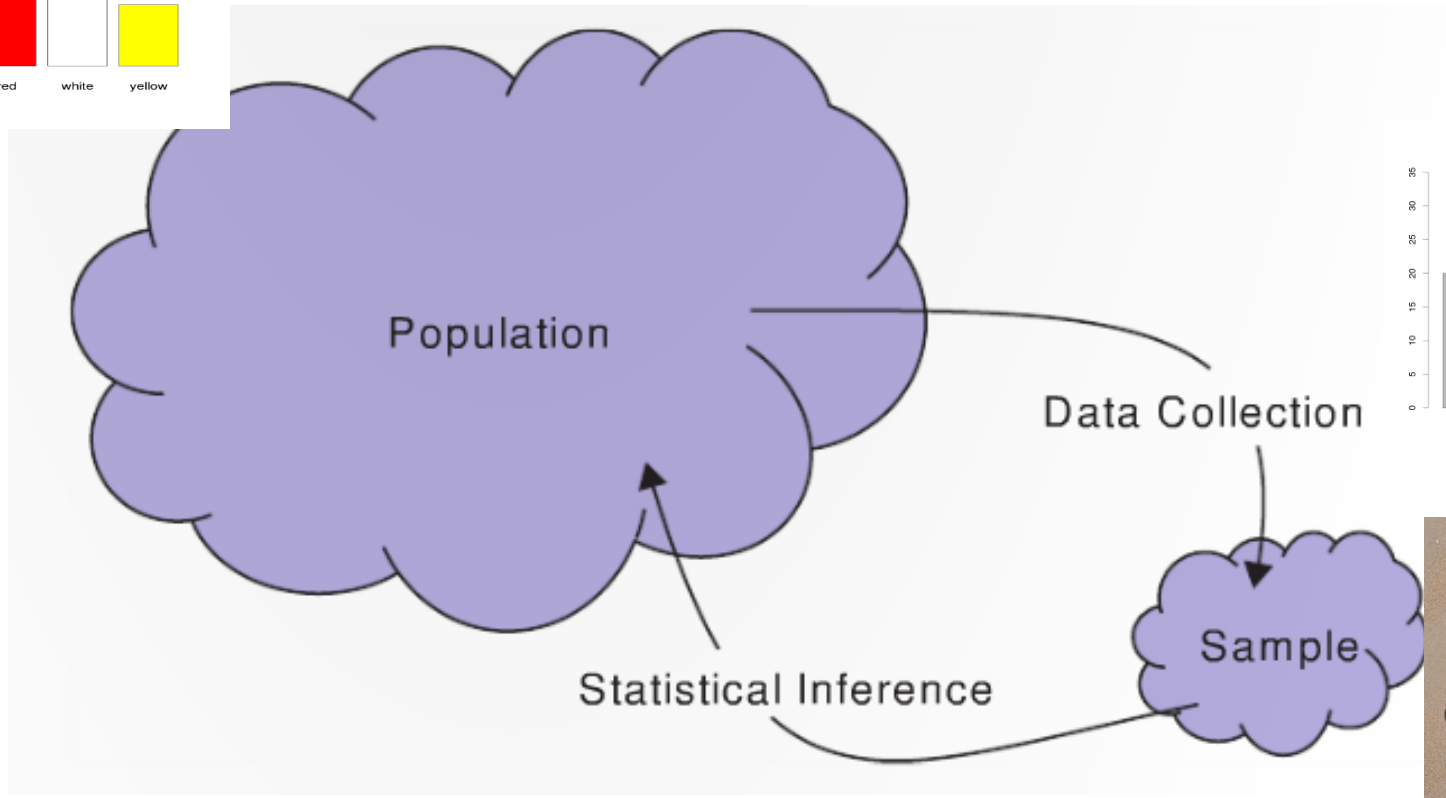
R Markdown and the next worksheet

Categorical variables: Sample vs. Population proportion

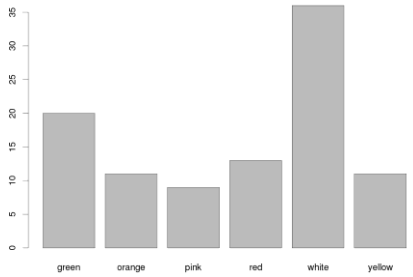


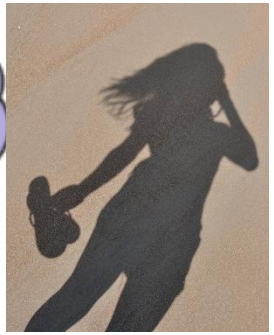
π_{red}



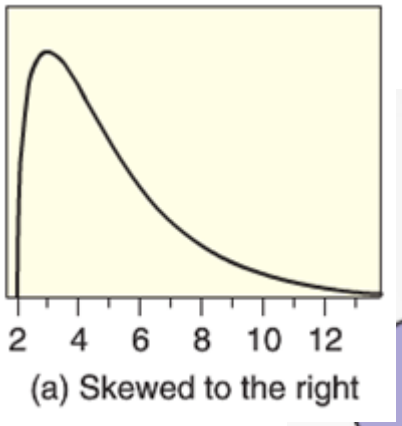


$\hat{p}_{red}$

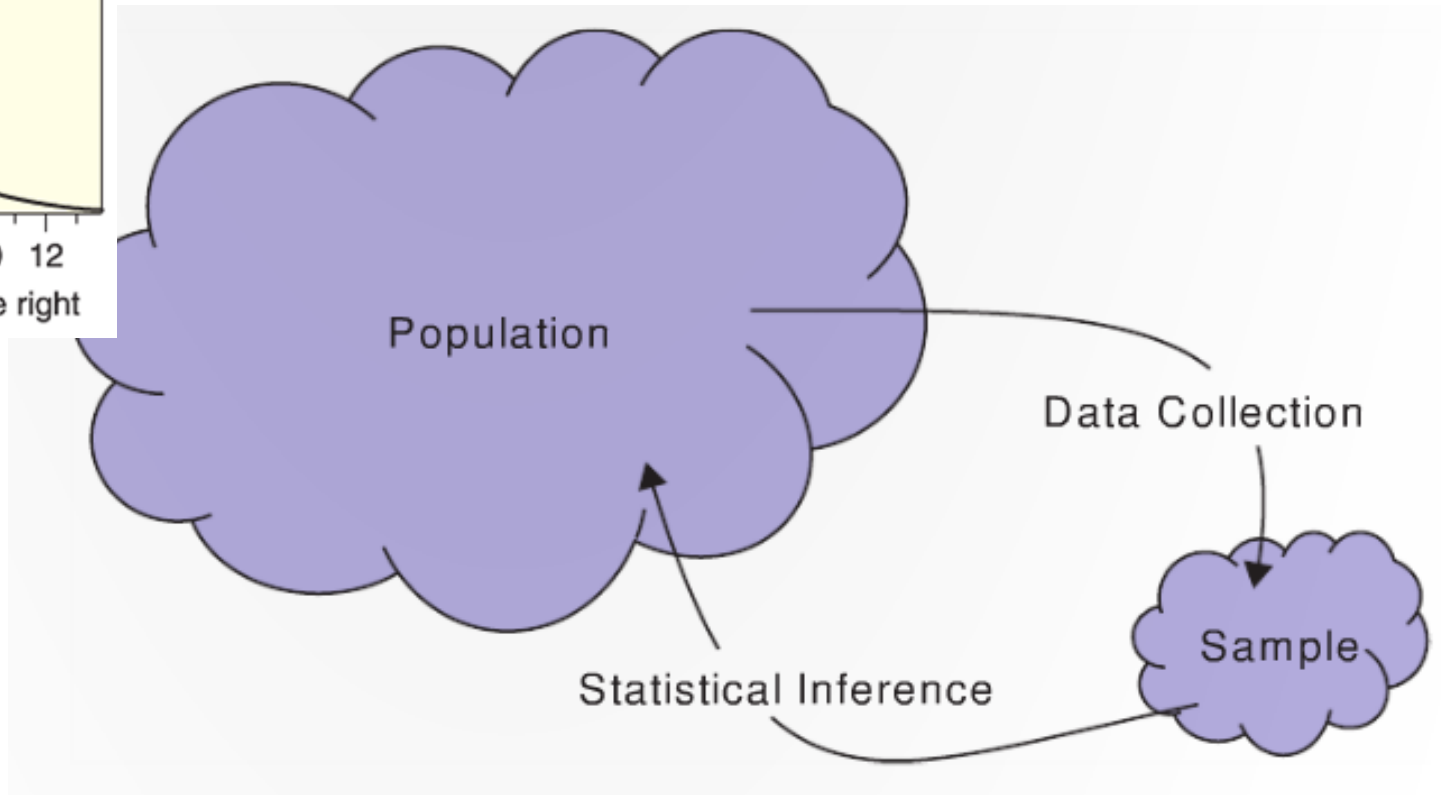


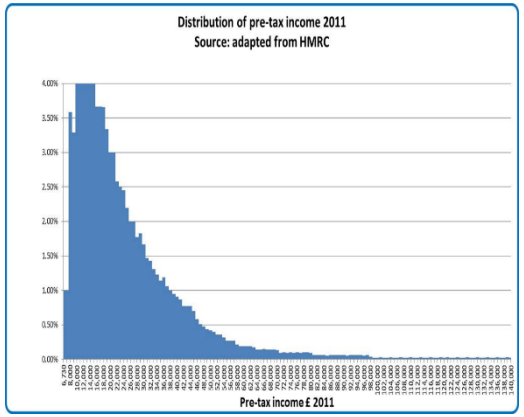


Quantitative variables: Sample vs. Population means

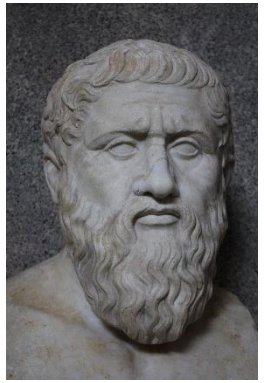


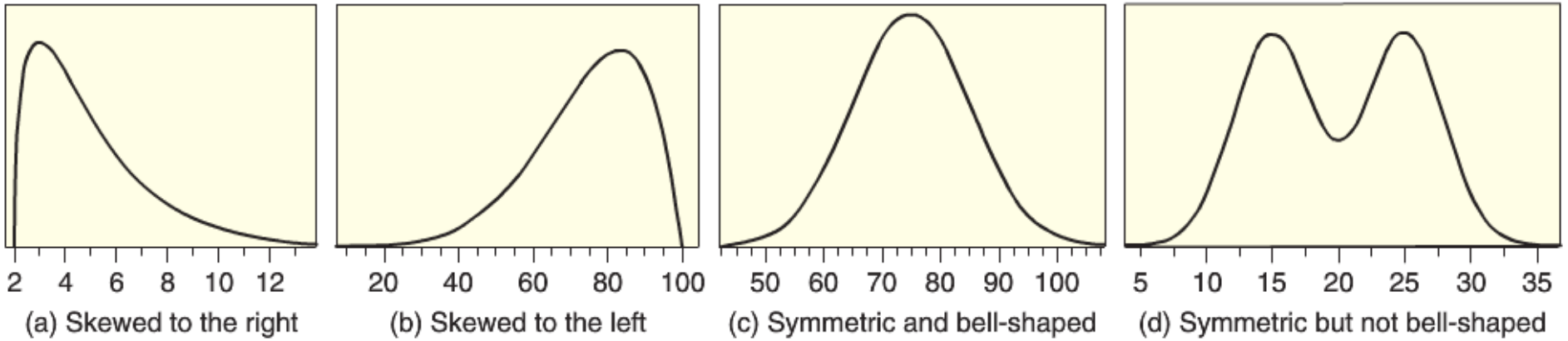
(a) Skewed to the right



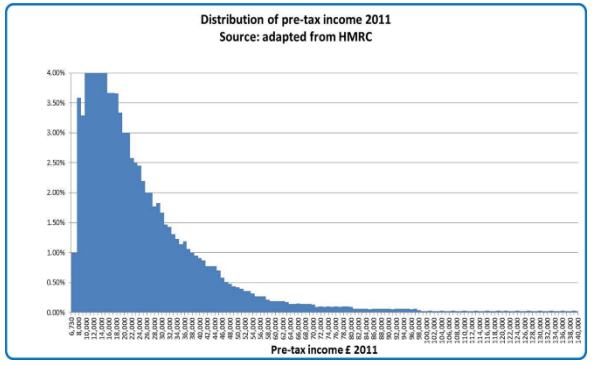


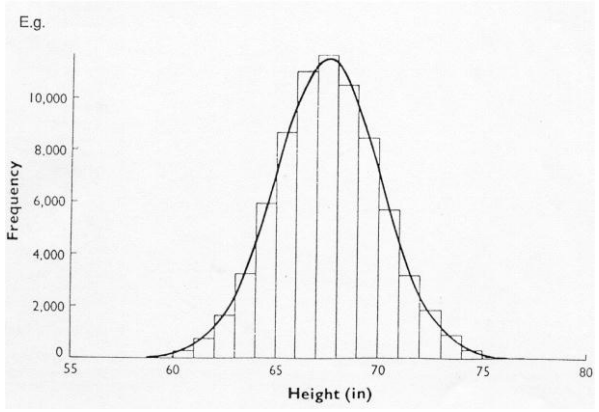
Plato and shadows: distributions and histograms

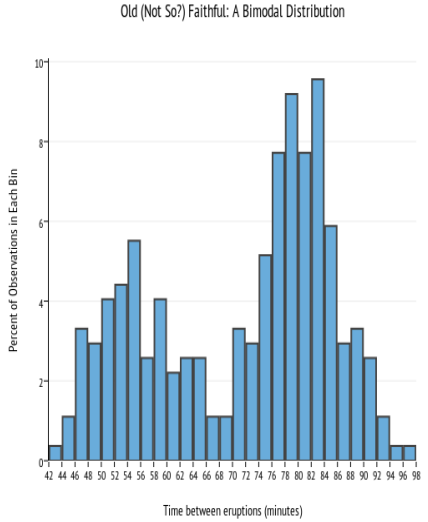




Income distribution

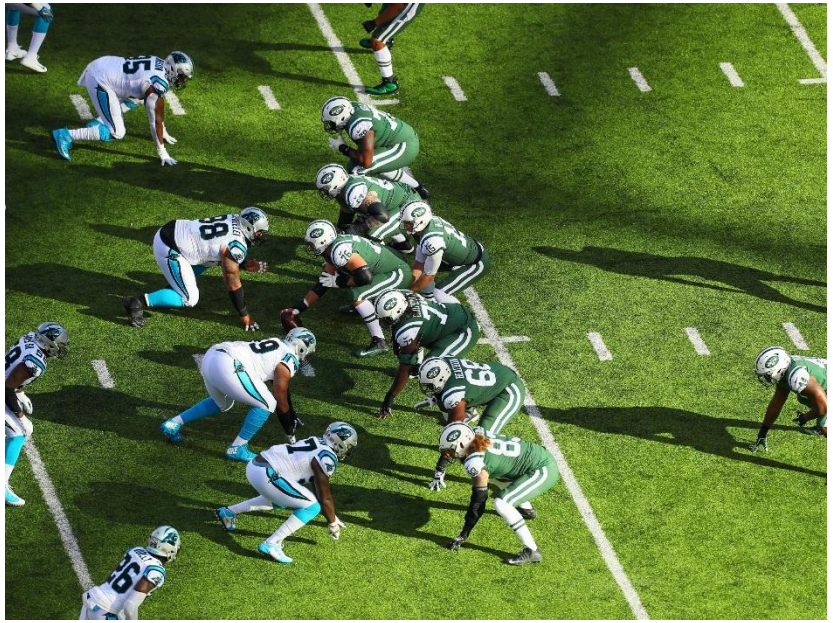


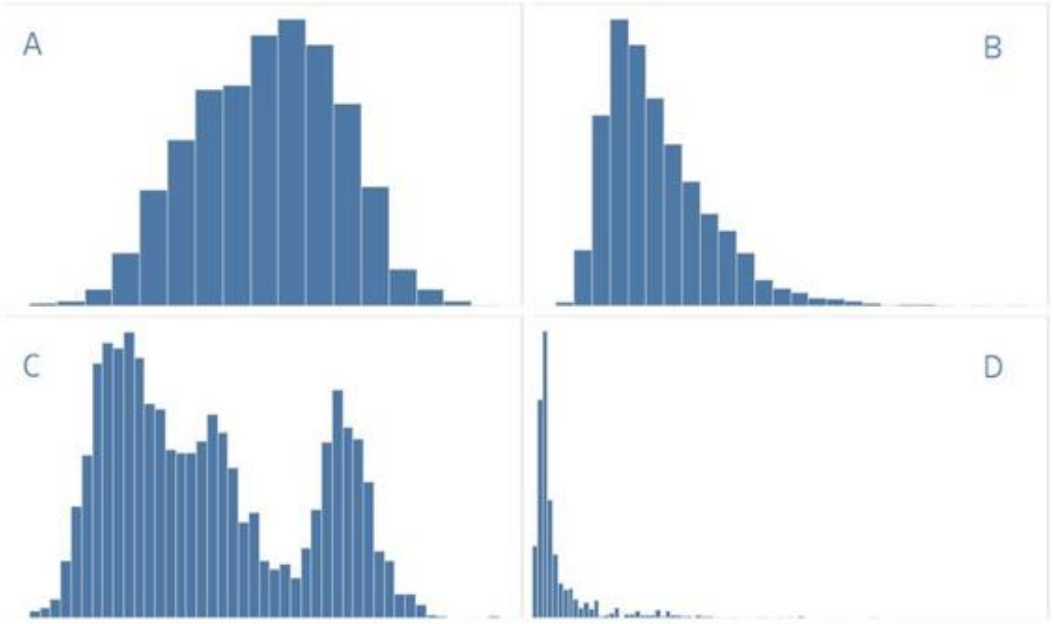




Neat facts - average NFL player is:

- 1. About 25 years old (age)
- 2. Just over 6'2" in height (height)
- 3. Weighs a little more than 244lbs (weight)
- 4. Makes slightly less than $1.5M in salary per year (salary)





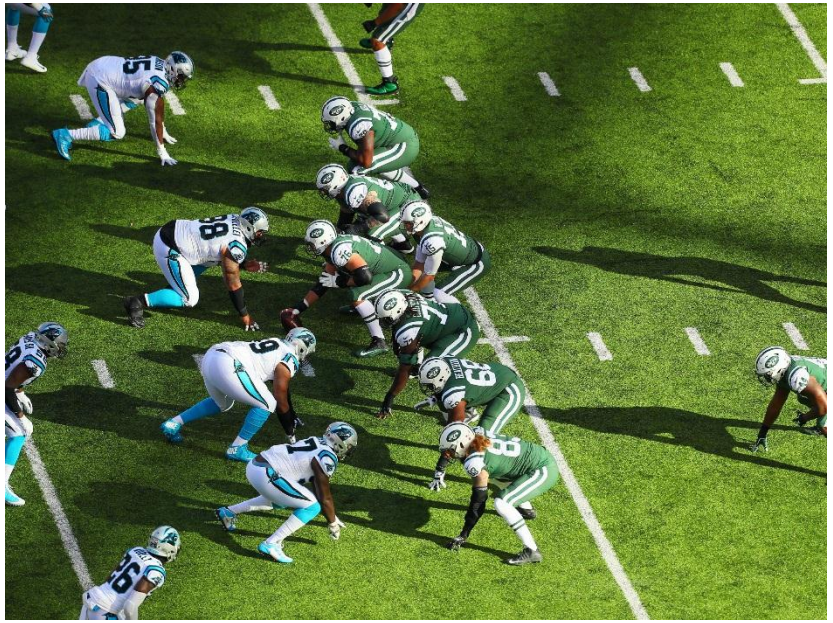
Question: Can you tell which histogram goes with which trait?

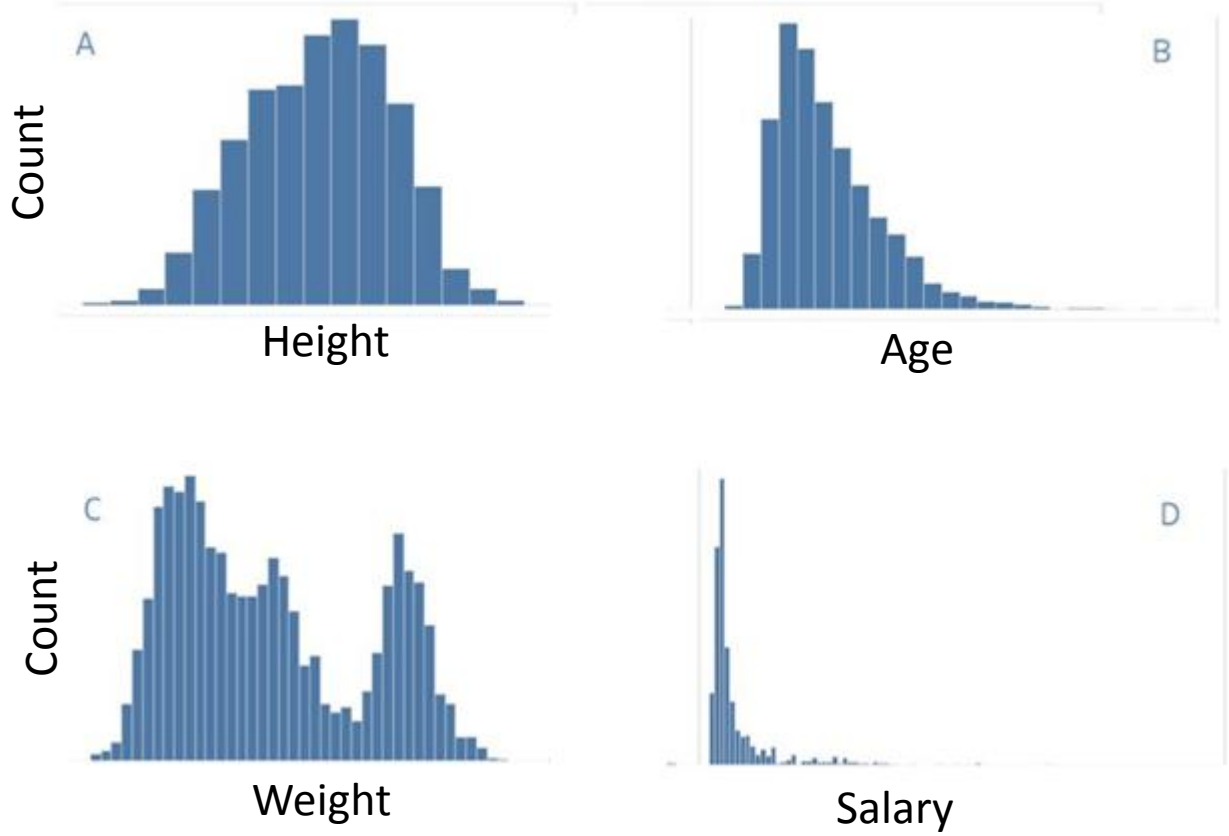
Task is to add the labels: age, height, weight, and salary

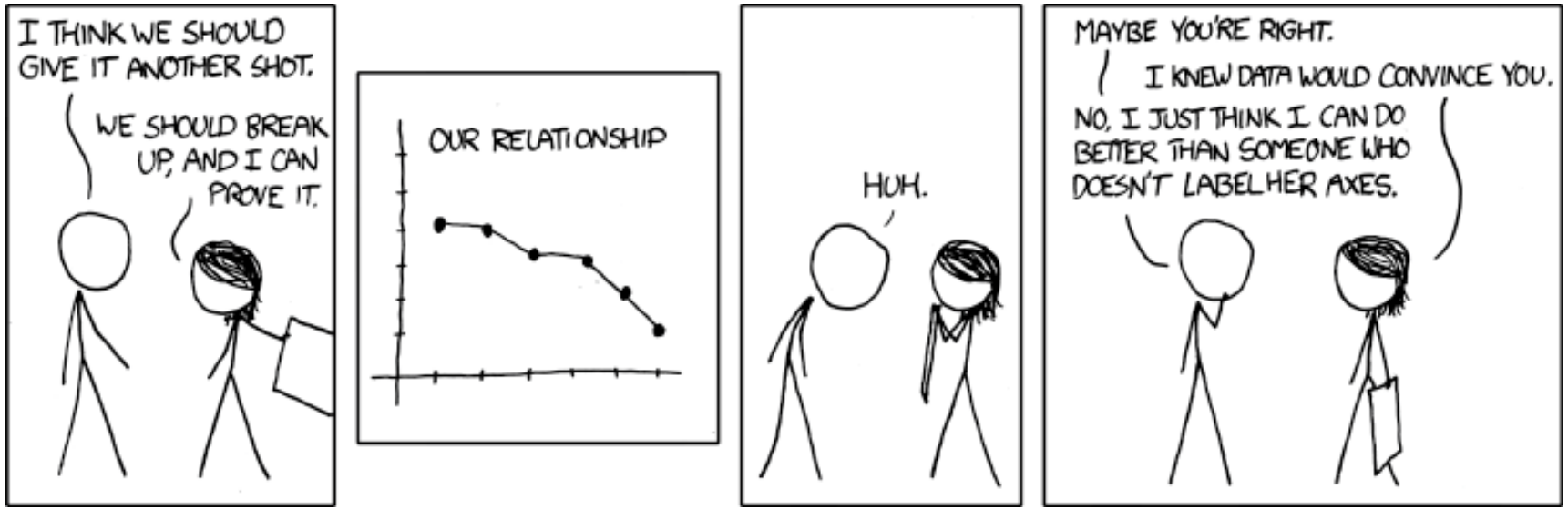
- Hint: There are a wide range of positions in football that have very different roles
  - E.g., placekickers only play for small factions of the game, while quarterbacks are essentially to a team's success

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

- A: Frequency or count







If you don’t want an ex, label you axes!

Back to the Gapminder data…

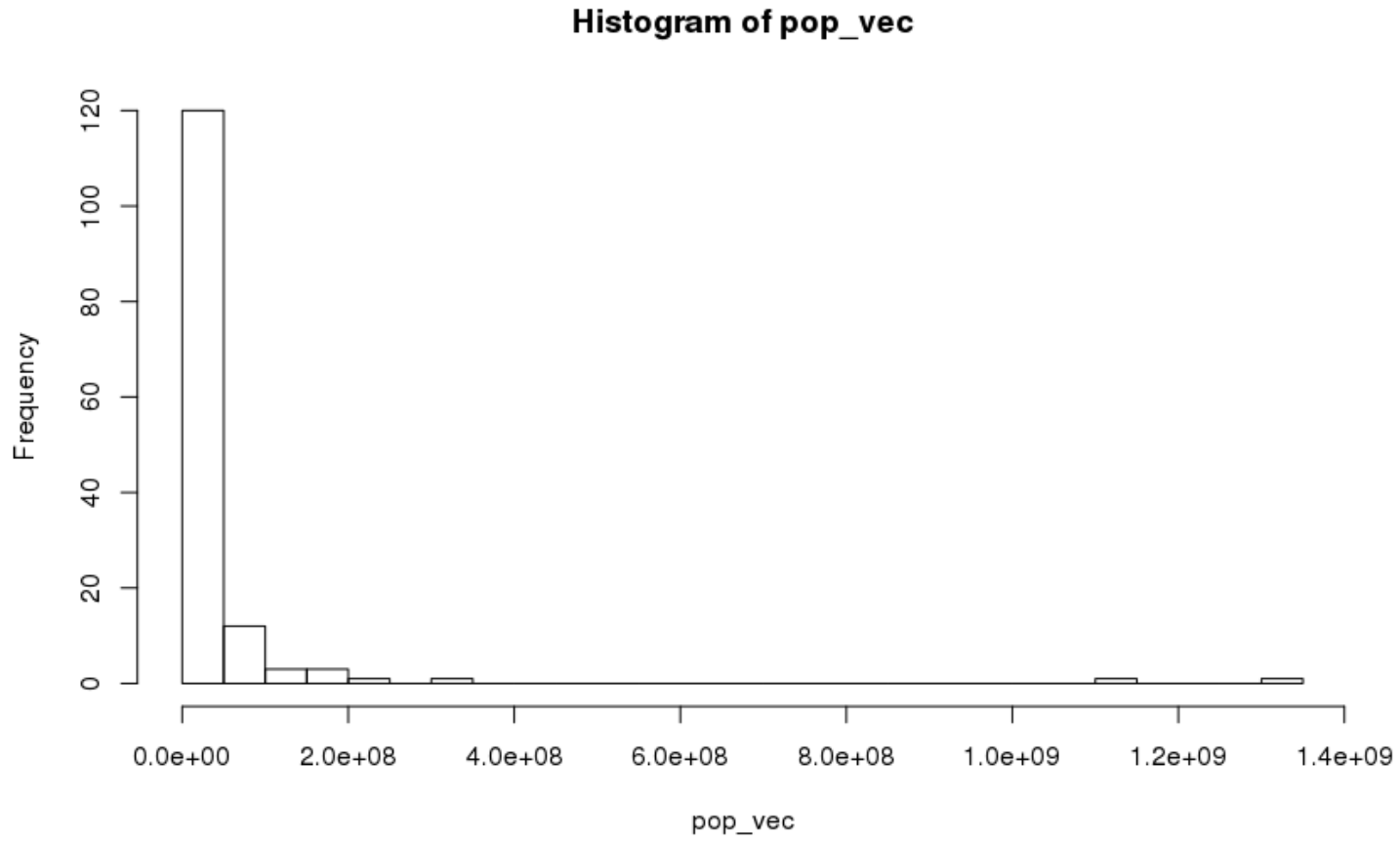
```
# get a data frame with information about the countries in the world
> source("/home/shared/intro_stats/cs206_functions.R")
> country_data <- get_cs206_data("gapminder_2007")
```

	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 <- country_data$pop   # first create a vector with the population of each country
> hist(pop_vec, n = 20)      # then create the histogram
```

What is missing from this histogram?



Axes labels could be more informative!

Labeling axes

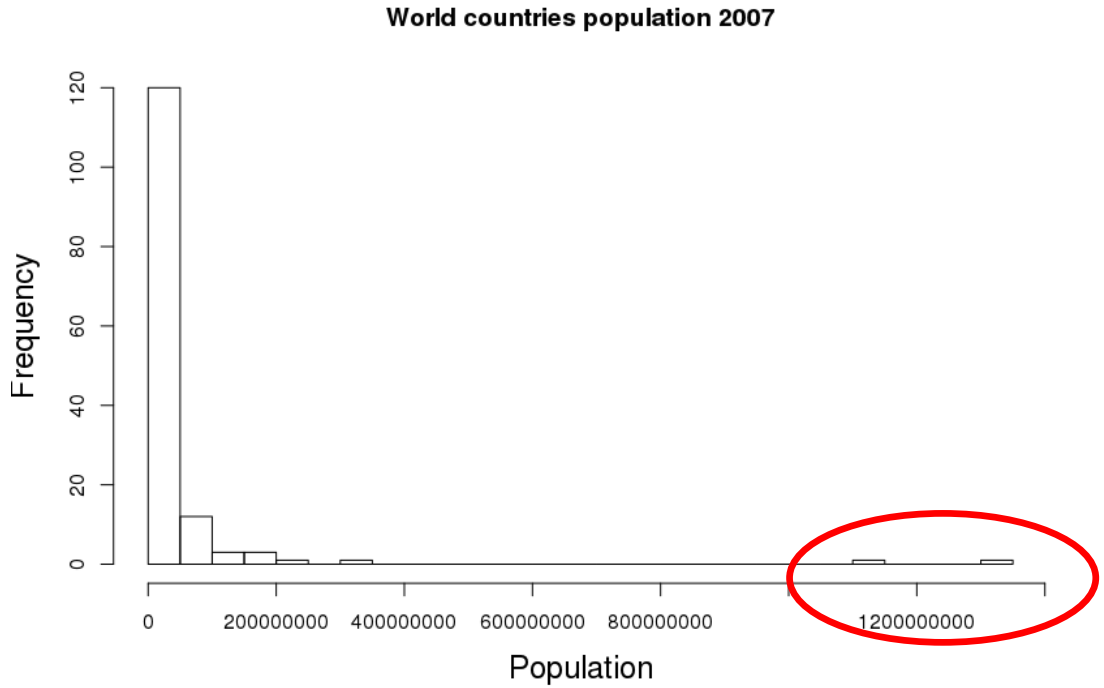
Can you figure out how to label the axes?

- A: xlab and ylab!

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

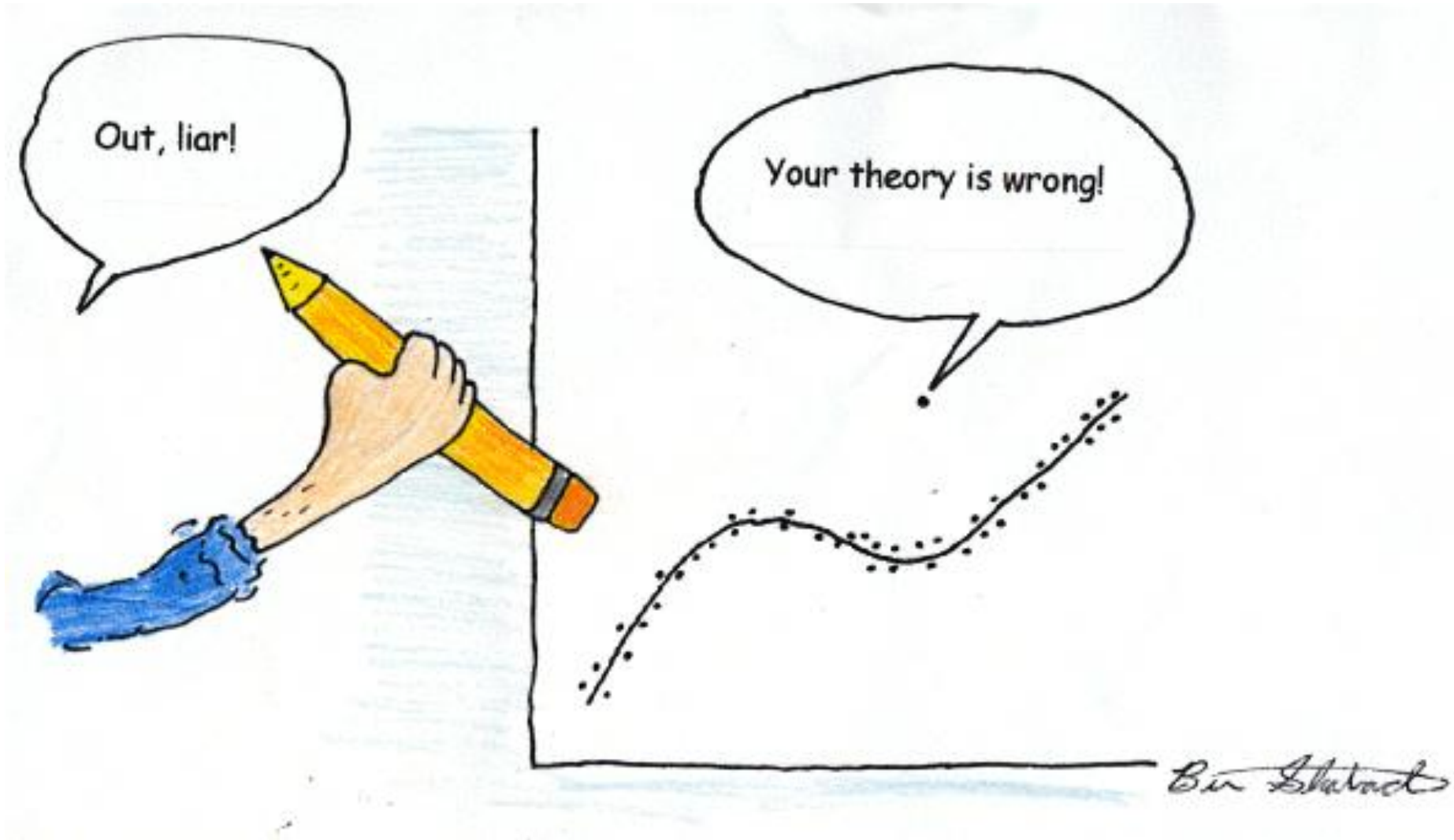
Outliers

An **outlier** is an observed value that is notably distinct from the other values in a dataset by being much smaller or larger than the rest of the data.



Outliers can potentially have a large influence on the statistics you calculate

- One should examine outliers in more detail to understand what is causing them


Out, liar!
Your theory is wrong!

Descriptive statistics for the center of a distribution

Graphs are useful for visualizing data to get a sense of what of what the data look like

We can also summarize data numerically

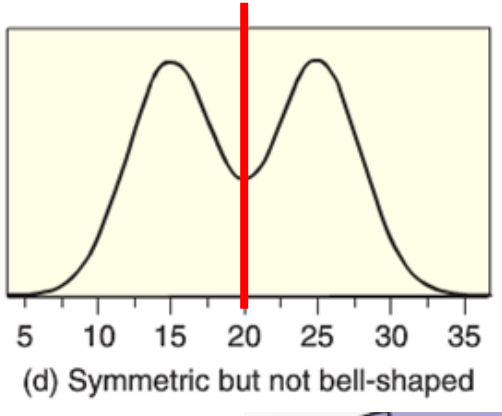
Question: what is a numerical summary of a sample of data called?

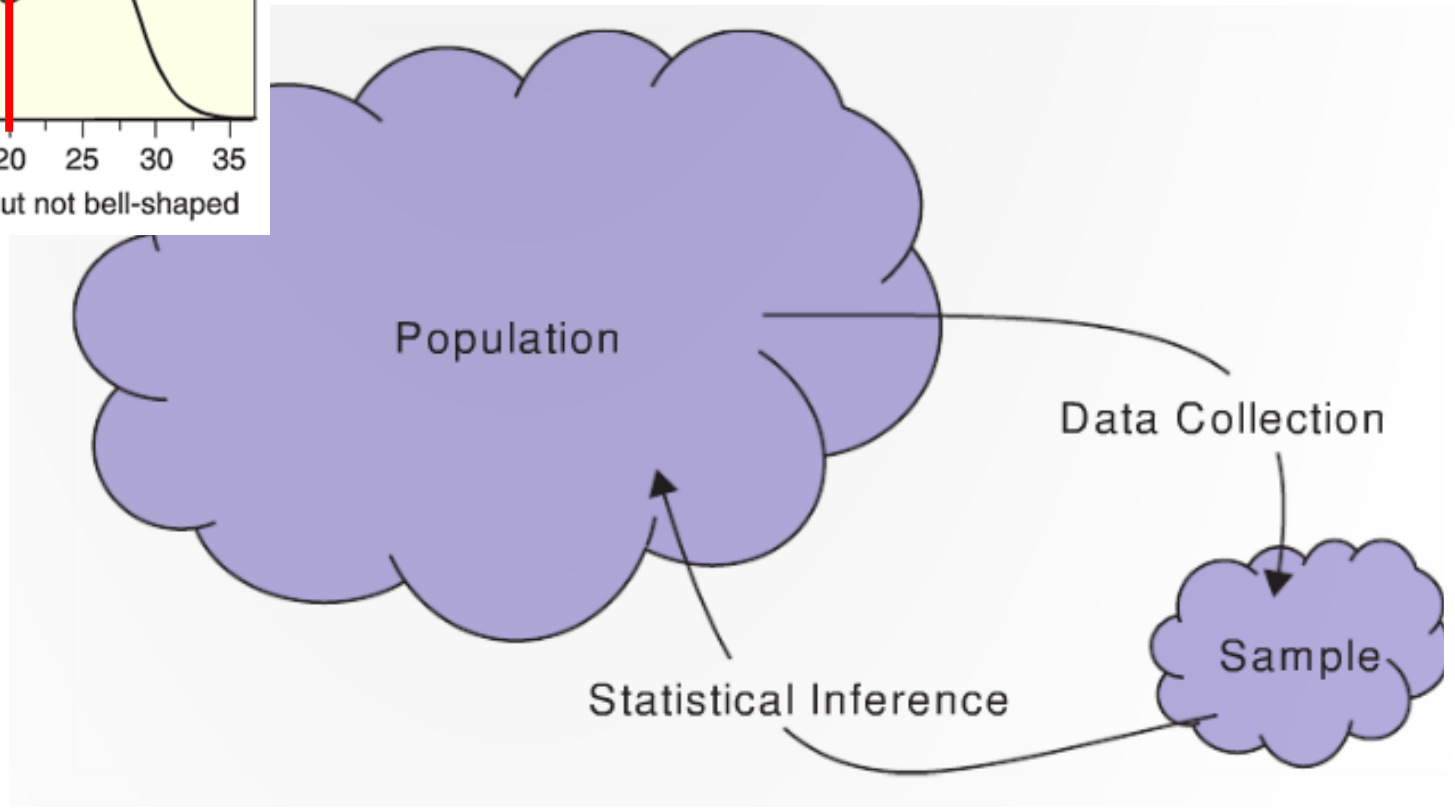
A: a statistic!

Two important statistics that can be used to describe the center of the data are the **mean** and the **median**

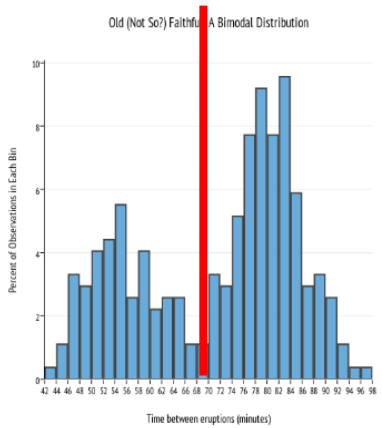
Sample and population mean

μ





$\bar{x}$



The mean

$$\text{Mean} = \frac{\text{Sum of all data values}}{\text{Number of data values}}$$

$$\text{Mean} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n} = \frac{\sum_i^n x_i}{n}$$

```
R: mean(x)
R: mean(x, na.rm = TRUE)
```

Give the proper notation: μ vs. $\bar{x}$?

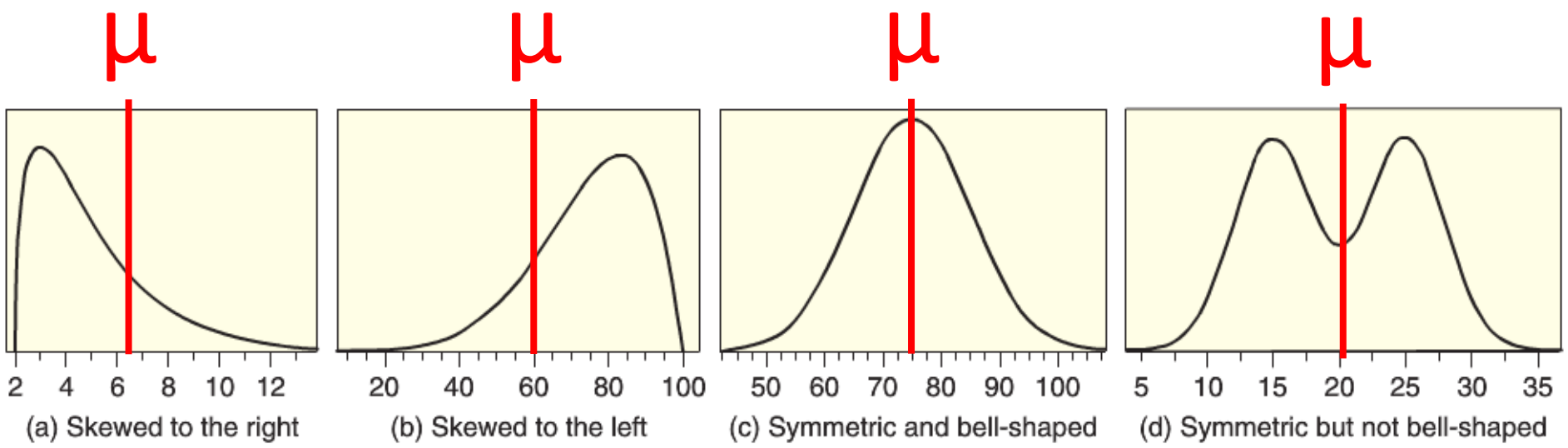
We measure the height of 50 randomly chosen Hampshire students

We measure the height of all Hampshire students

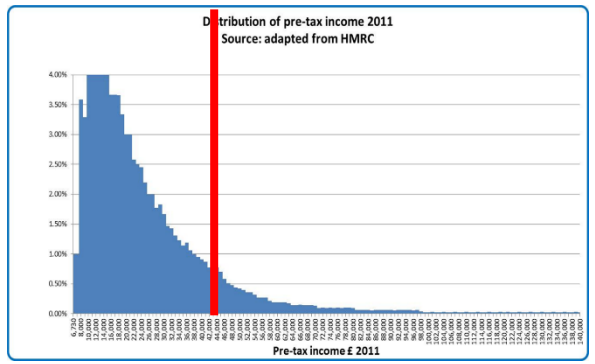
Can you calculate the mean of the countries life expectancy in R?

```
> life_expectancy <-  country_data$lifeExp
> mean(life_expectancy)
```

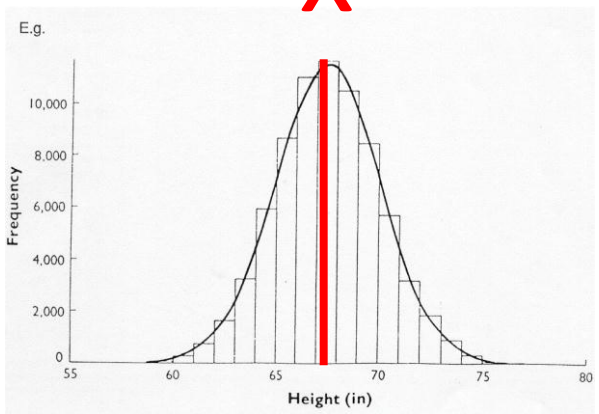
Histograms: a way to plot quantitative data



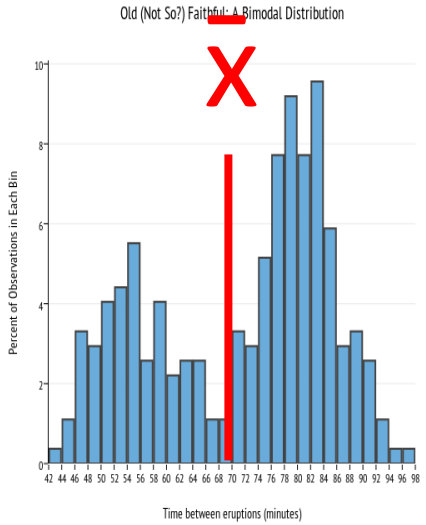
$\bar{x}$



$\bar{x}$



$\bar{x}$



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

Example of calculating the mean and median

When a perspective Hampshire student visit a Hampshire webpage a ‘ping’ is generated

Below is a random sample of ping counts from 7 perspective students who pinged the site at least once:

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

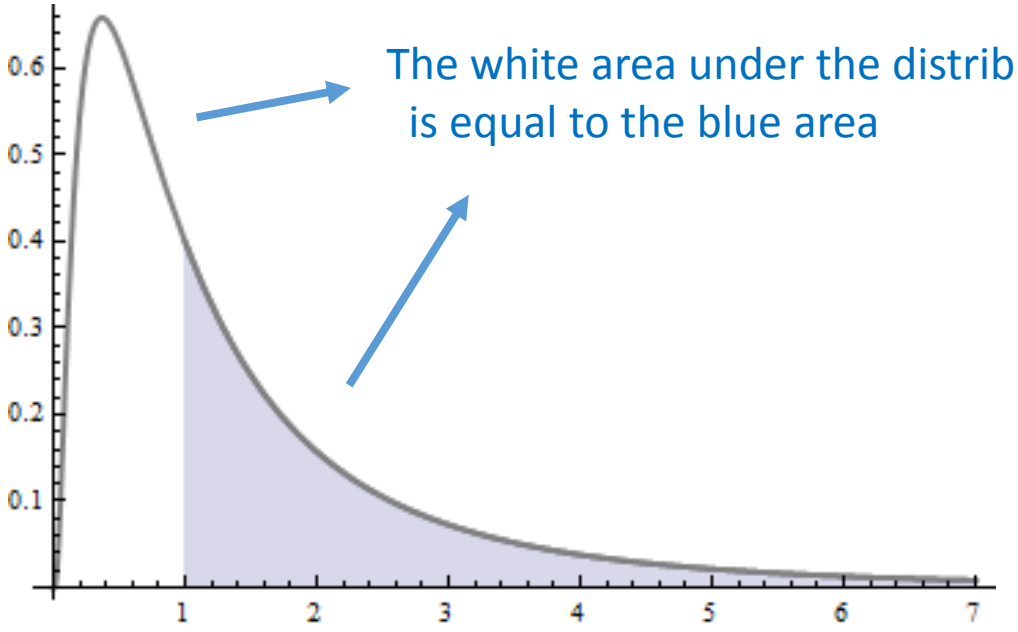
Q: What is the mean and median ping count in this sample?

A: mean = 42

median = 12

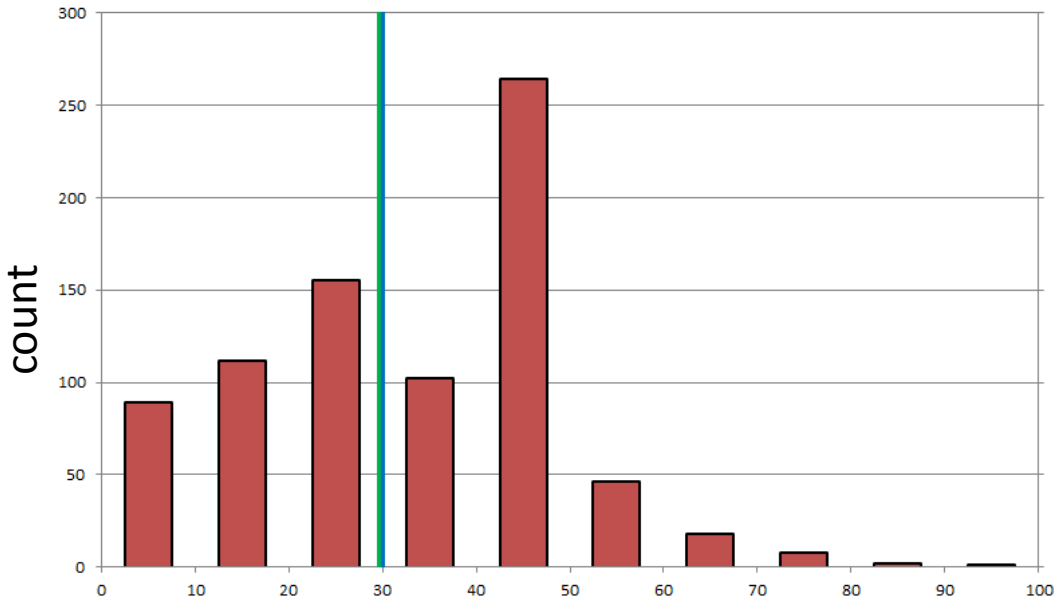
Mean = $\frac{\sum_i^n x_i}{n}$

The median



The white area under the distribution is equal to the blue area

The sum of the heights of the bars on the left is equal to the sum of the heights of the bars on the right



R: `median(v)`

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

Resistance

We say that a statistics is **resistant** if it is relatively unaffected by extreme values (outliers).

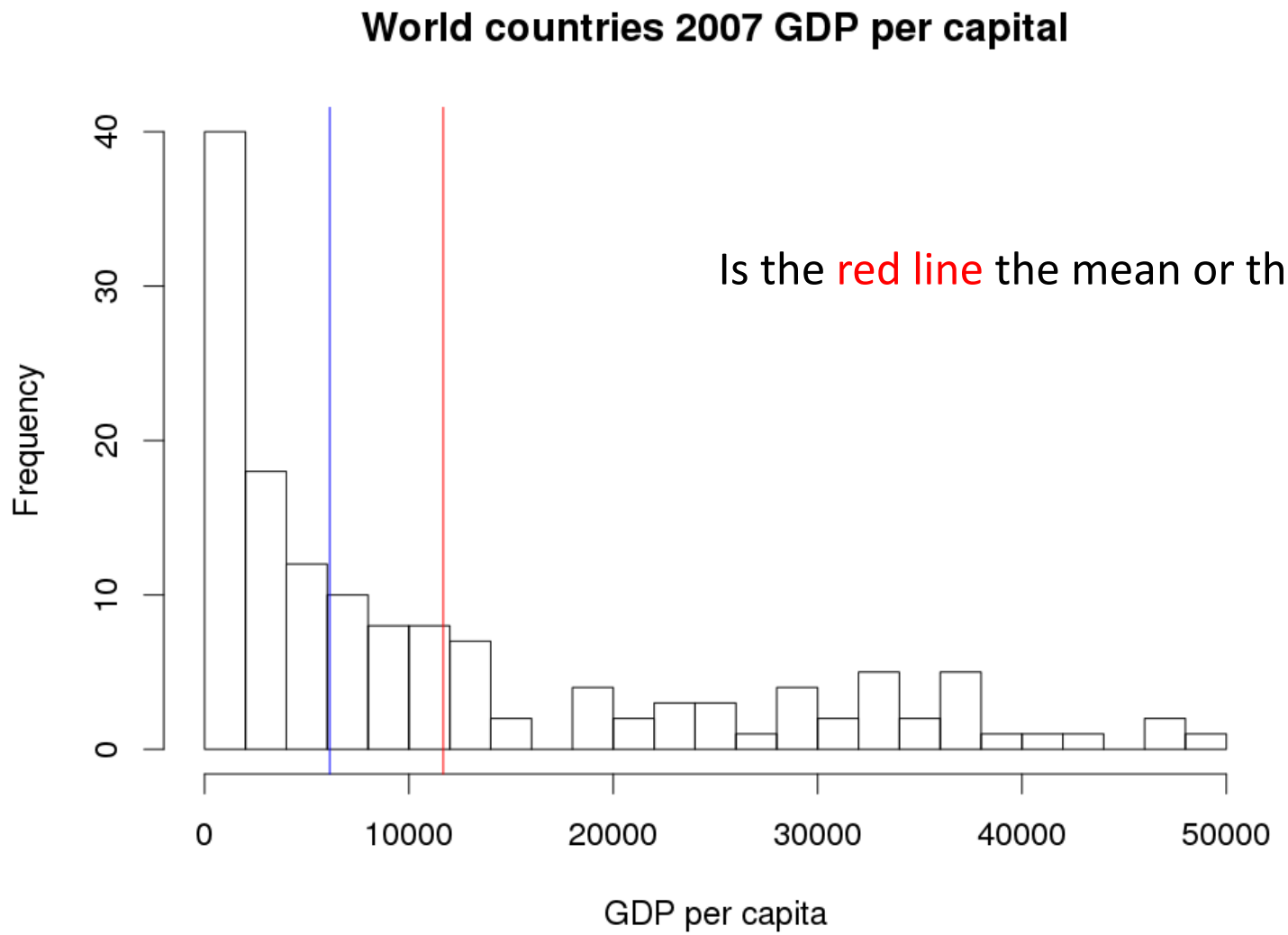
The median is resistant when the mean is not

Example:

Mean US salary = $72,641

Median US salary = $51,939

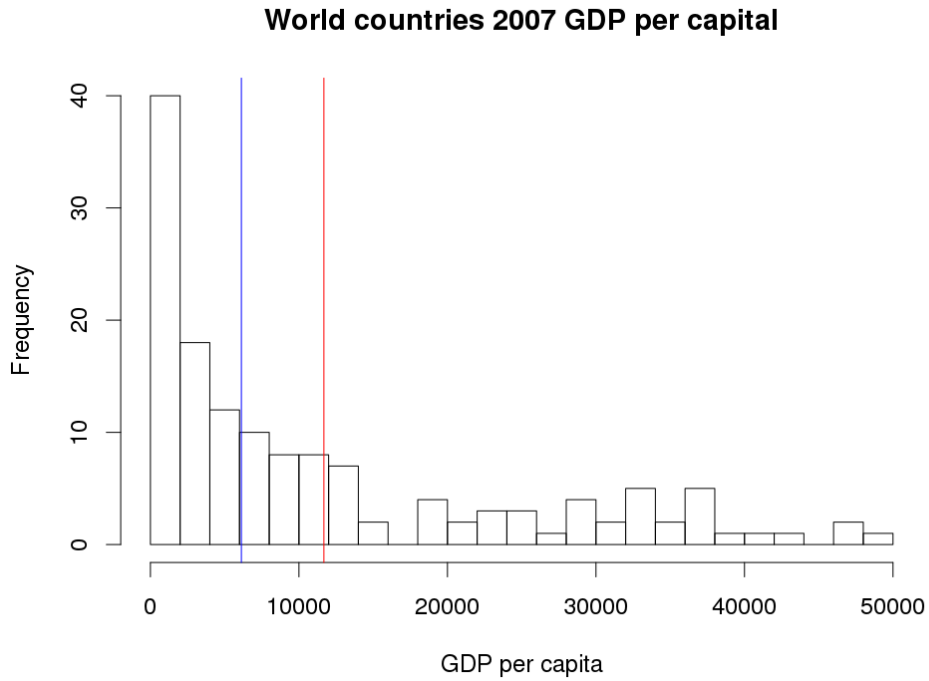
Measure of central tendency: mean and median



Is the red line the mean or the median?

Characterizing the spread

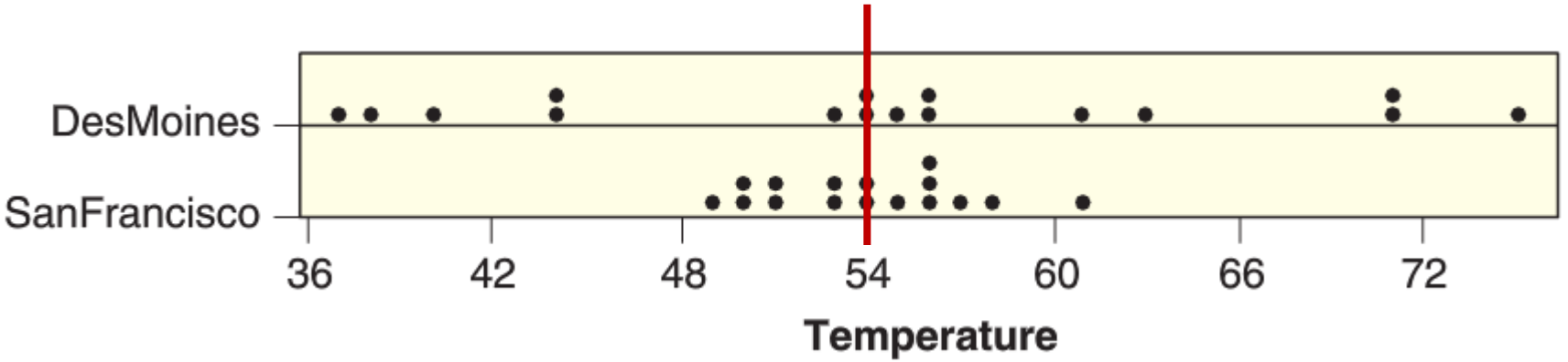
The mean and median are numbers that tell us about the center of a distribution



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:

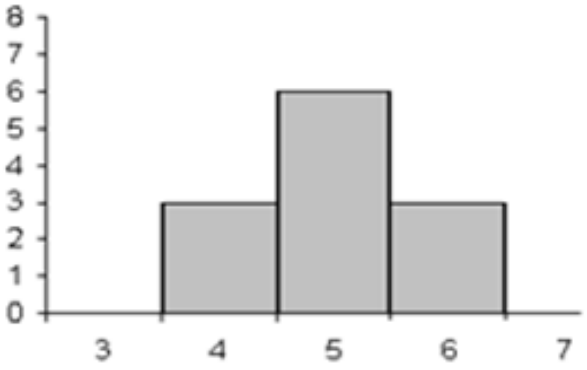


Mean temperature (°F): Des Moines = 54.49 San Fran = 54.01

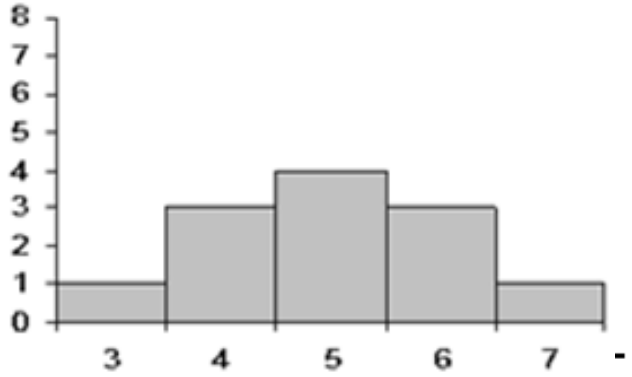
The standard deviation

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

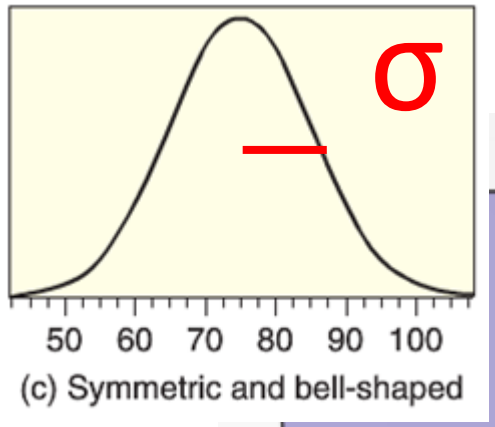
The standard deviation of a ***sample*** is denoted **s**

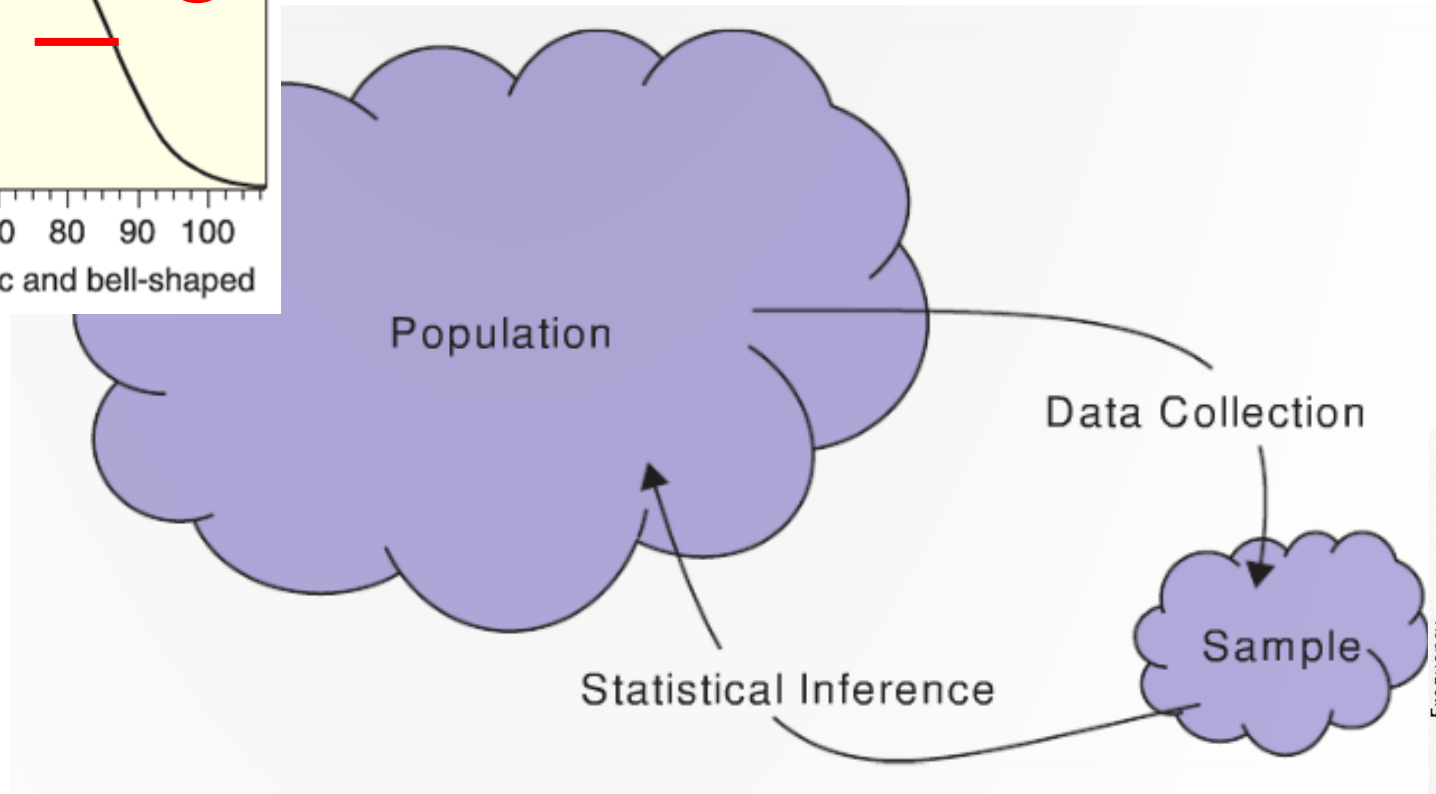
- It measure the spread of the data from the sample mean

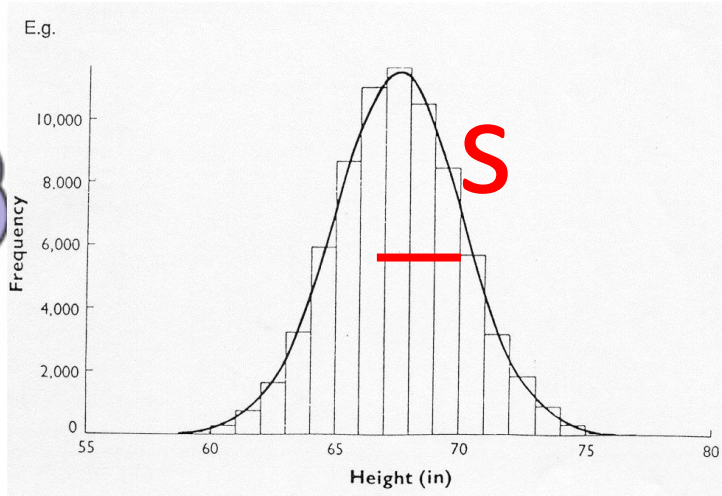
The standard deviation of the ***population*** is denoted **σ**

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

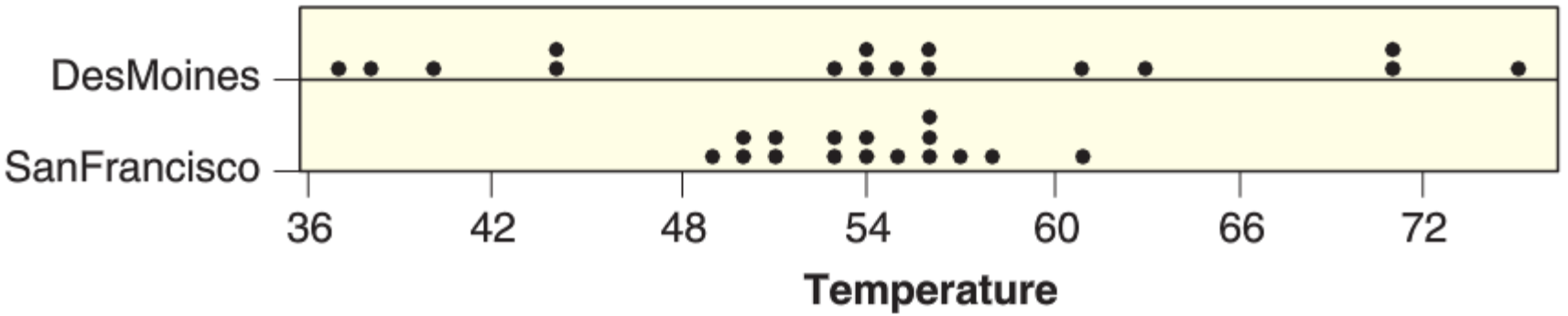
Population and sample standard deviation







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

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

Part 1: Calculate the mean and standard deviation for the number of hot dogs eaten!

Worksheet 2: examining cars sold

Lock5 questions

Worksheet 2!

A. Accessing R with the Hampshire server: https://asterius.hampshire.edu

B. Go to the console and download a file using the following command:

```
> source('/home/shared/intro_stats_2018/cs206_functions.R')
> get_worksheet(2)
```