Measures of central tendency and spread

Overview

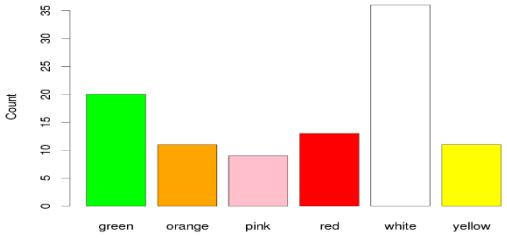
Quick review and continuation of shapes distributions

Outliers

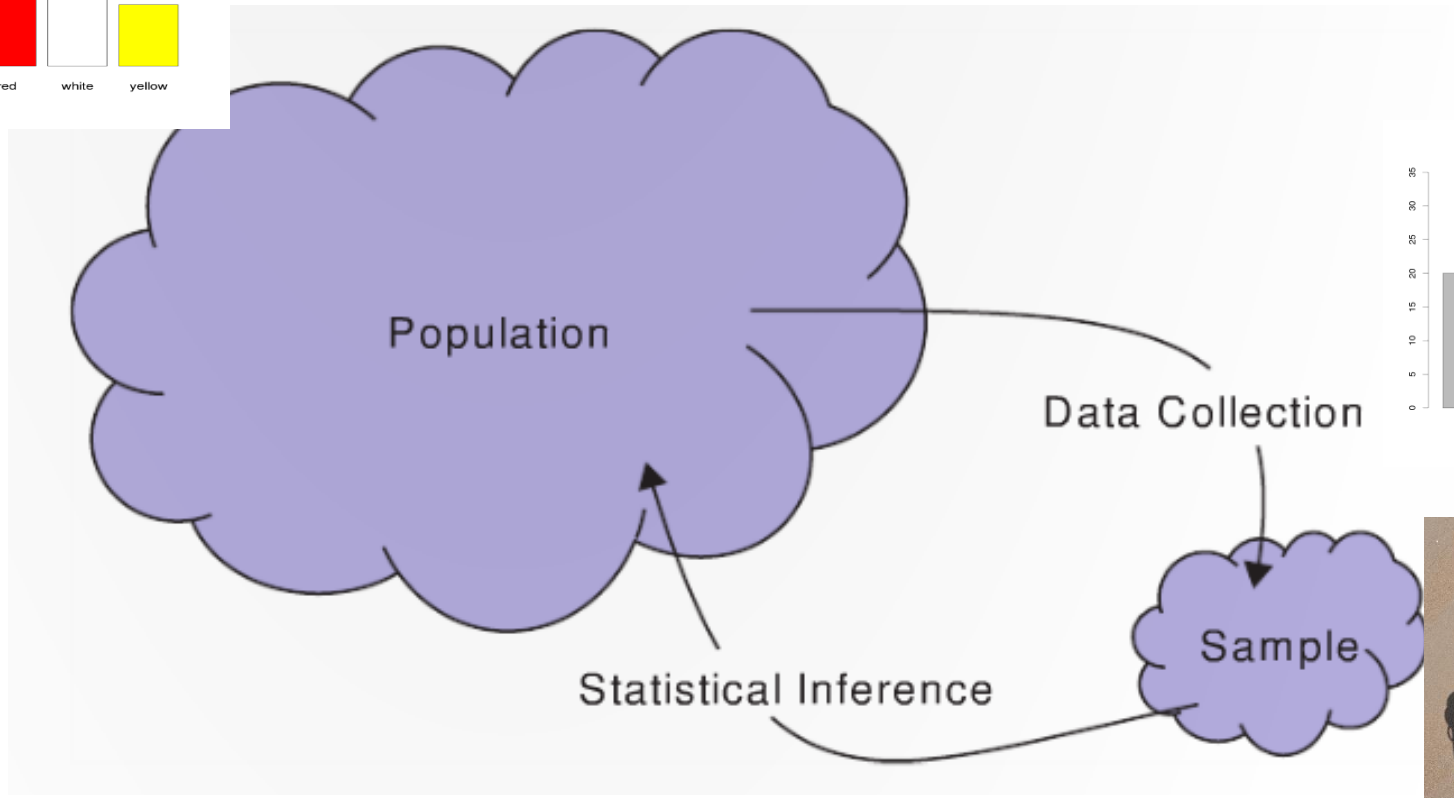
The mean and median

The standard deviation

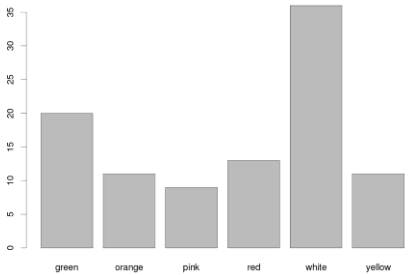
Review: Categorical variables

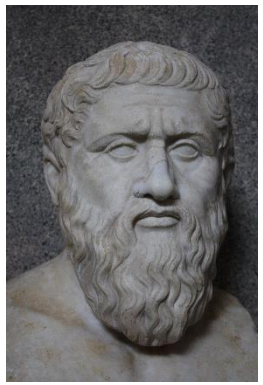


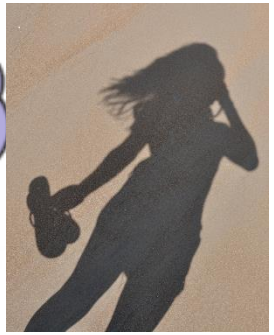
π_{red}



$\hat{p}_{red}$

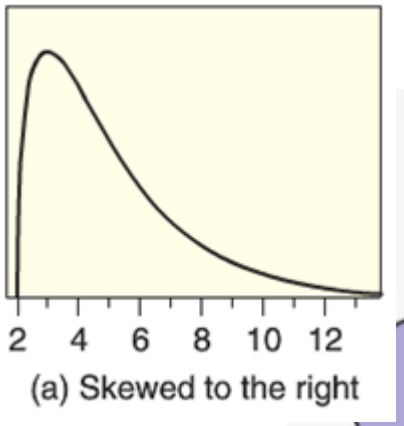




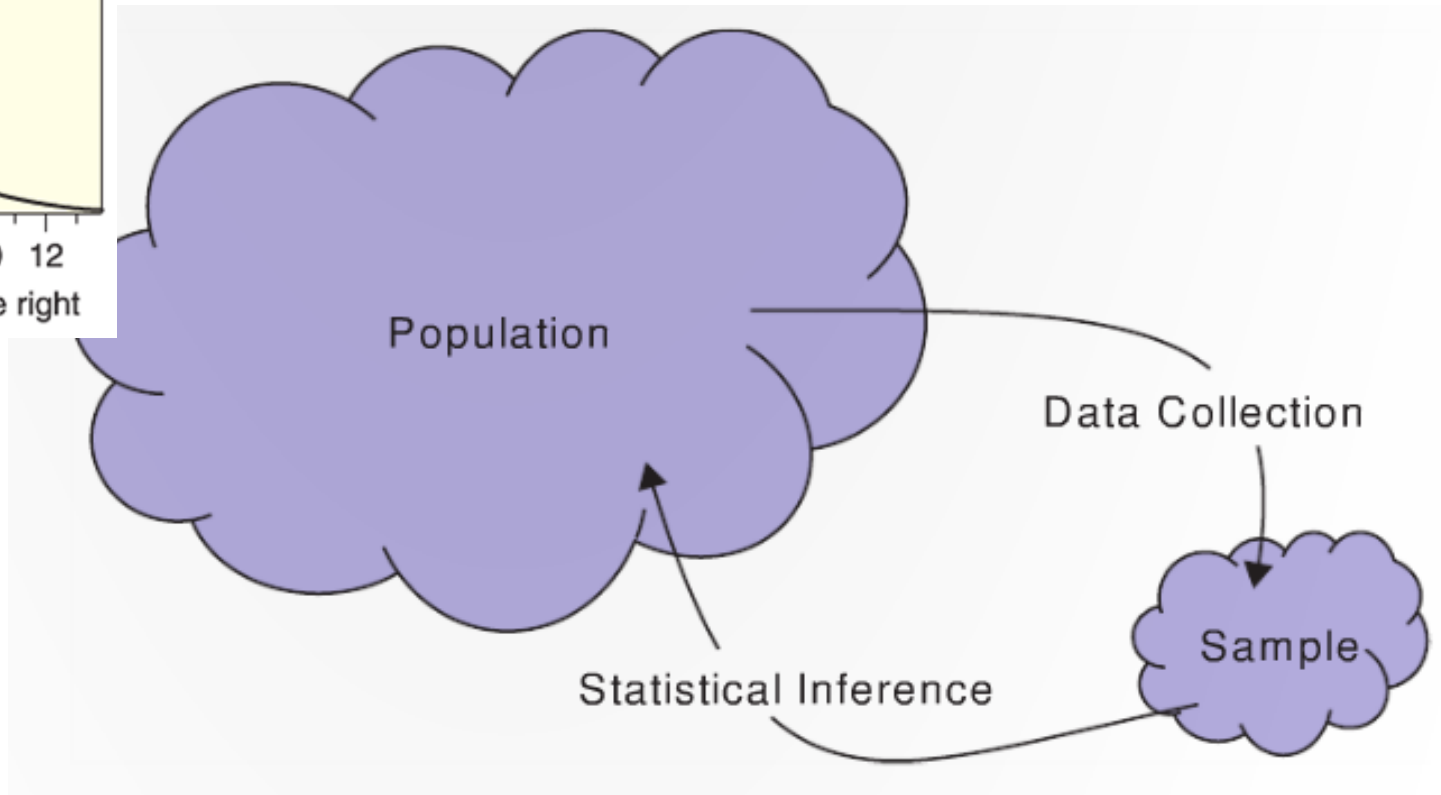


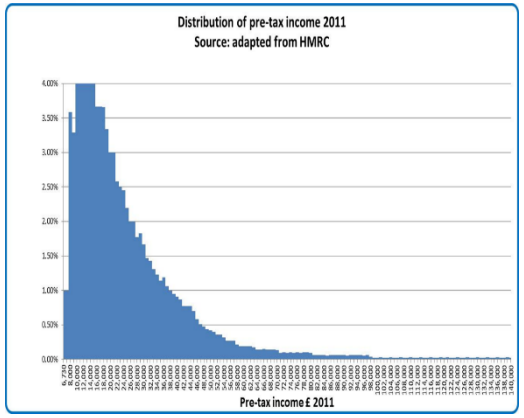
Last class we started talking about…

Quantitative variables: Sample vs. Population means

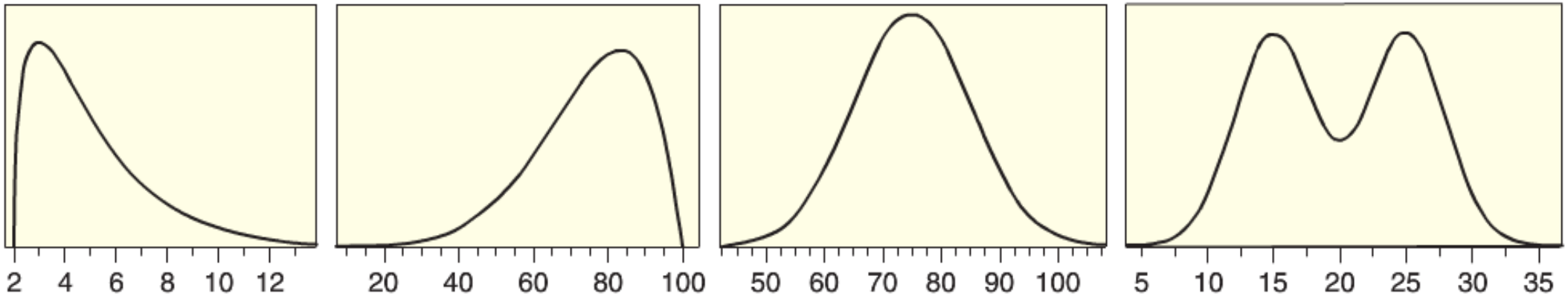


(a) Skewed to the right

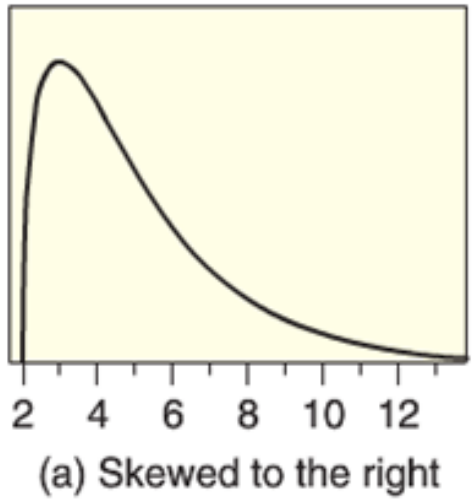




Review: Common shapes for distributions

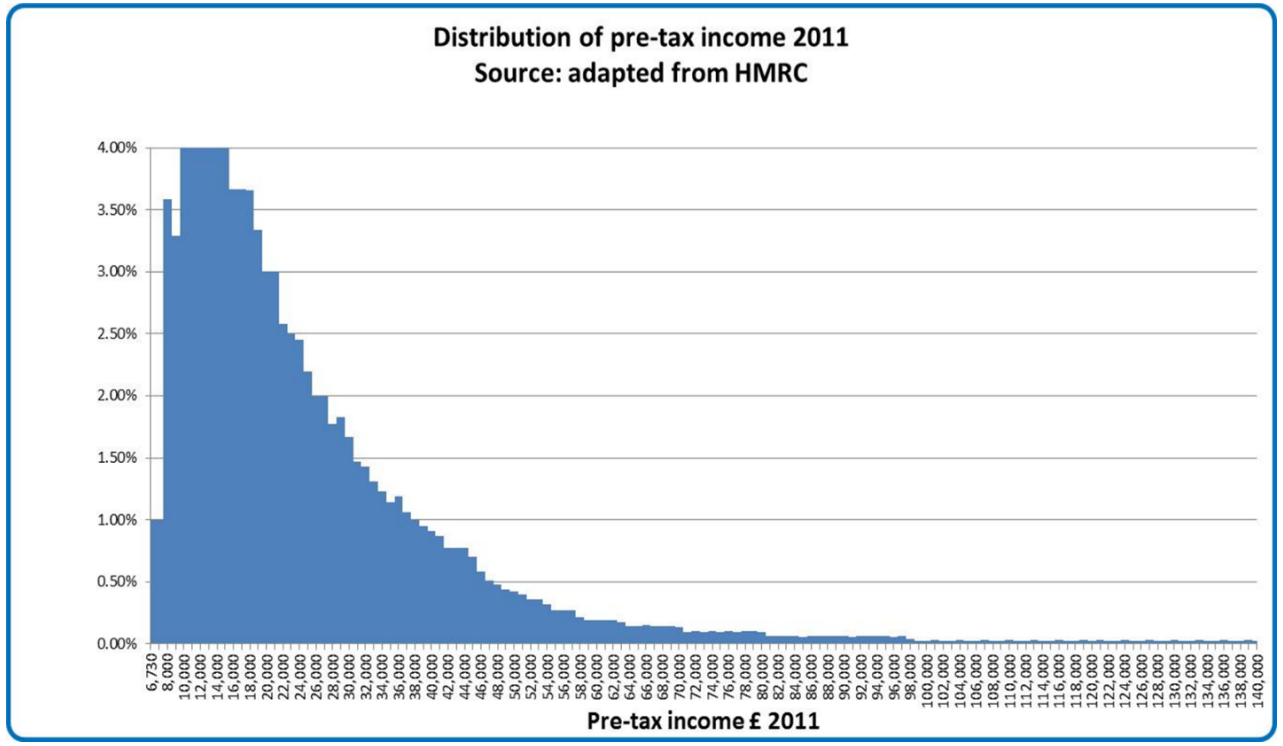


Review: Can you think of a distribution that is right skewed?

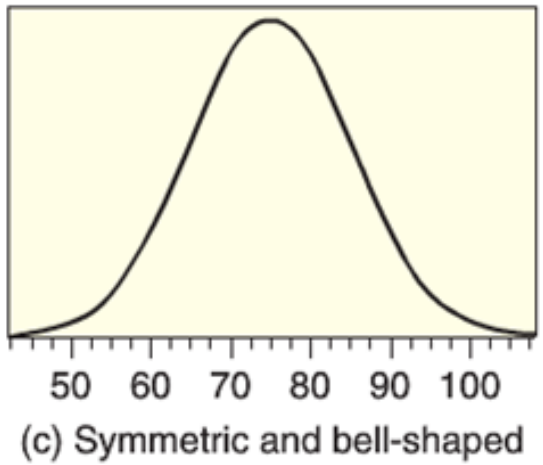


(a) Skewed to the right

Income distribution

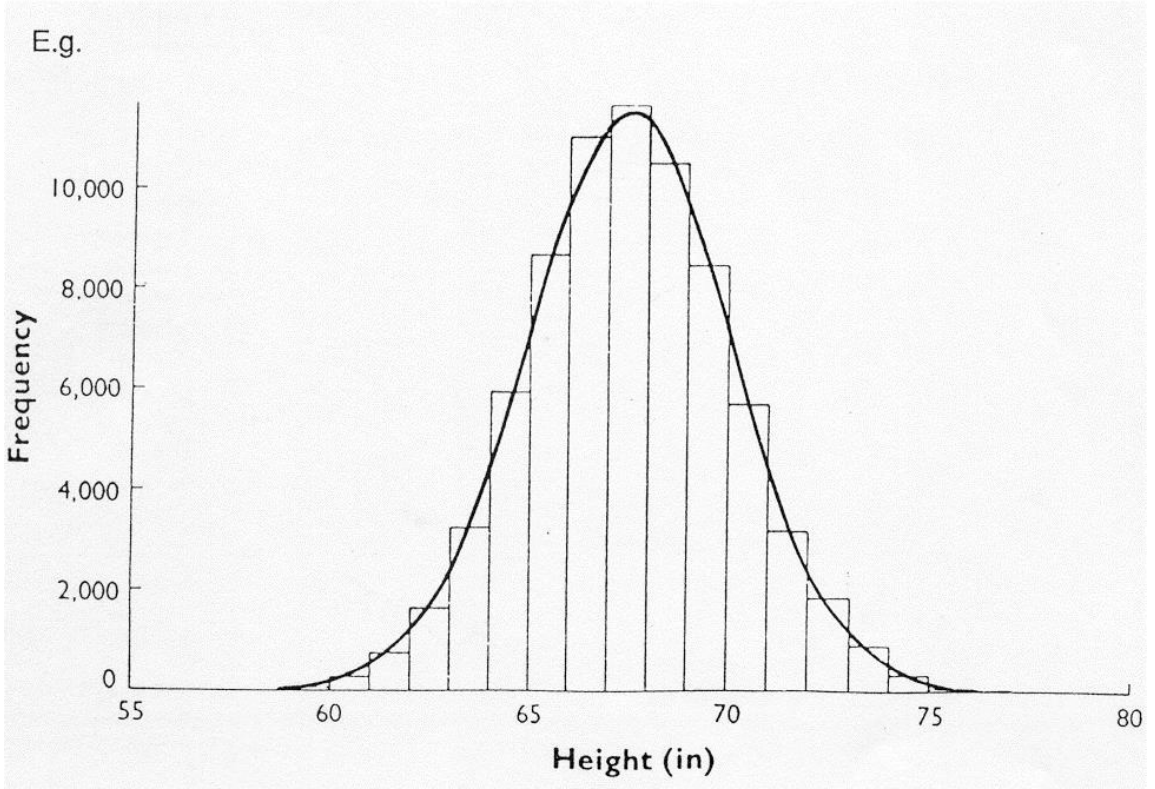


Review: Can you think of a distribution that is symmetric and bell-shaped?

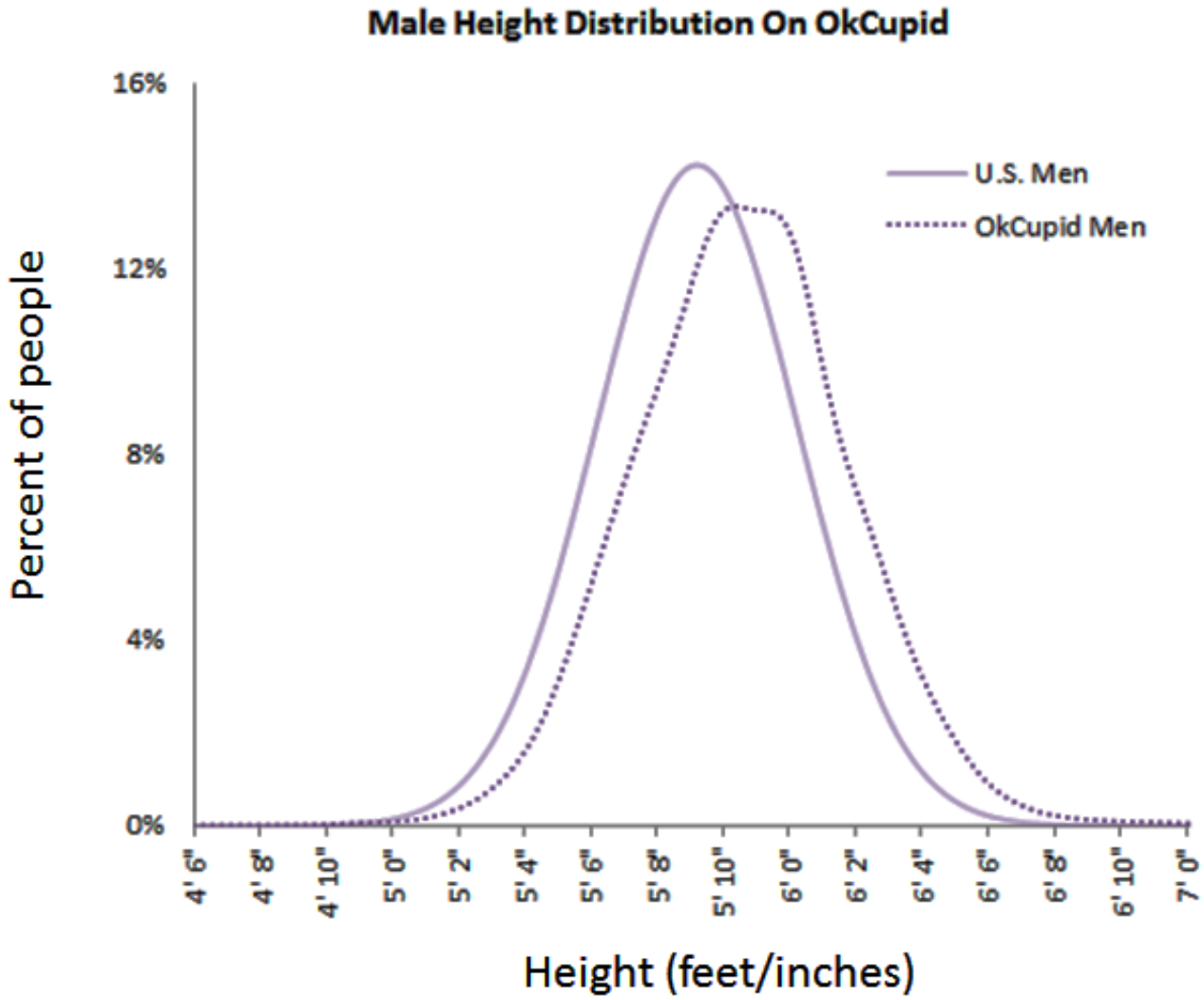


(c) Symmetric and bell-shaped

Young adult male heights (Martin, 1949)

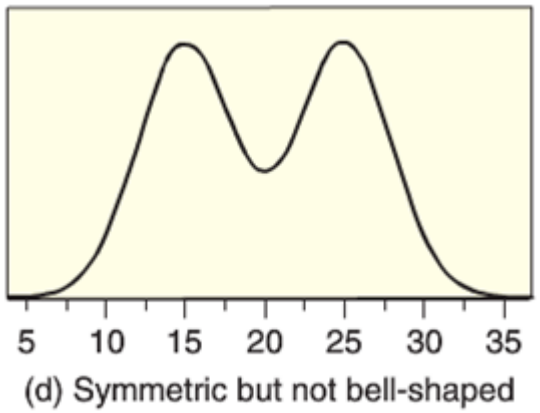


Review: Men on OkCupid are taller!



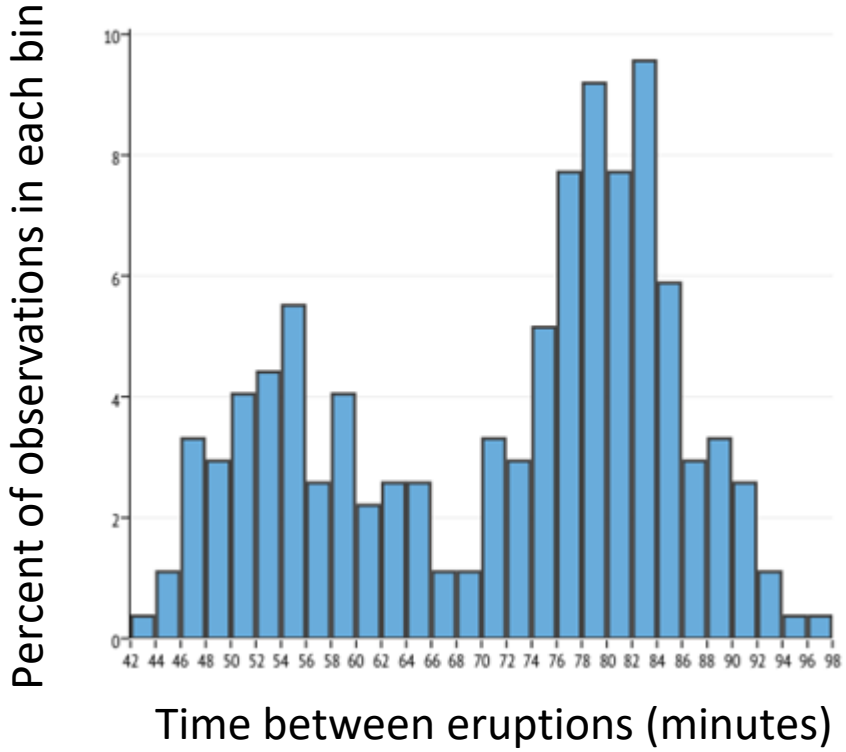
Bias?

Can you think of a distribution that is symmetric but not bell-shaped?

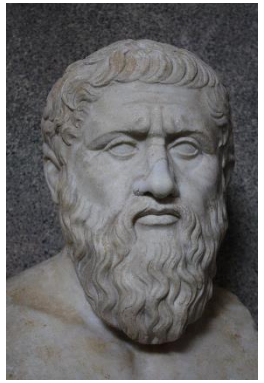


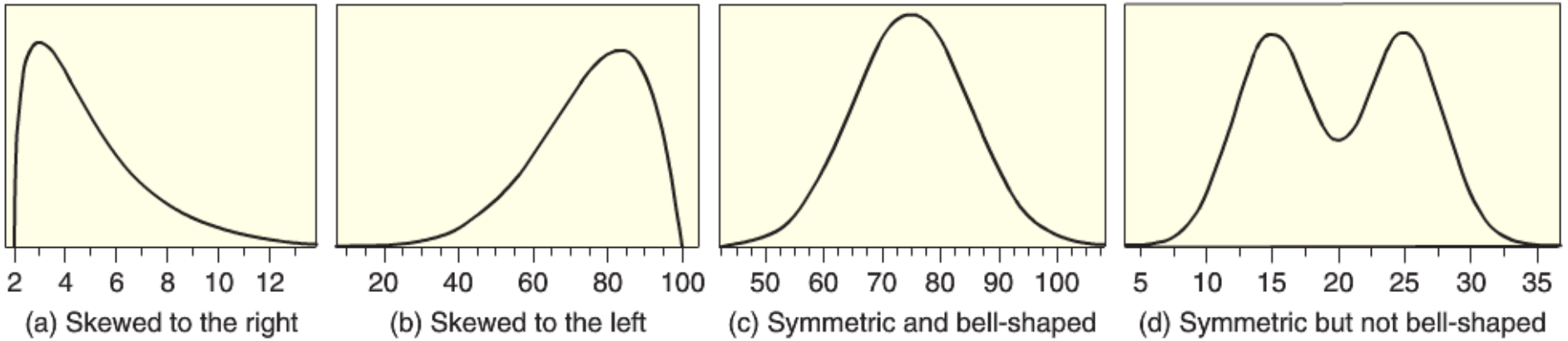
(d) Symmetric but not bell-shaped

Old Faithful eruption times

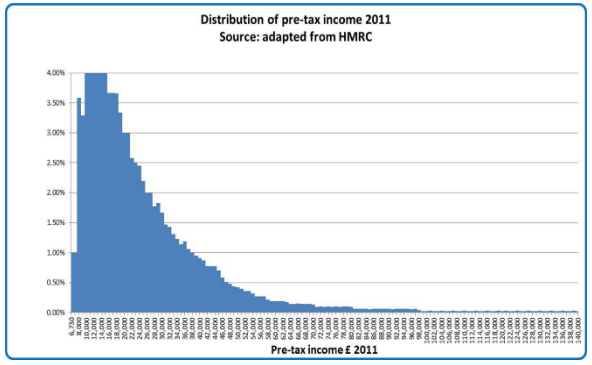


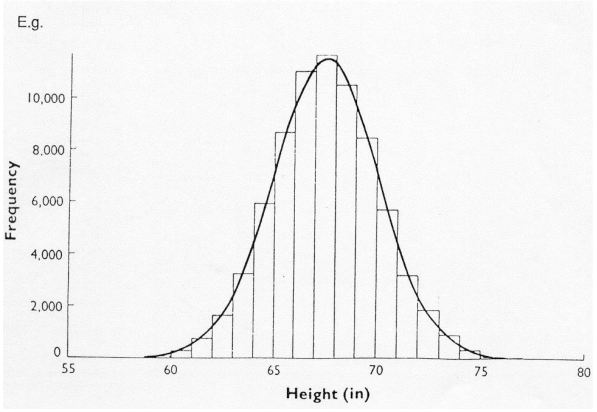
Plato and shadows: distributions and histograms

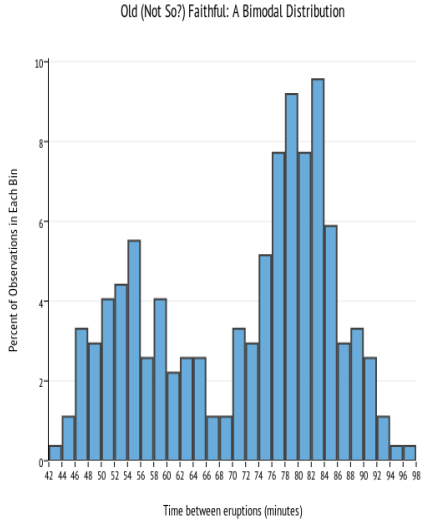




Income distribution

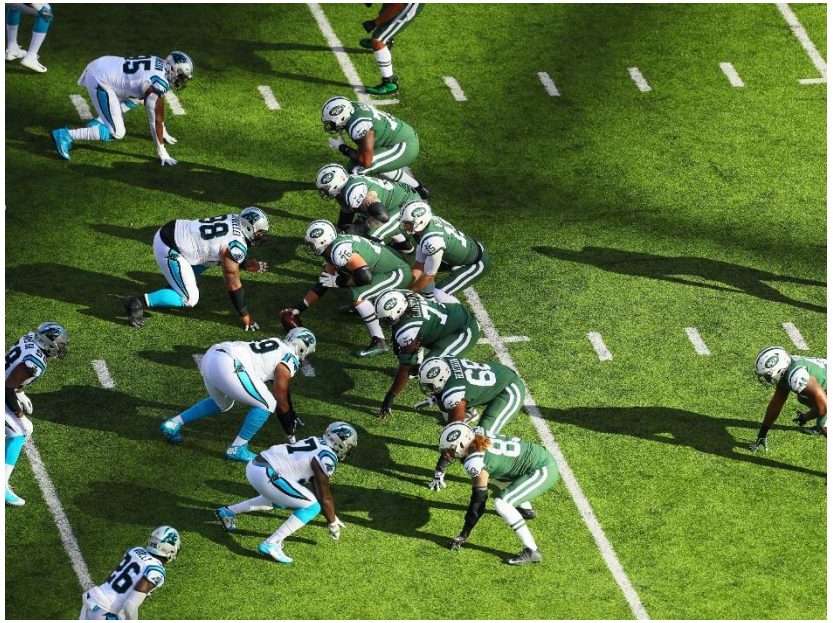


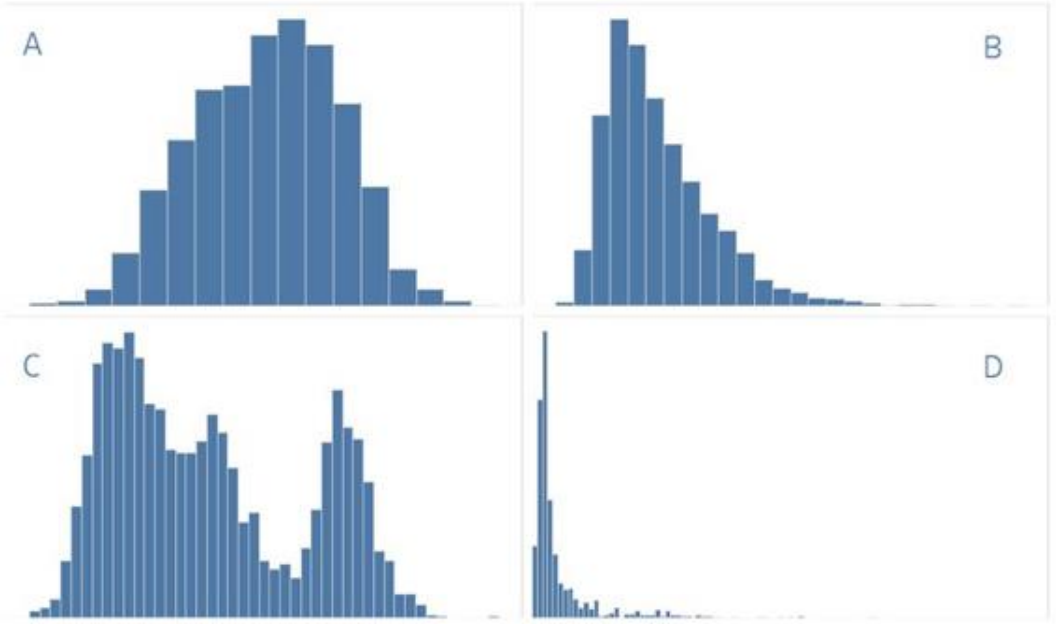




Neat facts - average NFL player is:

- 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





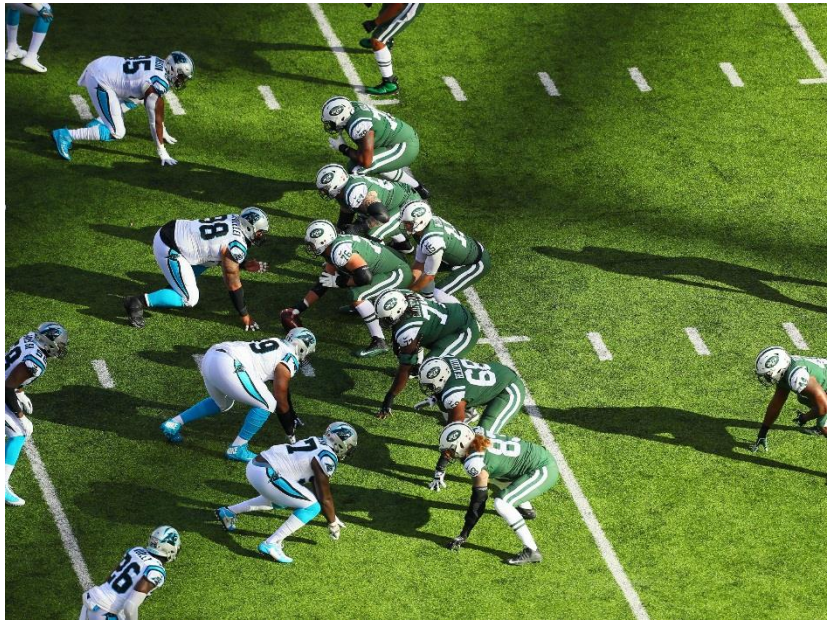
Question: Can you tell which histogram goes with which trait?

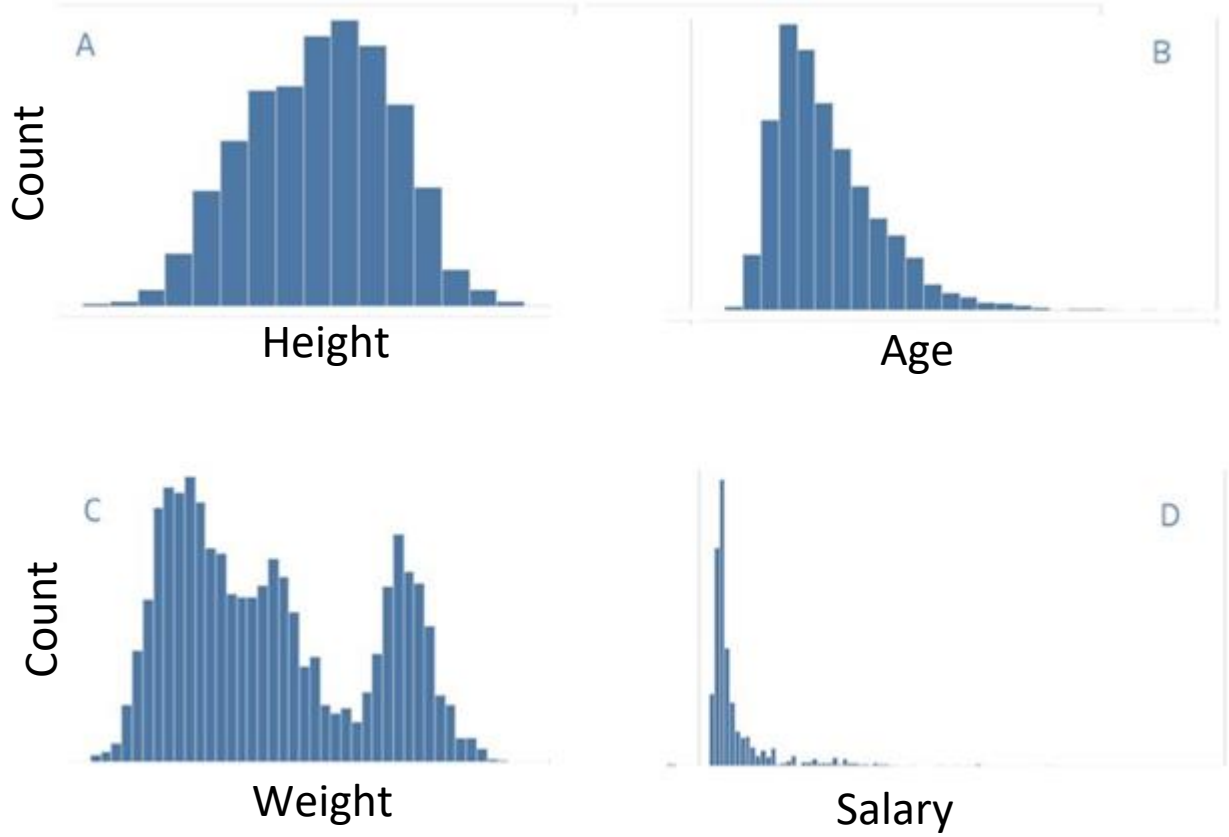
Task is to add the labels: **Age, Height, Weight, and Salary**

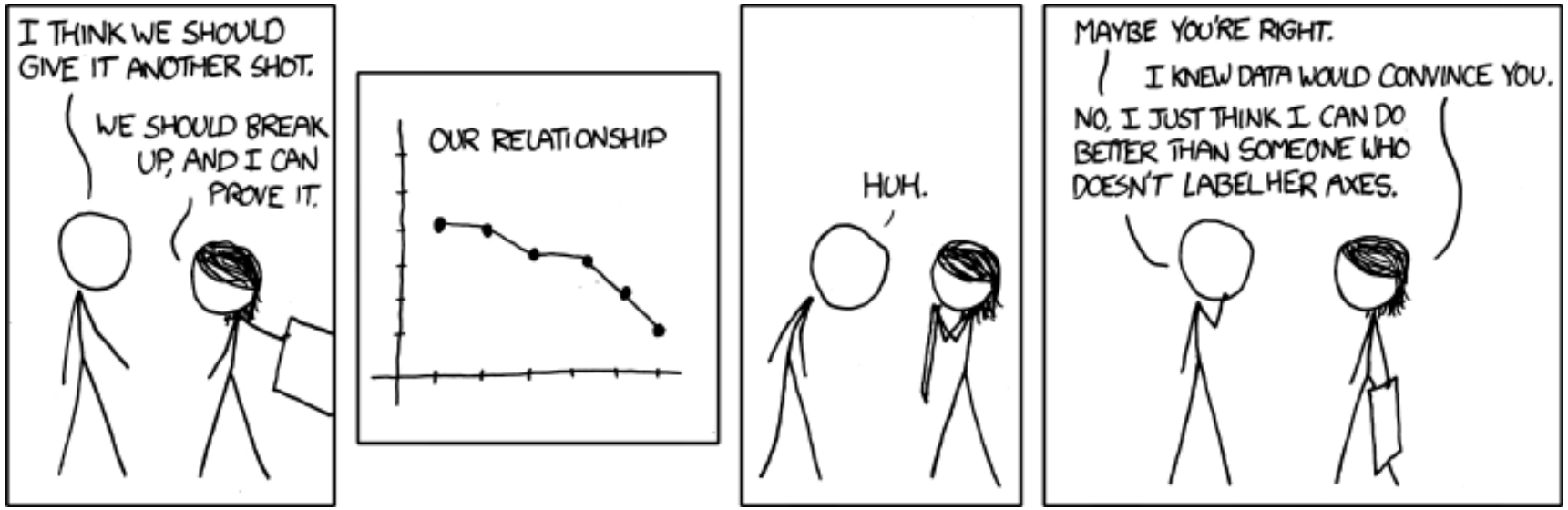
- Hint: There are a wide range of positions in football that have very different roles
  - E.g., placekickers only play for small factions of the game, while quarterbacks are essentially to a team's success

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

- A: Frequency or count







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

Back to the Gapminder data…

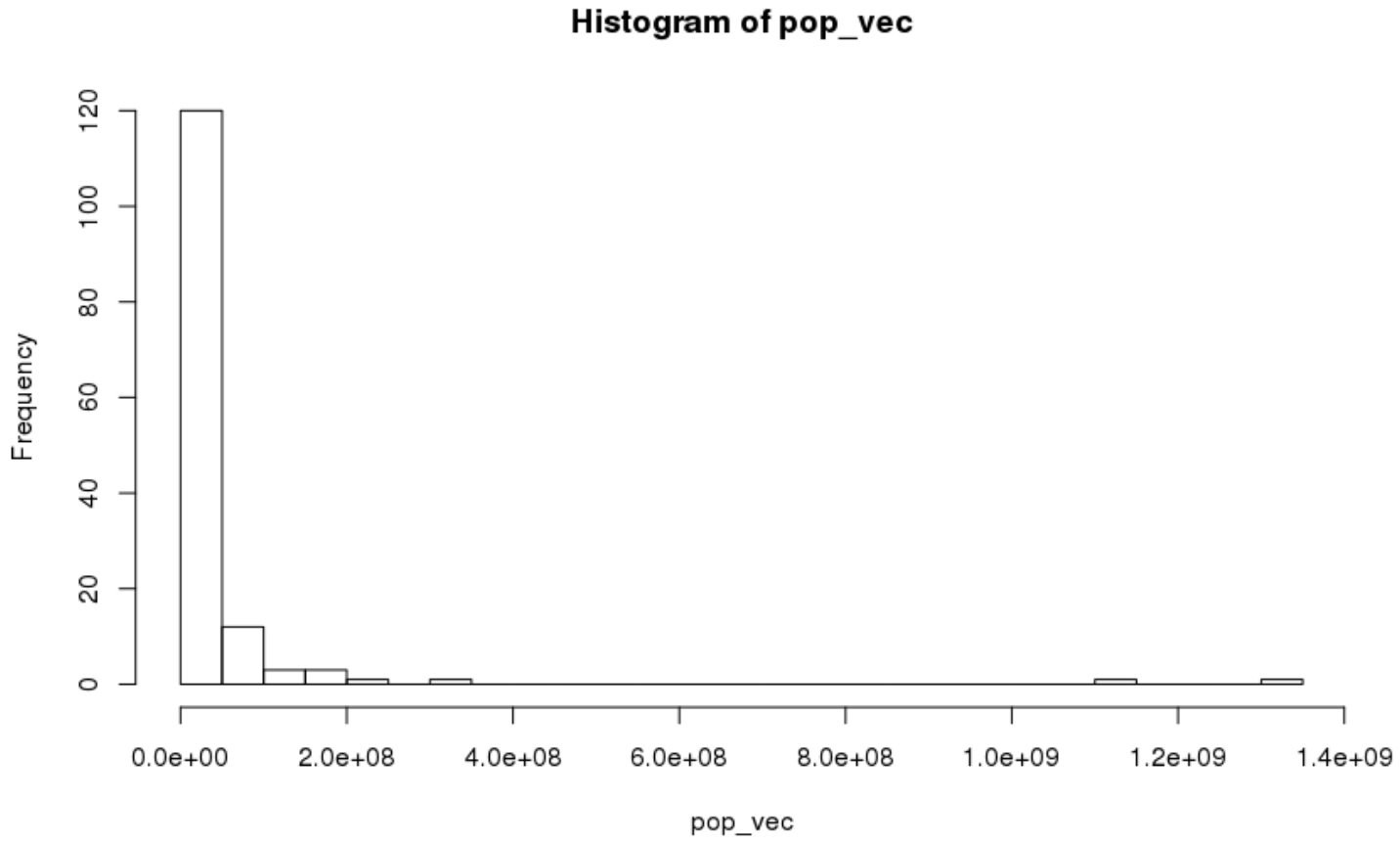
```
# get a data frame with information about the countries in the world
> download_class_data("gapminder_2007.Rda")   # only need to run this once
>  load("gapminder_2007.Rda")
```

	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?

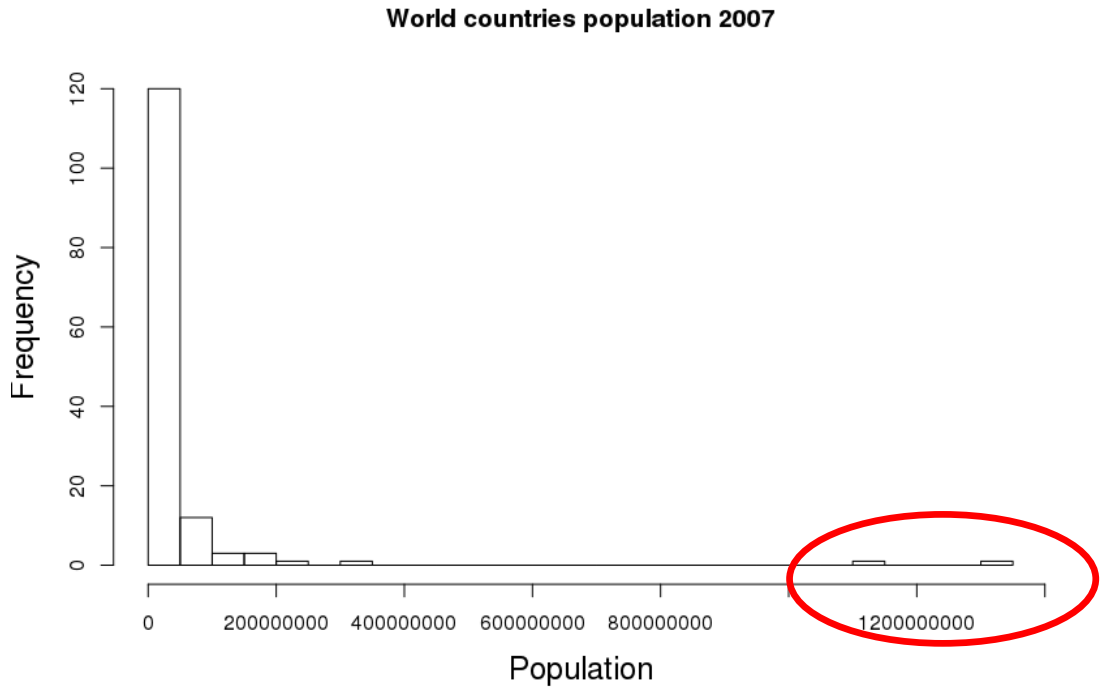
- Answer: xlab and ylab!

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

Outliers

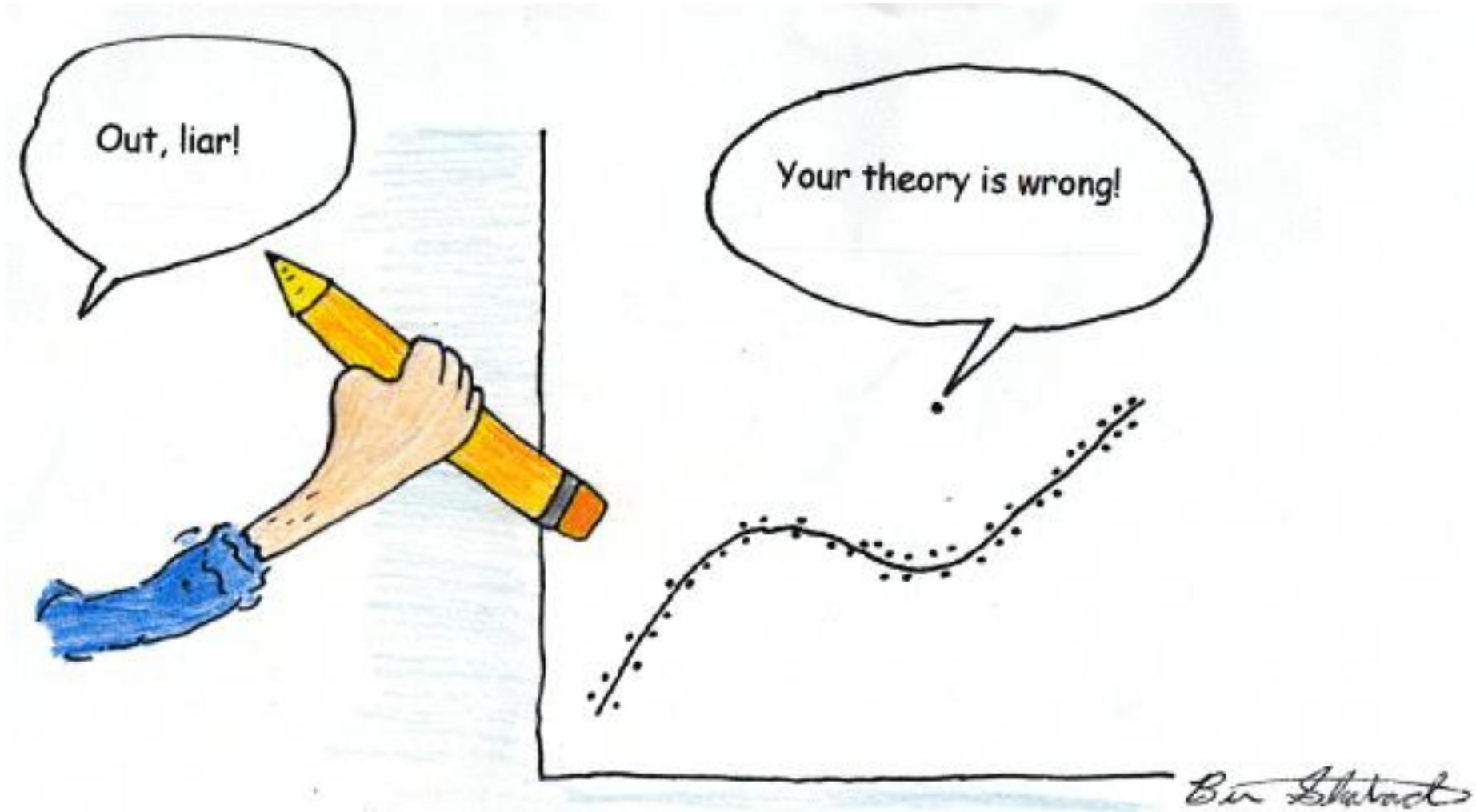
Question: What is an outlier?

A: An **outlier** is an observed value that is notably distinct from the other values



Question: what should you do if there is an outlier in your data?

- 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

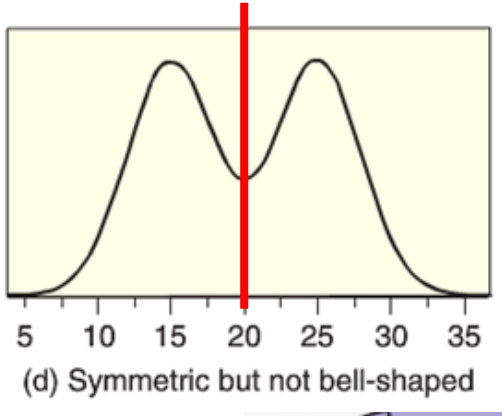
Question: what is a numerical summary of a sample of data called?

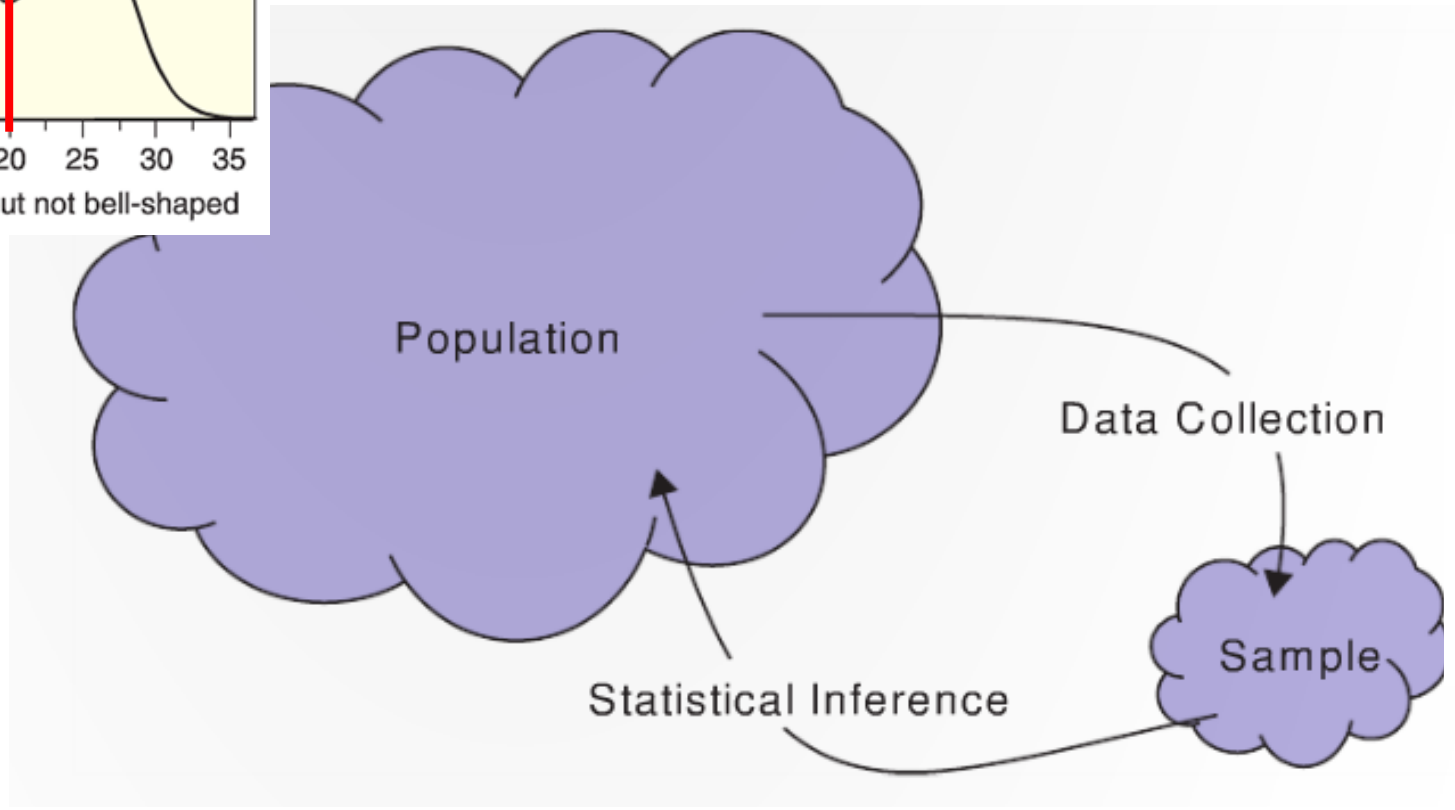
A: a statistic!

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

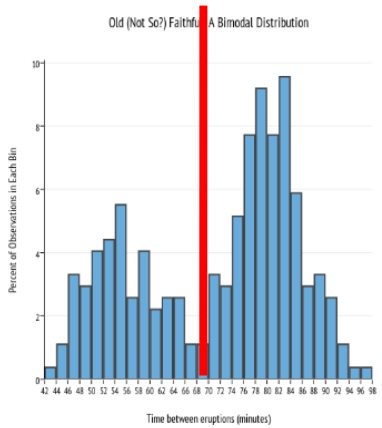
Sample and population mean

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. $\bar{x}$?

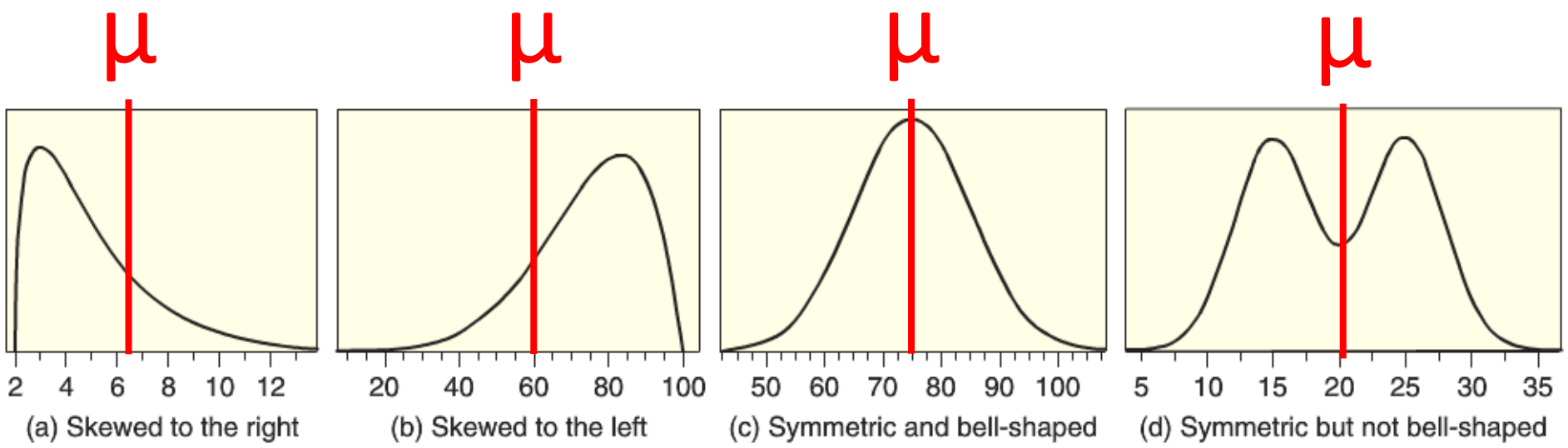
We measure the height of 50 randomly chosen Yale students

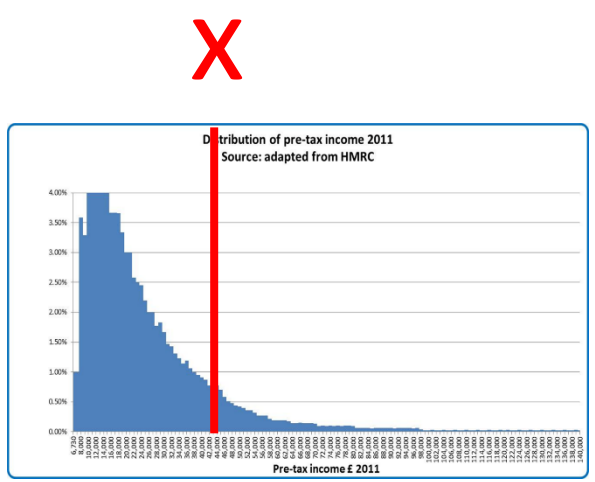
We measure the height of all Yale students

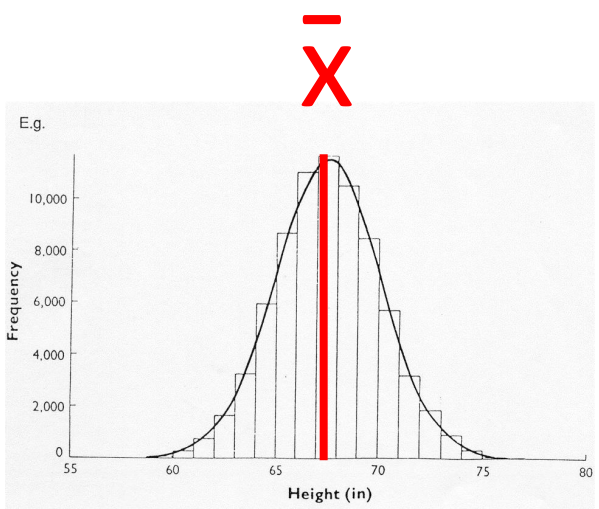
Can you calculate the mean of the countries life expectancy in R?

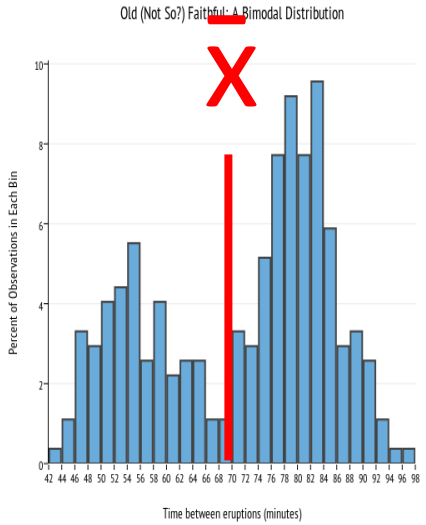
```
> life_expectancy <-  gapminder_2007$lifeExp
> mean(life_expectancy)
```

Means for differently shaped distributions









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 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
```

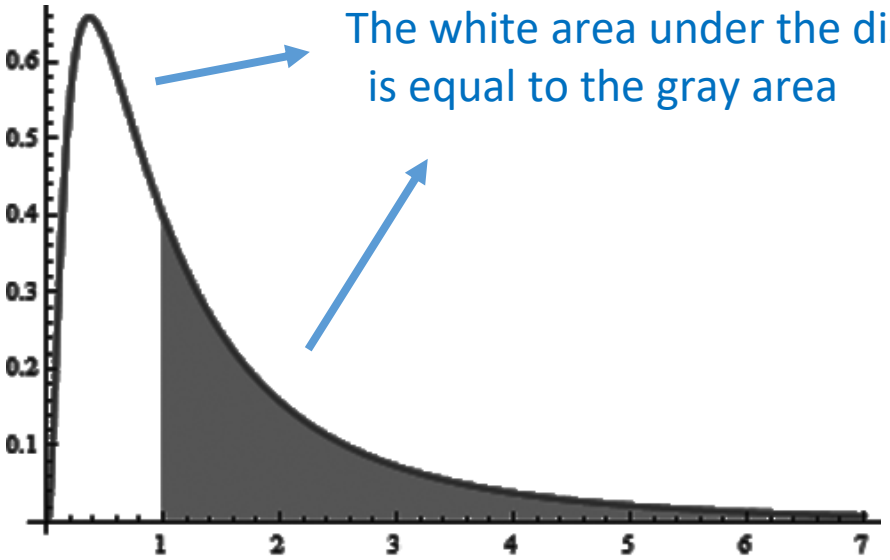
Question: What is the mean and median ping count in this sample?

A: mean = 42

median = 12

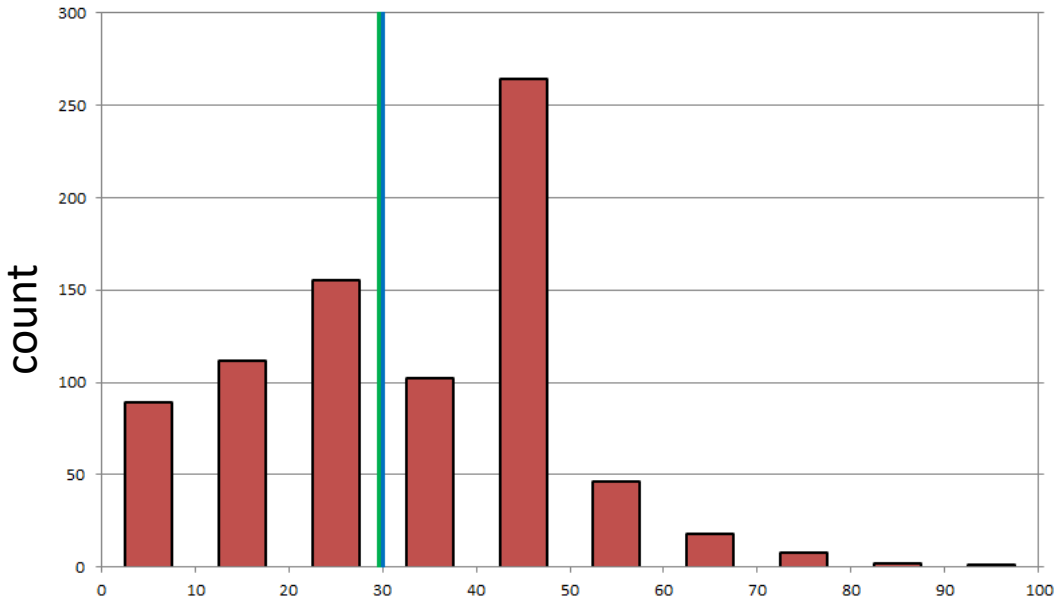
Mean = $\frac{\sum_i^n x_i}{n}$

The median



The white area under the distribution is equal to the gray 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).

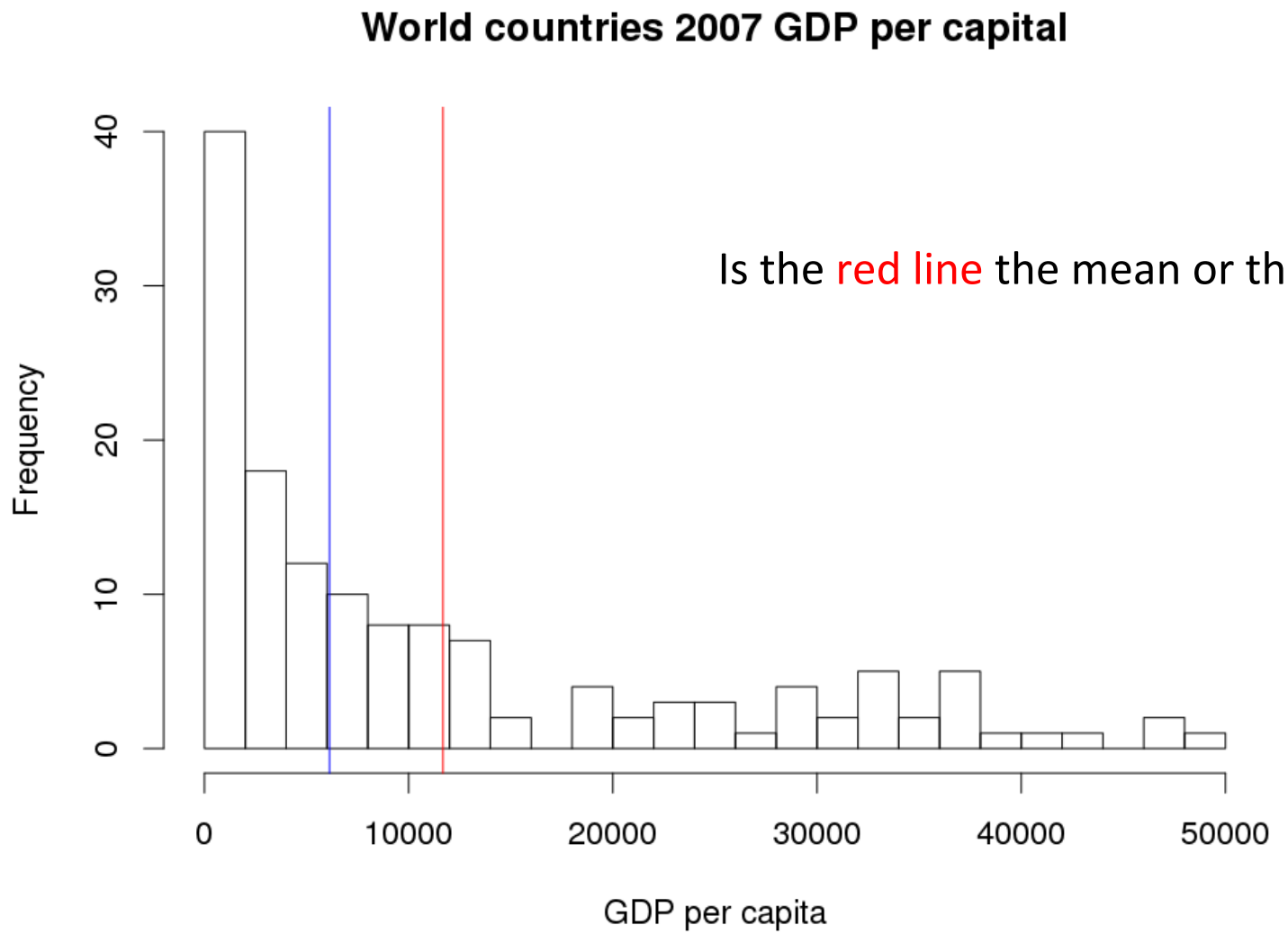
The median is resistant when the mean is not

Example:

Mean US salary = $72,641

Median US salary = $51,939

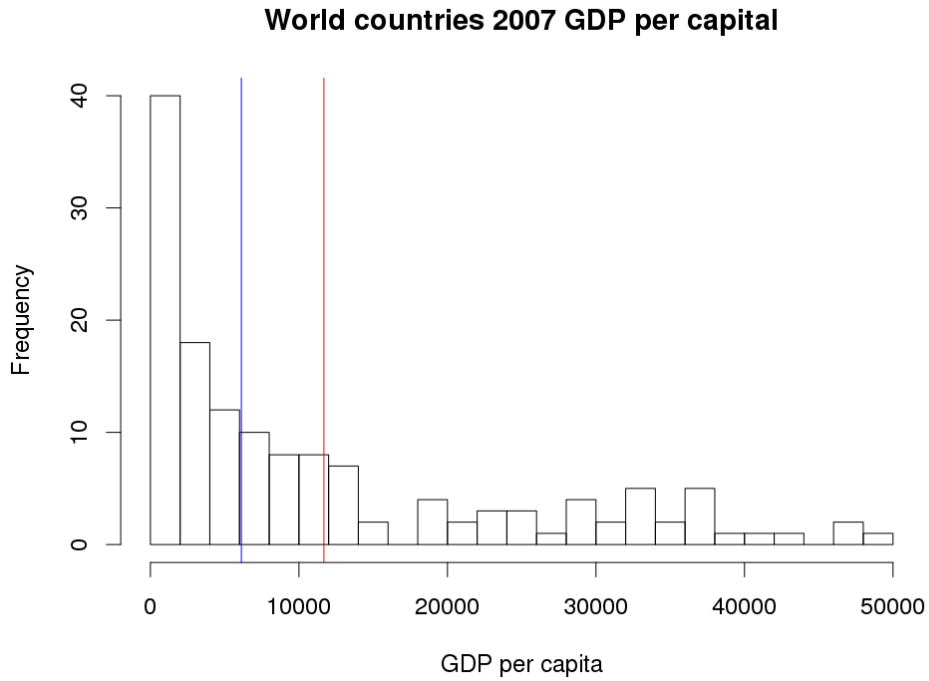
Measure of central tendency: mean and median



Is the red line the mean or the median?

Characterizing the spread

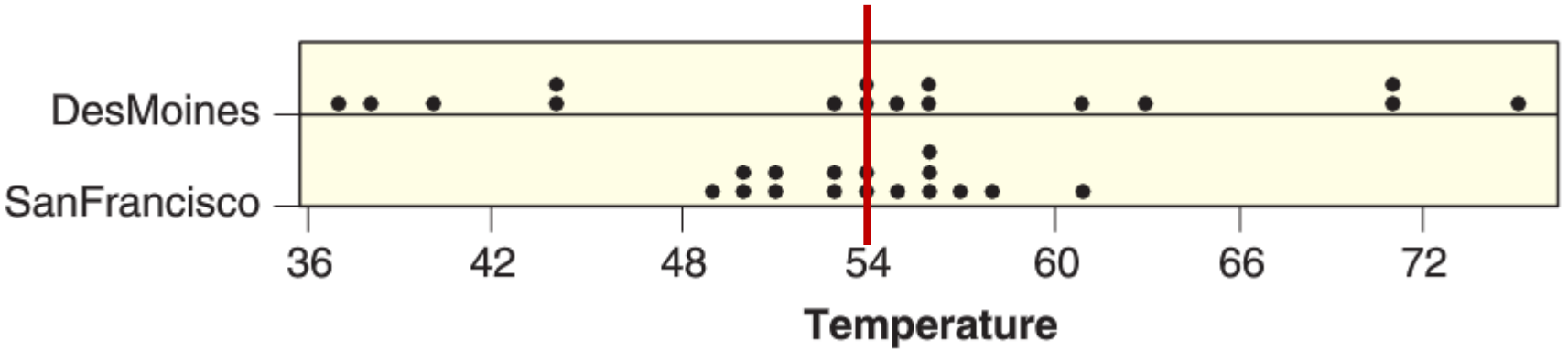
The mean and median are numbers that tell us about the center of a distribution



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:

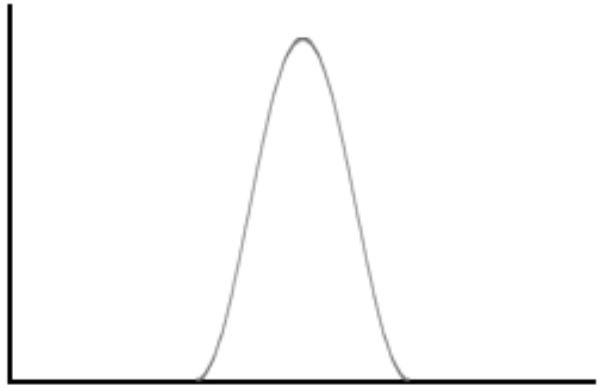


Mean temperature (°F): Des Moines = 54.49 San Fran = 54.01

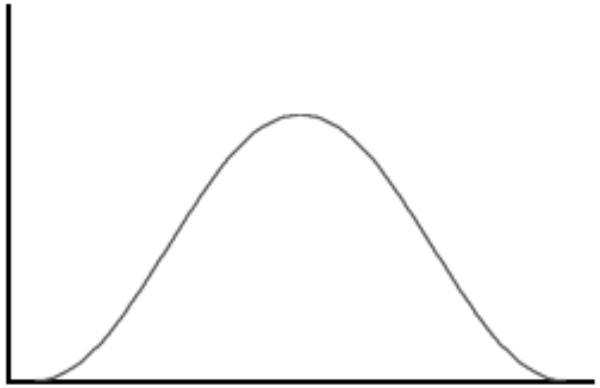
The standard deviation

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

Smaller standard deviation



Larger standard deviation

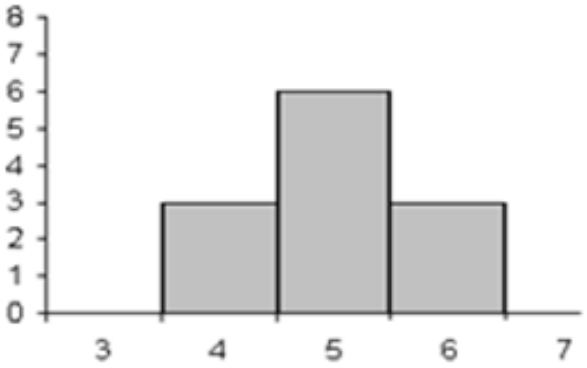


It gives a rough estimate for a typical distance a point is from the center

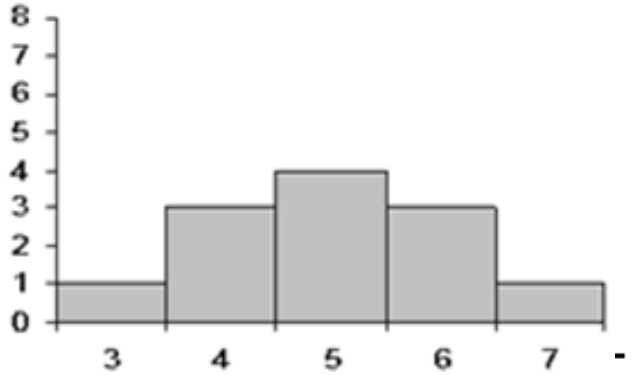
The standard deviation

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

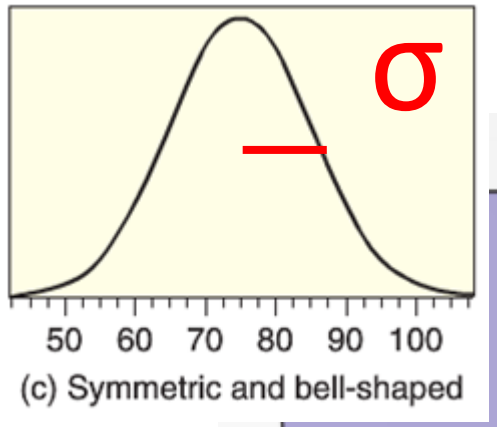
The standard deviation of the ***population*** is denoted **σ**

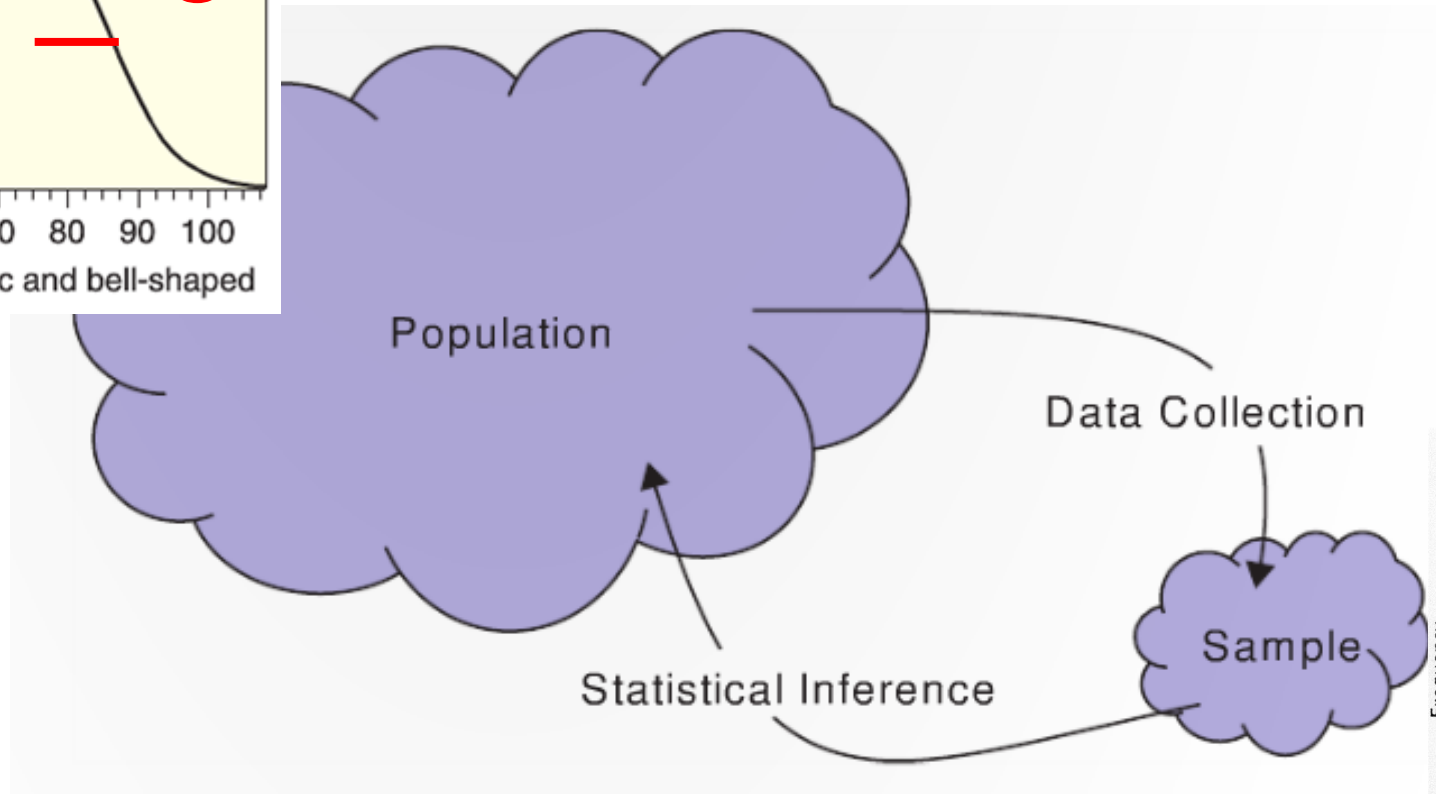
- It measure the spread of the data from the population mean

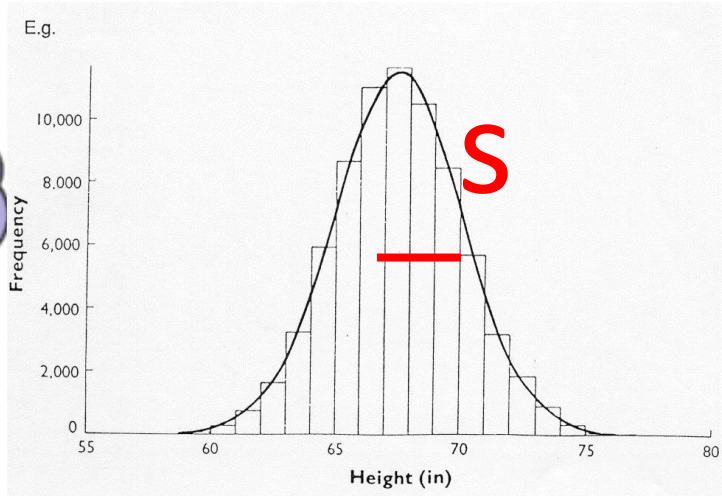
The standard deviation of a ***sample*** is denoted **s**

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

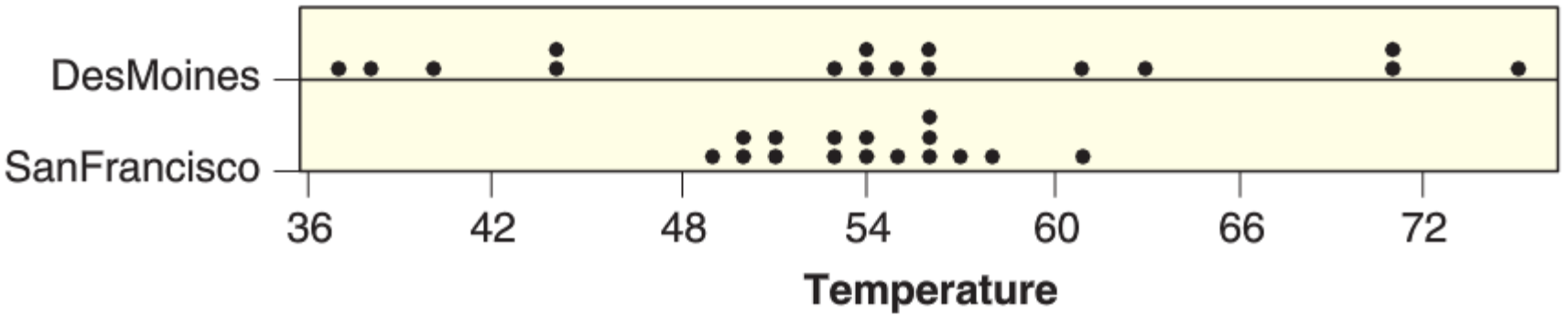
Population and sample standard deviation







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

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

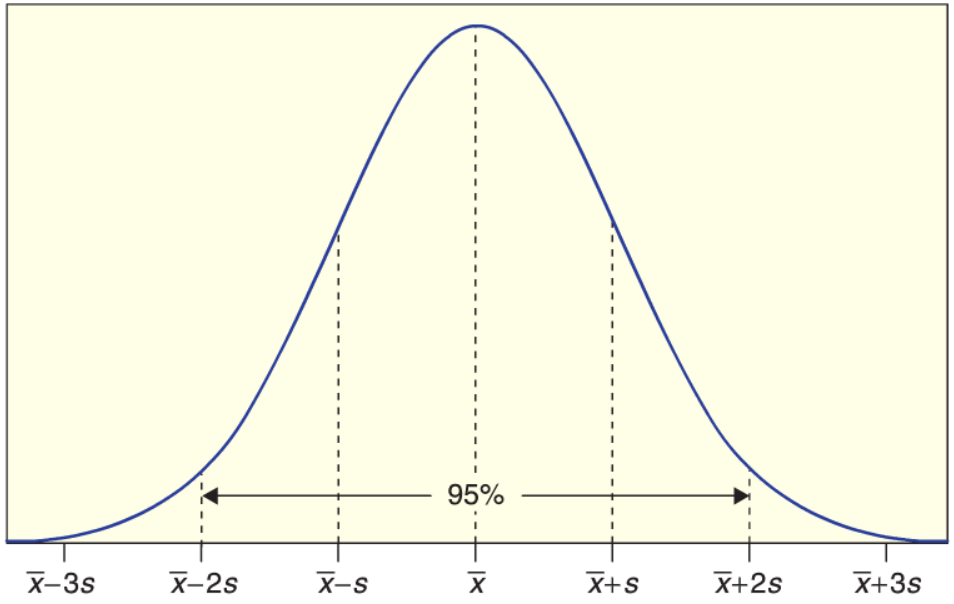
Homework part 2: Calculate the mean and standard deviation for the number of hot dogs eaten by the winners. Due on Sunday at 11:30pm on Gradescope.

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

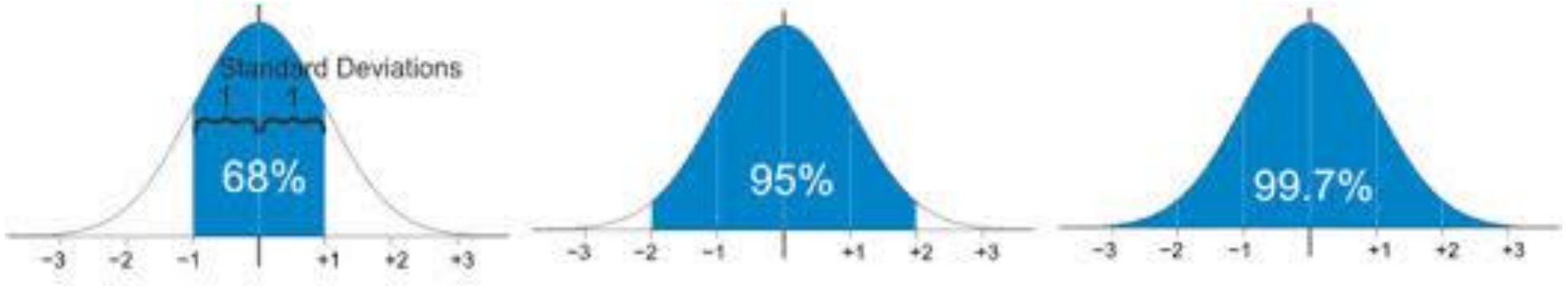
If a distribution of data is approximately normally distributed, about 95% of the data should fall within two standard deviations of the mean

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?

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



Homework 1

Homework 1 is due at 11:30pm on Sunday January 26th

Use Piazza 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. Hand in the standard deviation calculation under Homework 1 standard deviation deviation

Overall should be relatively short and hopefully not too hard