

Quantitative data: shape, measures of central tendency, and spread



# Overview

Very quick review of categorical data analysis

Quantitative data

- Graphing the shape: histograms and outliers

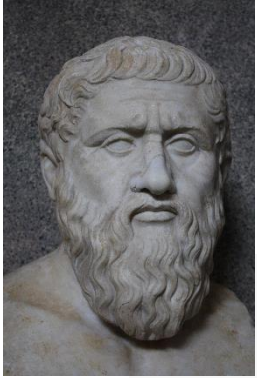
- Measures of the central tendency: mean and median

If there is time

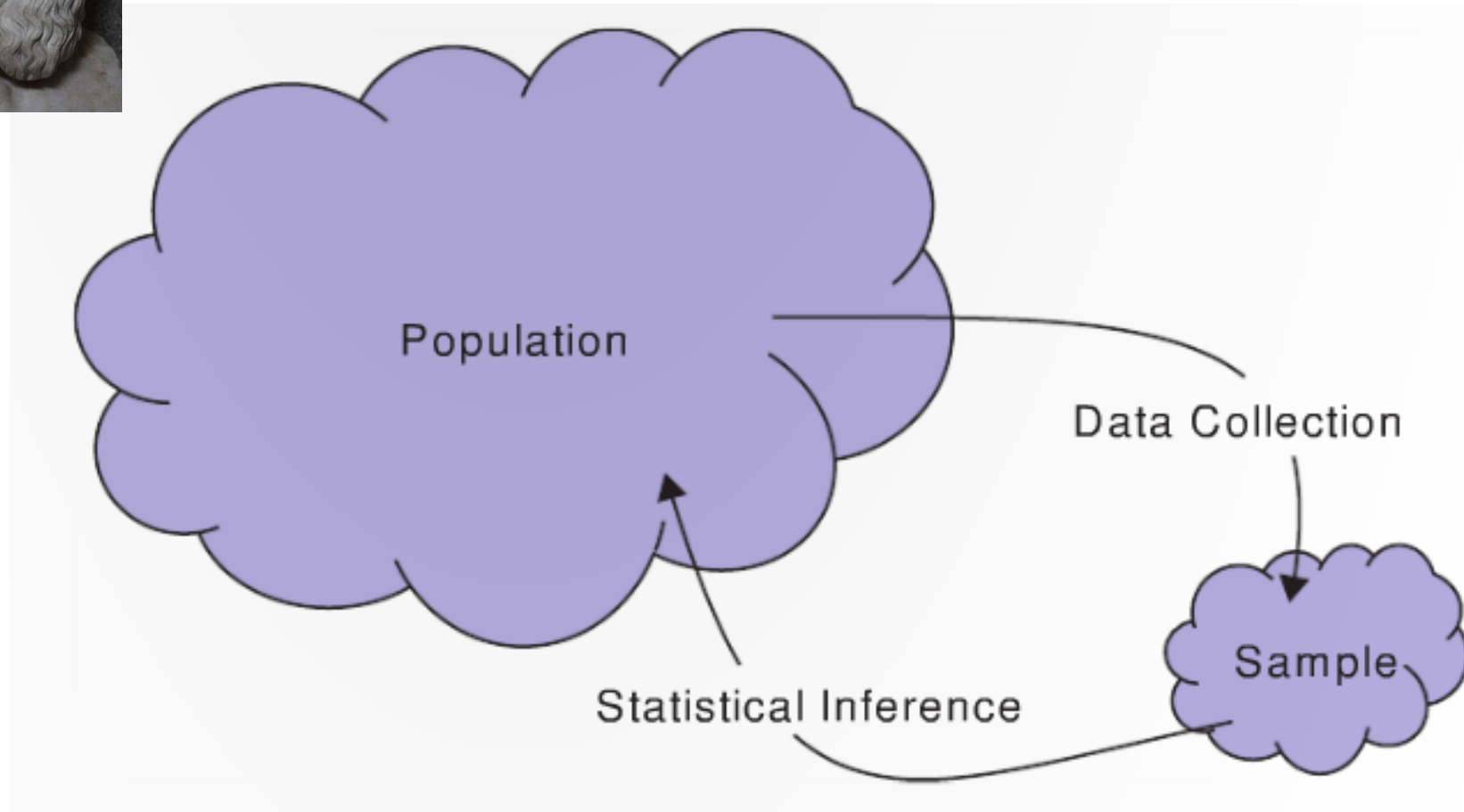
- Measure of spread: The variance and standard deviation

# Review

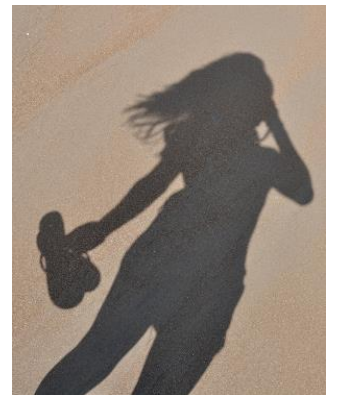
## Categorical variables



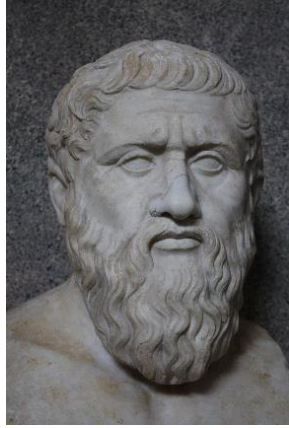
parameter:  $\pi$



statistic:  $\hat{p}$



# 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)

# Gapminder data

	country	continent	year	lifeExp
1	Afghanistan	Asia	2007	43.828
2	Albania	Europe	2007	76.423
3	Algeria	Africa	2007	72.301
4	Angola	Africa	2007	42.731
5	Argentina	Americas	2007	75.320

**Data frames** are the way R represents structured data

Data frames can be thought of as collections of related vectors

- Each vector corresponds to a variable (column) in the structured data

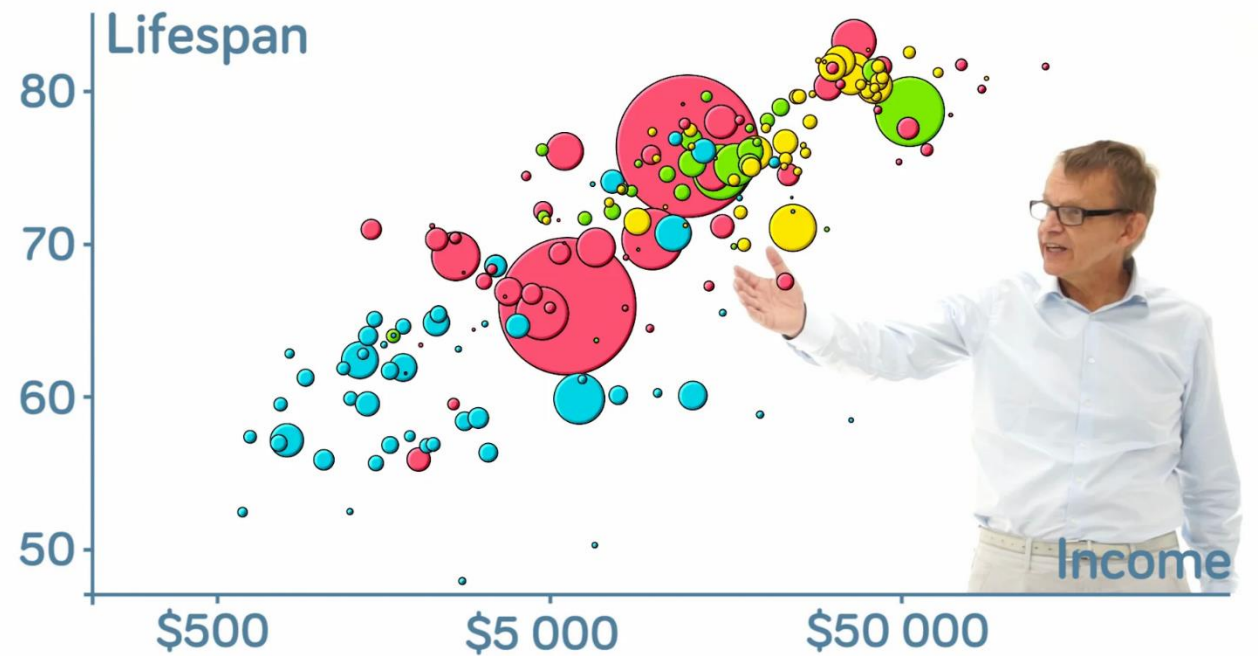
The [gapminder](#) data, was create by Hans Rosling's and contains information about countries in the world.

# loading the gapminder data for information about countries in 2007

```
load("gapminder_2007.rda")
```

```
View(gapminder_2007)
```

Let's do one more quick practice of analyzing categorical data in R...



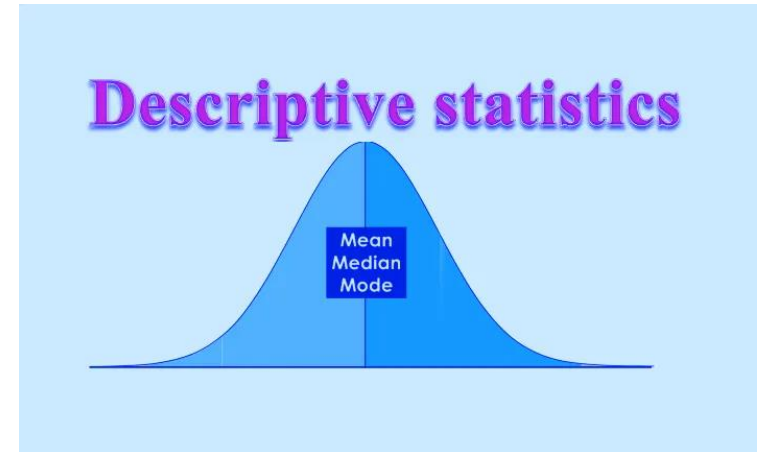
# Quantitative variables



# Descriptive statistics for one quantitative variable

We will be looking at:

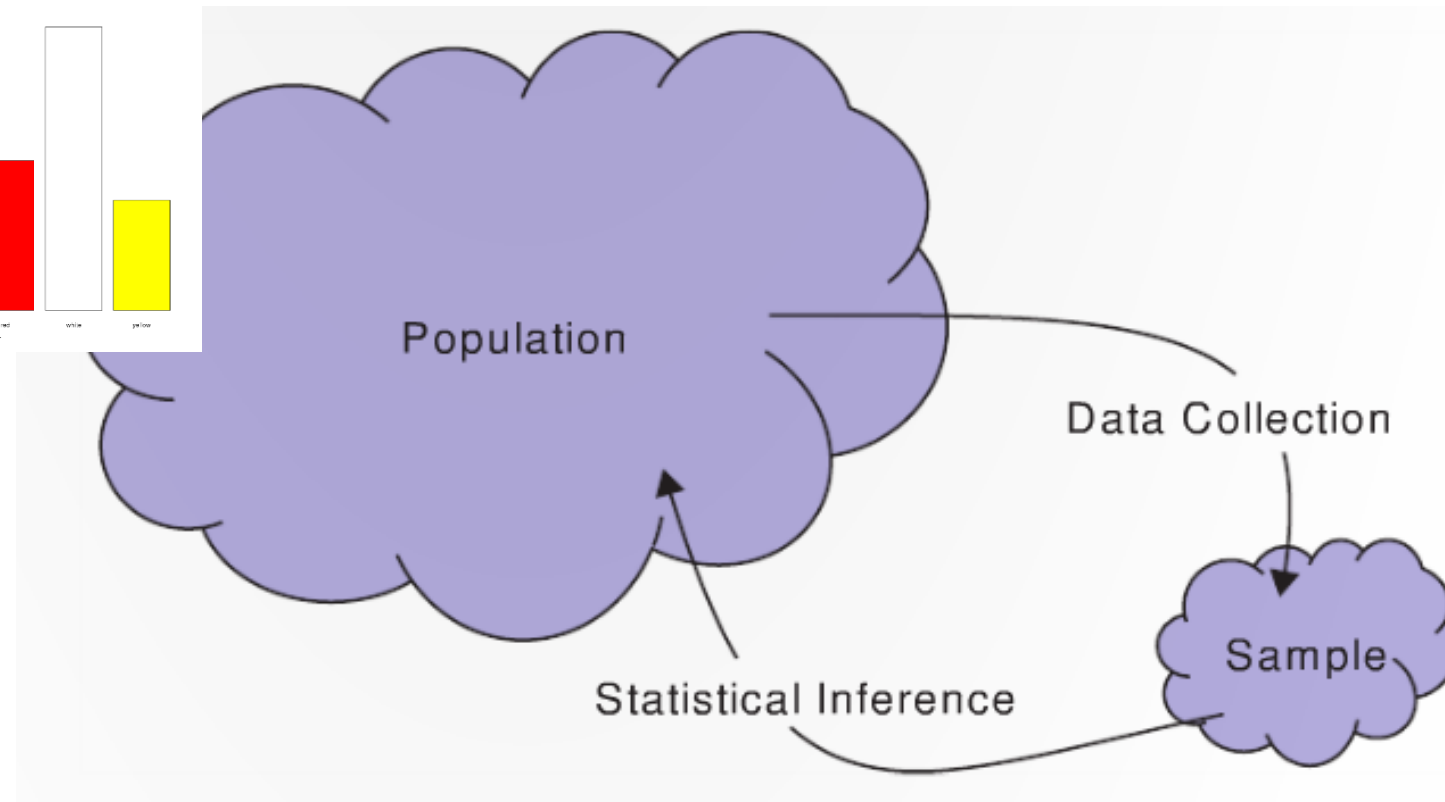
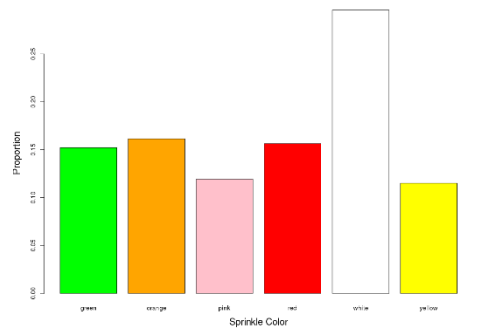
- What is the general 'shape' of the data
- Where are the values centered
- How do the data vary



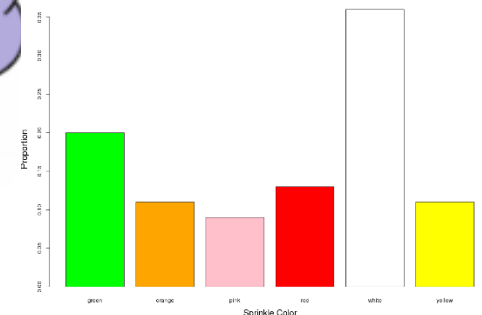
There are all properties of how the data is ***distributed***

# For categorical data we had...

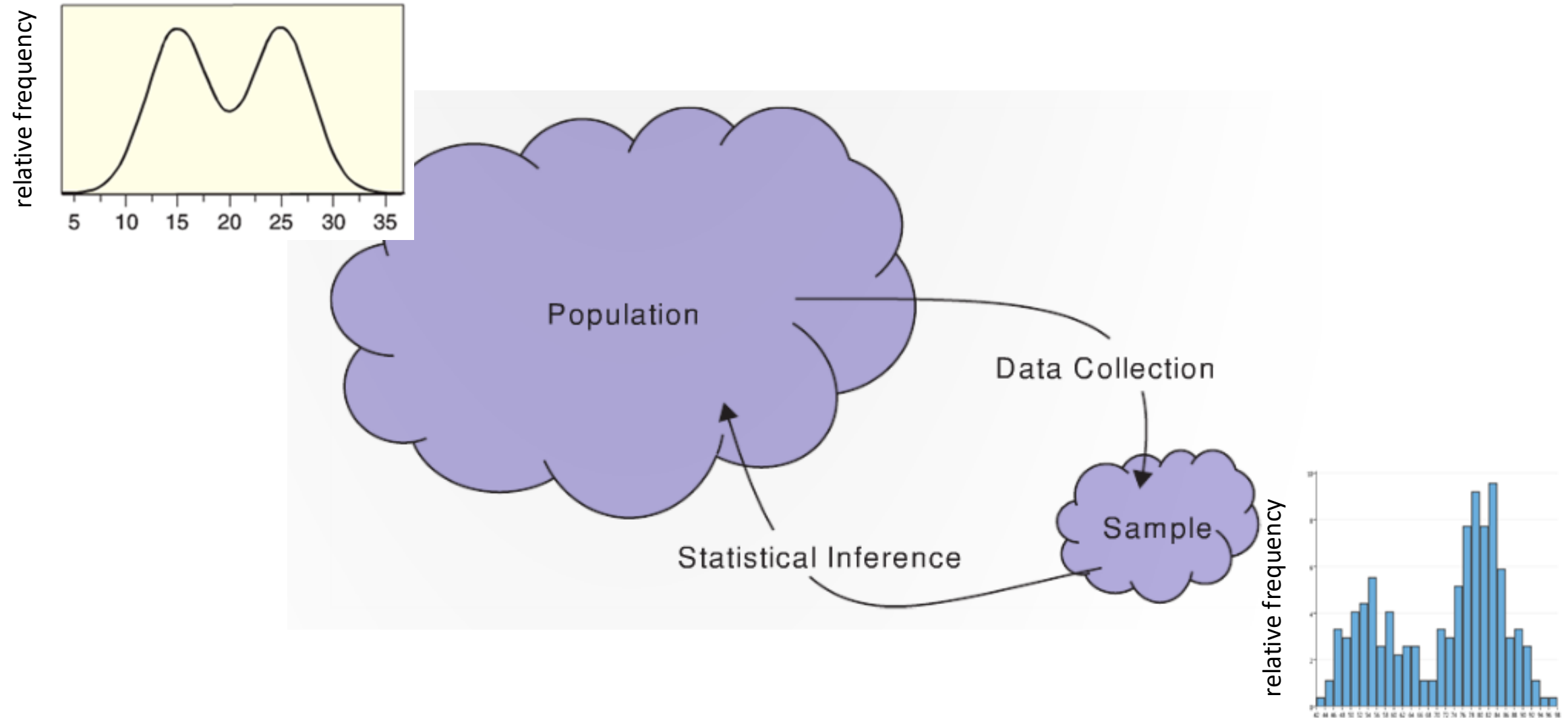
Categorical  
Distribution ( $\pi$ )



Bar chart ( $\hat{p}$ )



# Population distributions and sample histograms

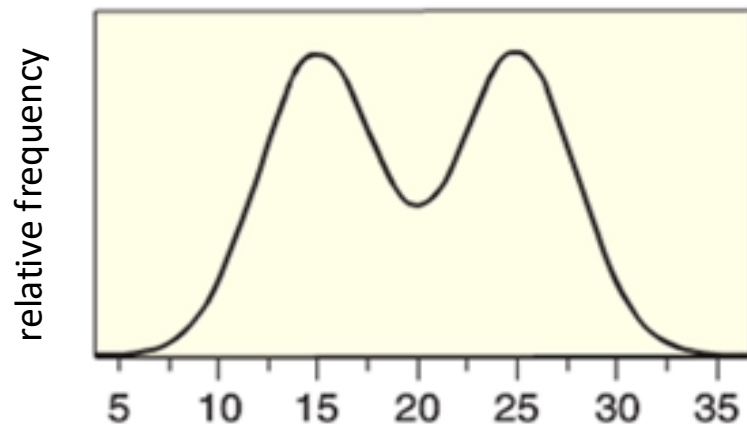


# Histograms

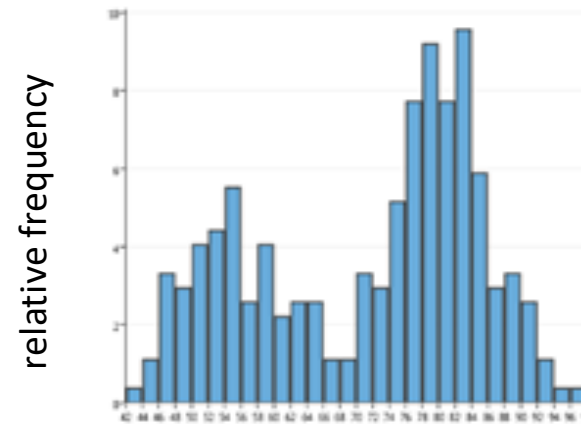
Histograms are a way of visualizing a sample of quantitative data

- They are similar to bar charts but for quantitative variables
- They aim to give a picture of how the data is distributed

Continuous distribution



Histogram



# Gapminder: life expectancy in different countries

Let's look at the life expectancy in different countries, which is a quantitative variable

# pull a vector of life expectancies from the data frame

```
life_expectancy <- gapminder_2007$lifeExp
```

# Histograms – countries life expectancy in 2007

Life expectancy for different countries for 142 countries in the world:

- 43.83, 72.30, 76.42, 42.73, ...

To create a histogram we create a set of intervals

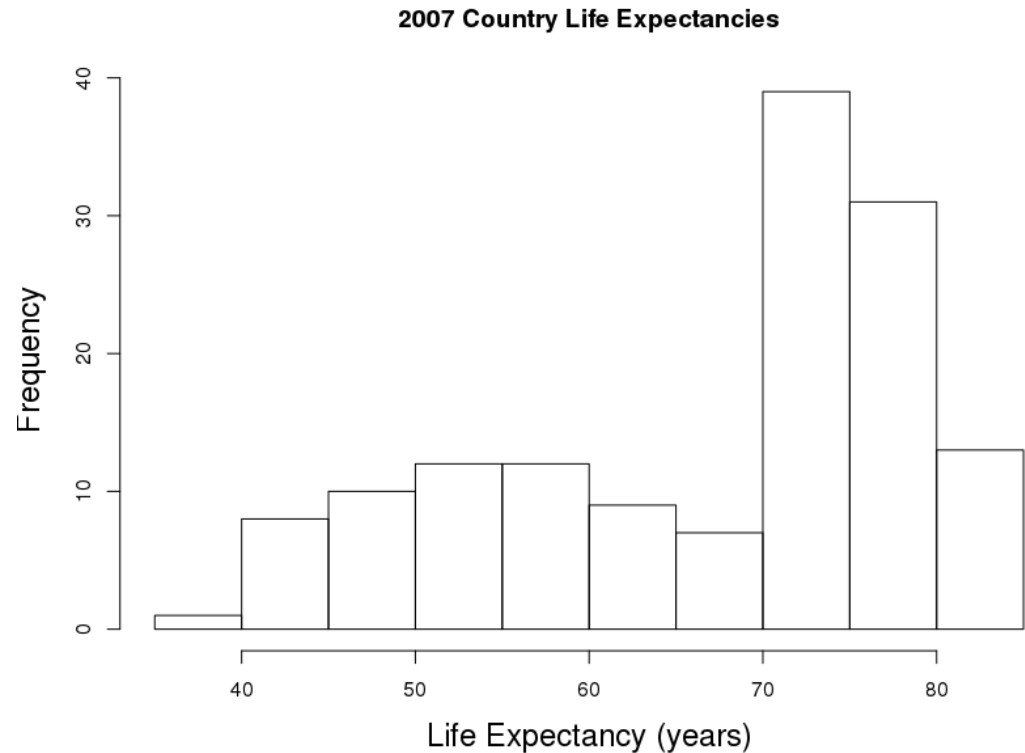
- 35-40, 40-45, 45-50, ... 75-80, 80-85

We count the number of points that fall in each interval

We create a bar chart with the counts in each bin

# Histograms – countries life expectancy in 2007

Life Expectancy	Frequency Count
(35 – 40]	1
(40 – 45]	8
(45 – 50]	10
(50 – 55]	12
(55 – 60]	12
(60 – 65]	9
(65 – 70]	7
(70 – 75]	39
(75 – 80]	31
(80 – 85]	13



R: `hist(v)`

# Gapminder: life expectancy in different countries

Let's create a histogram of the life expectancy in different countries using the `hist()` function

```
# pull a vector of life expectancies from the data frame
```

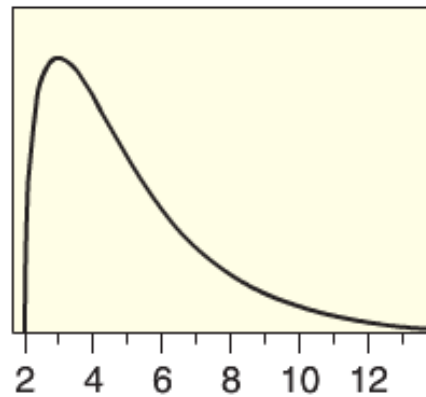
```
life_expectancy <- gapminder_2007$lifeExp
```

```
# create the histogram
```

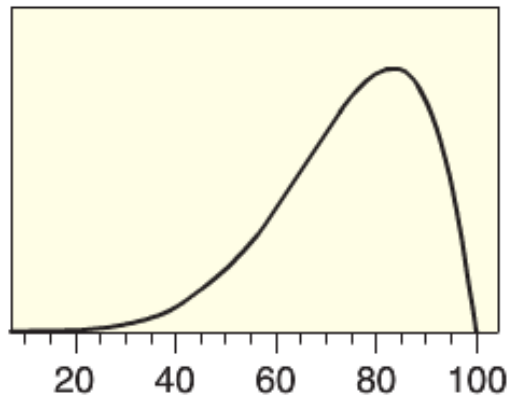
```
hist(life_expectancy,  
     xlab = "Life expectancy")
```



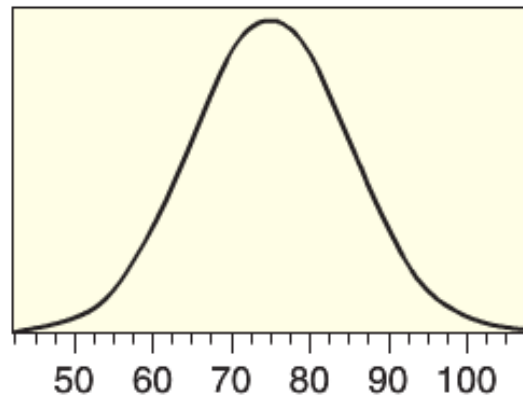
# Common shapes for distributions



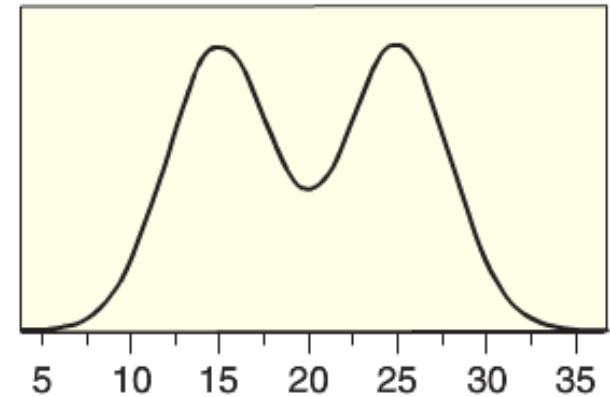
(a) Skewed to the right



(b) Skewed to the left

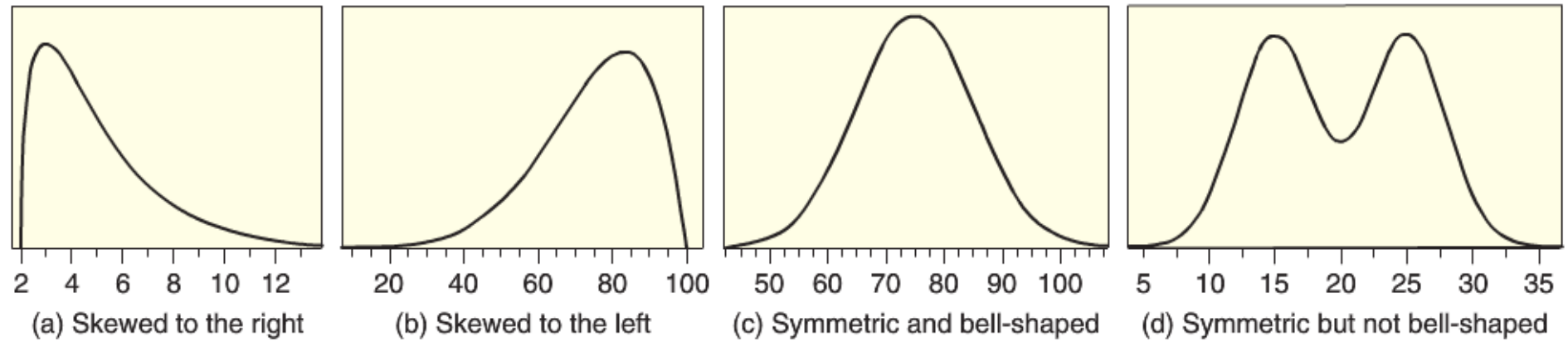
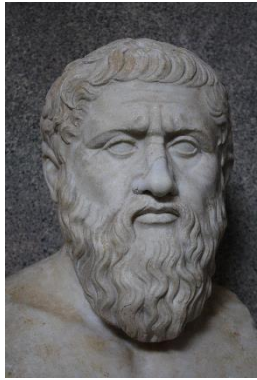


(c) Symmetric and bell-shaped

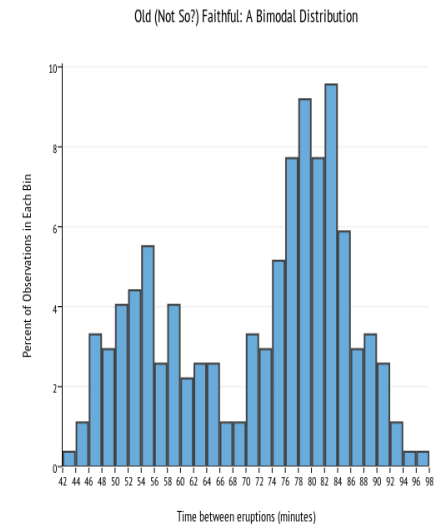
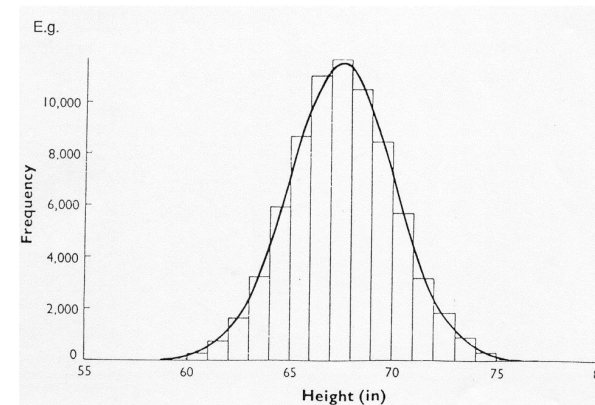
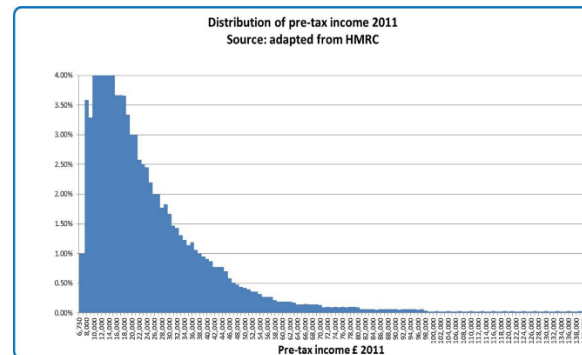


(d) Symmetric but not bell-shaped

# Plato and shadows: distributions and histograms

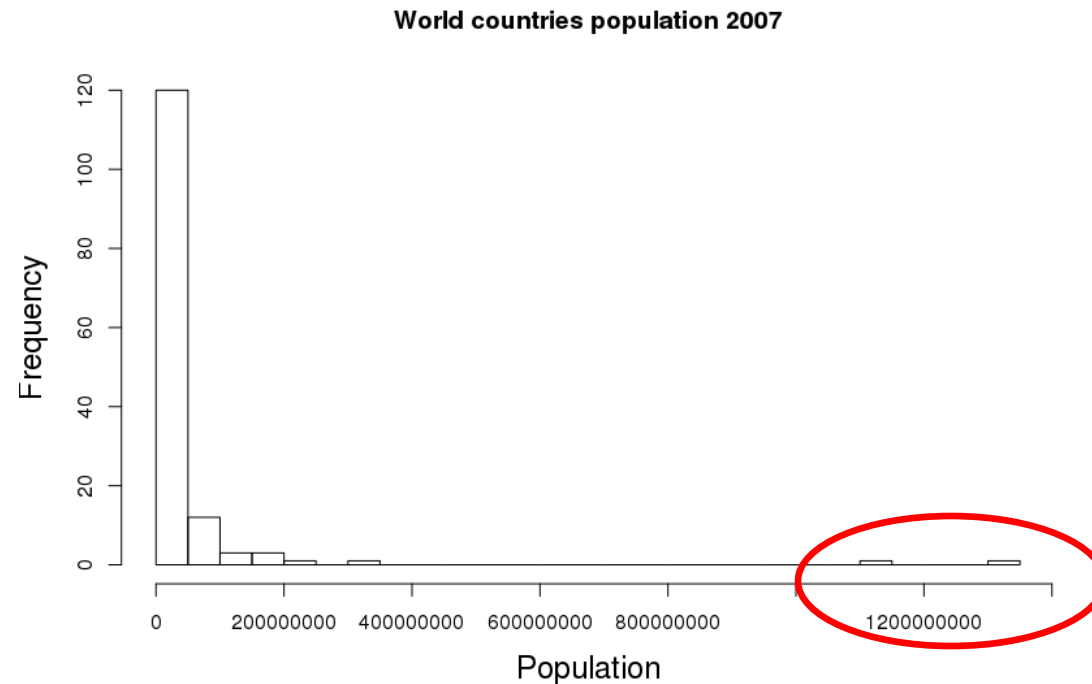


## Income distribution



# 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.



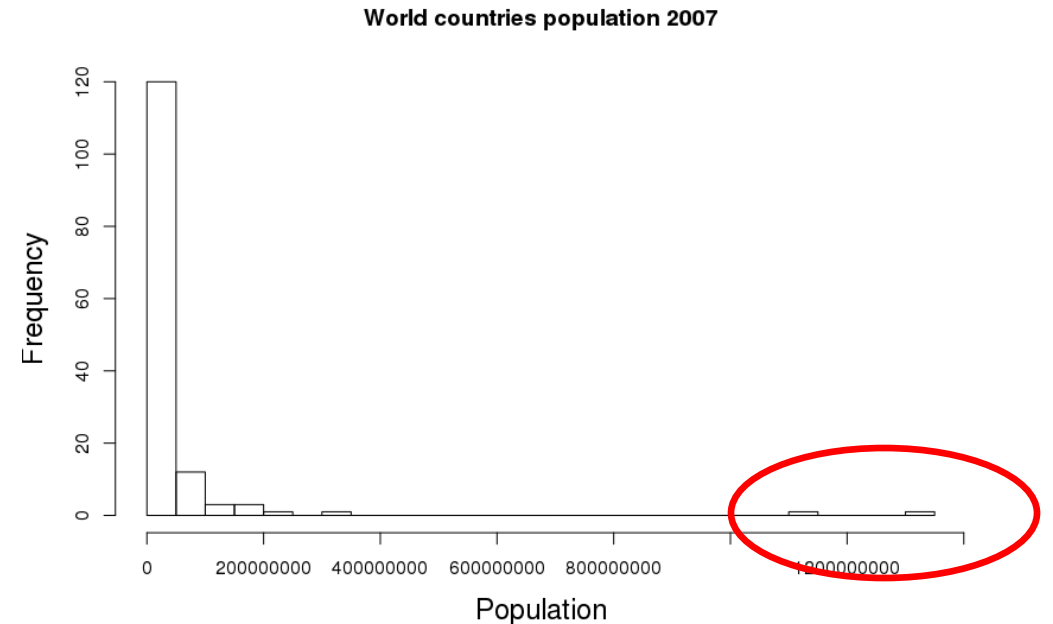
Outliers can potentially have a large influence on the statistics you calculate!

# Outliers

**Q:** What should you do if you have outliers in your data?

**A:** See if we can tell why the outliers exist by examining the data!

- If the outliers are due to a mistakes, one can remove them
- If they are not due to mistakes, one should explain why they exist, and potentially try the analyses with and without the outliers to see if the analysis is affected



Descriptive statistics for the center of a distribution

# Descriptive statistics for the center of a distribution

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

We can also summarize data numerically

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

# The mean

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

$$\text{Mean} = \frac{x_1 + x_2 + x_3 + \dots + x_n}{n} = \frac{1}{n} \sum_{i=1}^n x_i$$

R: `mean(x)`

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

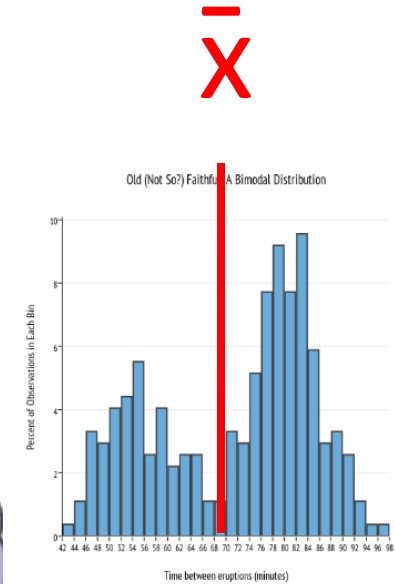
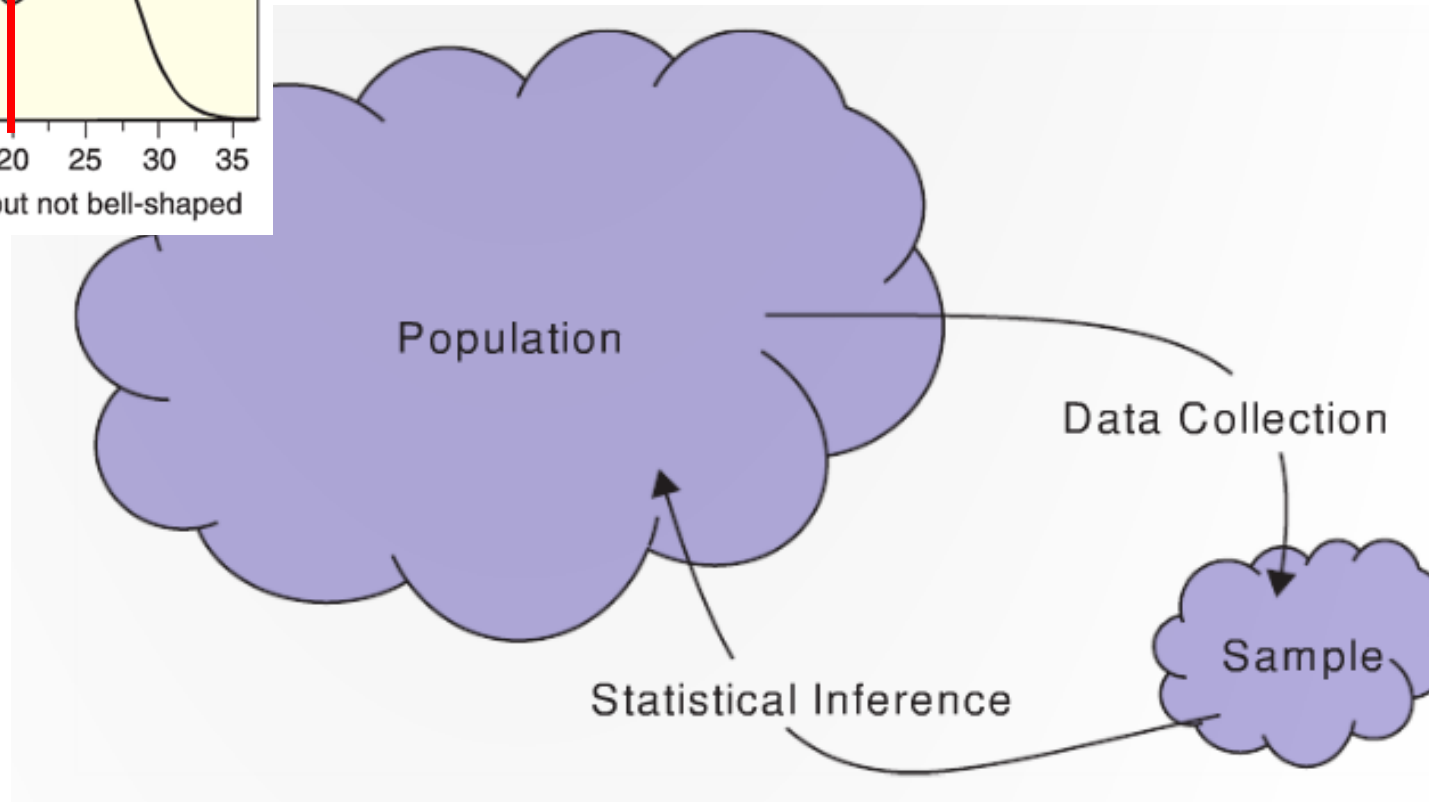
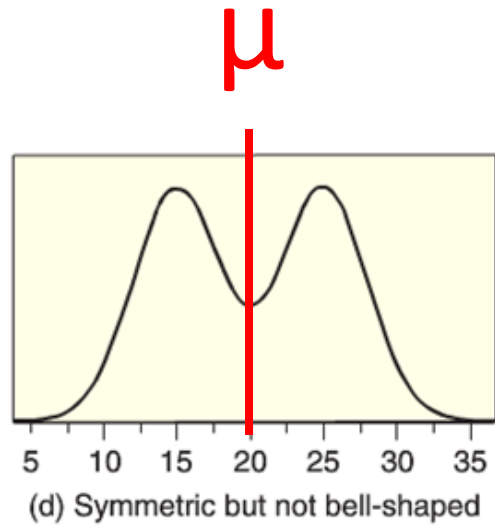
# Notation

The mean of the ***population*** is denoted  $\mu$

The mean of a ***sample*** is denoted  $\bar{x}$



# Sample and population mean



Give the proper notation:  $\mu$  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)
```

# The median

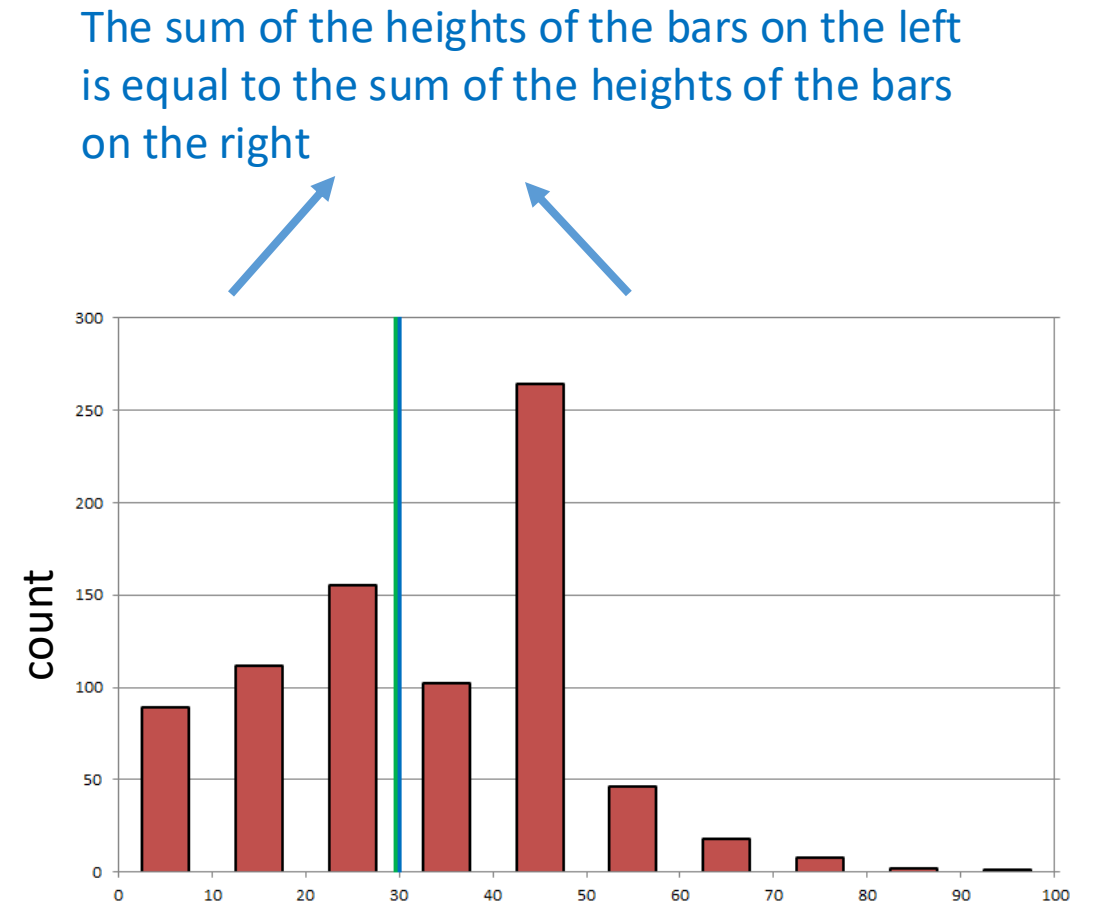
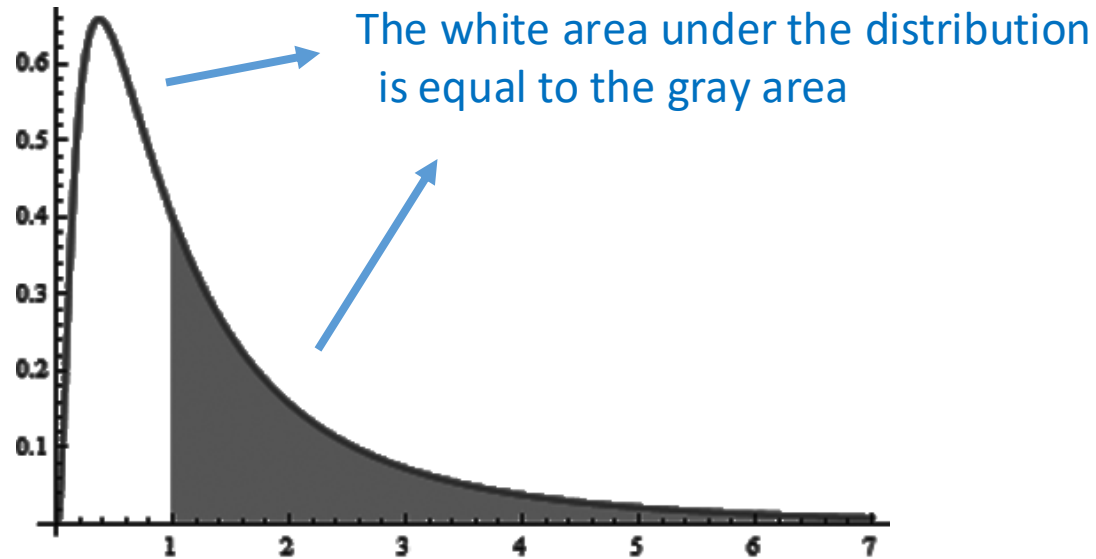
The **median** is a value that splits the data in half

- 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)
```

# 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

# 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?

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

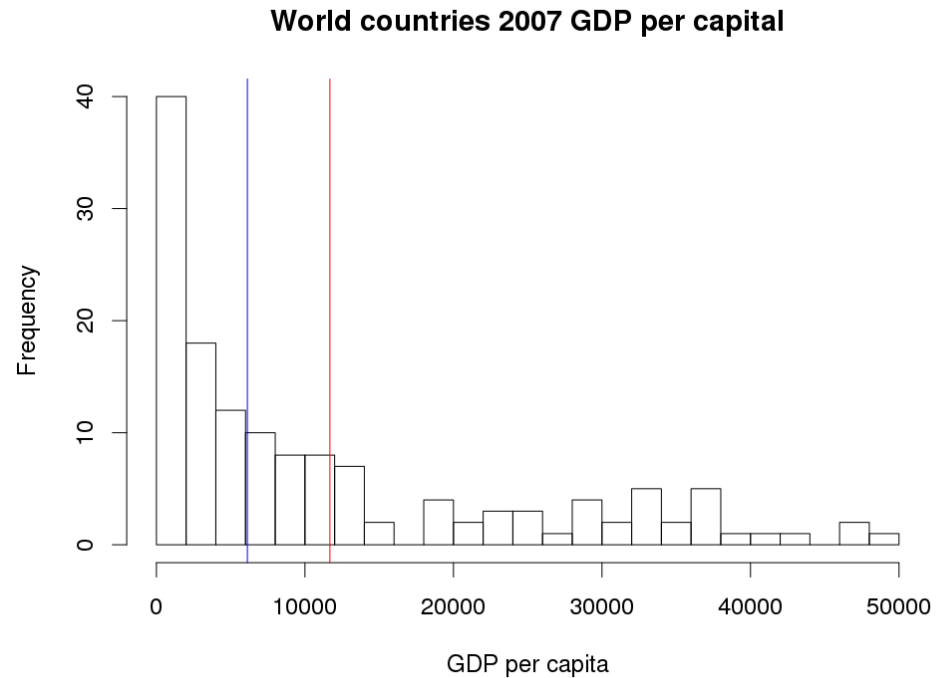


Let's explore calculating the mean and median in R!

# Measures of spread

# Characterizing the spread

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

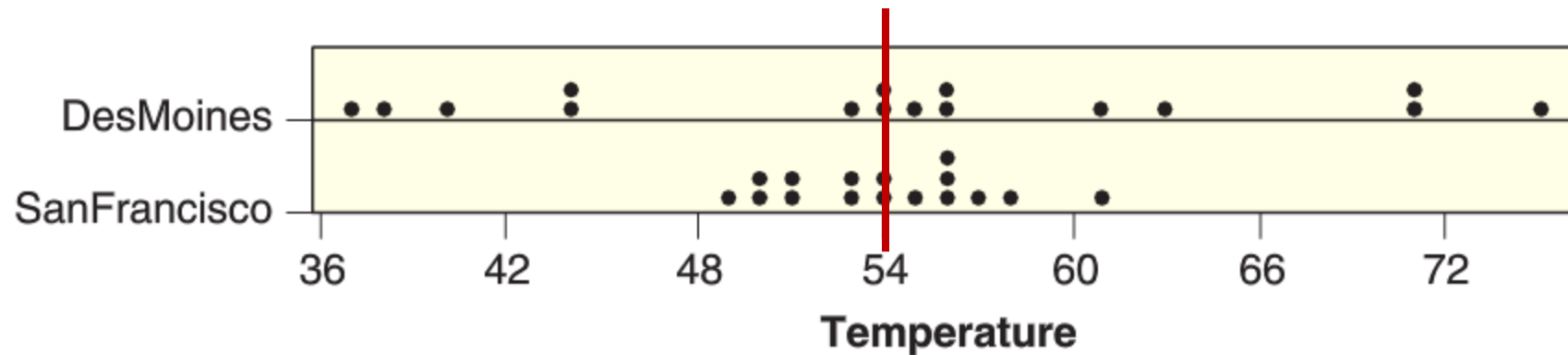


We can also use numbers to characterize how data is spread



# Average monthly temperature: Des Moines vs. San Francisco

Data measured on April 14<sup>th</sup> from 1997 to 2010:

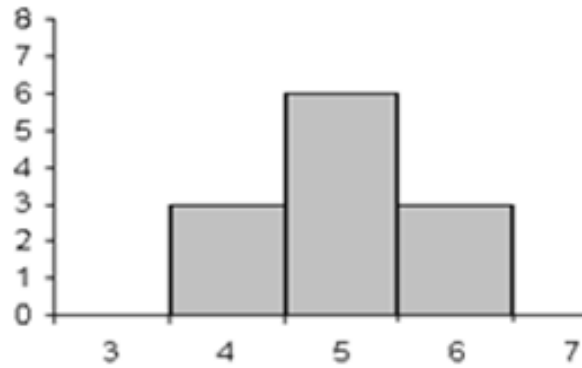


Mean temperature (°F): Des Moines = 54.49 San Fran = 54.01

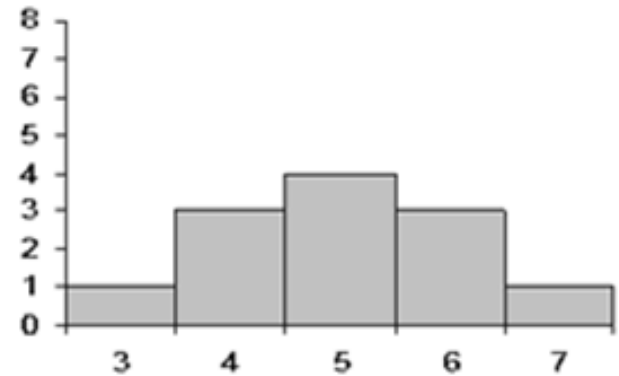
# The standard deviation and variance

The **standard deviation** measures 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

The **variance** is the standard deviation squared

# Notation

The standard deviation of the ***population*** is denoted  $\sigma$

- It measure the spread of the data from the population mean  $\mu$

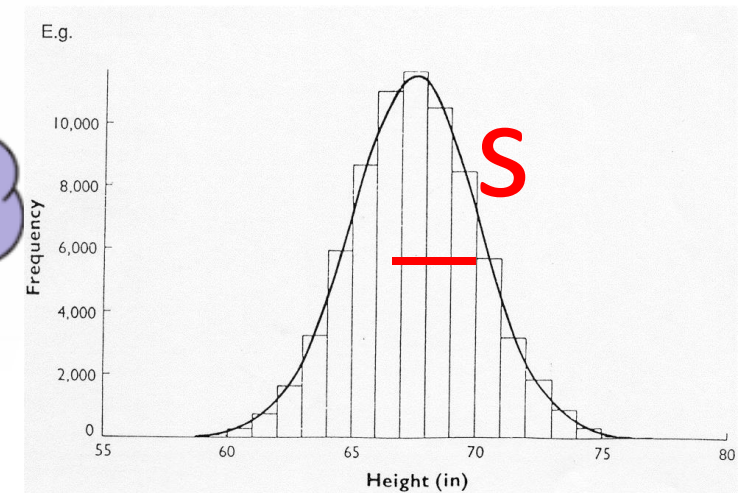
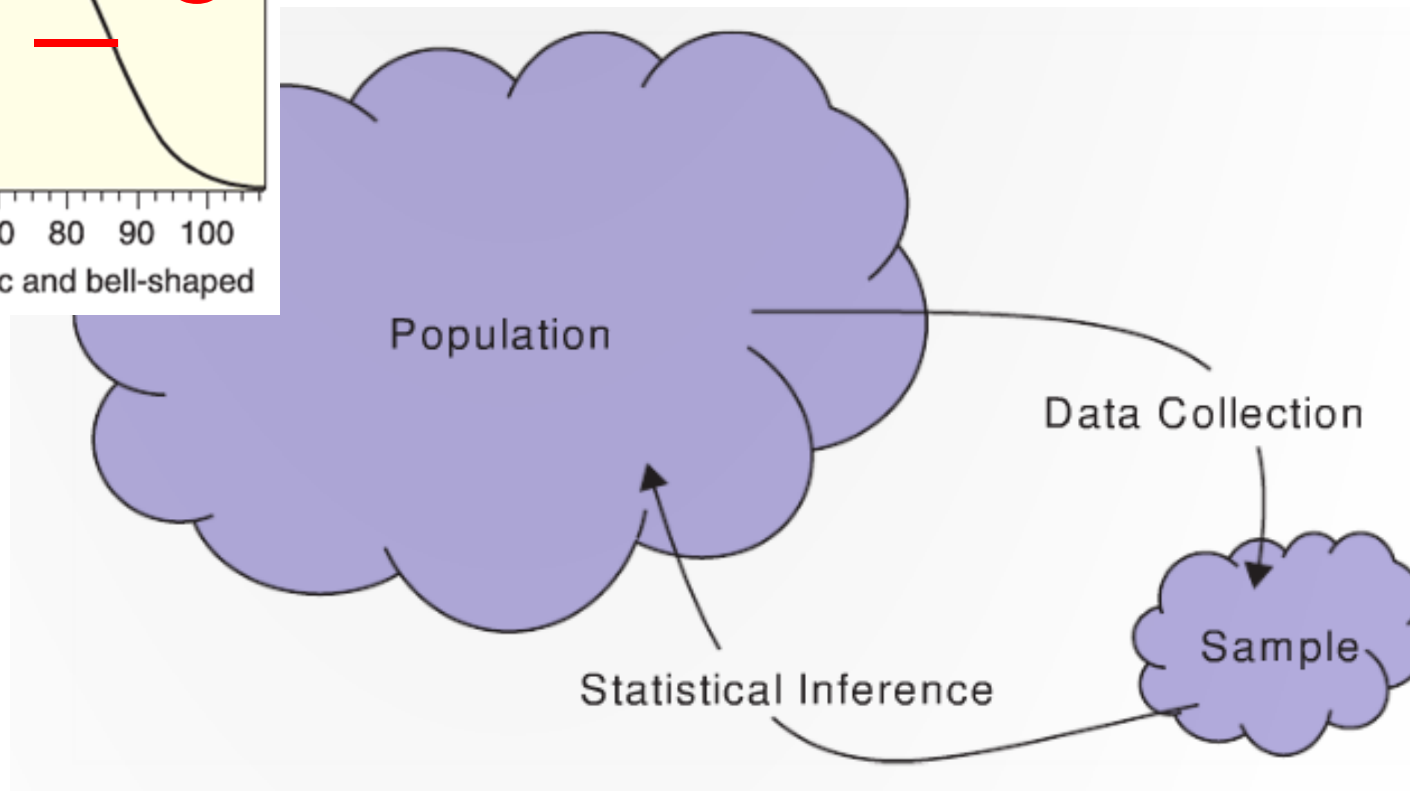
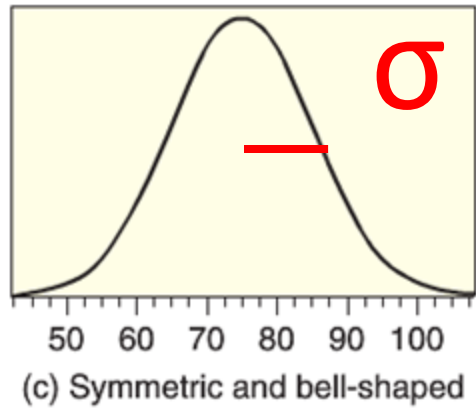
The standard deviation of a ***sample*** is denoted  $s$

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

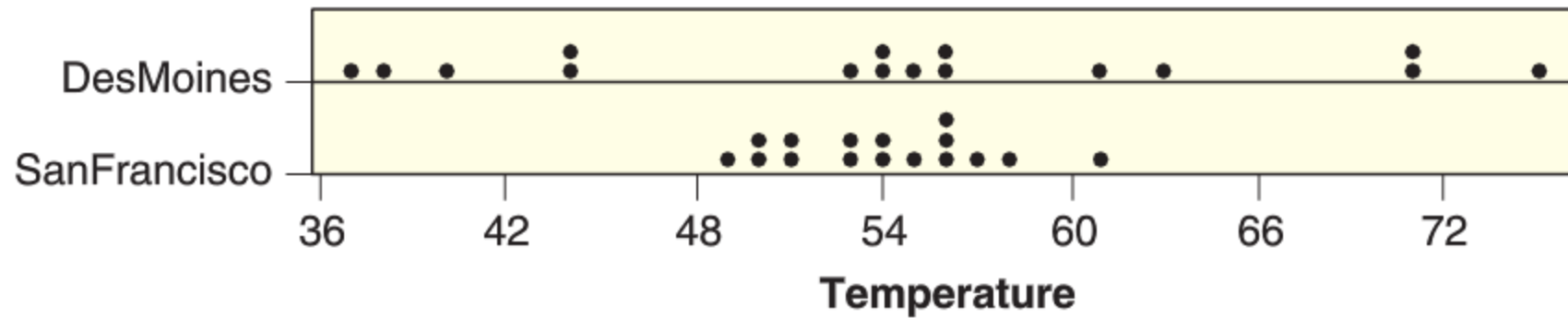
For the variance we use the notation:

- Population variance:  $\sigma^2$
- Sample variance:  $s^2$

# 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 variance and standard deviation

The **variance** can be computed using the formula:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \quad \text{R: } \text{var}(v)$$

The **standard deviation** can be computed using the following formula:

- i.e., the standard deviation is the square root of the variance

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2} \quad \text{R: } \text{sd}(v)$$

# 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 4<sup>th</sup> 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



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

**Homework 2:** Calculate the standard deviation “by hand” for the number of hot dogs eaten by the winners.

Let's calculating some variances and standard deviations in R!



# Summary of concepts

1. A **probability distribution** shows the **relative likelihood** that we will get a data point in the population with a particular value

- (for a more precise definition take a class in probability)

2. Distributions can have different shapes

- E.g., left skewed, right skewed, bell shaped, etc.

3. The **mean** is one measure of central tendency

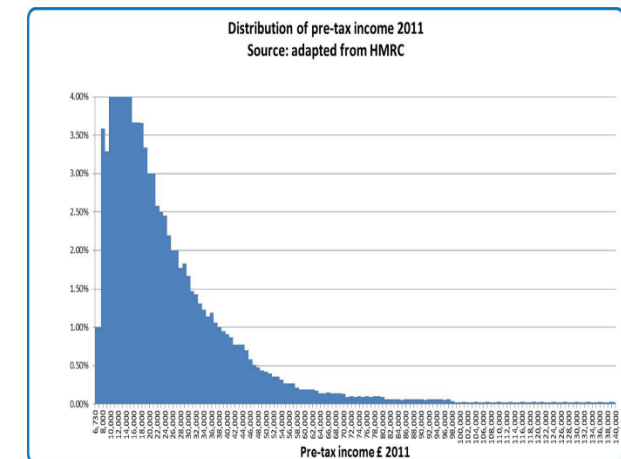
- Sample mean is denoted  $\bar{x}$  (statistic)
- Population mean is denoted  $\mu$  (parameter)

4. The **median** is another measure of central tendency

- The median is resistant to outliers while the mean is not

5. Two measures of variability are the **variance** ( $\sigma^2$ ,  $s^2$ ) and the **standard deviation** ( $\sigma$ ,  $s$ )

## Income distribution



# Summary of R

**Data frames** contain structured data

- We can extract vectors from a data frame using:

```
my_vec <- my_data_frame$my_var
```

We can get a sense of how quantitative data is distributed by creating a histogram

```
hist(my_vec)
```

We can calculate measures of central tendency using:

```
mean(my_vec)
```

```
median(my_vec)
```

We can calculate measures of variability using:

```
var(my_vec)
```

```
sd(my_vec)
```