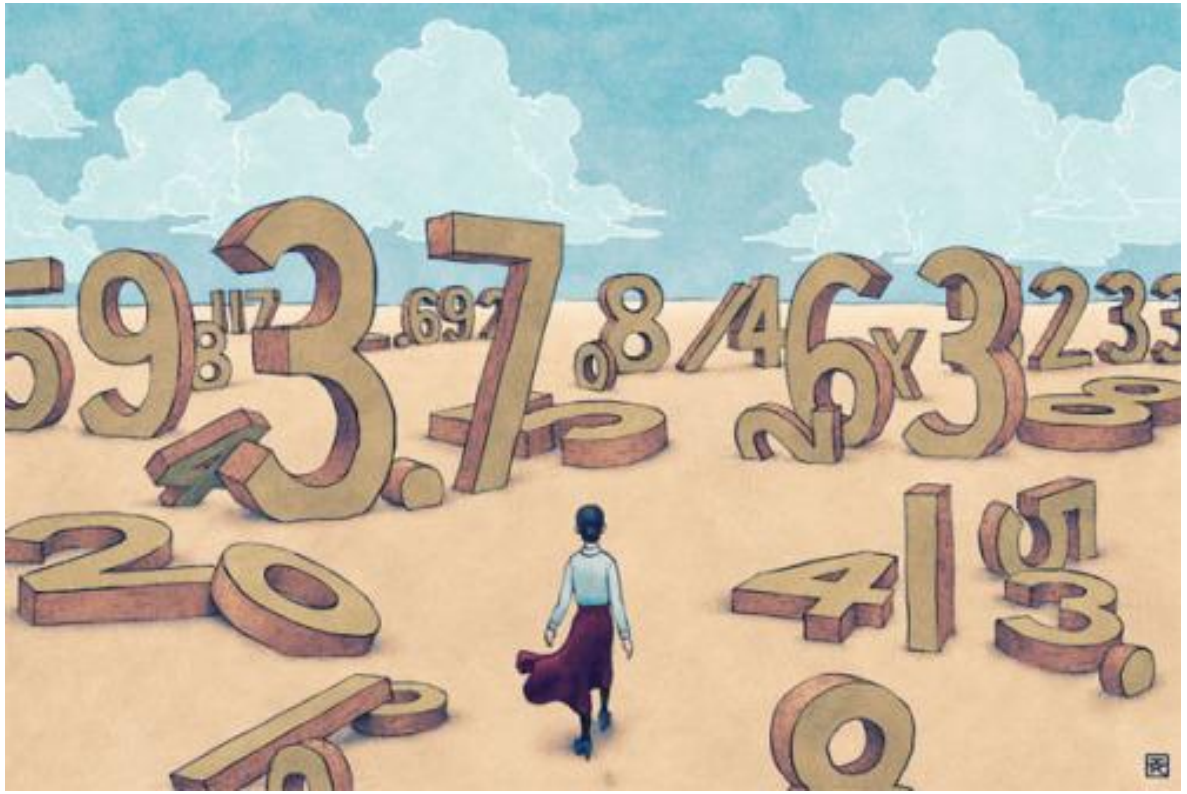
Quantitative data: shape, measures of central tendency, and spread



Overview

Warmup: categorical data analysis in R

Quantitative data

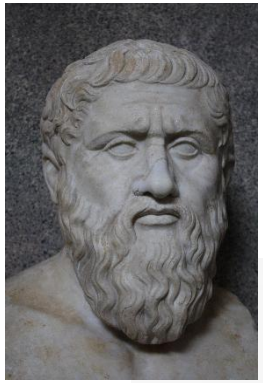
- Graphing the shape: histograms and outliers
- Measures of the central tendency: mean and median

If there is time

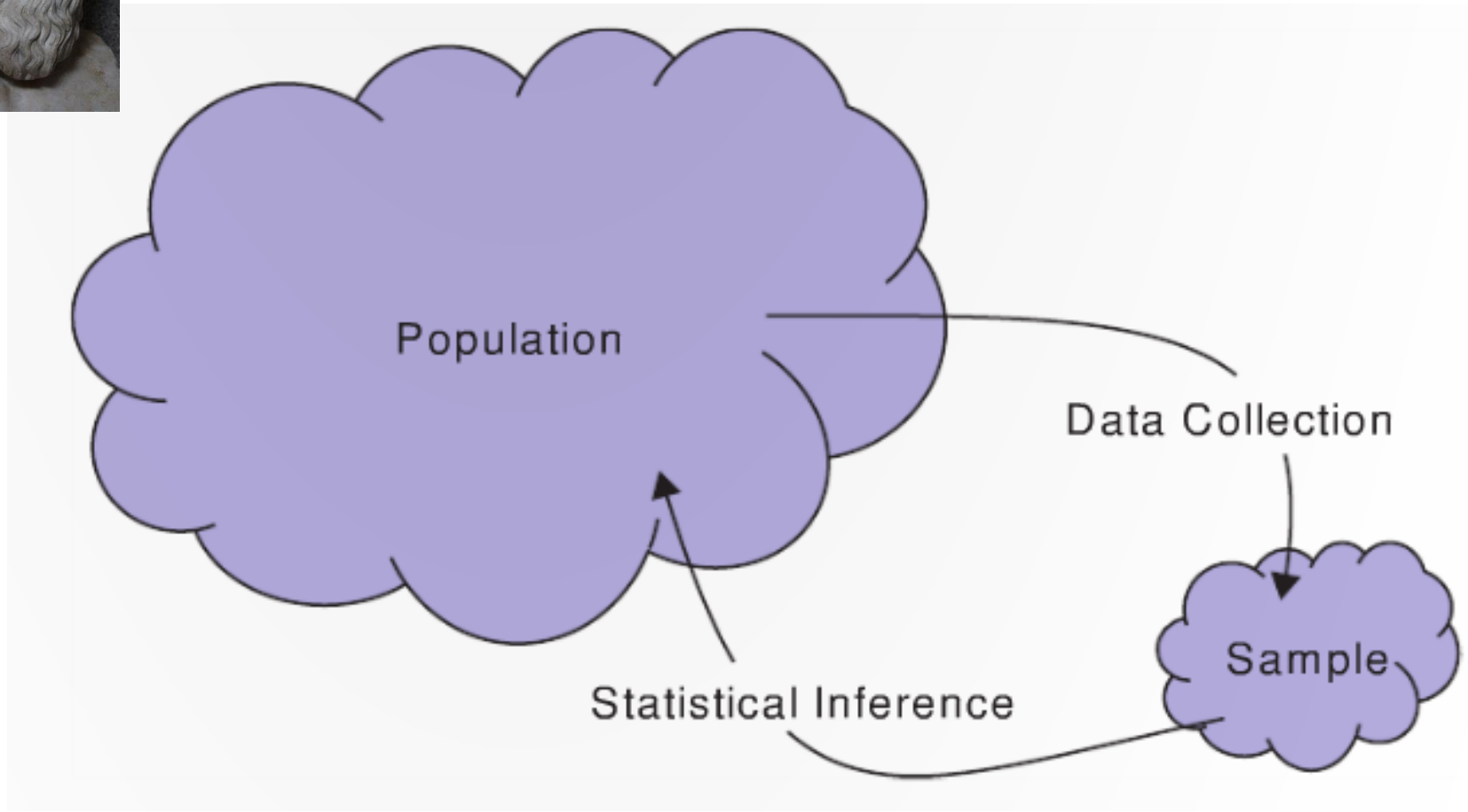
- Measure of spread: The variance and standard deviation

Review

Categorical variables

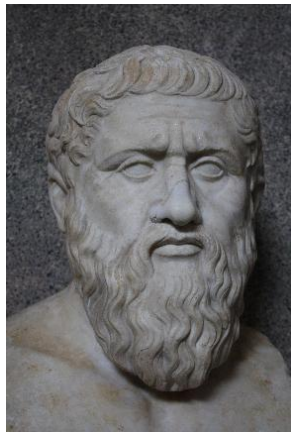


parameter: π



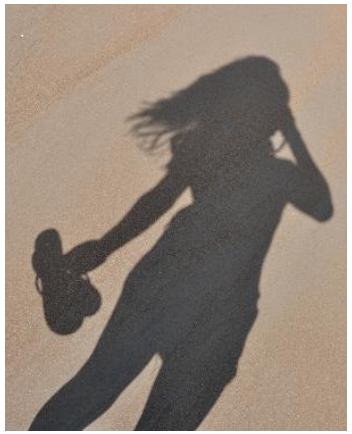
statistic: $\hat{p}$

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



P-Truth

- Population or process
- Parameter π, …
- Plato (Greek symbols)



S-shadows

- Sample
- Statistic $\hat{p}$, …
- Shadow (Latin symbols)

Gapminder data

Data frames are the way R represents structured data

Data frames can be thought of as collections of related vectors

- i.e., each column in the DataFrame can be thought of as a vector of data

```
# load data into R
load("gapminder_2007.rda")
```

	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

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

The gapminder_2007 data frame contains information about countries in the world

We can access individual vectors of data using the $ symbol

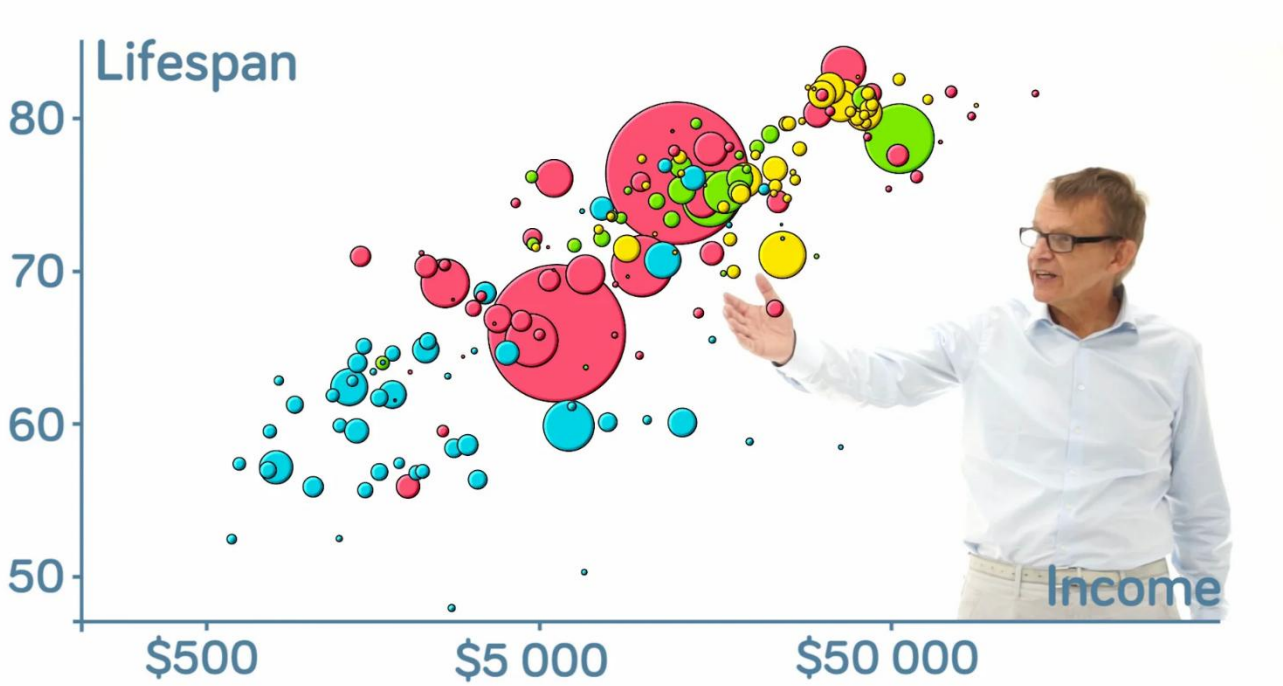
```
continents <- gapminder_2007$continent    # same as using  c("Asia", "Europe", etc.
```

Since this is categorical data we could create frequency tables, bar plots, etc.

```
continent_table <- table(continents)

barplot(continent_table)
```

Let’s do one more quick practice of analyzing categorical data in R…

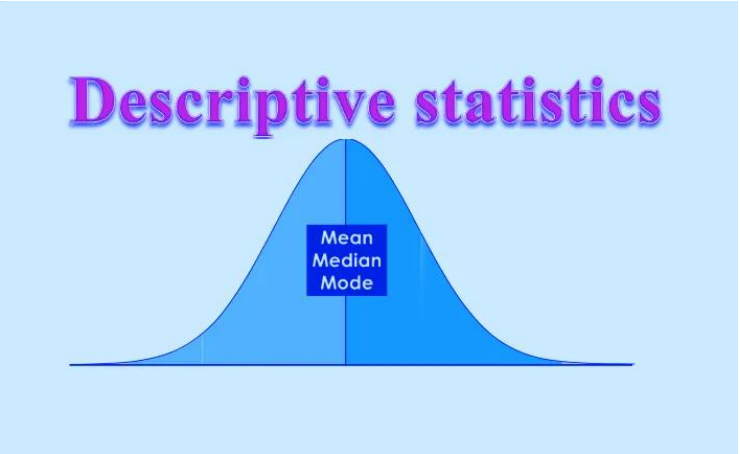


Quantitative variables

Descriptive statistics for one quantitative variable

We will be looking at:

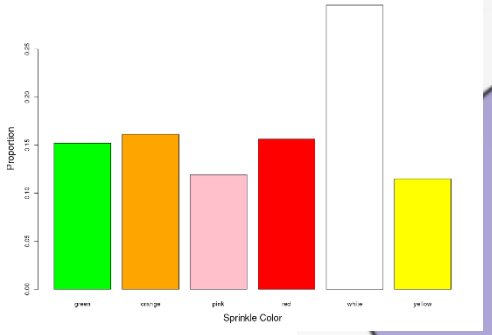
- What is the general ‘shape’ of the data
- Where are the values centered
- How do the data vary

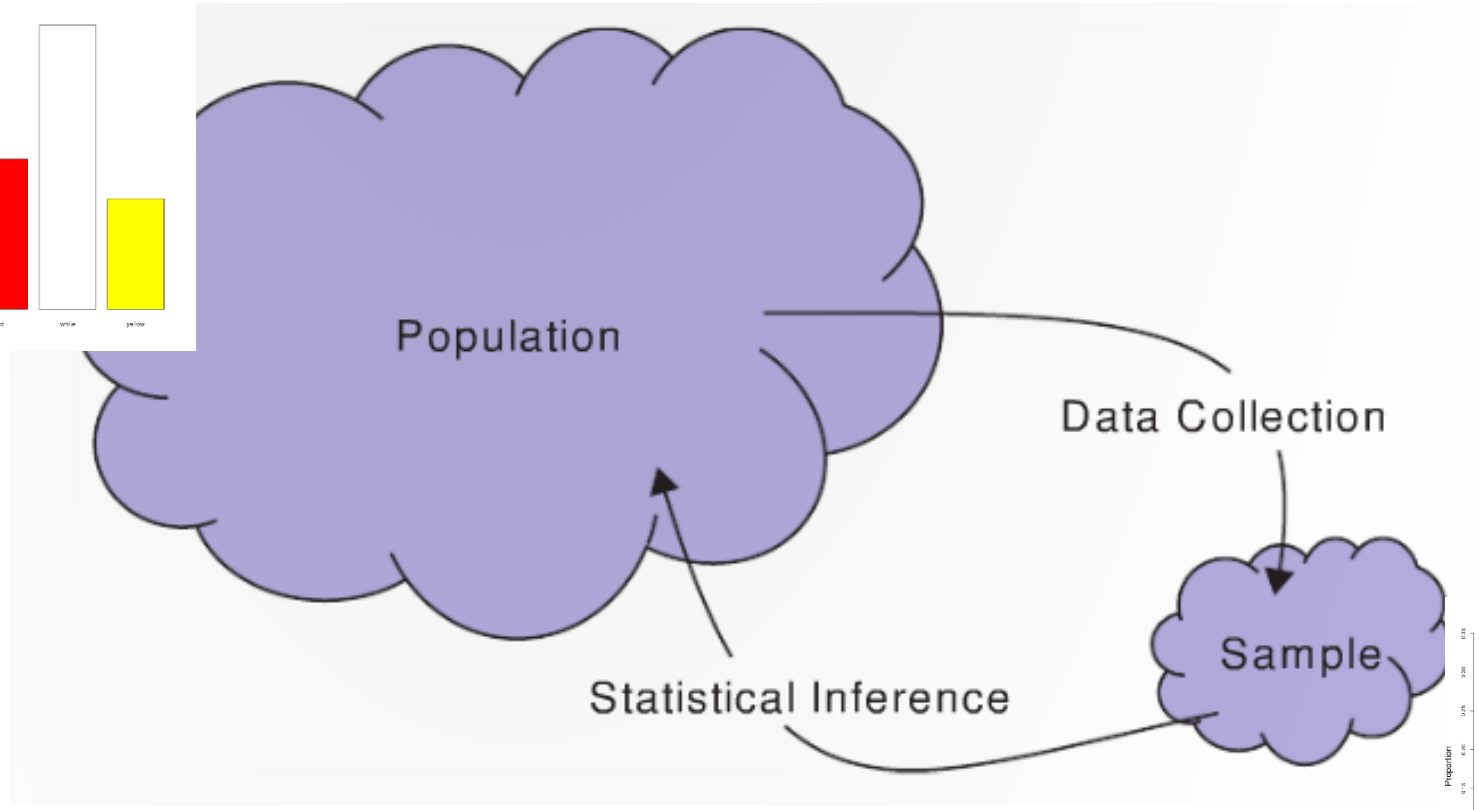


There are all properties of how the data is ***distributed***

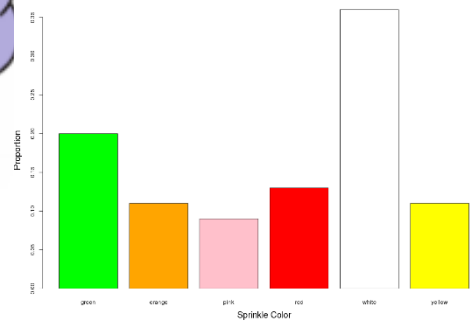
For categorical data we had…

Categorical Distribution (π)

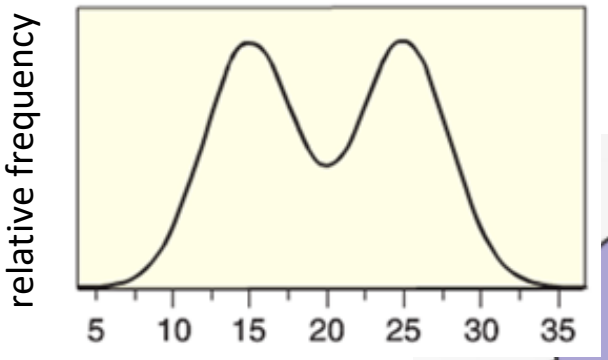


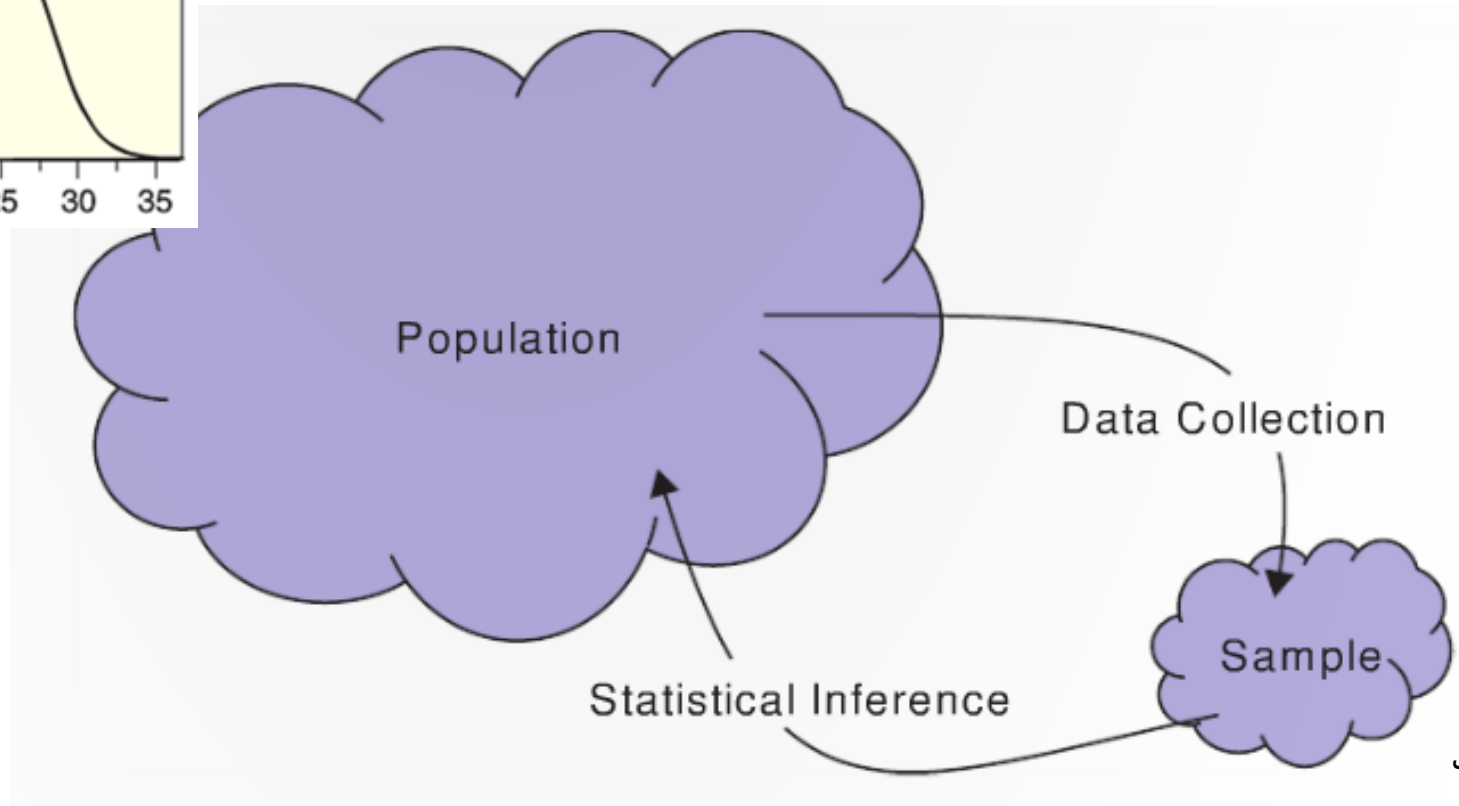


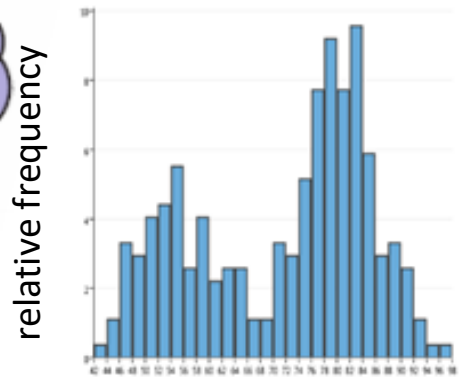
Bar chart ($\hat{p}$)



Population distributions and sample histograms





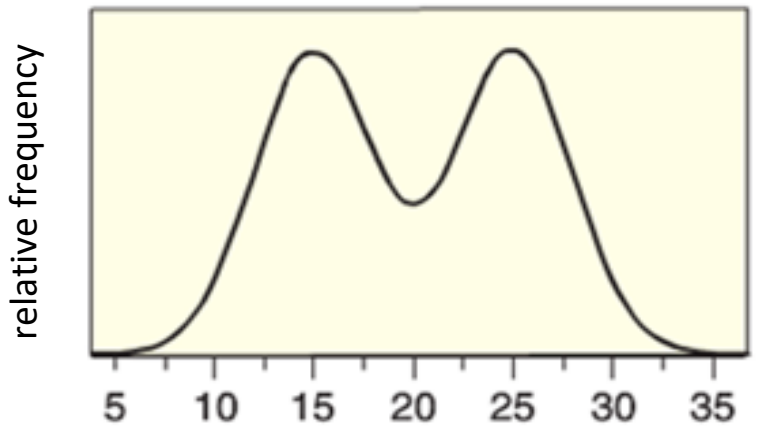


Histograms

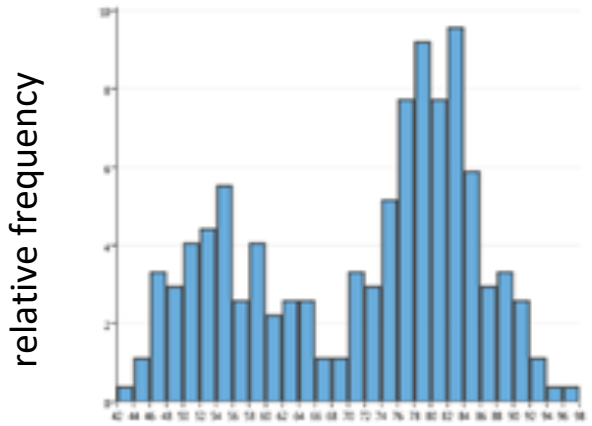
Histograms are a way of visualizing a sample of quantitative data

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

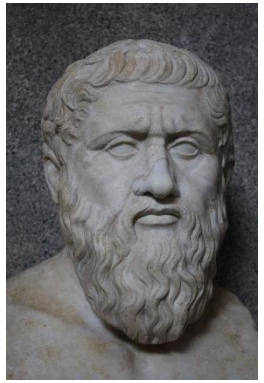
Continuous distribution

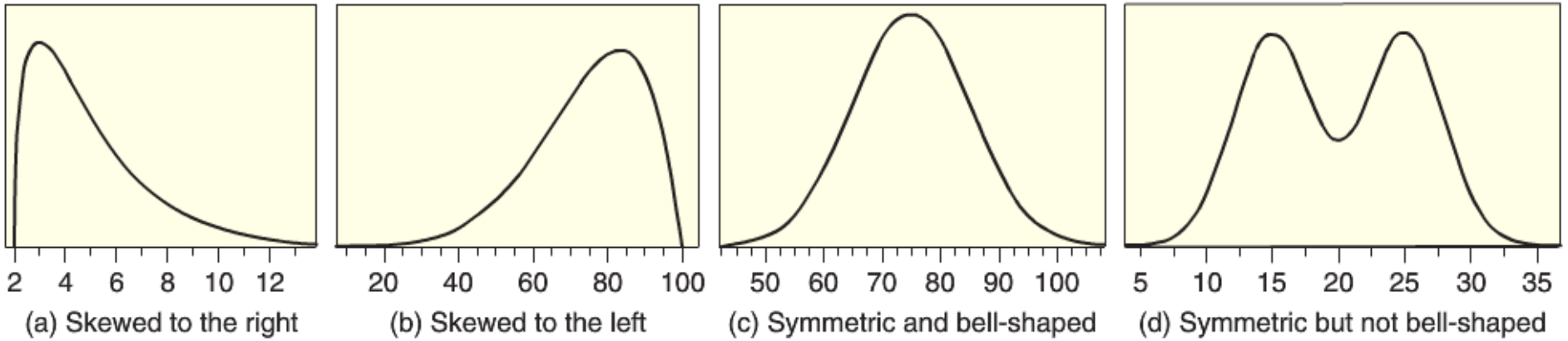


Histogram

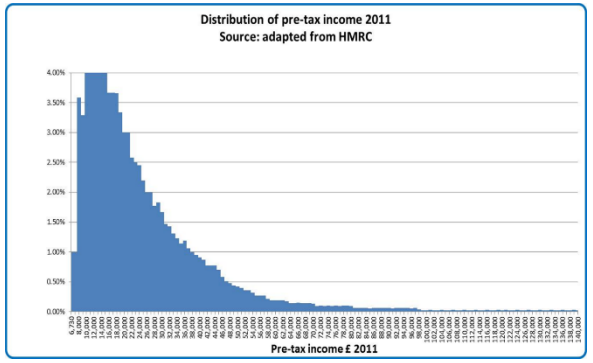


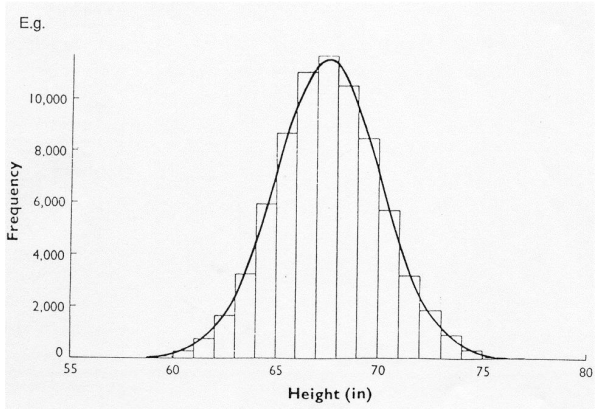
Plato and shadows: distributions and histograms

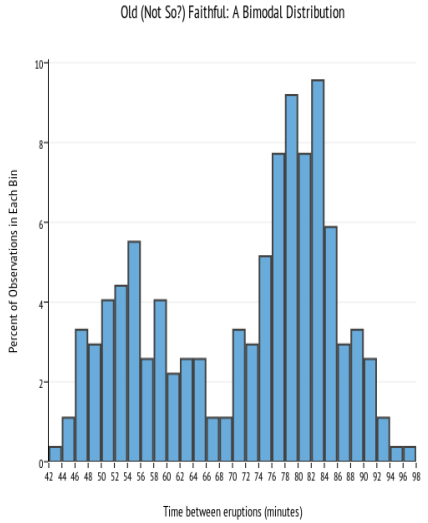




Income distribution

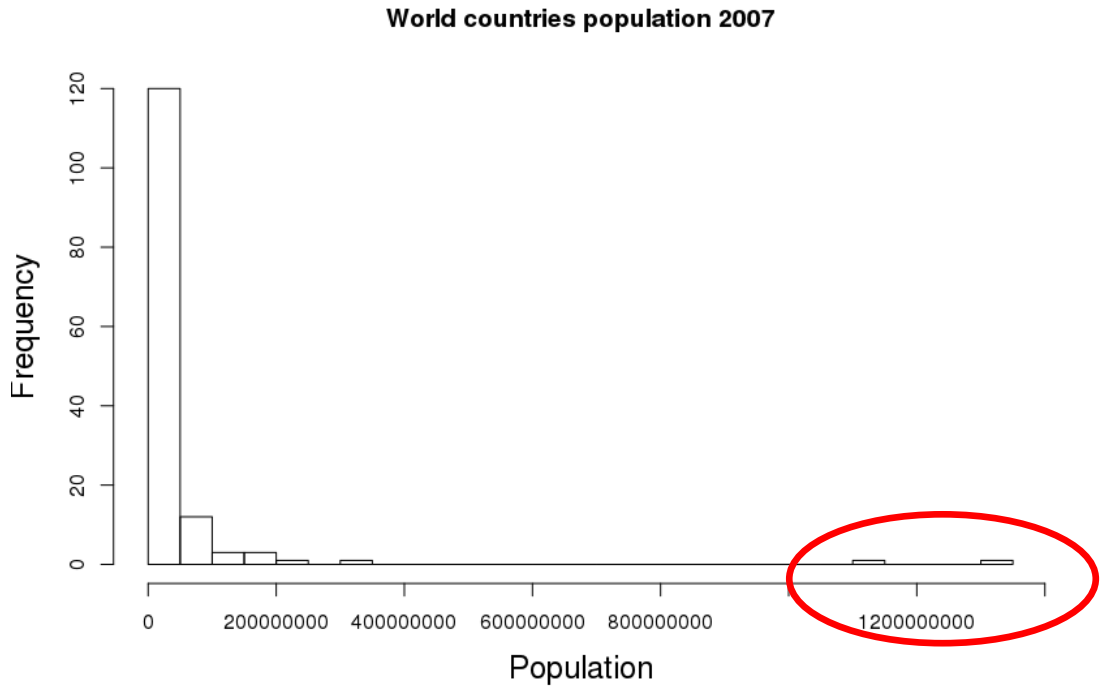






Outliers

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



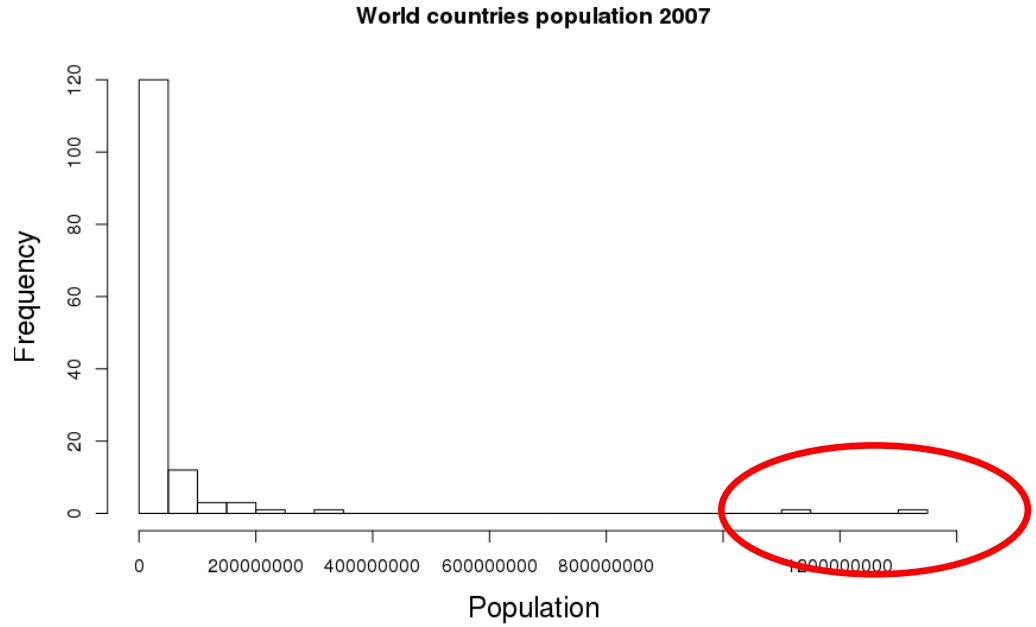
Outliers can potentially have a large influence on the statistics you calculate!

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



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
```

Let's try it in R!

	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

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 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**

The mean

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

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

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

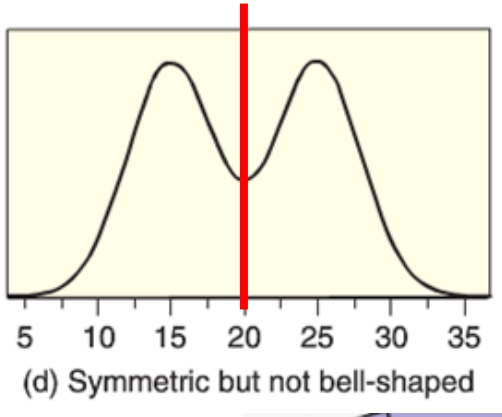
Notation

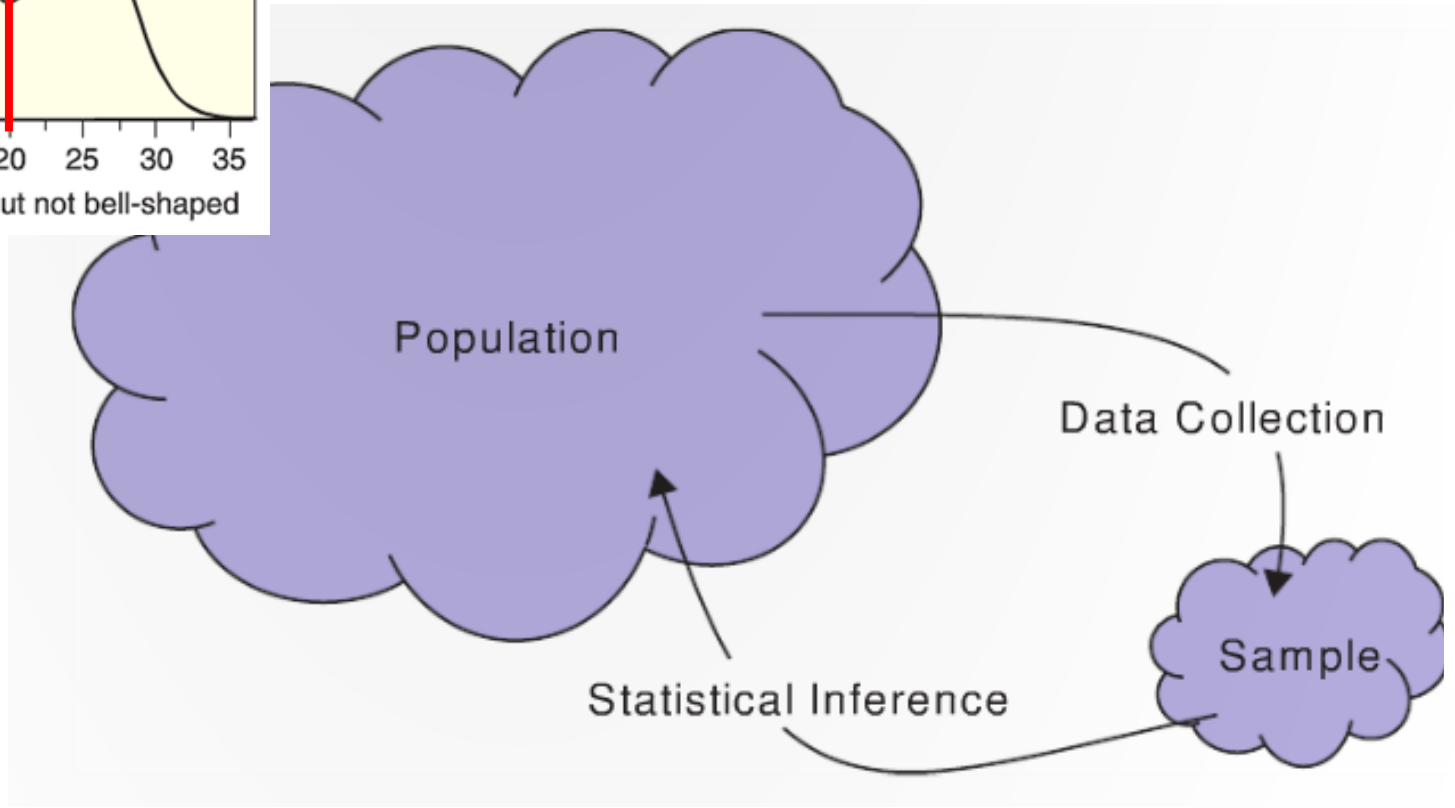
The mean of the ***population*** is denoted μ

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

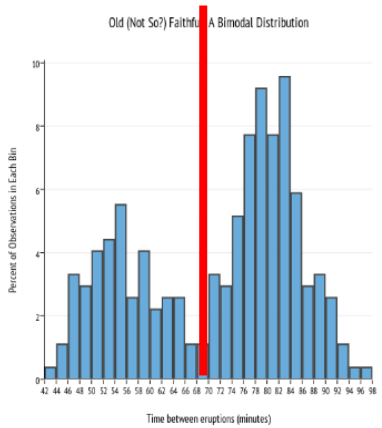
Sample and population mean

μ





$\bar{x}$



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)
```

The median

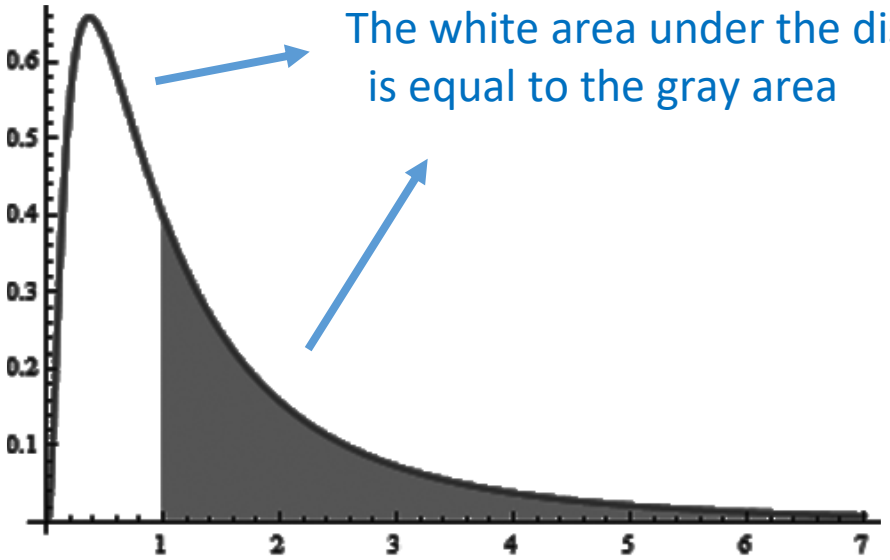
The **median** is a value that splits the data in half

- 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:

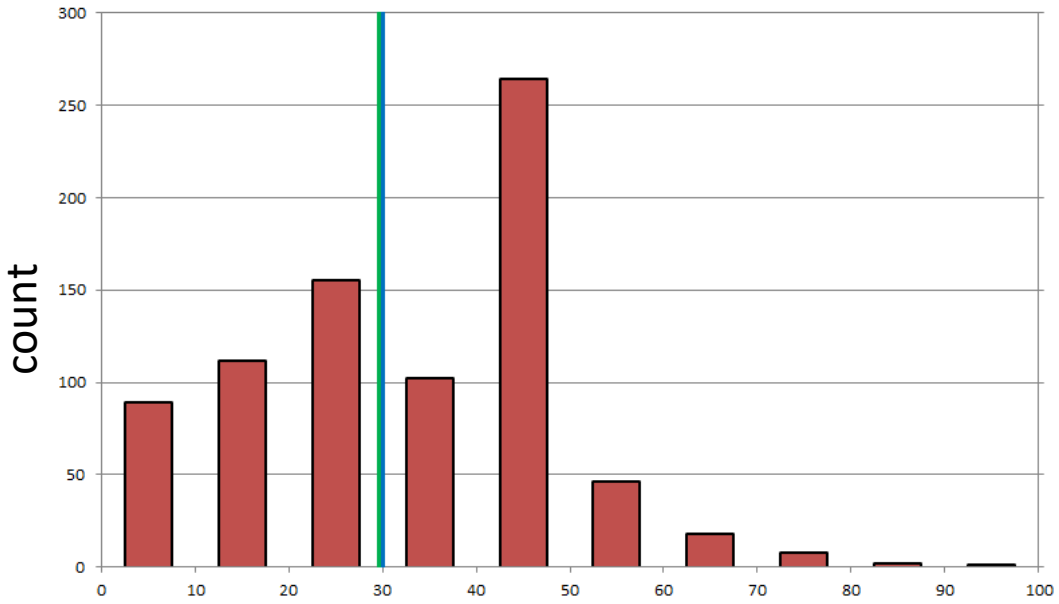
- If n is odd: The middle value of the sorted data
- If n is even: The average of the middle two values of the sorted data

The median



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

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?

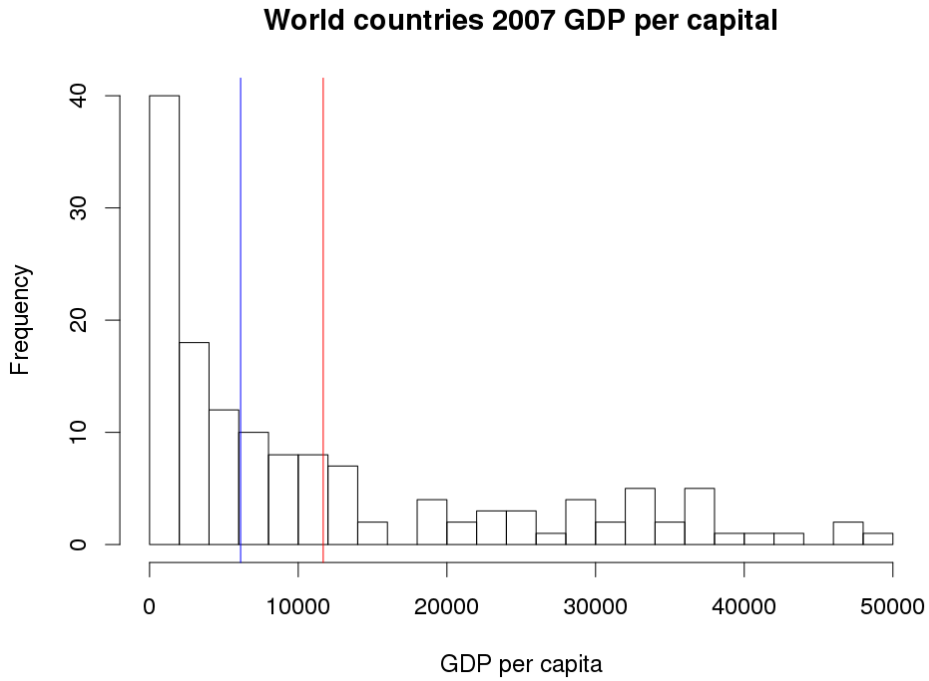
$$\bar{x} = \frac{1}{n}\sum_{i}^{n} x_i$$

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

Measures of spread

Characterizing the spread

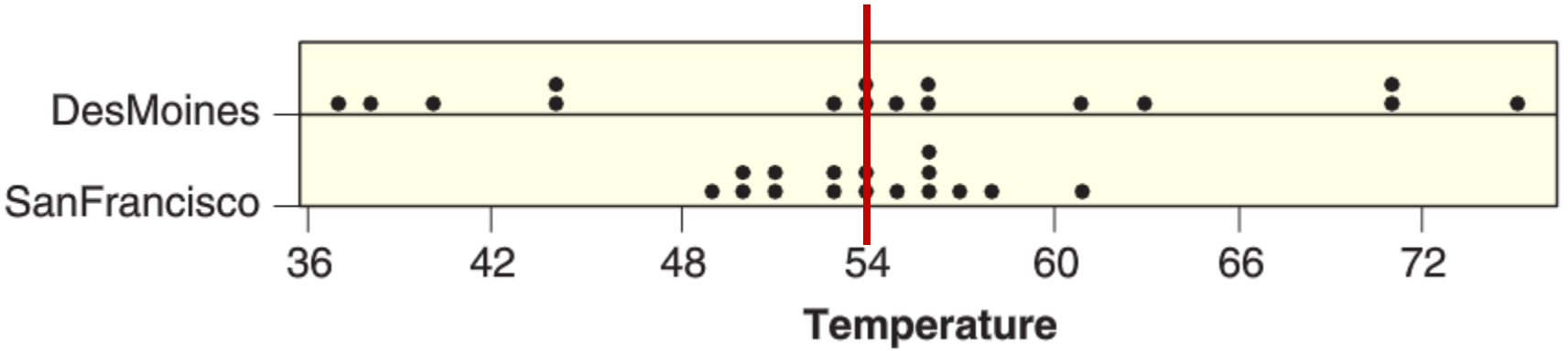
The mean and median are numbers that tell us about the center of a distribution



We can also use numbers to characterize how data is spread

Average monthly temperature: Des Moines vs. San Francisco

Data measured on April 14th from 1997 to 2010:

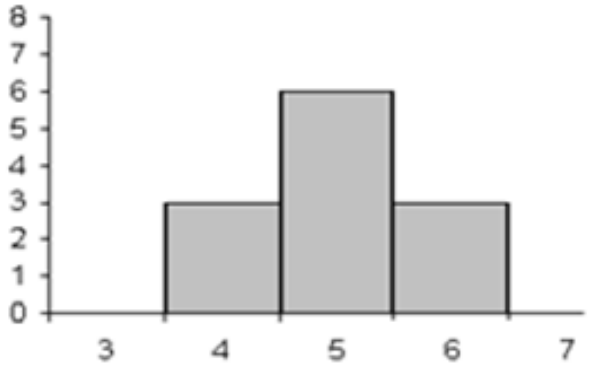


Mean temperature (°F): Des Moines = 54.49 San Fran = 54.01

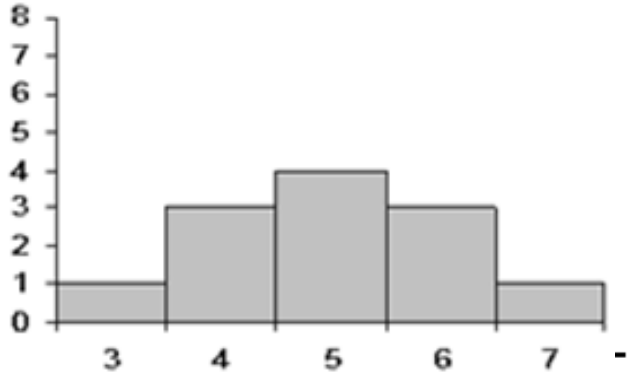
The standard deviation and variance

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 **σ**

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

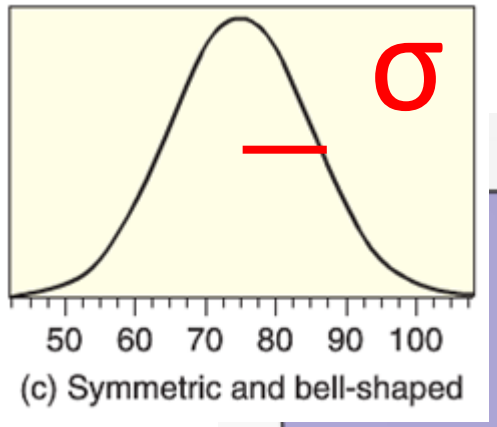
The standard deviation of a ***sample*** is denoted **s**

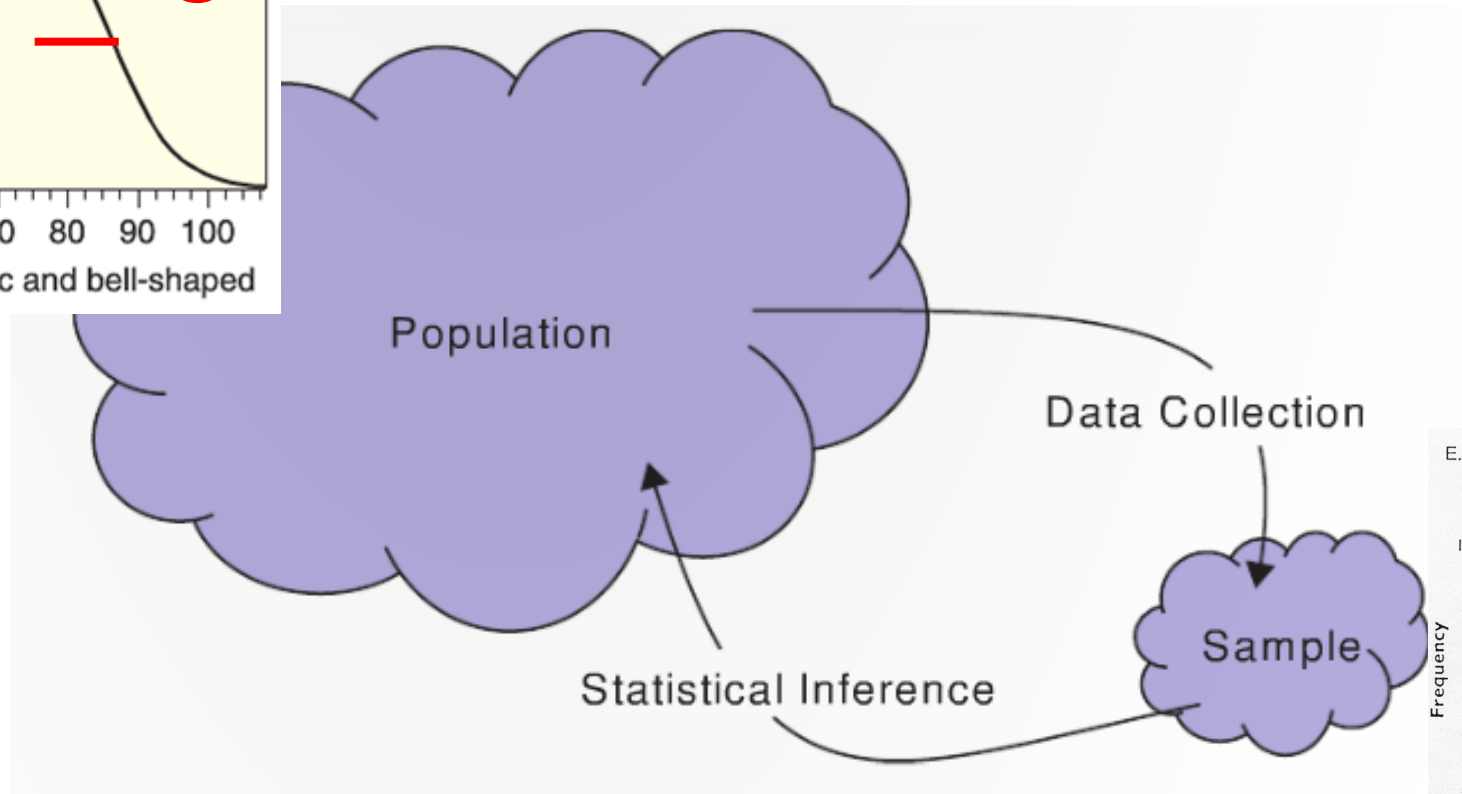
- It measure the spread of the data from the sample mean $\bar{x}$

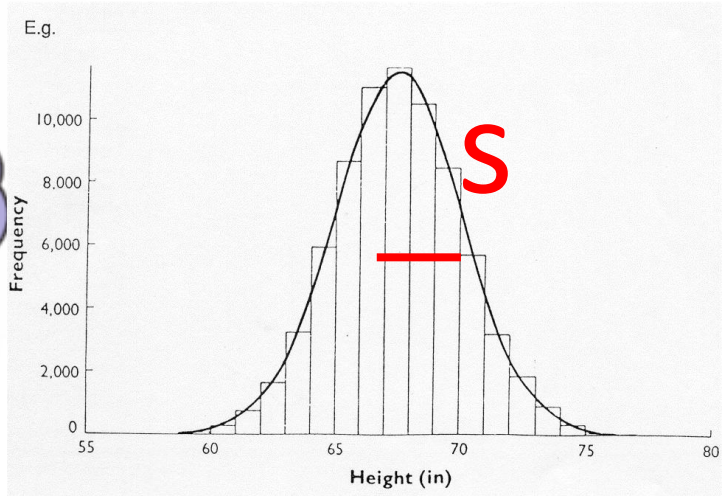
For the variance we use the notation:

- Population variance: $\boldsymbol{\sigma^2}$
- Sample variance: $\boldsymbol{s^2}$

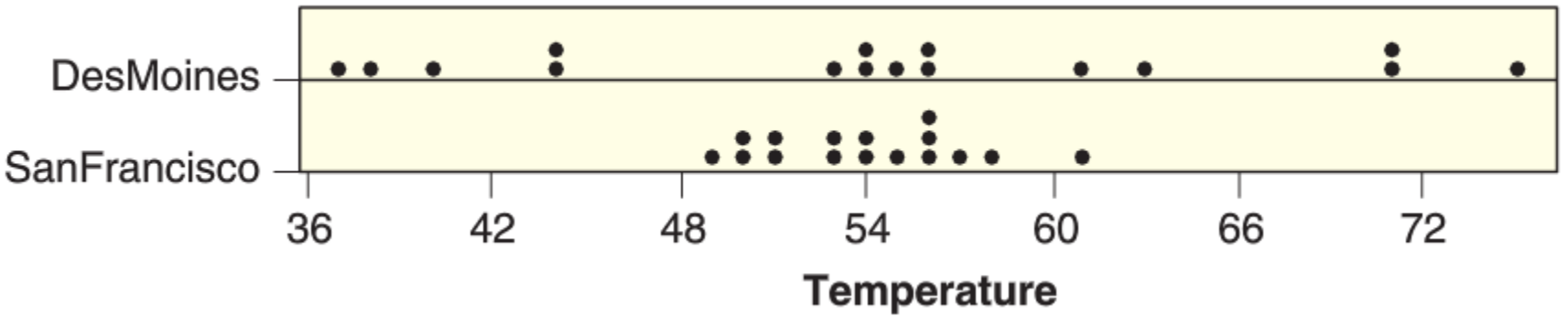
Population and sample standard deviation







Which has the larger standard deviation?



s_{DM} = 11.73 °F s_{SF} = 3.38 °F

The variance and standard deviation

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

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

R: `var(v)`

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

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

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

R: `sd(v)`

Example: computing the variance

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:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad = \tfrac{1}{4} \cdot (x_1 + x_2 + x_3 + x_4) \quad = \tfrac{1}{4} \cdot (8 + 2 + 6 + 4) \quad = 5$$

The variance can be computed using the formula:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \quad = \quad 1/3 \cdot ((8 - 5)^2 \; + \; (2 - 5)^2 \; + \; (6 - 5)^2 \; + \; (4- 5)^2 \;) \quad = \; 20/3$$

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

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

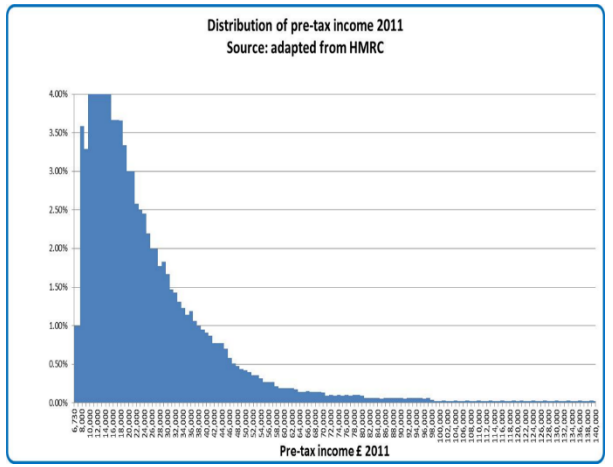
Homework 2: Calculate the standard deviation "by hand" for the number of hot dogs eaten by the winners.

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)

Income distribution



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 $\boldsymbol{\mu}$ (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** ($\boldsymbol{\sigma^2}$, $\mathbf{s^2}$) and the **standard deviation** ($\boldsymbol{\sigma}$, $\mathbf{s}$)

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)
```