

# **Big Data Analytics**

## **Session 2**

## **Basic Statistics**

# Where were we last week



- Introduction to Big Data Analysis
  - Big: 4V dimension of Data
  - Data: Turning data to data products
  - Analysis: Statistical learning
    - Why estimate  $f$ ?
    - How do we estimate  $f$ ?
    - The trade-off between prediction accuracy and model interpretability
    - Supervised vs. unsupervised learning
    - Regression vs. classification problems

# What is Statistics?



Main purpose of statistics, among others, is to

develop and apply methodology for

extracting useful knowledge from data.

# Statistical data analysis



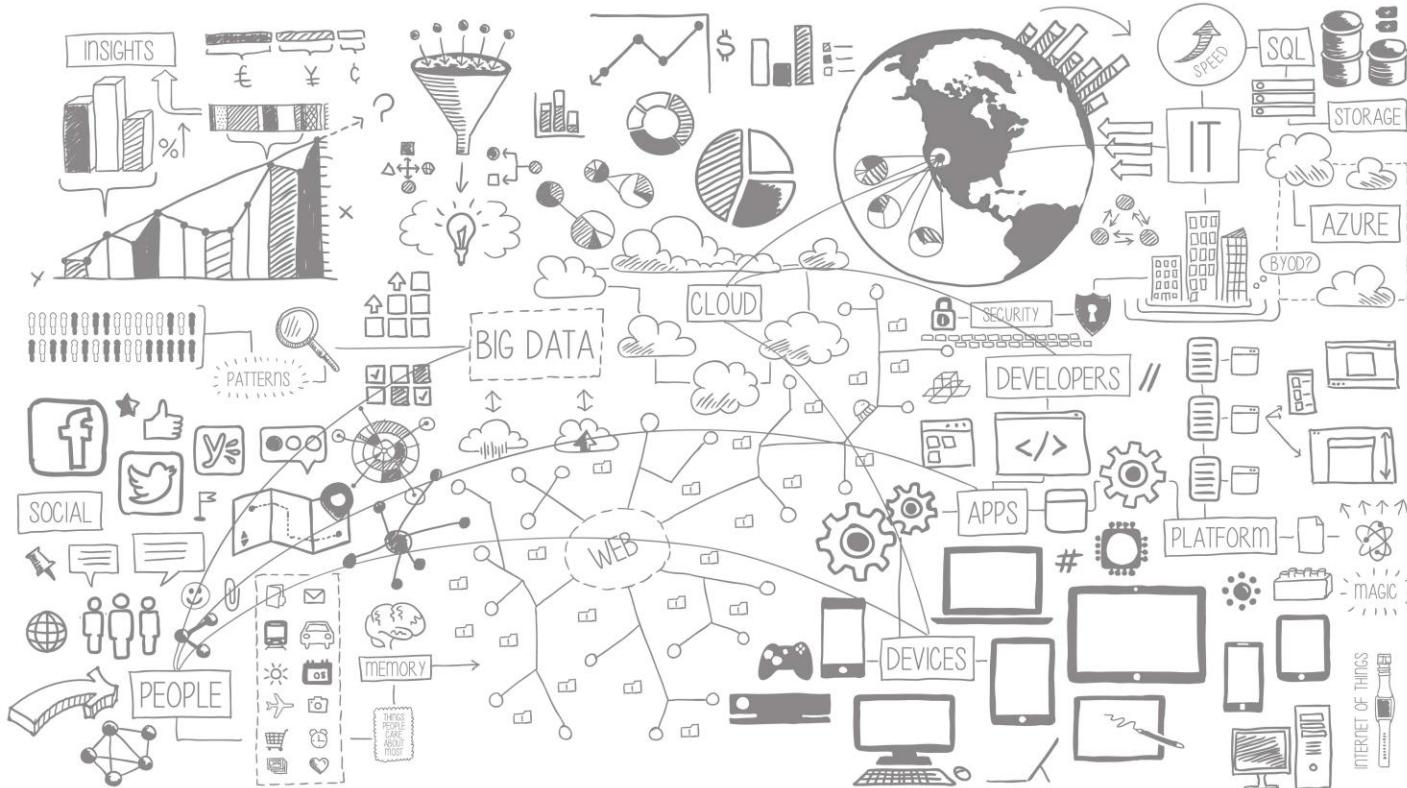
- Data
  - Nominal, Ordinal, Interval, and Ratio
- Descriptive statistics
  - Exploring, visualising, and summarising data without fitting the data to any models
- Inferential statistics
  - Identification of a suitable model
  - Testing either predictions or hypotheses of the model
  - Will be covered in the following sessions

# Statistical data analysis



- Data
  - Nominal, Ordinal, Interval, and Ratio
- Descriptive statistics
  - Exploring, visualizing, and summarizing data without fitting the data to any models
- Inferential statistics
  - Identification of a suitable model
  - Testing either predictions or hypotheses of the model
  - Will be covered in the following sessions

# Data



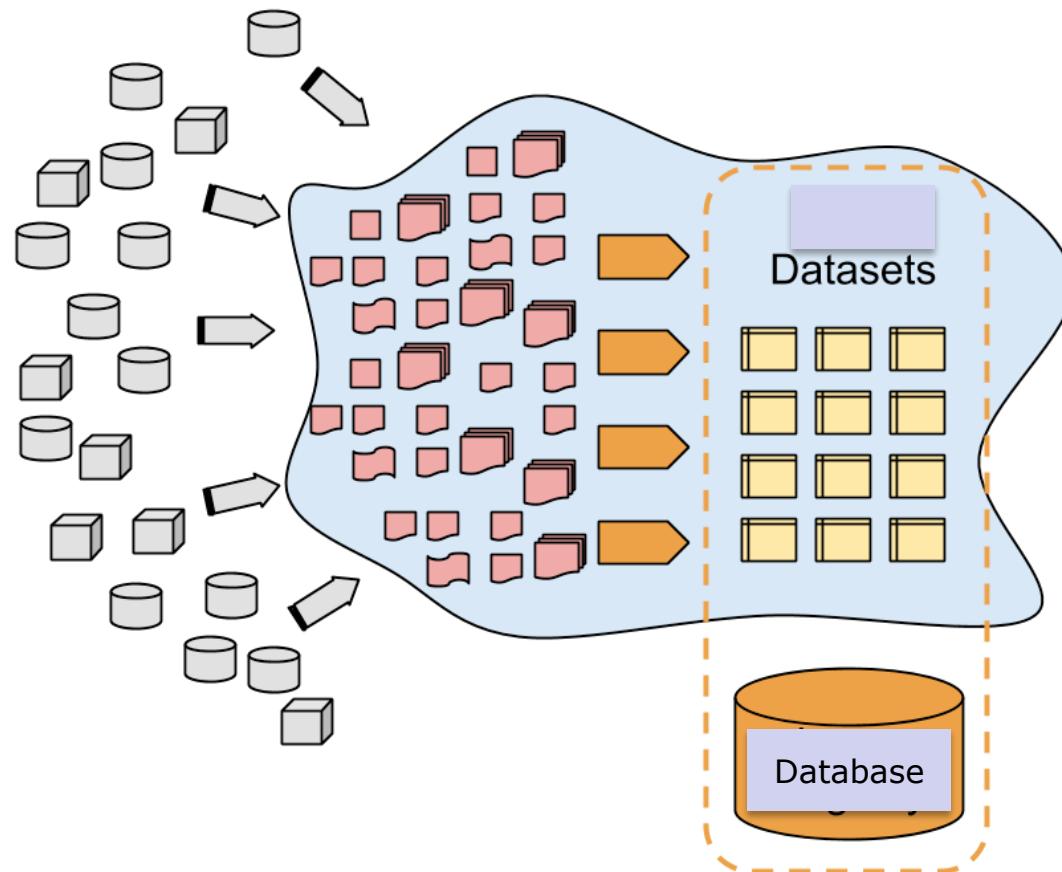
- **Data** are the **facts and figures** collected, summarised, analysed, and interpreted.

# Data



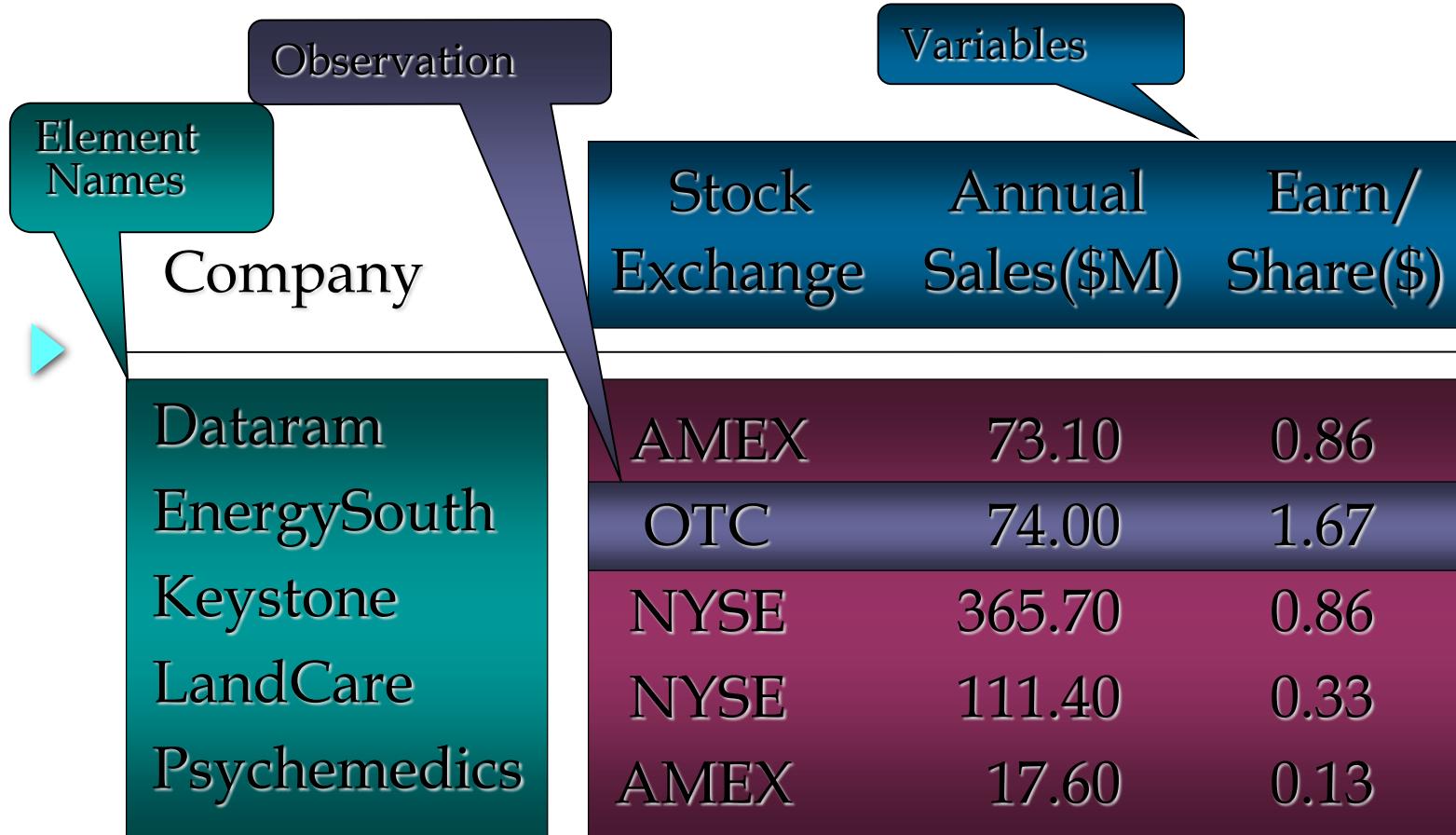
- **Data** are the results of measurements and can be the basis of graphs, images, or observations of a set of variables.

# Data Sets



- The data collected in a particular study are referred to as the **data set**.

# Data Sets



| Element Names | Observation    | Variables         |                 |
|---------------|----------------|-------------------|-----------------|
| Company       | Stock Exchange | Annual Sales(\$M) | Earn/ Share(\$) |
| Dataram       | AMEX           | 73.10             | 0.86            |
| EnergySouth   | OTC            | 74.00             | 1.67            |
| Keystone      | NYSE           | 365.70            | 0.86            |
| LandCare      | NYSE           | 111.40            | 0.33            |
| Psychomedics  | AMEX           | 17.60             | 0.13            |

Data Set

# Scales of Measurement



- Scales of measurement include:
  - Nominal
  - Ordinal
  - Interval
  - Ratio
- The scale determines the **amount of information** contained in the data.
- The scale indicates the **data summarisation** and **statistical analysis** that are most appropriate.

# Scales of Measurement



- Nominal: You cannot count them.
  - Data are **labels or names** to identify an attribute of the element
  - A non-numeric label or numeric code may be used
  - Example:
    - Naming Schools in a University: Education, Business, Humanities, etc
    - Alternatively, using numeric code: 1 for Education, 2 for Business, etc
- Ordinal: You can count and order, but not add or subtract them.
  - The ordinal type allows for **rank order** by which data can be sorted
  - But it still does **not** allow for **relative degree of difference** between them.
  - Example:
    - Measuring opinion: completely agree, mostly agree, mostly disagree, etc.
    - Alternatively, using numeric code: 1 for completely agree, 2 mostly agree, etc

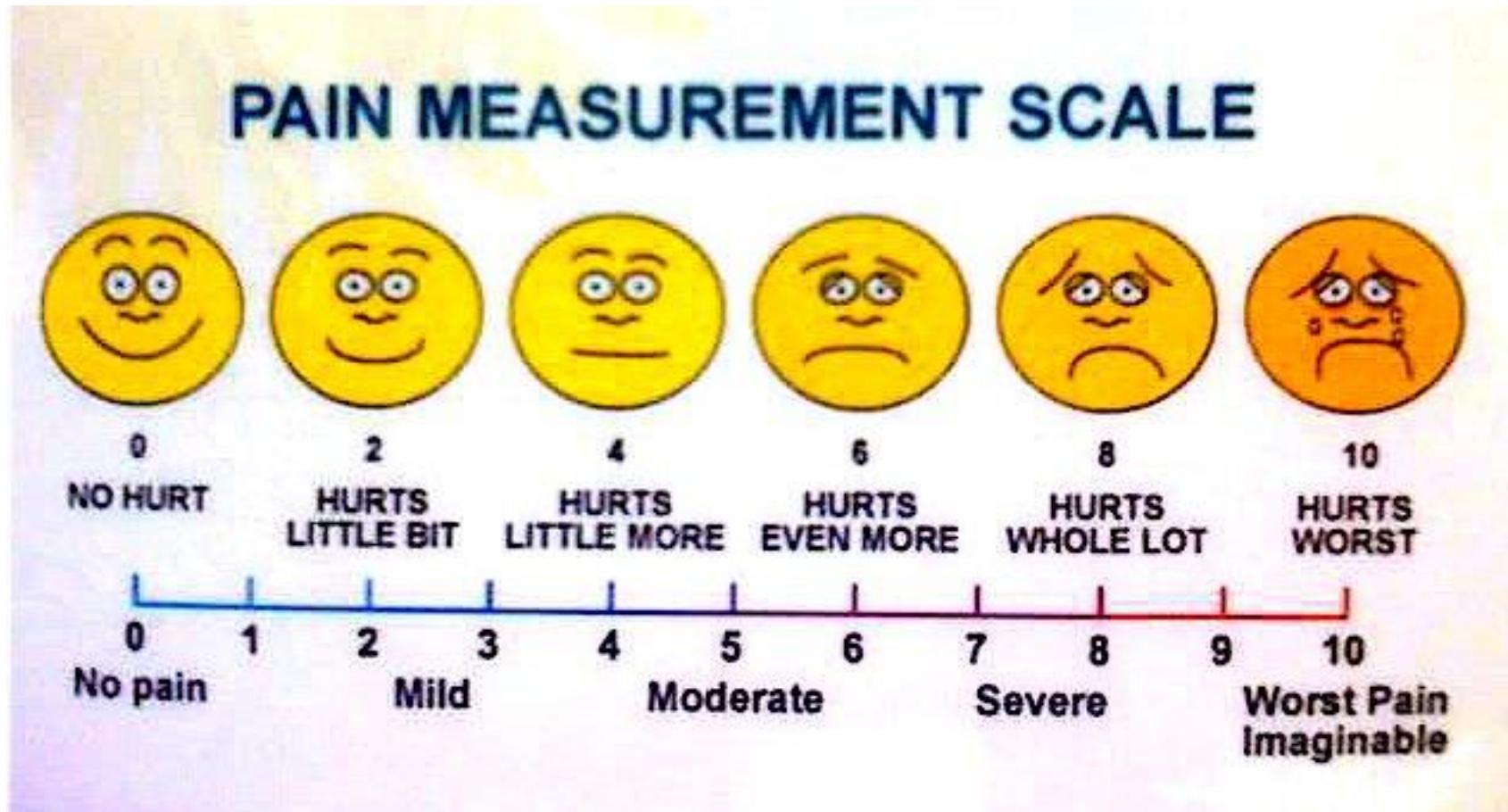
# Scales of Measurement



- Interval: can be added or subtracted, but not multiplied or divided
  - The interval type allows for the degree of difference between items, but not the ratio between them.
  - Interval data are always numeric.
  - Example:
    - Measuring temp.:  $20^{\circ}\text{C}$ ,  $10^{\circ}\text{C}$ . We cannot say  $20^{\circ}\text{C}$  is twice as hot as  $10^{\circ}\text{C}$ .
- Ratio: can be multiplied or divided, has zero value
  - The ratio of two values is meaningful.
  - Variables such as distance, height, weight and time use the ratio scale.
  - A ratio scale possesses a unique and non-arbitrary zero value.
  - Example:
    - Measuring length: 10cm is twice as long as 5cm.

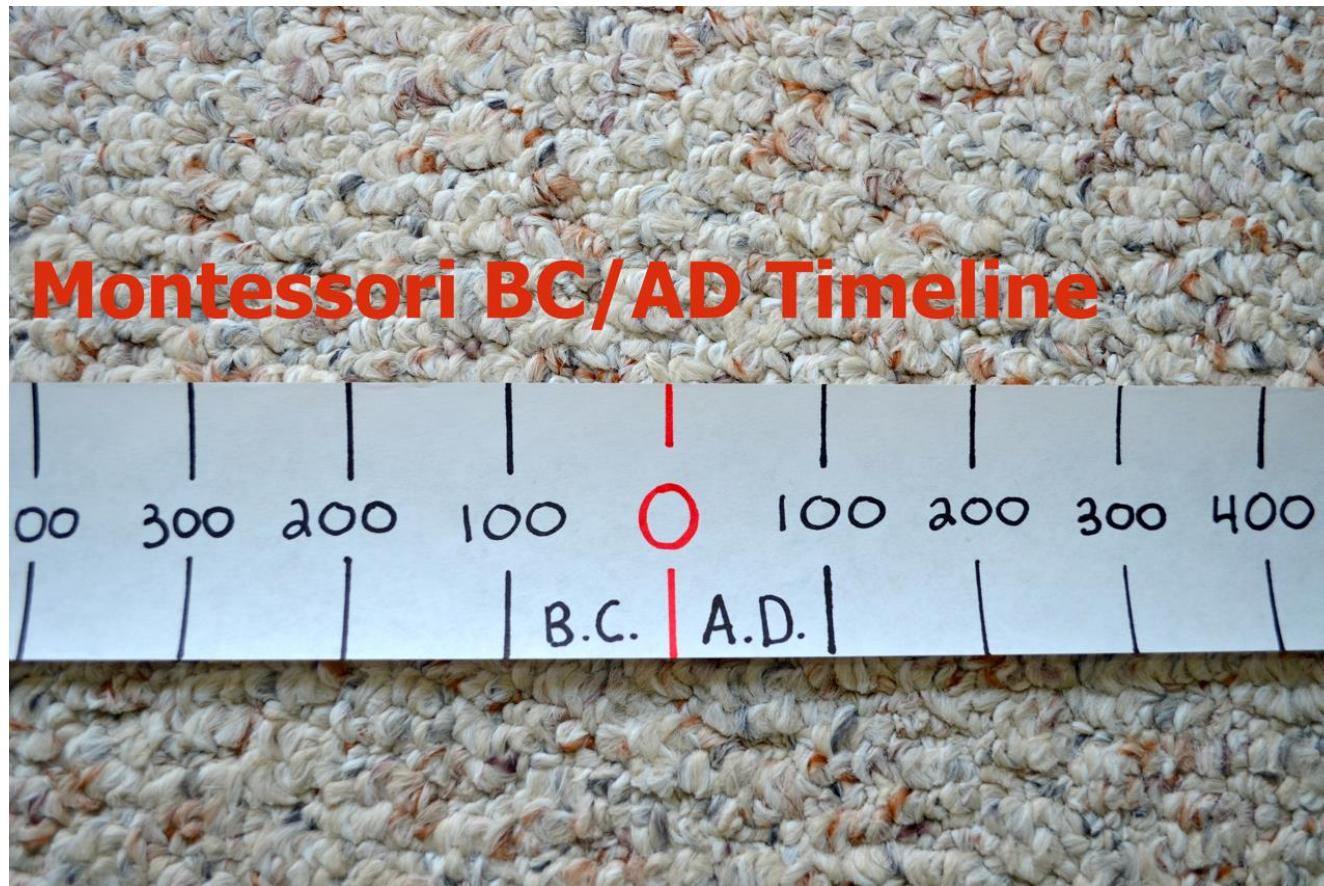
# Scales of Measurement

- Examples: pain measurement



# Scales of Measurement

- Examples: date



# Home Exercise



- Give your own examples of the following levels of measurements
  - Nominal
  - Ordinal
  - Interval
  - Ratio

# Qualitative and Quantitative Data



- Data can be further classified as being qualitative and quantitative.
- The **statistical analysis** that is appropriate **depends on** whether the data for the variable are **qualitative or quantitative**.
  - Qualitative data → qualitative analysis
  - Quantitative data → quantitative analysis
- In general, there are more alternatives for statistical analysis when the data are quantitative.

# Qualitative vs. Quantitative Data



I can give 9.5/10 for her looks,  
9/10 for her choice of dress and  
10/10 for her dressing sense !!



Quantitative Analysis !!  
(Purely "Objective")

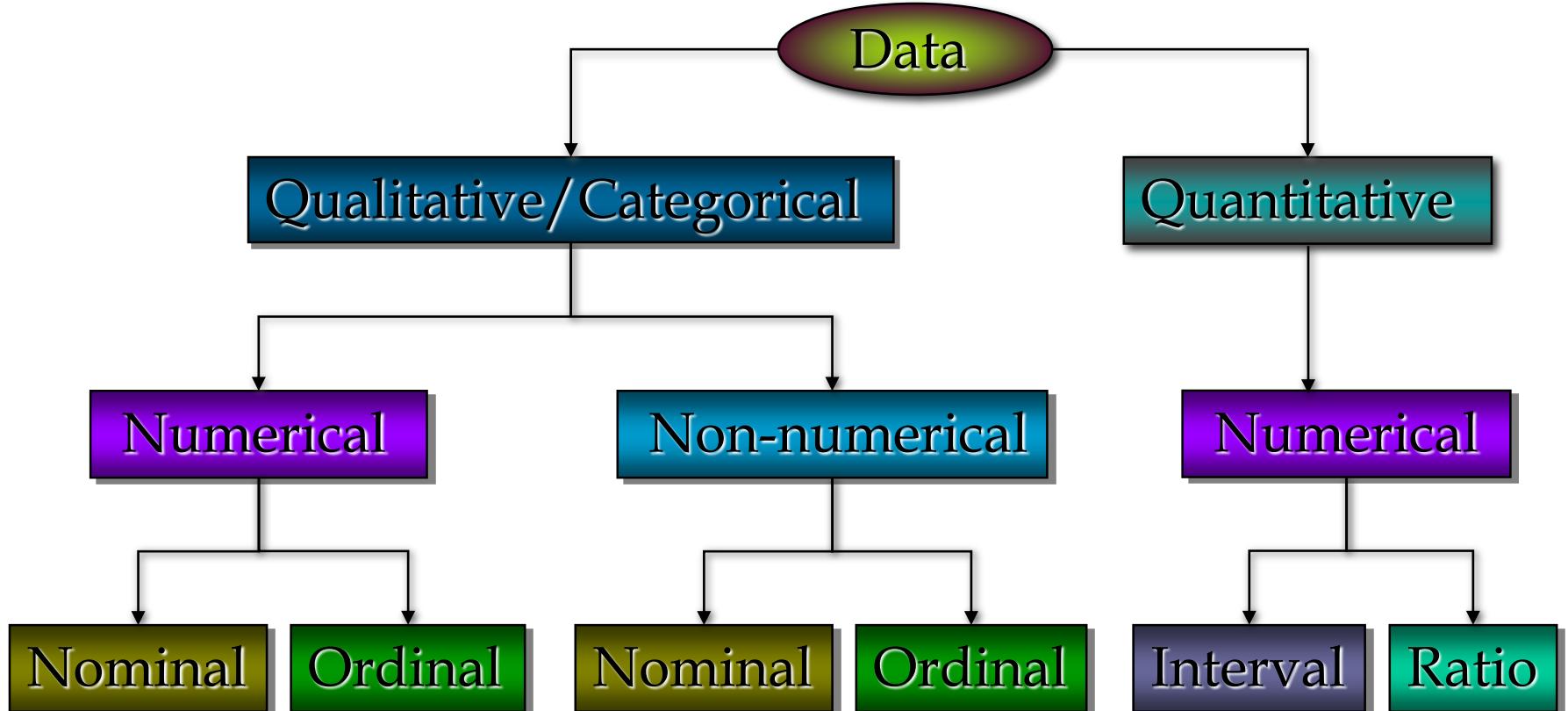
She looks 'beautiful'..  
her dress is 'awesome'.. and  
She dresses 'very well' !!

WWW.MYBIZMASTER.COM



Qualitative Analysis !!  
( Purely "Subjective" )

# Data



1 (Education), 1 (Completely agree),  
2 (Business), 2 (Mostly agree),  
...

Education,  
Business,  
...

Completely agree,  
Mostly agree,  
...

10°C, 20°C  
10cm, 20cm  
...

# Statistical data analysis



- Data
  - Nominal, Ordinal, Interval, and Ratio
- Descriptive statistics
  - Exploring, visualising, and summarising data without fitting the data to any models
- Inferential statistics
  - Identification of a suitable model
  - Testing either predictions or hypotheses of the model
  - Will be covered in the following sessions

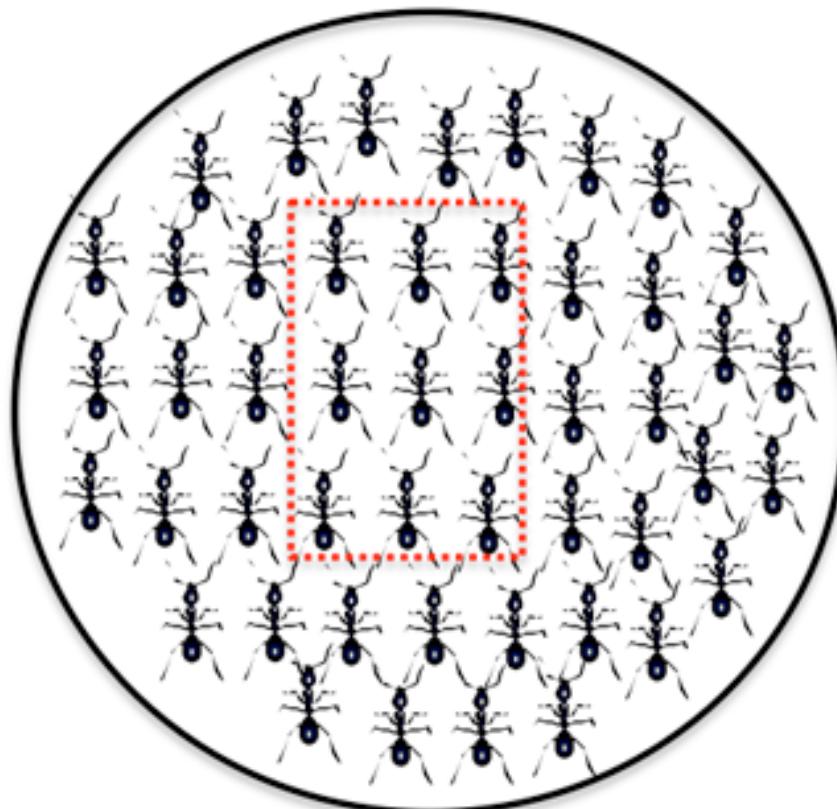
# Descriptive Statistics



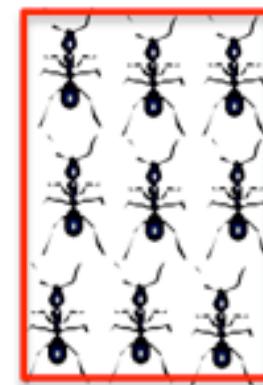
- Numerical measures
- Tabular and graphical presentation
  - Frequency distribution table
  - Histogram
  - Box plot
  - Scatter diagram

# Sample and Population

Population ( $N$ )



Sample ( $n$ )



# Numerical Measures



- If the measures are computed for data from a sample, they are called sample statistics.
- If the measures are computed for data from a population, they are called population parameters.
- A sample statistic is referred to as the **point estimator** of the corresponding population parameter.

# Descriptive Analysis



- **Univariate analysis:** describing the distribution of a **single variable**
  - Measures of central tendency
    - Mean, Median, Mode
  - Measures of spread
    - Variance, Standard Deviation
  - Measures of dispersion
    - Range, Quartiles, Interquartile Range
- **Bivariate analysis:** describing the relationship between **pairs of variables**
  - Quantitative measures of dependence
    - Correlation, Covariance

# Descriptive Analysis



- **Univariate analysis:** describing the distribution of a single variable
  - Measures of central tendency
    - Mean, Median, Mode
  - Measures of spread
    - Variance, Standard Deviation
  - Measures of dispersion
    - Range, Quartiles, Interquartile Range
- **Bivariate analysis:** describing the relationship between pairs of variables
  - Quantitative measures of dependence
    - Correlation, Covariance

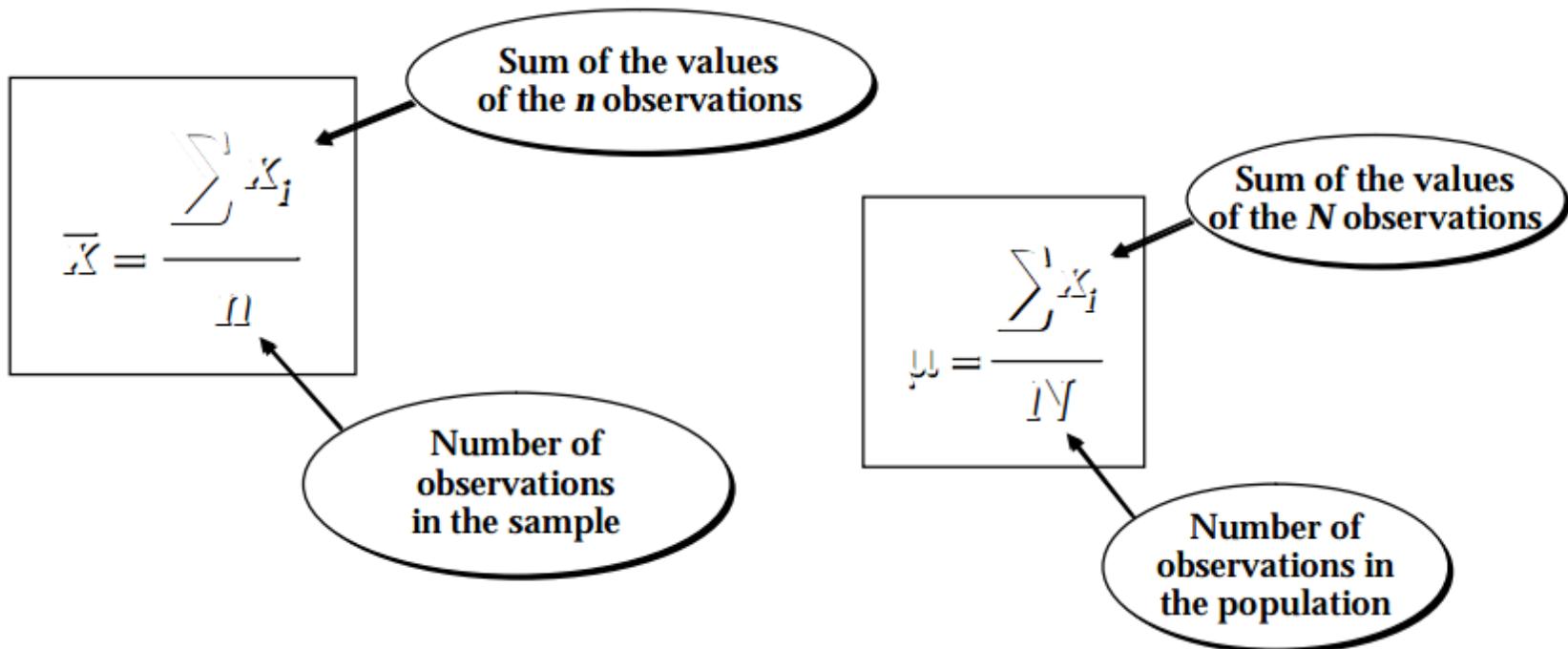
# Measures of Central Tendency



- It identifies the central position within a set of data.
- As such, measures of central tendency are sometimes called measures of central location.
  - Mean
  - Median
  - Mode

# Mean

- The **mean** of a data set is the arithmetic average of all the data values.
- The sample mean  $\bar{x}$  is the **point estimator** of the population mean  $\mu$
- The sample mean is a **statistic** and the population mean is a **parameter**



# Median

- The **median** is the middle observation in a group of data when the data are ranked in order of magnitude
  - Odd number of observations: **the middle one**

(11)

|    |    |    |    |    |    |    |    |    |    |    |
|----|----|----|----|----|----|----|----|----|----|----|
| 65 | 55 | 89 | 56 | 35 | 14 | 56 | 55 | 87 | 45 | 92 |
| 14 | 35 | 45 | 55 | 55 | 56 | 56 | 65 | 87 | 89 | 92 |

Median = 56

- Even number of observations: **the average of the middle two**

(10)

|    |    |    |    |    |    |    |    |    |    |
|----|----|----|----|----|----|----|----|----|----|
| 65 | 55 | 89 | 56 | 35 | 14 | 56 | 55 | 87 | 45 |
| 14 | 35 | 45 | 55 | 55 | 56 | 56 | 65 | 87 | 89 |

Median =  $(55+56)/2 = 55.5$

# Mean or Median?



- Consider data set with an outlier (extreme value)
  - Example: graduate salary

27K, 29K, 33K, 34K, 35K, 39K, **500K** (an outlier)

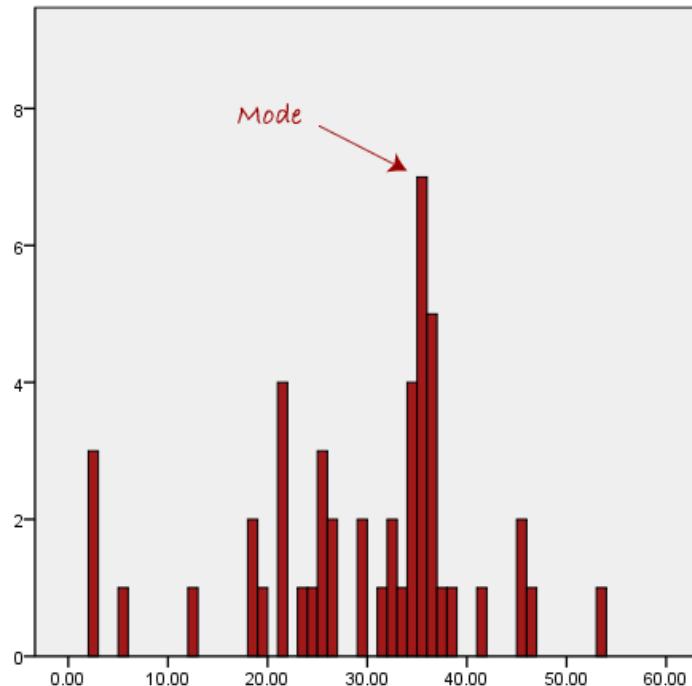
Mean = 99.6K, Median = 34K

Which is better as a representative of the central location?

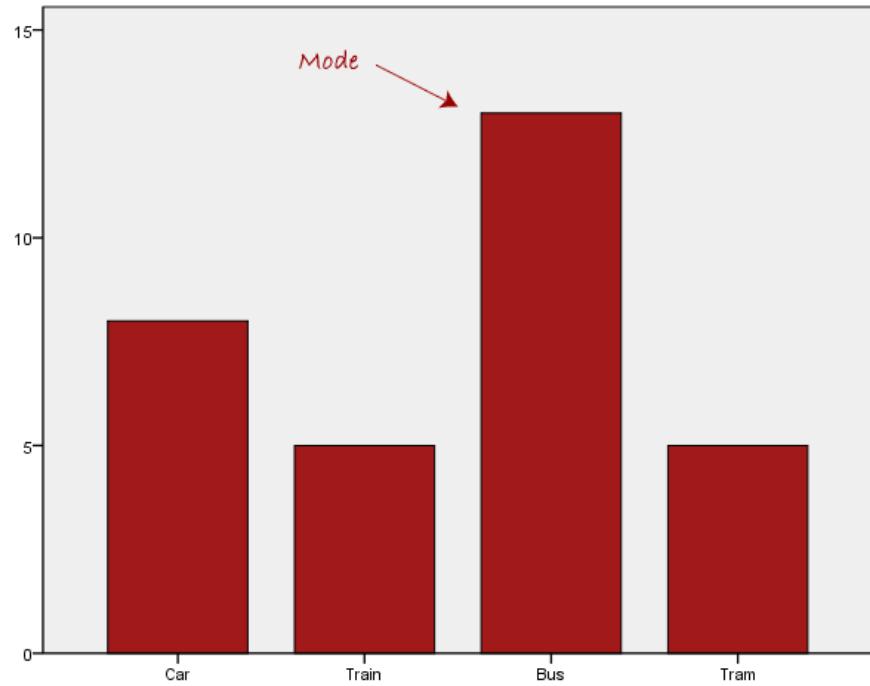
- Mean is highly influenced by one or two oddly high or low values.
- Whenever a data set has extreme values, the median is the preferred measure of central location.

# Mode

- The Mode is the most frequent value in a data set
  - The mode describes the most popular option (categorical data)



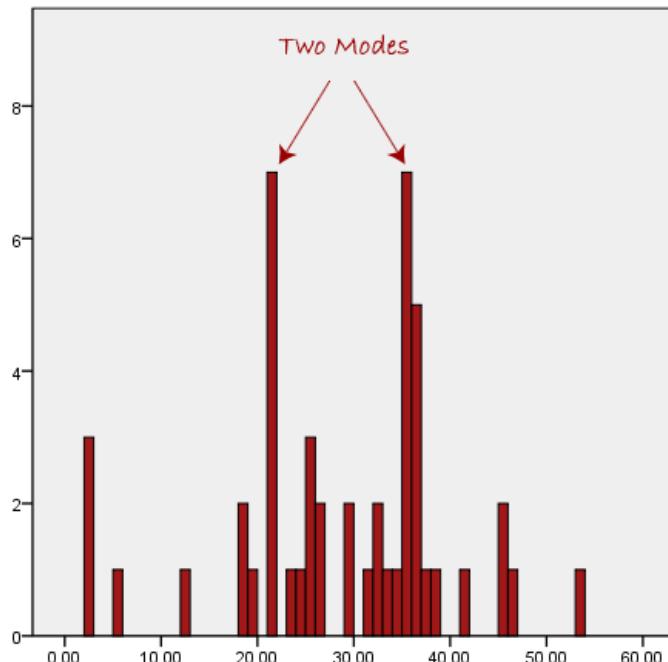
Highest bar in a [Histogram](#)



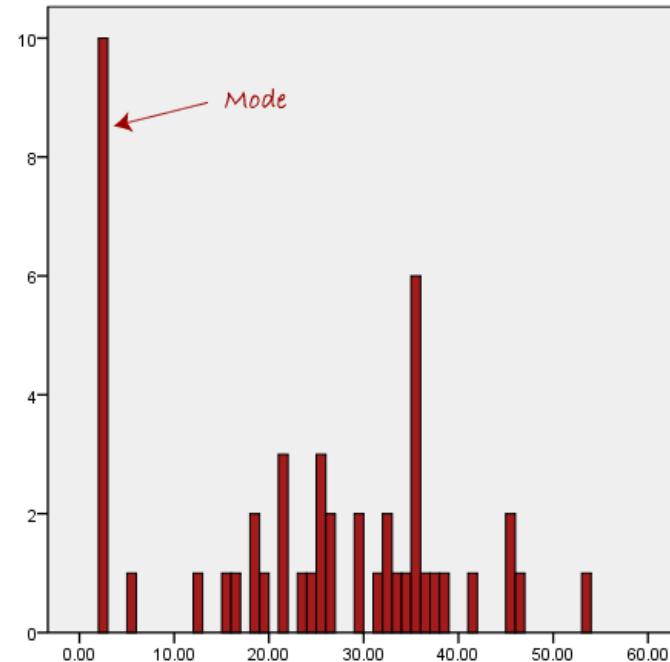
Bus being the most popular means of transportation

# Mode

- The mode is rare in the continuous data
- A data set might have more than one mode
- The limitation of using the mode



Two modes, which is more representative?



The mode that is far away from the rest of the data

# Mean, Median and Mode - Exercise

- What is the mean, median and mode for weight and height?

| Observations | variables |        |
|--------------|-----------|--------|
|              | weight    | height |
| student1     | 145       | 170    |
| student2     | 170       | 190    |
| student3     | 155       | 172    |
| student4     | 122       | 180    |
| student5     | 167       | 187    |
| student6     | 160       | 174    |
| student7     | 143       | 174    |
| student8     | 142       | 166    |
| student9     | 139       | 164    |
| Student10    | 165       | 182    |

- Consider height

- Mean

$$(170+190+172+\dots+182)/10 = 175.9$$

- Median

164 166 170 172 174 174 180 182 187 190

- Mode

174

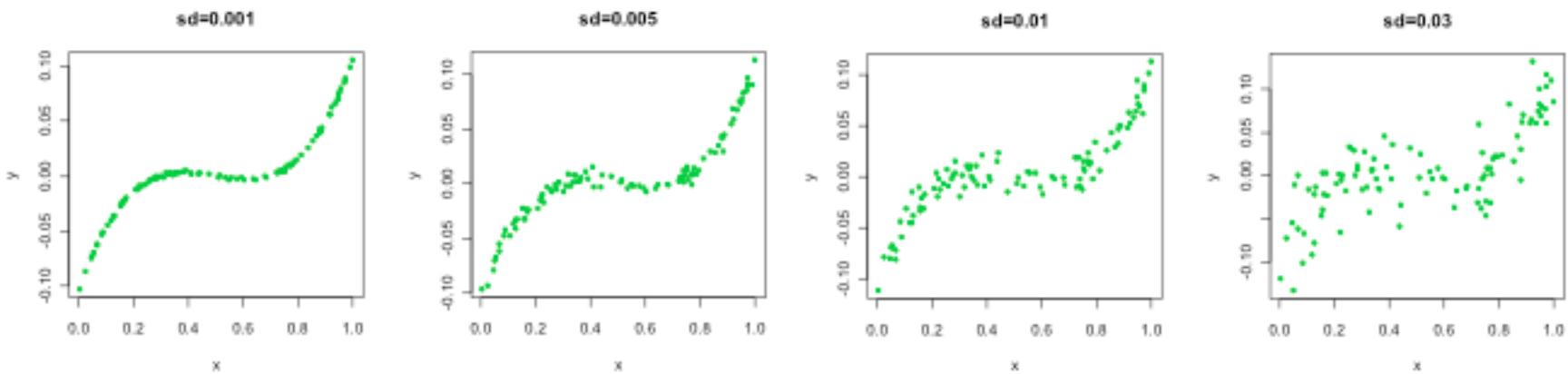
# Descriptive Analysis



- Univariate analysis: describing the distribution of a single variable
  - Measures of central tendency
    - Mean, Median, Mode
  - Measures of spread
    - Variance, Standard Deviation
  - Measures of dispersion
    - Range, Quartiles, Interquartile Range
- Bivariate analysis: describing the relationship between pairs of variables
  - Quantitative measures of dependence
    - Correlation, Covariance

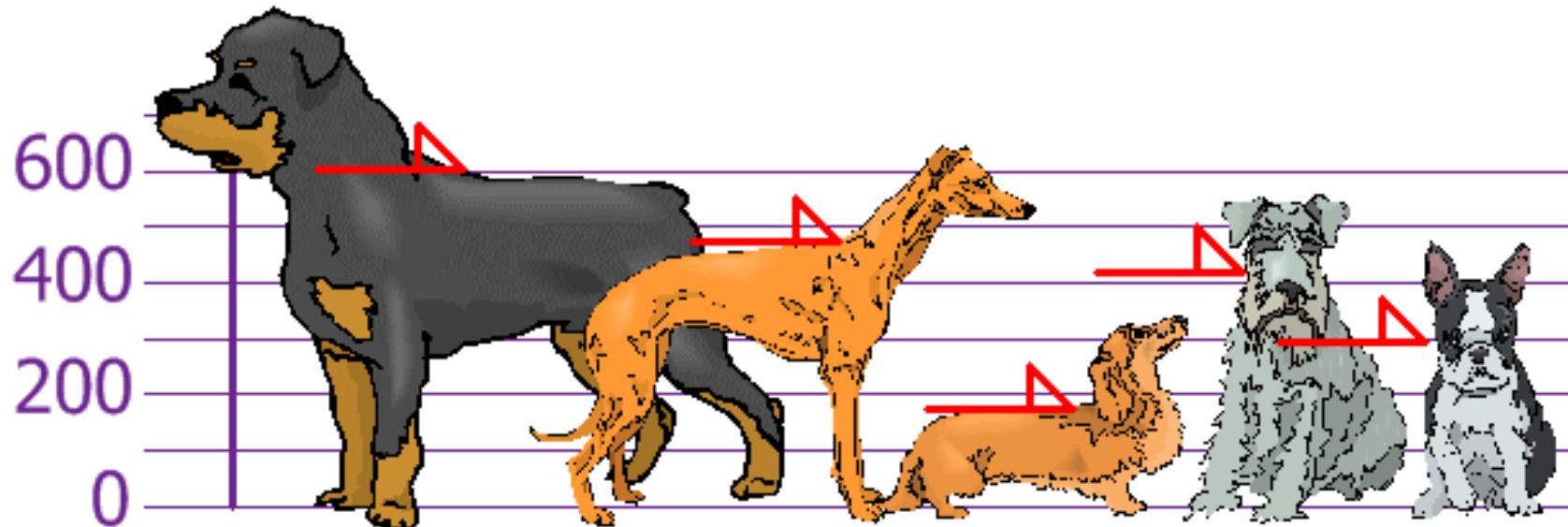
# Measures of spread

- It tells us how spread out numbers are.
  - Variance ( $s^2(\text{sample})$ ;  $\sigma^2(\text{population})$ )
    - The average of the squared differences from the mean
  - Standard Deviation ( $s(\text{sample})$ ;  $\sigma(\text{population})$ )
    - The square root of variance



# Example – variance & standard deviation

You and your friends have just measured the heights of your dogs (in millimeters):



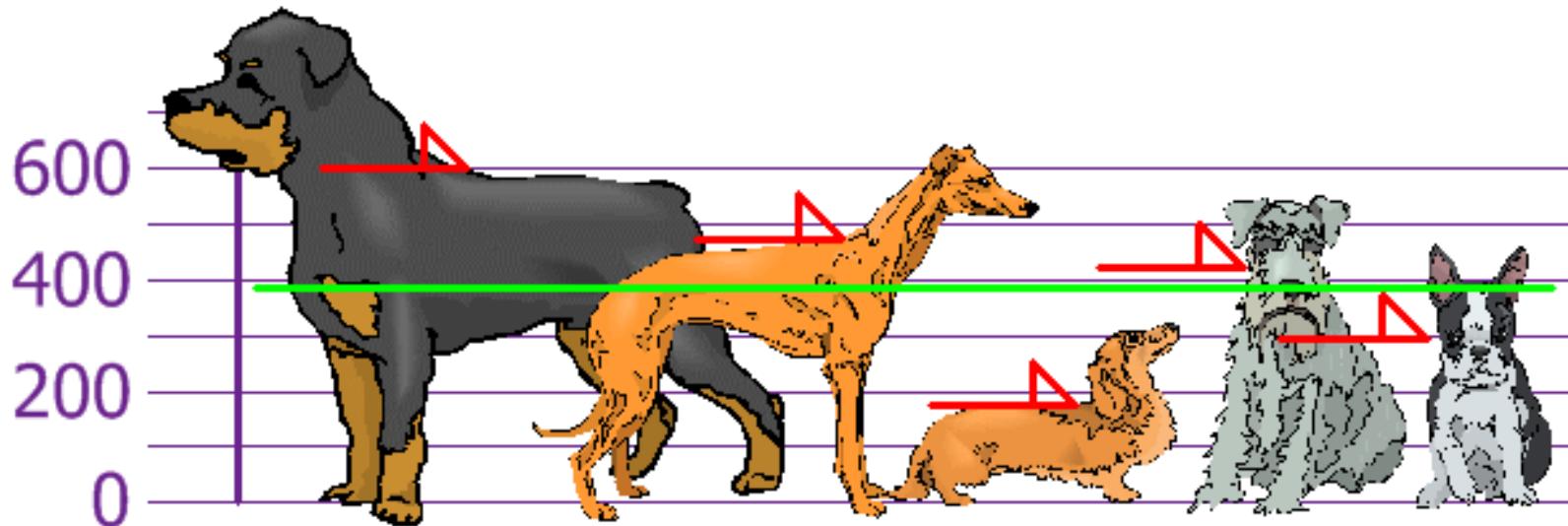
The heights (at the shoulders) are: 600mm, 470mm, 170mm, 430mm and 300mm.

Find out the Mean, the Variance, and the Standard Deviation.

# Example – variance & standard deviation

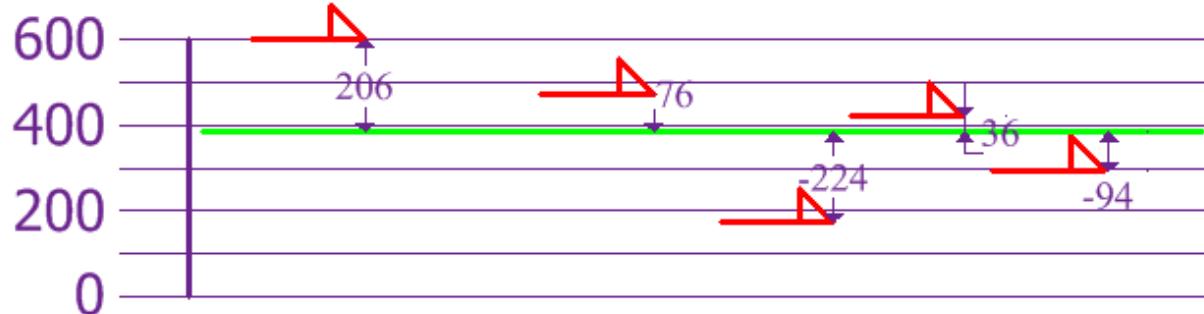
$$\text{Mean} = \frac{600 + 470 + 170 + 430 + 300}{5} = \frac{1970}{5} = 394$$

so the mean (average) height is 394 mm. Let's plot this on the chart:



# Example – variance & standard deviation

Now, we calculate each dogs difference from the Mean:



To calculate the Variance, take each difference, square it, and then average the result:

$$\begin{aligned}\text{Variance: } \sigma^2 &= \frac{206^2 + 76^2 + (-224)^2 + 36^2 + (-94)^2}{5} \\ &= \frac{42,436 + 5,776 + 50,176 + 1,296 + 8,836}{5} \\ &= \frac{108,520}{5} = 21,704\end{aligned}$$

So, the Variance is **21,704**.

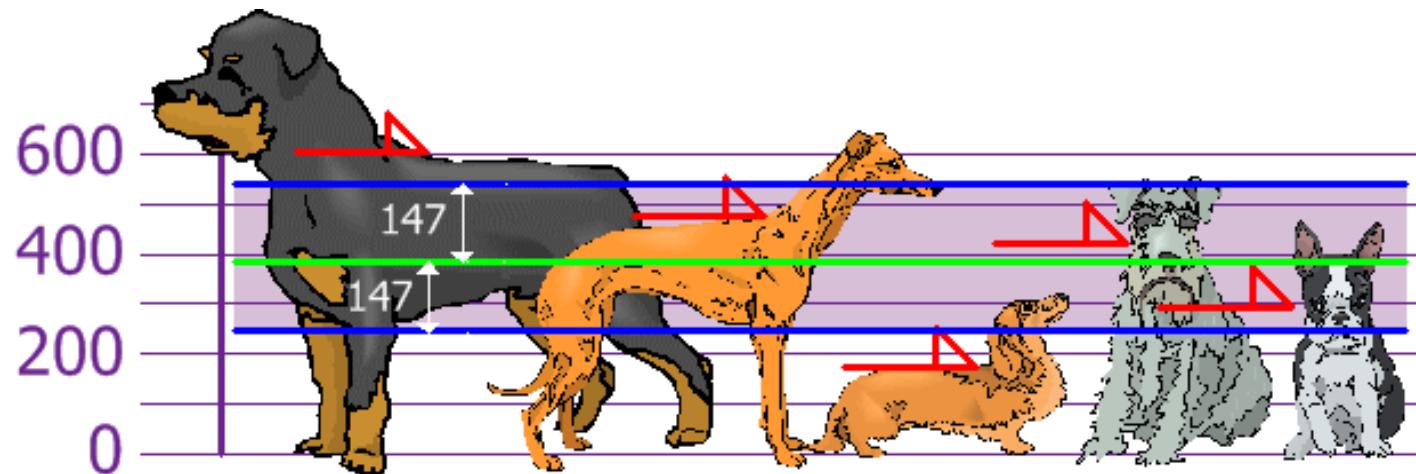
Note that SD has the same unit as mean

And the Standard Deviation is just the square root of Variance, so:

Standard Deviation:  $\sigma = \sqrt{21,704} = 147.32\dots = 147$  (to the nearest mm)

# Standard Deviation

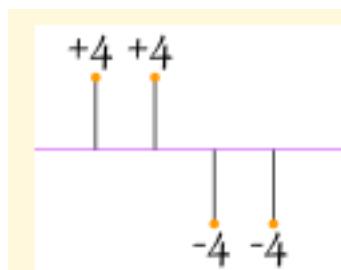
- Why study standard deviation (SD)?
  - From the point of view of one data:
    - A ‘standard’ way of knowing what is normal, and what is extra large or extra small.



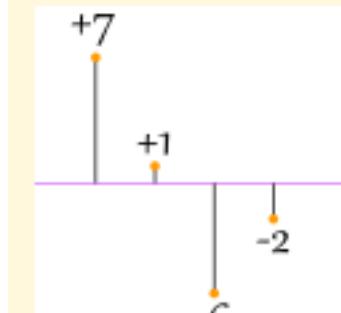
Rottweilers **are** tall dogs. And Dachshunds **are** a bit short ... but don't tell them!

# Standard Deviation

- Why study standard deviation (SD)?
  - From the point of view of the data set:
    - A low SD indicates that the data points tend to be very close to the mean;
    - A high SD indicates that the data points are spread out over a large range of values.



$$\sqrt{\frac{4^2 + 4^2 + 4^2 + 4^2}{4}} = \sqrt{\frac{64}{4}} = 4$$



$$\sqrt{\frac{7^2 + 1^2 + 6^2 + 2^2}{4}} = \sqrt{\frac{90}{4}} = 4.74\dots$$

# SD for Sample & Population



The "**Population** Standard Deviation":  $\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2}$

# SD for Sample & Population



The "**Population** Standard Deviation":  $\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2}$

The "**Sample** Standard Deviation":  $s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$

- Example:

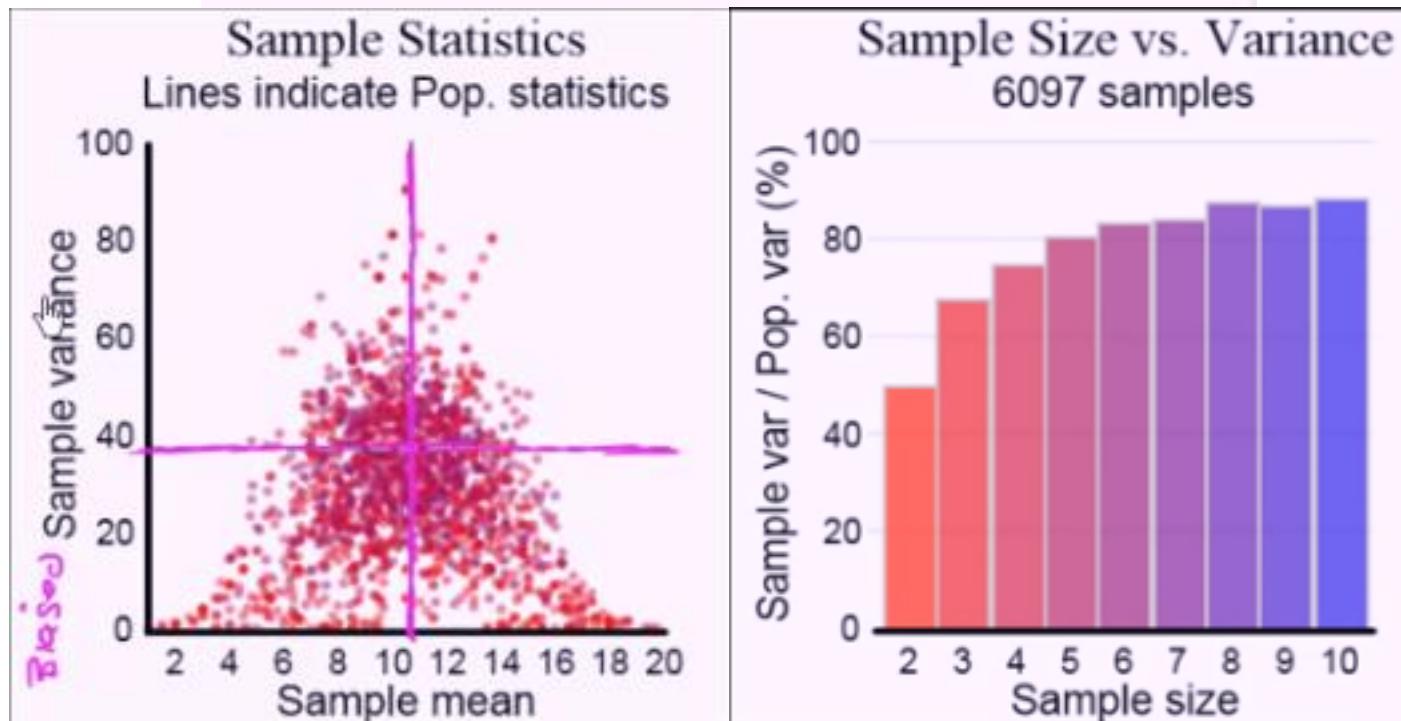
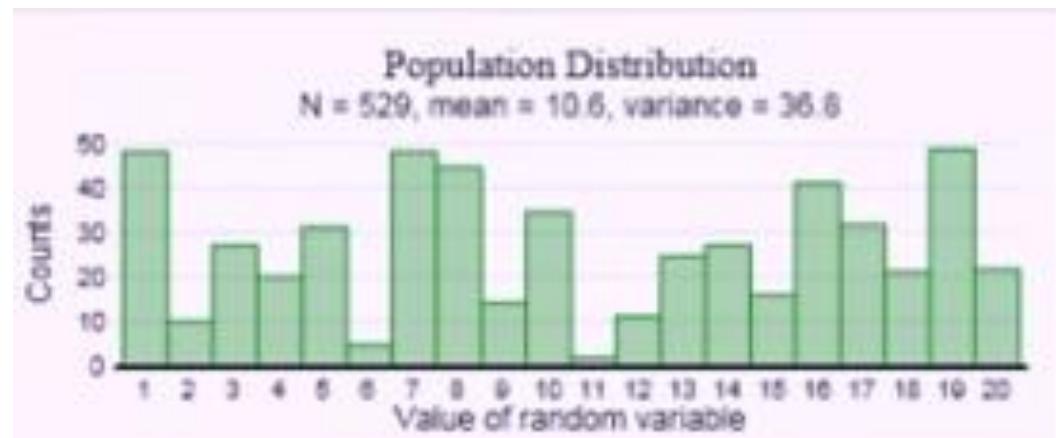
If our 5 dogs were just a **sample** of a bigger population of dogs, we would divide by **4 instead of 5** like this:

- Sample Variance =  $108,520 / 4 = 27,130$
- Sample Standard Deviation =  $\sqrt{27,130} \approx 164$

- Why divided by  $n-1$ ?

# Why divided by $n-1$ ?

Simulation showing bias in sample variance



# Descriptive Analysis - Exercise

- Compute the variance and standard deviation of height and weight

| Observations | variables |        |
|--------------|-----------|--------|
|              | weight    | height |
| student1     | 145       | 170    |
| student2     | 170       | 190    |
| student3     | 155       | 172    |
| student4     | 122       | 180    |
| student5     | 167       | 187    |
| student6     | 160       | 174    |
| student7     | 143       | 174    |
| student8     | 142       | 166    |
| student9     | 139       | 164    |
| Student10    | 165       | 182    |

- Height

- Mean

$$(170+190+172+\dots+182)/10 = 175.9$$

- Variance

$$[(170-175.9)^2 + (190-175.9)^2 + \dots + (182-175.9)^2]/10 = 67.29$$

- Standard deviation

$$\sqrt{67.29} \approx 8.20$$

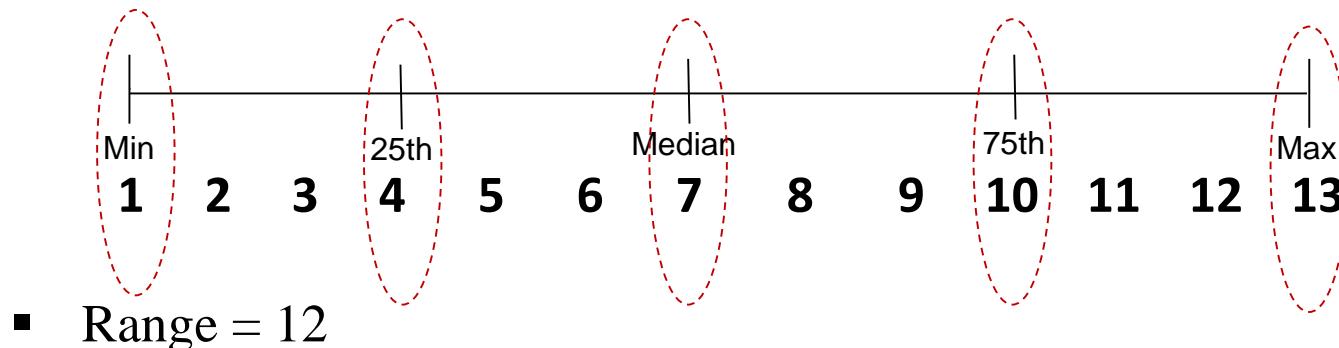
# Descriptive Analysis



- Univariate analysis: describing the distribution of a single variable
  - Measures of central tendency
    - Mean, Median, Mode
  - Measures of spread
    - Variance, Standard Deviation
  - Measures of dispersion (variability)
    - Range, Quartiles, Interquartile Range
- Bivariate analysis: describing the relationship between pairs of variables
  - Quantitative measures of dependence
    - Correlation, Covariance

# Measures of dispersion

- Range: difference between maximum and minimum value
  - Min: the lowest, or minimum value in variable
  - Max: the highest, or maximum value in variable
- Q1: the first (or 25<sup>th</sup>) quartile
- Q2: the second (or 50<sup>th</sup>) quartile – the Median
- Q3: the third (or 75<sup>th</sup>) quartile
- Box Plot



# Home Exercises



Use the following observations of variable  $x$  to find the values below.

10, 2, 15, 6, 4, 9, 12, 11, 3, 0, 12, 10, 9, 7, 11, 10, 8, 5, 10, 6

- Q1:  $n$  (number of observations)
- Q2: sum of  $x$ 's
- Q3:  $X$  (mean)
- Q4: Median and Mode
- Q5: Five number summary
  - Min, Q1, M, Q3, Max
- Q6:  $s^2$  (variance)
- Q7:  $s$  (standard deviation)

# Solution in R



```
x=c(10, 2, 15, 6, 4, 9, 12, 11, 3, 0, 12, 10, 9, 7, 11, 10, 8, 5, 10, 6)
```

```
Q1: length(x)
```

```
Q2: sum(x)
```

```
Q3: mean(x)
```

```
Q4 (mean): median(x)
```

```
Q5: summary(x)
```

```
Q6: var(x)
```

```
Q7: sd(x)
```

```
Q4 (mode): names(sort(-table(x)))[1]
```

```
sort(x)
```

```
table(x)
```

```
sort(-table(x))
```

# Descriptive Analysis



- **Univariate analysis:** describing the distribution of a single variable
  - Measures of central tendency
    - Mean, Median, Mode
  - Measures of spread
    - Variance, Standard Deviation
  - Measures of dispersion
    - Range, Quartiles, Interquartile Range
- **Bivariate analysis:** describing the relationship between pairs of variables
  - Quantitative measures of dependence
    - Correlation, Covariance

# Covariance and Correlation



- Variables may change in relation to each other.
- Both quantify relationship.
- Difference:
  - Covariance is a dimensional quantity
    - The value depends on the units of the data
    - ➔ difficult to compare covariances among data sets that have different scales.
  - Correlation is a dimensionless quantity
    - Always between -1 and 1
    - ➔ facilitates the comparison of different data sets

# Variance

- First recall: variance of one variable

| Case        | X         | X - Avg          | (X - Avg) <sup>2</sup> |
|-------------|-----------|------------------|------------------------|
| A           | 3         | -1               | 1                      |
| B           | 1         | -3               | 9                      |
| C           | 3         | -1               | 1                      |
| D           | 9         | 5                | 25                     |
| <b>Sum:</b> | <b>16</b> | <b>Sum:</b>      | <b>36</b>              |
| <b>Avg:</b> | <b>4</b>  | <b>Variance:</b> | <b>9</b>               |

$$\text{Variance} = \sum (x-\text{Avg})^2 / N = 36/4 = 9$$

- X: 4, 4, 4, 4; variance = 0
- X: 1, 1, 1, 13; variance =  $[-3]^2 + [-3]^2 + [-3]^2 + 9^2] / 4 = 108/4 = 27$

# Covariance

- Variance of one variable
- Covariance of two variables

| Case | X  | Y  | (X - Xavg) | (Y - YAvg) | Multiplied |                        |
|------|----|----|------------|------------|------------|------------------------|
| A    | 3  | 4  | -1         | -2         | 2          |                        |
| B    | 1  | 4  | -3         | -2         | 6          |                        |
| C    | 3  | 8  | -1         | 2          | -2         |                        |
| D    | 9  | 8  | 5          | 2          | 10         | covariance coefficient |
| Sum: | 16 | 24 |            |            | 16         |                        |
| Avg. | 4  | 6  |            |            | 4          |                        |

$$\text{Covariance} = \Sigma (X_i - X_{\text{avg}})(Y_i - Y_{\text{avg}}) / N = (2+6-2+10)/4 = 4$$

- X: 4, 4, 4, 4; covariance = 0
- X: 1, 1, 1, 13; covariance = 6
- X: 13, 1, 1, 1; covariance = -6

We write covariance of X and Y as  $\sigma_{XY}$

$\sigma_{XX}$  is the covariance of X with itself  
→ Variance of X:  $\sigma_X^2$

# Covariance

- Formally, covariance coefficient can be calculated as:

$$s_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

for  
samples

$$\sigma_{xy} = \frac{\sum (x_i - \mu_x)(y_i - \mu_y)}{N}$$

for  
populations

# Covariance



- Covariance measures how much the movement in one variable predicts the movement in a corresponding variable
  - Positive covariance indicates that higher than average values of one variable tend to be paired with higher than average values of the other variable.
  - Negative covariance indicates that higher than average values of one variable tend to be paired with lower than average values of the other variable.
- In other words, it measures the degree of linkage between two variables that covary.

# From Covariance to Correlation

- Covariance is a dimensional quantity
  - The value depends on the units of the data
  - difficult to compare covariances among data sets that have different scales.
- We need a dimensionless quantity to facilitate comparison → correlation
  - Always between -1 and 1
- The correlation of X and Y, denoted  $\rho_{XY}$ , is simply calculated as:

$$\text{correlation of X and Y} = \frac{\text{covariance of X and Y}}{\text{standard deviation of X} * \text{standard deviation of Y}}$$

$$\rho_{XY} = \sigma_{XY} / (\sigma_X * \sigma_Y)$$

# Correlation

$$\text{correlation of X and Y} = \frac{\text{covariance of X and Y}}{\text{standard deviation of X} * \text{standard deviation of Y}}$$

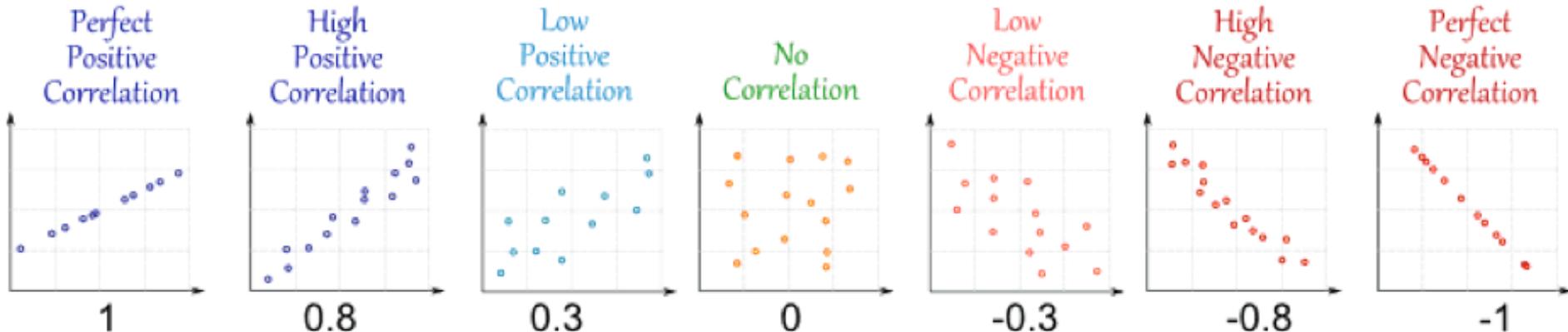
- Example:

| Case | X  | Y  | (X - Xavg) | (Y - YAvg) | Multiplied |
|------|----|----|------------|------------|------------|
| A    | 3  | 4  | -1         | -2         | 2          |
| B    | 1  | 4  | -3         | -2         | 6          |
| C    | 3  | 8  | -1         | 2          | -2         |
| D    | 9  | 8  | 5          | 2          | 10         |
| Sum: | 16 | 24 |            | Sum:       | 16         |
| Avg. | 4  | 6  |            | Avg:       | 4          |

- SD of X:  $\sigma_X=3$
- SD of Y:  $\sigma_Y=2$
- Covariance of X and Y:  $\sigma_{XY}=4$
- Correlation of X and Y:  $\rho_{XY}=4/(2*3) = 0.67$

# Correlation Coefficient

- When the two sets of data are strongly linked together, we say they have a **high correlation**.
- Correlation is **positive** when the values increase together, and
- Correlation is **negative** when one value decreases as the other increase.
- The coefficient can take on values between -1 and +1.
- Values **near +1** indicate a **strong positive** linear relationship.
- Values **near -1** indicate a **strong negative** linear relationship.



# Correlation Coefficient

- The correlation coefficient is computed as follows:

▷ 
$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

**for  
samples**

◁ 
$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

**for  
populations**

- It is obtained by dividing the covariance of the two variables by the product of their standard deviations.

- The functions are most straightforward in R

`x=c(3,1,3,9)`

`mean(x)`

`median(x)`

`var(x)`

`sd(x)`

`y=c(4,4,8,8)`

`cov(x,y)`

`cor(x,y)`

- Question: what is the `sd(x)` function computing? Unbiased or biased SD? How to calculate the other one?

# Correlation and Causation

- Correlation is a measure of linear association and not necessarily causation.
- Just because two variables are highly correlated, it does not mean that one variable is the cause of the other.



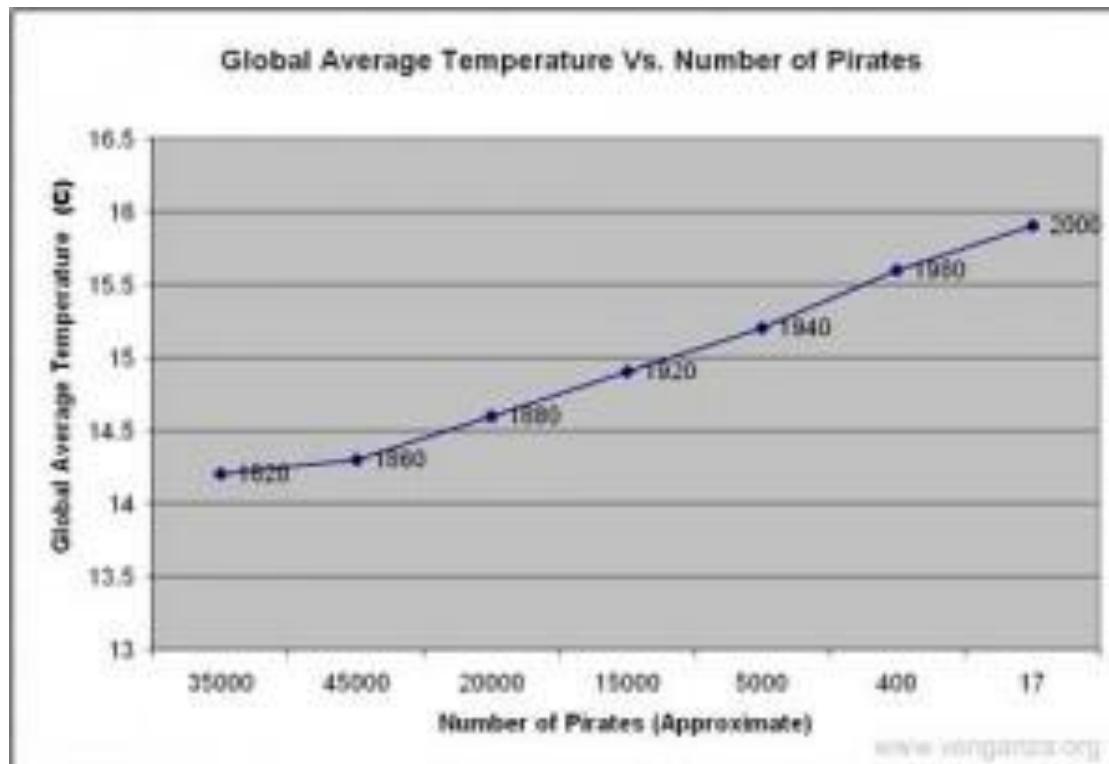
# Example

- Ice cream and homicide rates are positively correlated
- Do they have a causal relationship?
  - Does ice cream consumption turn harmless Joe into a murderous monster?



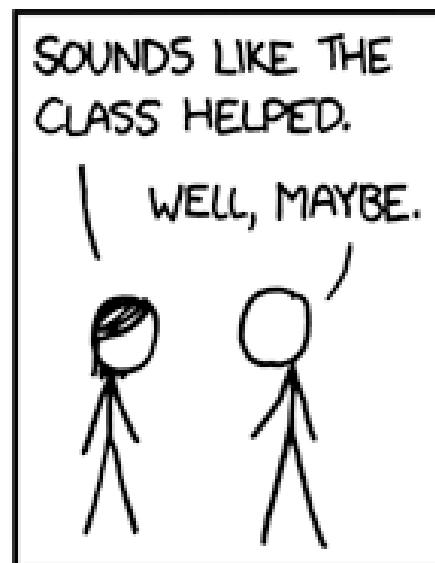
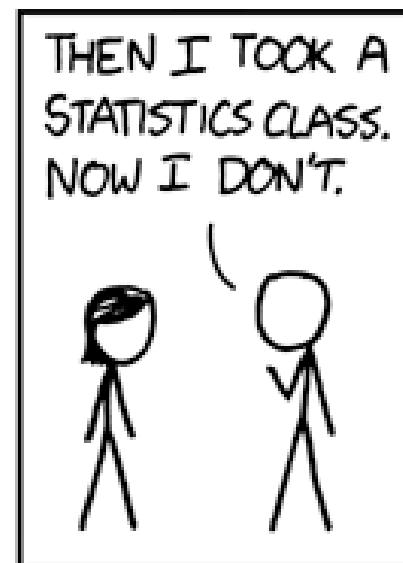
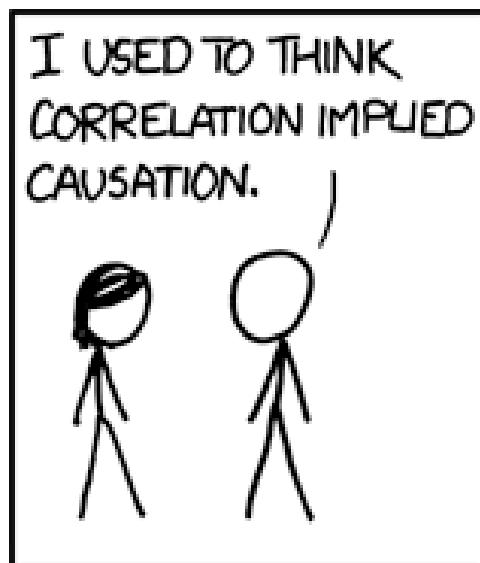
# Example

- An increase in both global temperature and number of pirates
  - That's a positive correlative relationship
- Do the pirates cause the global warming?



# Reflection

- Every correlation you have ever heard of can be questioned in your own mind.
  - Is there a cause and effect here, or
  - Is it just coincidence?
  - How are the two factors really related?



# Summary

|                    | Population (parameter)                                    | Sample (statistic)   |
|--------------------|---|--|
| mean               | $\mu = \frac{\sum x_i}{N}$                                | $\bar{x} = \frac{\sum x_i}{n}$                             |
| Variance           | $\sigma^2 = \frac{\sum (x_i - \mu)^2}{N}$                 | $s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$                 |
| Standard deviation | $\sigma = \sqrt{\sigma^2}$                                | $s = \sqrt{s^2}$   |
| Covariance         | $\sigma_{xy} = \frac{\sum (x_i - \mu_x)(y_i - \mu_y)}{N}$ | $s_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1}$ |
| Correlation        | $\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$       | $r_{xy} = \frac{s_{xy}}{s_x s_y}$                          |

# Scale of Measurement



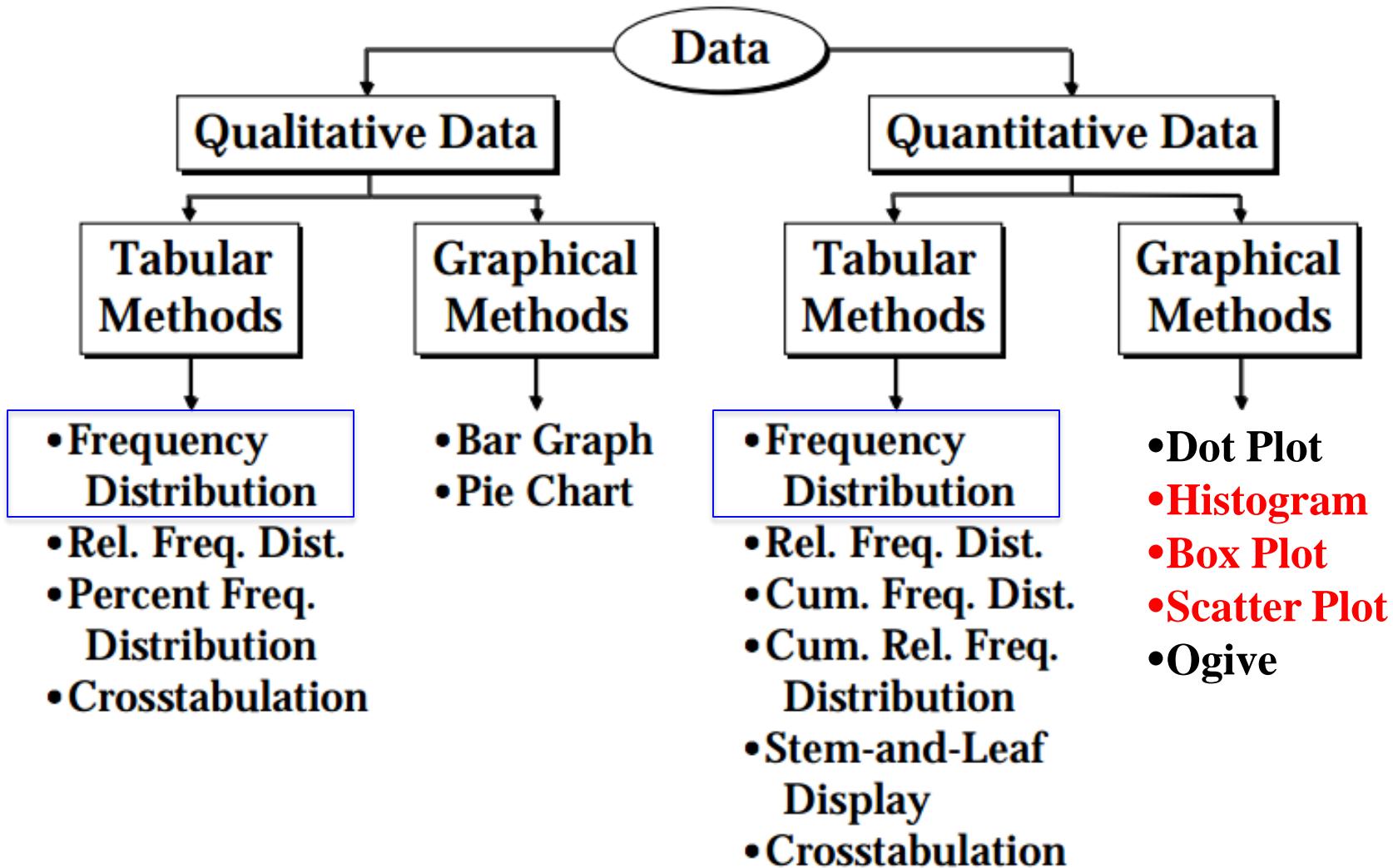
- Nominal
  - Mode
- Ordinal
  - Median, Mode
- Interval
  - Mode, Median, Mean
  - Range, Variance, Standard deviation
- Ratio
  - Mode, Median, Mean
  - Range, Variance, Standard deviation
  - And many more: geometric mean, harmonic mean, coefficient of variation, and all the other statistical measures

# Descriptive Statistics



- Numerical measures
- Tabular and graphical presentation
  - Frequency distribution
  - Histogram
  - Box plot
  - Scatter plot

# Tabular & Graphical Presentation



# Frequency distribution

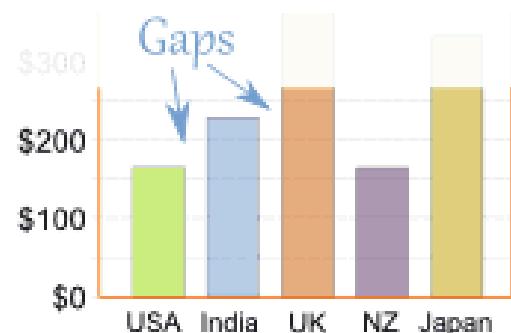
- A table that displays the frequency of various outcomes in a data set
  - Example:

Frequency Distribution for a Class of 25 M.B.A. Students

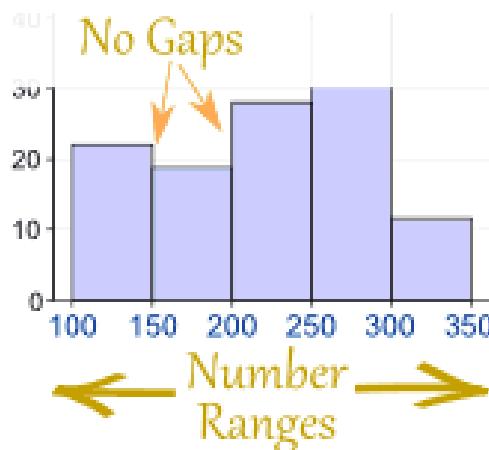
| Grade Scale    | Student/Grade Frequency | Relative Frequency |
|----------------|-------------------------|--------------------|
| A              | 5                       | 20%                |
| B              | 12                      | 48%                |
| C              | 4                       | 16%                |
| D              | 2                       | 8%                 |
| F              | 1                       | 4%                 |
| I (Incomplete) | 1                       | 4%                 |
| TOTAL          | 25                      | 100%               |

# Histogram

- A Histogram is a graphical display of data using bars of different heights.
- It is similar to a Bar Chart, but a histogram groups numbers into **ranges**. And you decide what ranges to use!
- Histograms are a great way to show results of continuous data, such as weight, height, how much time, etc.
- When the data is in **categories** (such as Country or Favorite Movie), we should use a Bar Chart.



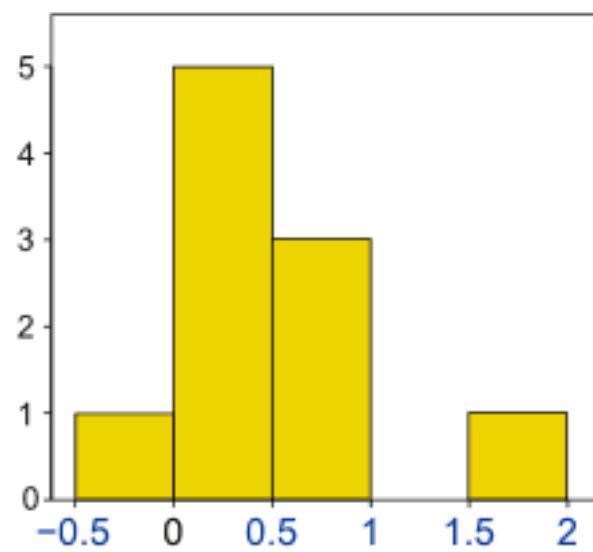
← Categories →  
**Bar Graph**



← Number Ranges →  
**Histogram**

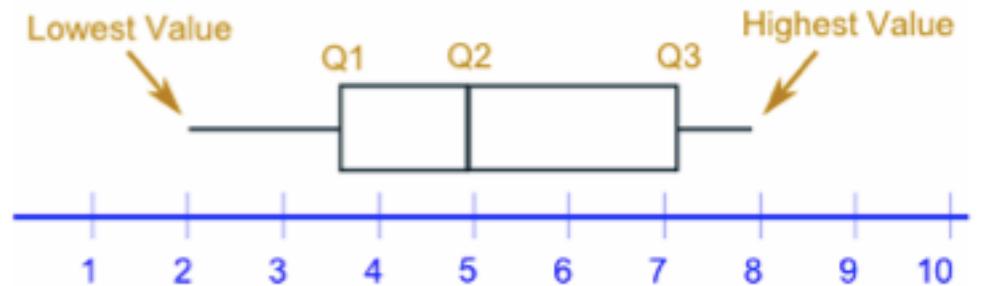
# Histogram

- Every month you measure how much weight your puppy has gained and get these results:  
 $0.5, 0.5, 0.3, -0.2, 1.6, 0, 0.1, 0.1, 0.6, 0.4$
- They vary from -0.2 (the puppy lost weight that month) to 1.6
- Put in order from lowest to highest weight gain:  
 $-0.2, 0, 0.1, 0.1, 0.3, 0.4, 0.5, 0.5, 0.6, 1.6$
- You decide to put the results into groups of 0.5:
  - The -0.5 to just below 0 range,
  - The 0 to just below 0.5 range,
  - etc ...



# Box Plot

- You can show all the important values in a “Box and Whisker Plot”, like this:



- Example: Box Plot and Interquartile Range for**

4, 17, 7, 14, 18, 12, 3, 16, 10, 4, 4, 11

- Put them in order:

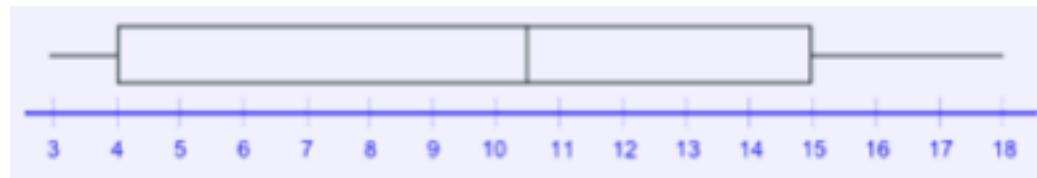
3, 4, 4, 4, 7, 10, 11, 12, 14, 16, 17, 18

- Cut it into quarters:

3, 4, 4 | 4, 7, 10 | 11, 12, 14 | 16, 17, 18

# Box Plot

- 3, 4, 4 | 4, 7, 10 | 11, 12, 14 | 16, 17, 18
- In this case all the quartiles are between numbers:
  - Quartile 1 (Q1) =  $(4+4)/2 = 4$
  - Quartile 2 (Q2) =  $(10+11)/2 = 10.5$
  - Quartile 3 (Q3) =  $(14+16)/2 = 15$
- Also:
  - The lowest value is 3,
  - The highest value is 18
- So now we have enough data for the Box and Whisker Plot:

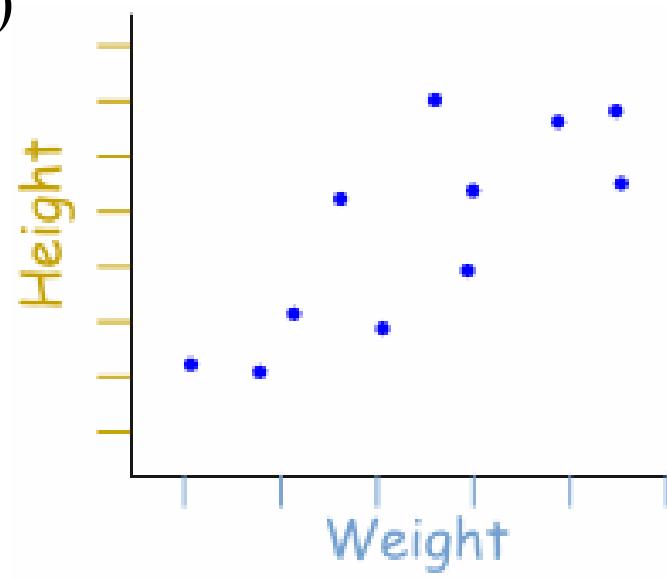


- And the Interquartile Range is:

$$Q3 - Q1 = 15 - 4 = 11$$

# Scatter Plots

- A graph of plotted points that show the relationship between two sets of data.
- In this example, each dot represents one person's weight versus their height.
- (The data is plotted on the graph as “Cartesian (x, y) Coordinates”)



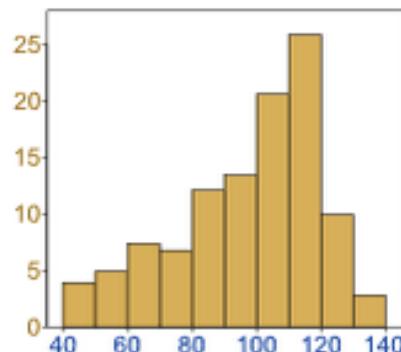


# LAB IN R

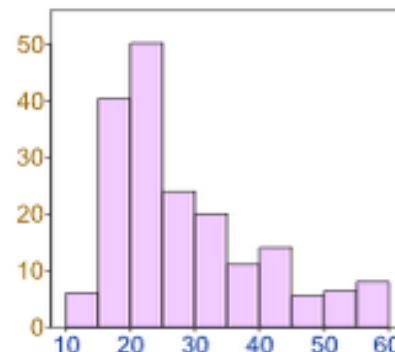
# Normal Distribution

- Data can be "distributed" (spread out) in different ways.

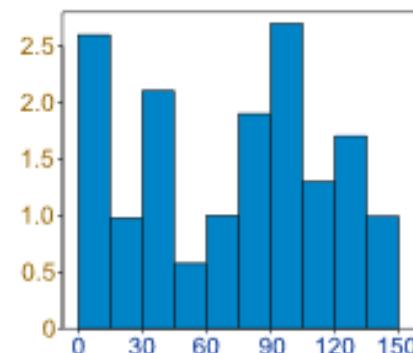
It can be spread out  
more on the left



Or more on the right



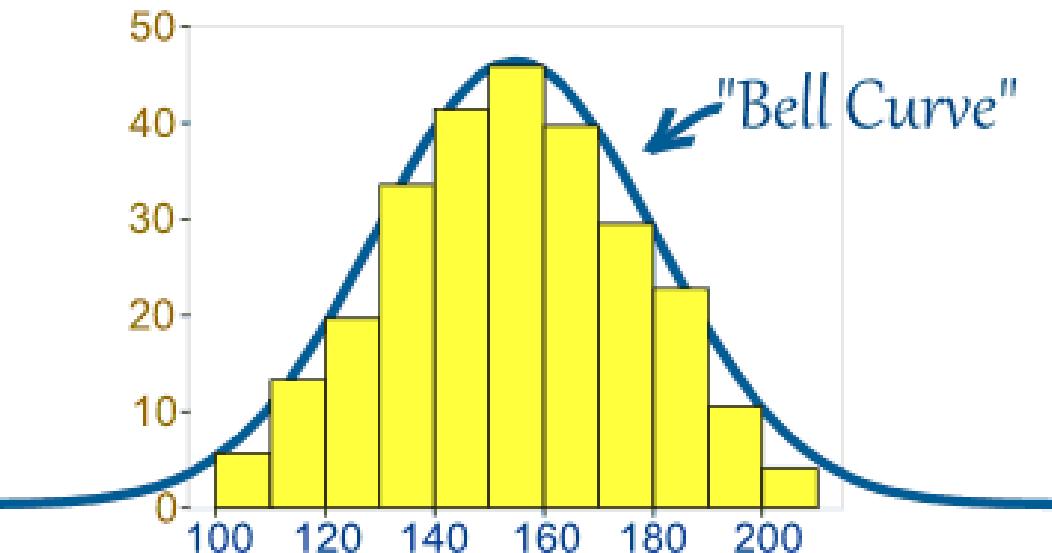
Or it can be all jumbled up



# Normal Distribution

- But there are many cases where the data tends to be
  - around a central value
  - with no bias left or right

and it gets close to a "Normal Distribution" like this:



Many things closely follow a Normal Distribution:

- heights of people
- size of things produced by machines
- errors in measurements
- blood pressure
- marks on a test

# Random Normal Distribution



- `rnorm(n, mean=0, sd=1)` generates a vector of random normal variables
  - n: sample size
  - default mean=0 and sd=1
  - each time different
- `set.seed(m)` reproduces the exact same set of random numbers as long as the arbitrary integer argument `m` stays the same.

```
> x=rnorm(50)
> y=x+rnorm(50,mean=50,sd=.1)
> cor(x,y)
[1] 0.995
```

```
> set.seed(3)
> y=rnorm(100)
> mean(y)
[1] 0.0110
> var(y)
[1] 0.7329
> sqrt(var(y))
[1] 0.8561
> sd(y)
[1] 0.8561
```

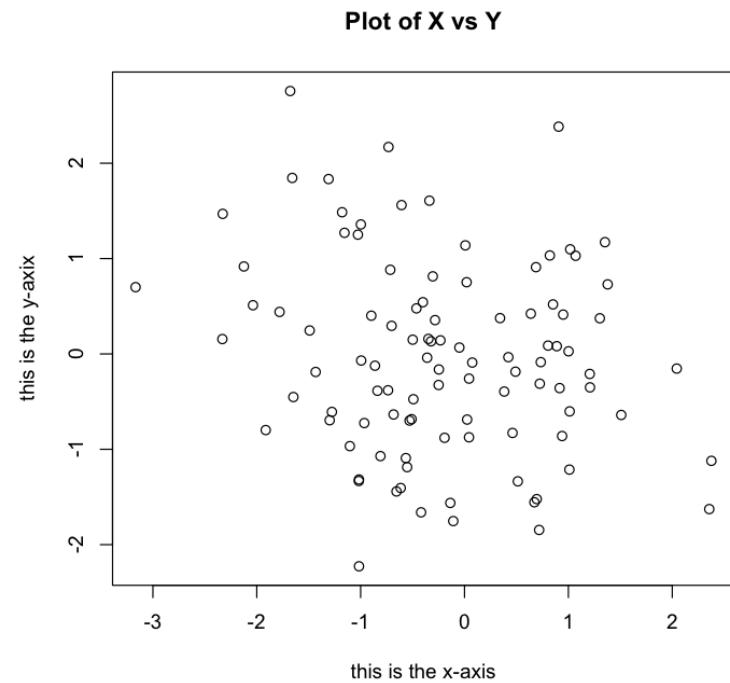
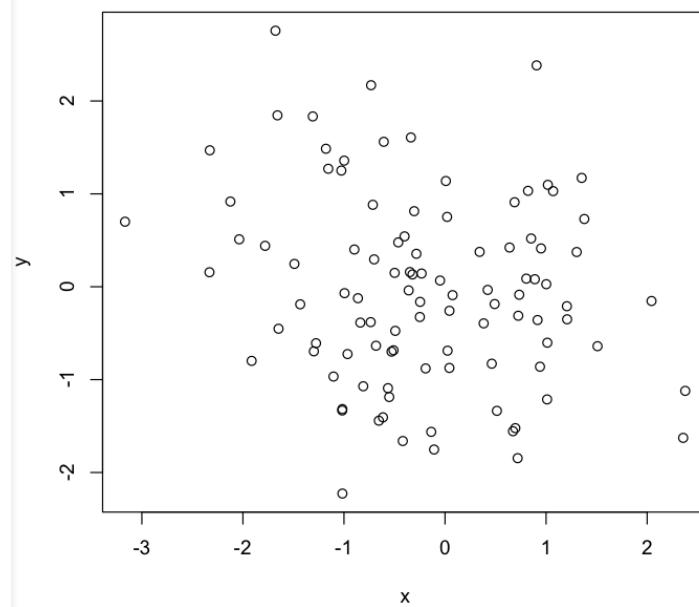
- `cor()`, `mean()`, `var()`, `sd()`

```
> set.seed(1303)
> rnorm(50)
[1] -1.1440  1.3421  2.1854  0.5364  0.0632  0.5022 -0.0004
. . .
```

# Basic Graphics

- `plot()`

```
> x=rnorm(100)
> y=rnorm(100)
> plot(x,y)
> plot(x,y,xlab="this is the x-axis",ylab="this is the y-axis",
      main="Plot of X vs Y")
```



# Exercises (using R)



- Investigate relationship between cigarette smoking and lung capacity
- Data: sample group response data on
  - smoking habits (number of years)
  - measured lung capacities (Spirometer)

| $N$ | Cigarettes<br>( $X$ ) | Lung Capacity<br>( $Y$ ) |
|-----|-----------------------|--------------------------|
| 1   | 0                     | 45                       |
| 2   | 5                     | 42                       |
| 3   | 10                    | 33                       |
| 4   | 15                    | 31                       |
| 5   | 20                    | 29                       |

Questions:

1. Try to plot the data
2. Calculate the sample covariance
3. Calculate the sample correlation
4. How are X and Y correlated?
5. Is there a causation between X and Y?