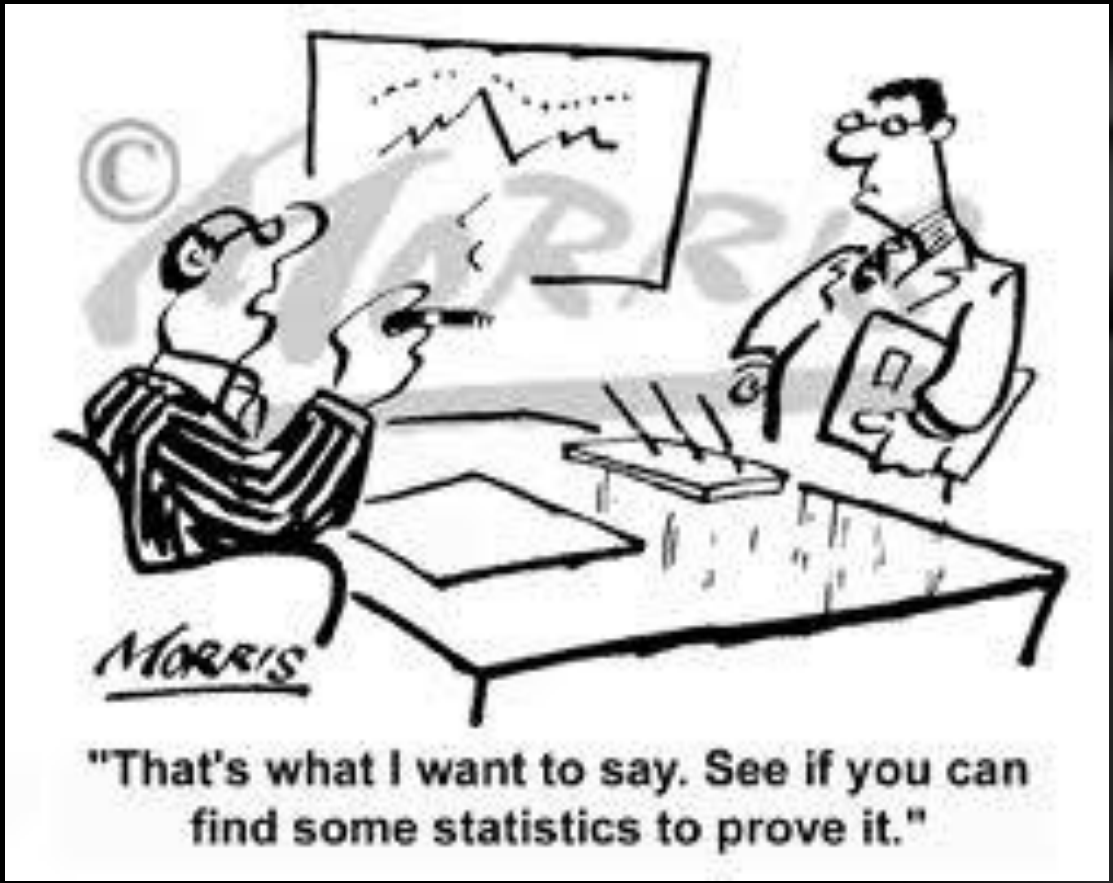
STATISTICS


MORRIS
"That's what I want to say. See if you can find some statistics to prove it."

MEASURES OF CENTRAL TENDENCY

1. MEAN

   The average value of the data set

$$\bar{x} = \frac{\sum x}{n}$$

2. MEDIAN

   The middle value of the ordered data set

3. MODE

   The data value that occurs most frequently

MEASURES OF SPREAD

1. RANGE

$$= Highest\ data\ value\ - Lowest\ data\ value$$

2. INTER-QUARTILE RANGE

$$IQR = Q_3 - Q_1$$

3. STANDARD DEVIATION

$$\boldsymbol{\sigma = \sqrt{\frac{\sum(x-\bar{x})^2}{n}}}$$

FIVE – NUMBER SUMMARY

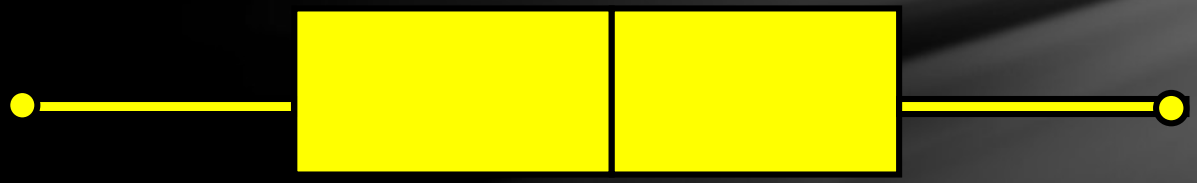
1. MINIMUM DATA VALUE
2. LOWER QUARTILE (Q1)
3. MEDIAN (Q2)
4. UPPER QUARTILE (Q3)
5. MAXIMUM DATA VALUE

Displayed by means of the Box-and-Whisker Plot, which in turn is a visual representation of the distribution of the data!

DISTRIBUTION OF DATA

1. SYMMETRICAL DISTRIBUTION

In a Box-and-Whisker Plot, symmetrical distribution can be seen when Q2 is in the middle of Q1 and Q3

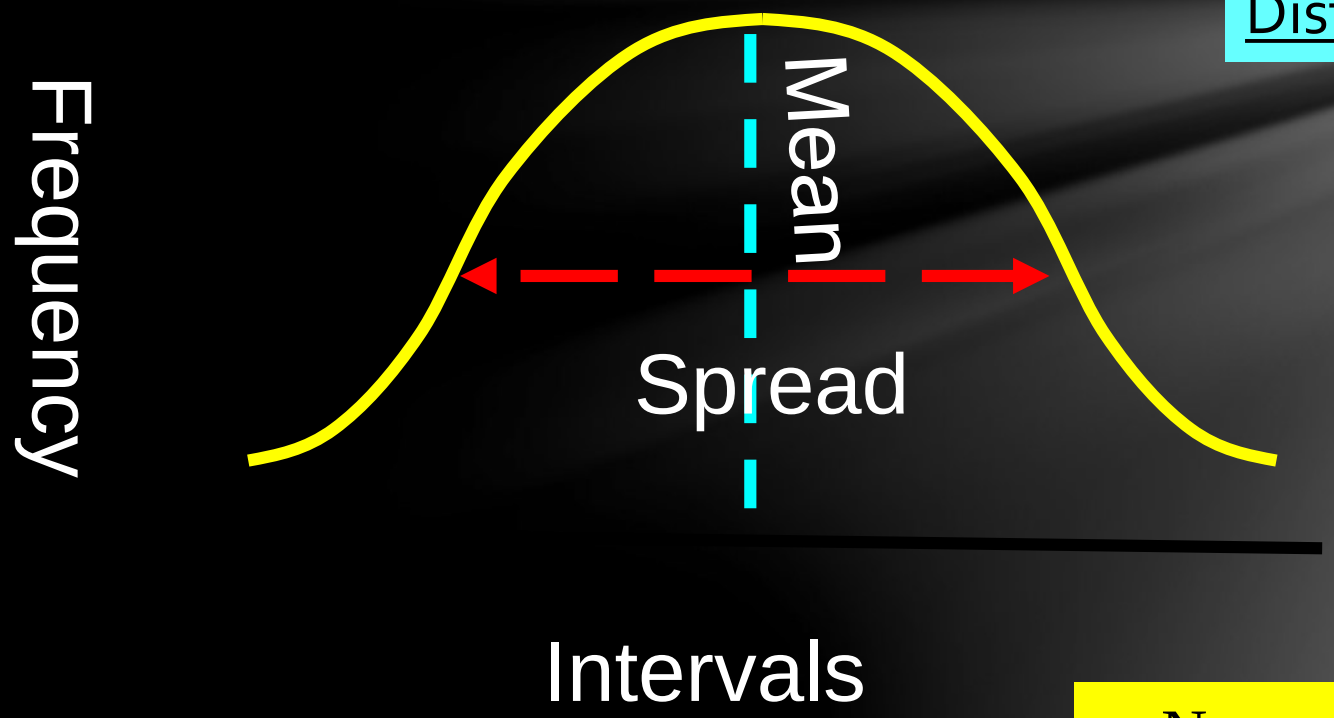


Minimum Q1 Q2 Q3 Maximum

Interpreting Box-and-Whisker Plots

In a Distribution Curve, normal distribution can be seen by the following shaped bell-curve:

Standard Deviation and Distribution Curves



Normal Distribution

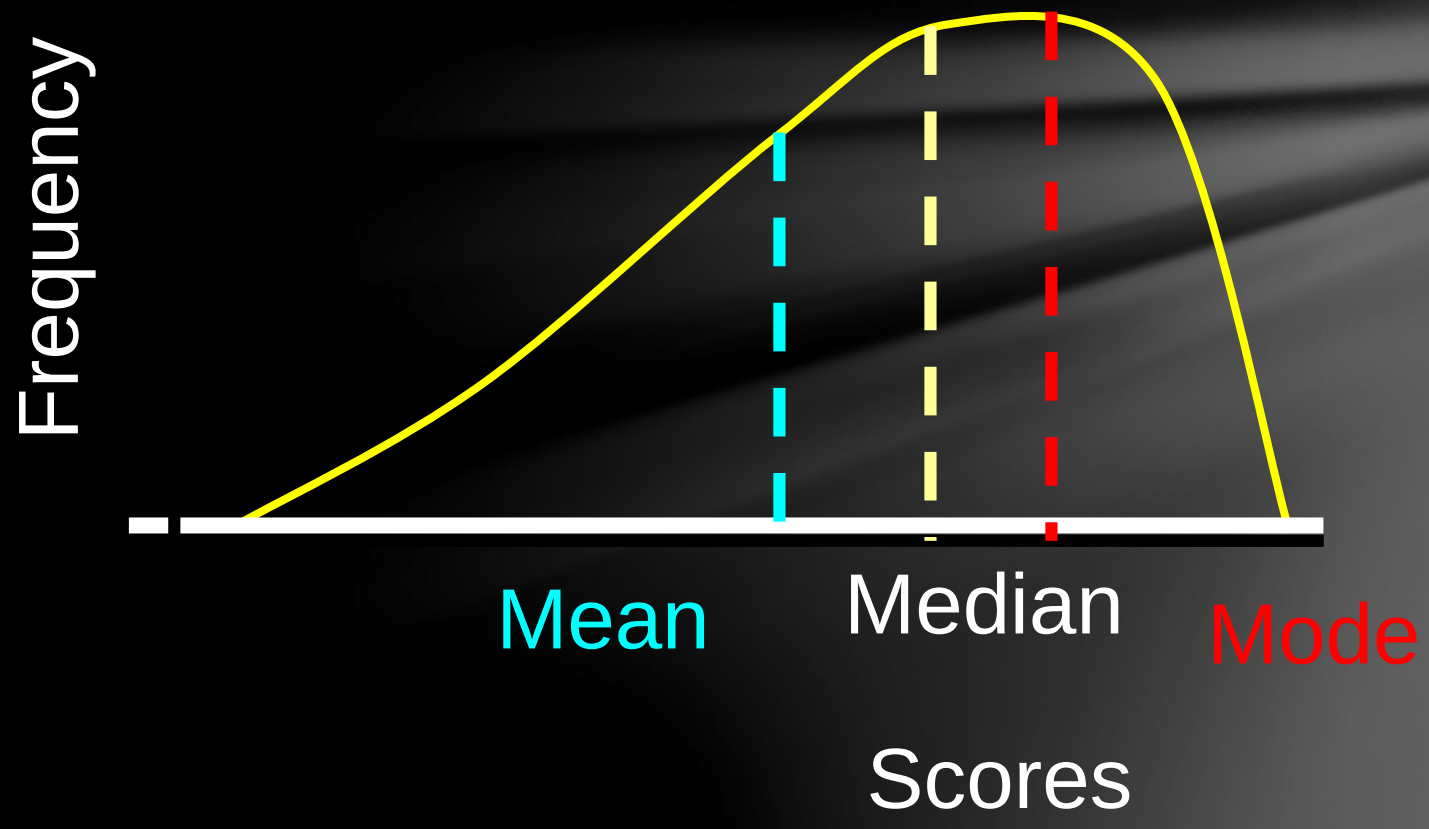
DISTRIBUTION OF DATA

2. SKEWED TO THE LEFT DISTRIBUTION

In a Box-and-Whisker Plot, data is skewed to the left when Q2 is closer to Q3

Minimum Q1 Q2 Q3 Maximum

In a Distribution Curve, data that is skewed to the left results in the following curve:



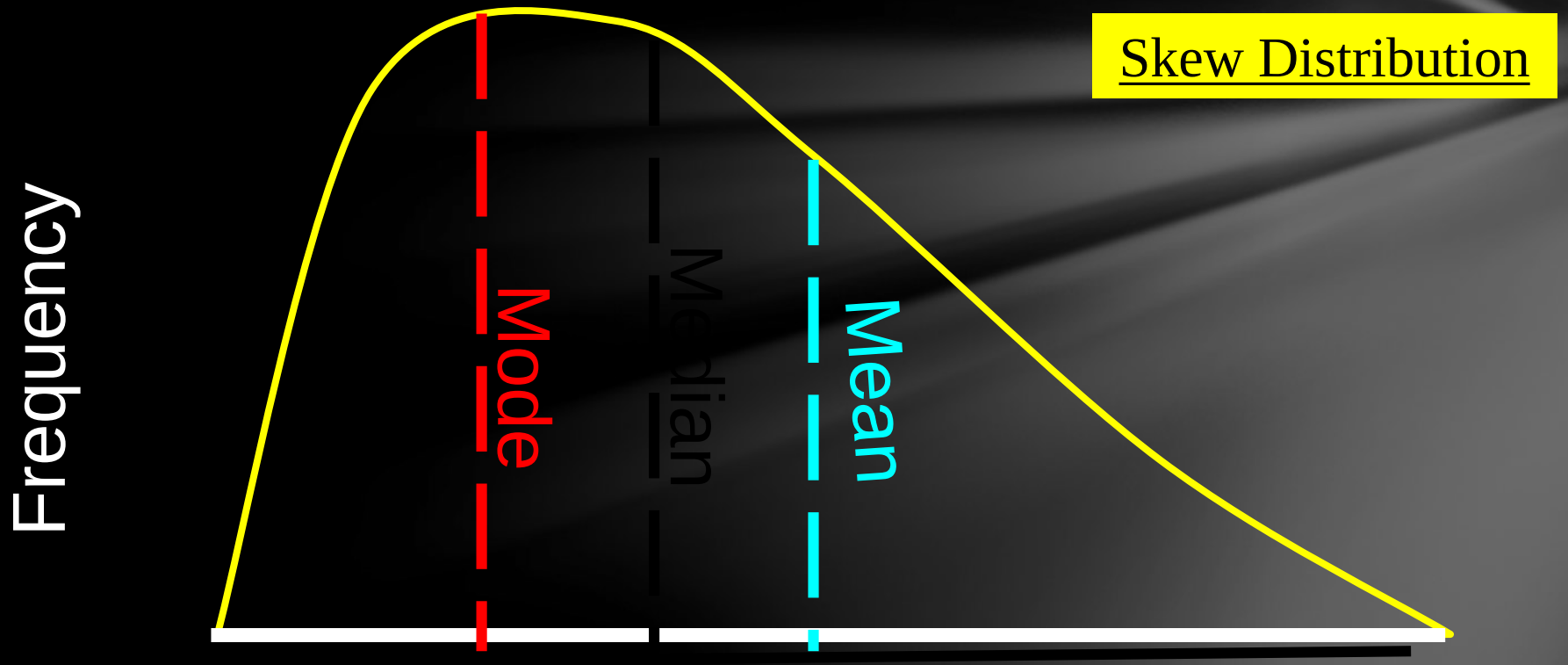
DISTRIBUTION OF DATA

3. SKEWED TO THE RIGHT DISTRIBUTION

In a Box-and-Whisker Plot, data is skewed to the right when Q2 is closer to Q1

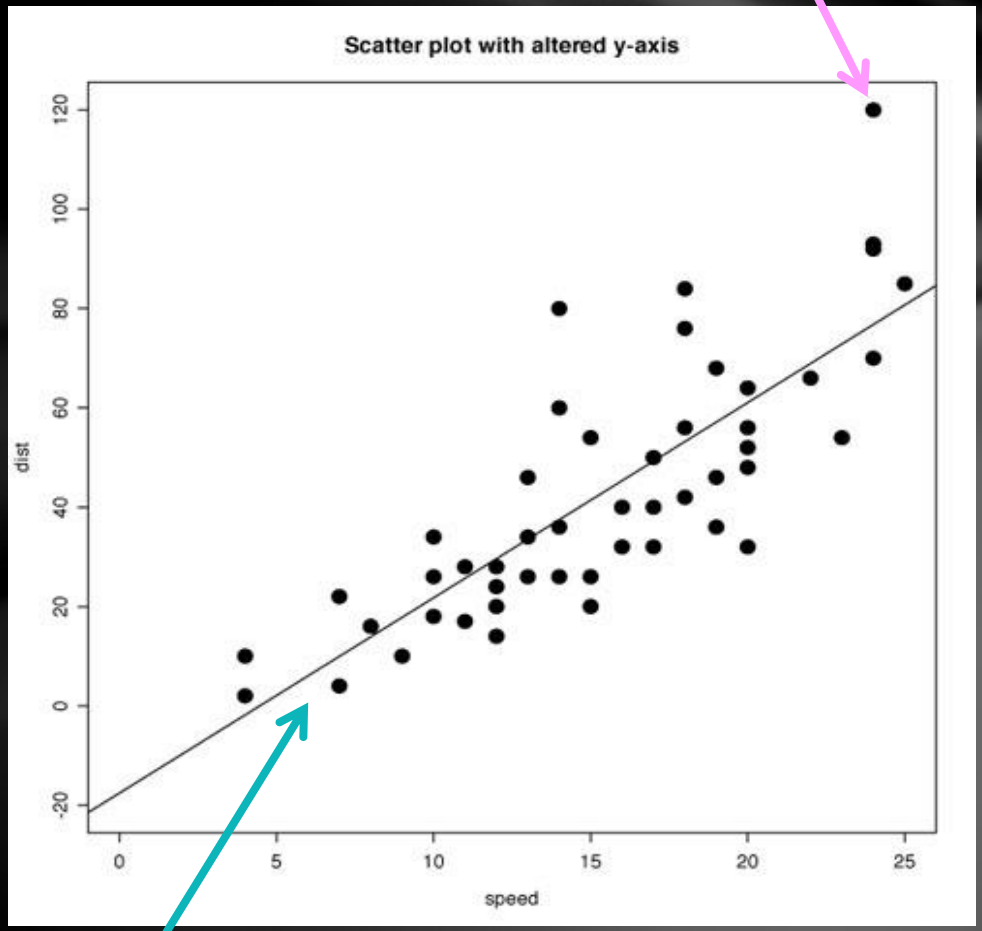
Minimum Q1 Q2 Q3 Maximum

In a Distribution Curve, data that is skewed to the right results in the following curve:

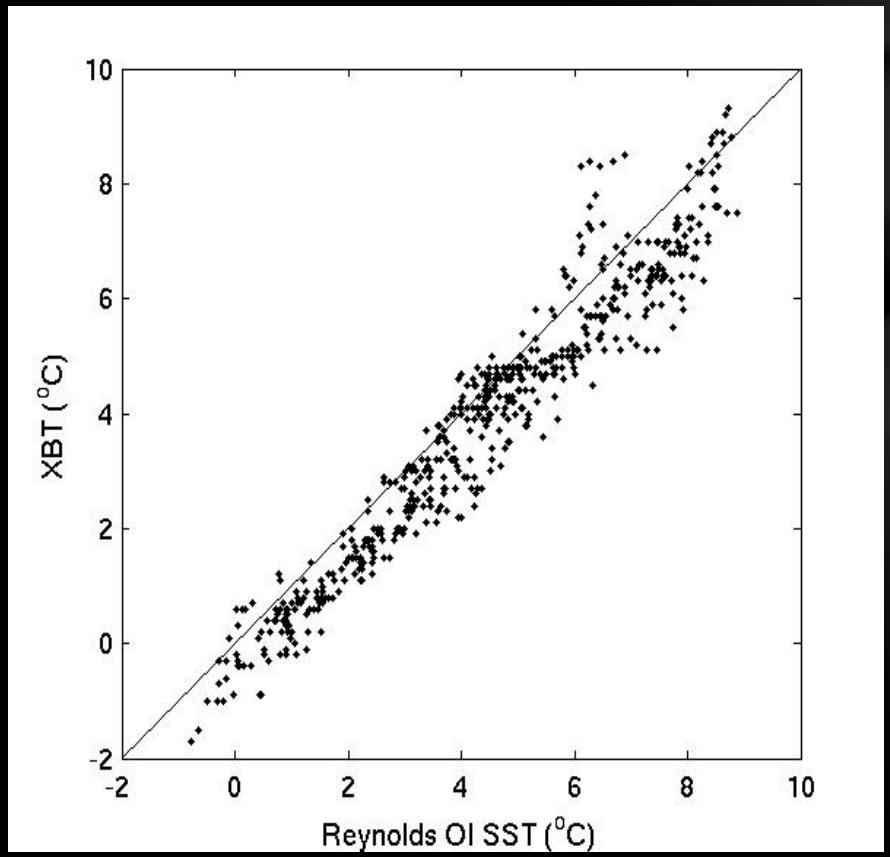


SCATTER PLOTS

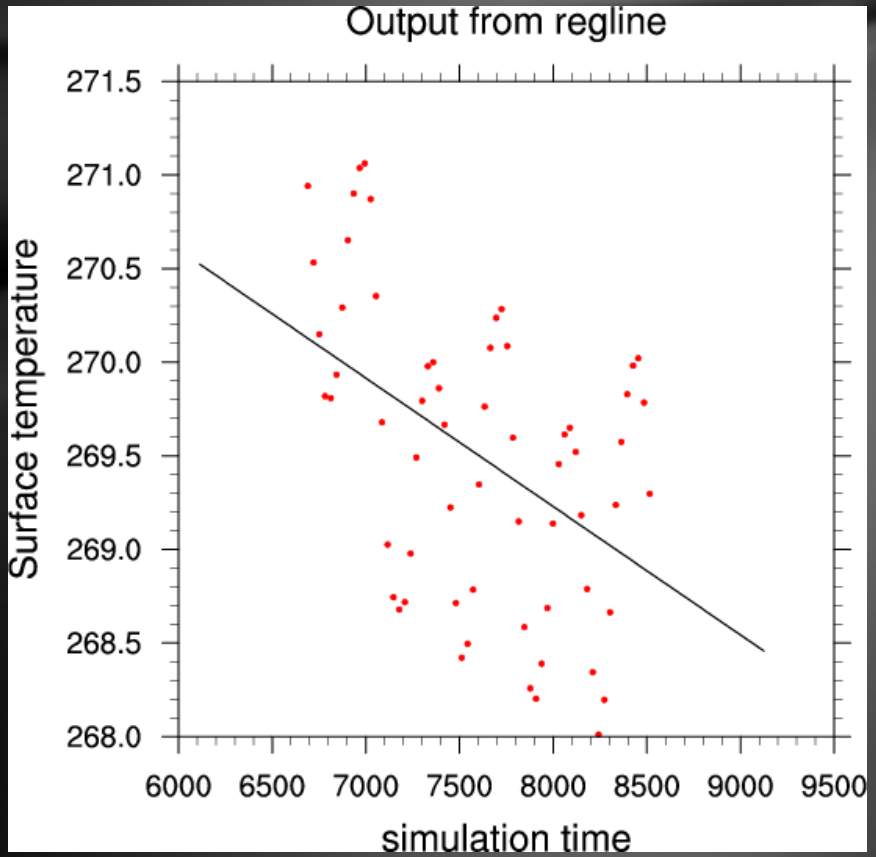
- Used to display bivariate data
- Shows a relationship or correlation between 2 variables
- Can draw a line of best fit
- Can identify outliers



A positive correlation exists when the line of best fit is a positive straight line

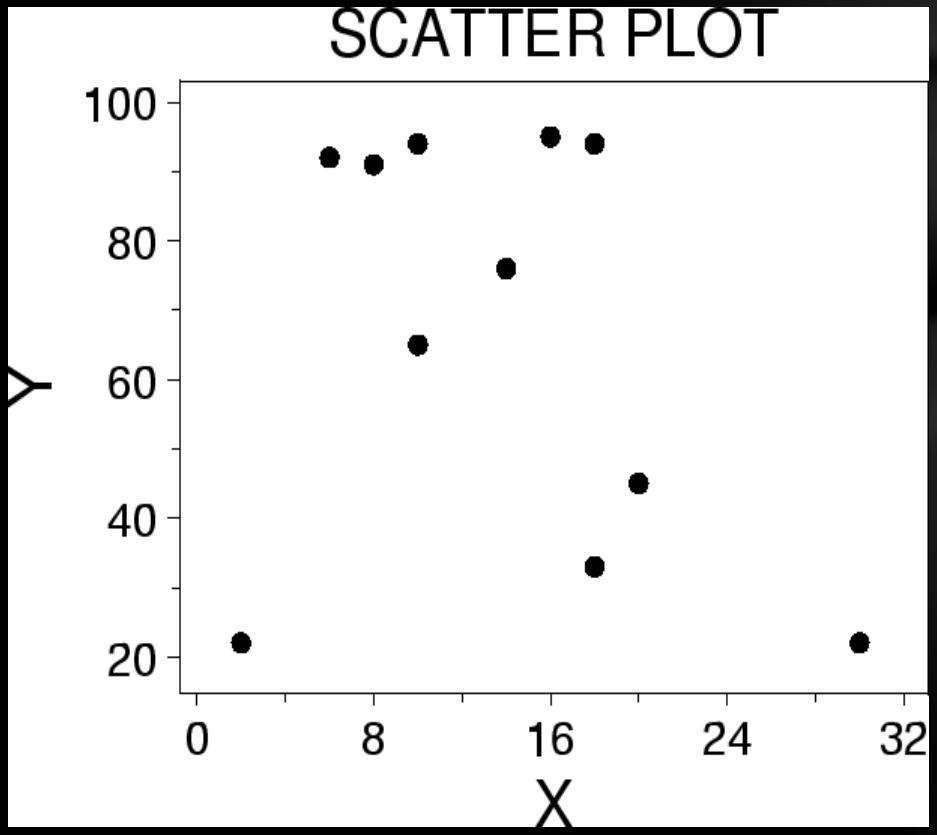


A negative correlation exists when the line of best fit is a negative straight line



No correlation

exists when one cannot draw a line of best fit!



Scatter plots & correlations

Estimating the line of best fit

Lines of best fit can now be more accurately drawn (than just by eye), by means of the least squares method in order to obtain the least squares regression line.

ENRICHMENT: DETERMINING THE LEAST SQUARES REGRESSION LINE

- The least squares regression line is the line of best fit that is positioned in such a way that the sum of the squared errors is a minimum

Squared error of regression line

Proof of minimizing the squared error of regression line (part 1)

Proof of minimizing the squared error of regression line (part 2)

Proof of minimizing the squared error of regression line (part 3)

LEAST SQUARES REGRESSION LINE

- The least squares regression line is the line of best fit that is positioned in such a way that the sum of the squared errors is a minimum
- Equation of least squares regression line is:

  $\hat{y} = a + bx$ where

  a = y-intercept; b = gradient
- Can be calculated manually or with a calculator
- NB! Outliers are excluded from the calculation

Manually determining the least squares regression line

- Calculate $\bar{x}$ and $\bar{y}$, using the formula:

$$\bar{x} = \frac{\sum x}{n} \text{ and } \bar{y} = \frac{\sum x}{n}$$

- Calculate the gradient (b) of the line, using the formula:

$$b = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2}$$

- Calculate the y-intercept (a) by substituting $(\bar{x}\ ;\ \bar{y})$ and b into the equation $\hat{y} = a + bx$

Manual regression line example 1

Manual regression line example 2

Determining the least squares regression line using a CASIO fx-82ES PLUS calculator

E.g. A coffee shop keeps a record of the number of cups of coffee sold over an 11 month period:
20; 22; 46; 10; 38; 74; 62; 88; 61; 86; 48; 55

1. Press [MODE] and then select [2: STAT]
2. Select [2: A+BX] for linear regression
3. Now enter the bivariate data, by entering the [X /Y] value and [=] for all x and y data points
4. Press [AC] to clear the screen and store the data values

E.g. A coffee shop keeps a record of the number of cups of coffee sold over an 11 month period:
20; 22; 46; 10; 38; 74; 62; 88; 61; 86; 48; 55

5. Press [SHIFT] [1] to get to the stats menu
6. Select [5: REG] for linear regression
7. To get the value of a (y-intercept), select [1: A] and [=].

Ans: $a = 21{,}47$

$\hat{y} = a + bx$

$\therefore \hat{y} = 21{,}47 + bx$

8. Press [AC] to clear the screen.

E.g. A coffee shop keeps a record of the number of cups of coffee sold over an 12 month period:
20; 22; 46; 10; 38; 74; 62; 88; 61; 86; 48; 55

9. Press [SHIFT] [1] to get to the stats menu
10. Select [5: REG] for linear regression
11. To get the value of b (gradient), select select [2: B] and [=].

Ans: $b = 4{,}52$

$$\hat{y} = a + bx$$
$$\therefore \hat{y} = 21{,}47 + 4{,}52x$$

PREDICTIONS USING THE LINE OF BEST FIT

Regression lines (line of best fit) are useful as we can make predictions about the given data set and beyond

- When we use the given x / y values of the line of best fit to make a prediction, we call it interpolation
- When we use x / y values outside of the line of best fit to make a prediction, we call it extrapolation

Interpolation & Extrapolation Example

Scatter Plot Real - Life Example

CORRELATION

- In the least squares regression line ($\hat{y} = a + bx$), b indicates whether the gradient is positive or negative, but not whether the association is strong or weak
- In order to determine the strength of the association between the bivariate data, we can calculate the Pearson's product moment correlation coefficient (r)

$$r = \frac{1}{n-1}\sum\left(\frac{x-\bar{x}}{s_x}\right)\left(\frac{y-\bar{y}}{s_y}\right)$$ where

n = no. data pairs ;
s_x / s_y = standard deviation of x /y-values

Determining the correlation coefficient using a CASIO fx-82ES PLUS calculator

E.g. A coffee shop keeps a record of the number of cups of coffee sold over an 11 month period:
20; 22; 46; 10; 38; 74; 62; 88; 61; 86; 48; 55

1. Press [MODE] and then select [2: STAT]
2. Select [2: A+BX] for linear regression
3. Now enter the bivariate data, by entering the [X /Y] value and [=] for all x and y data points
4. Press [AC] to clear the screen and store the data values

E.g. A coffee shop keeps a record of the number of cups of coffee sold over an 11 month period:
20; 22; 46; 10; 38; 74; 62; 88; 61; 86; 48; 55

5. Press [SHIFT] [1] to get to the stats menu
6. Select [5: REG] for linear regression
7. To get the value of r (correlation coefficient), select [3: r] and [=].

Ans: $r = 0{,}64$

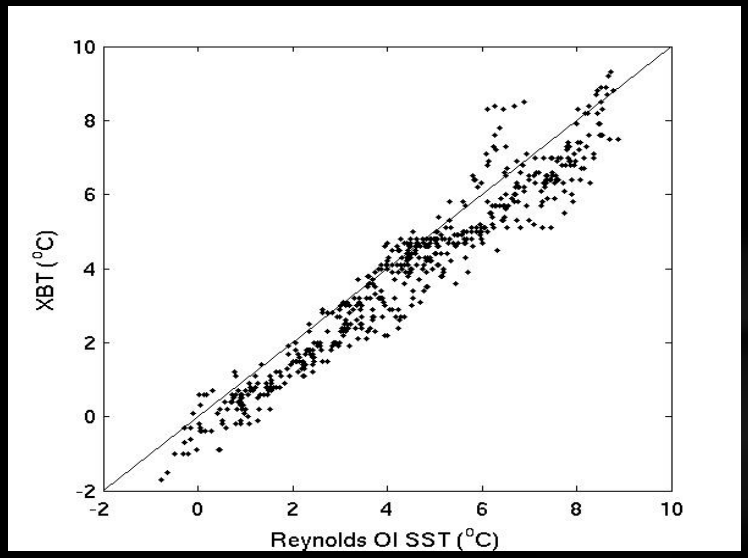
8. Press [AC] to clear the screen.

Now to interpret the value of *r* ...

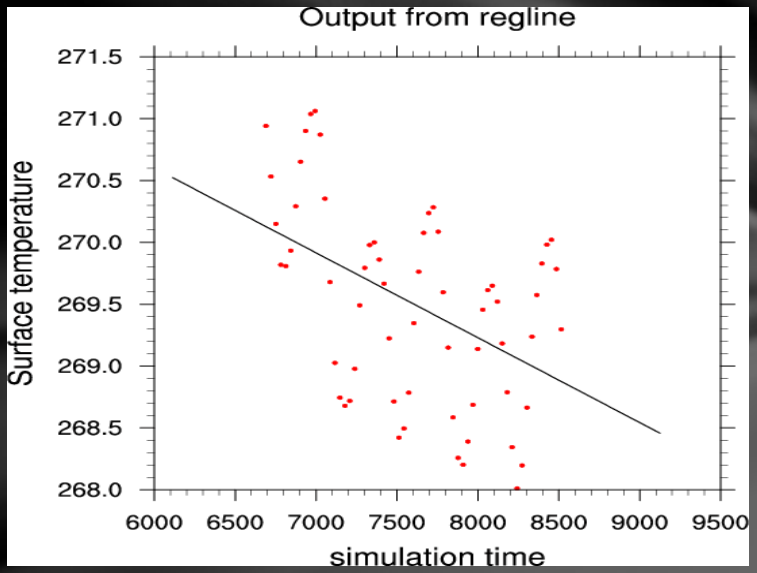
Interpreting the correlation coefficient (r)

- The correlation coefficient (r) indicates the strength of the association between the bivariate data points
- The correlation coefficient can assume the following values … $-1 \leq r \leq 1$ … where
  - -1 indicates a negative and strong correlation;
  - 0 indicates no correlation; and
  - 1 indicates a positive and strong correlation
- Let's evaluate some of the original scatter plots

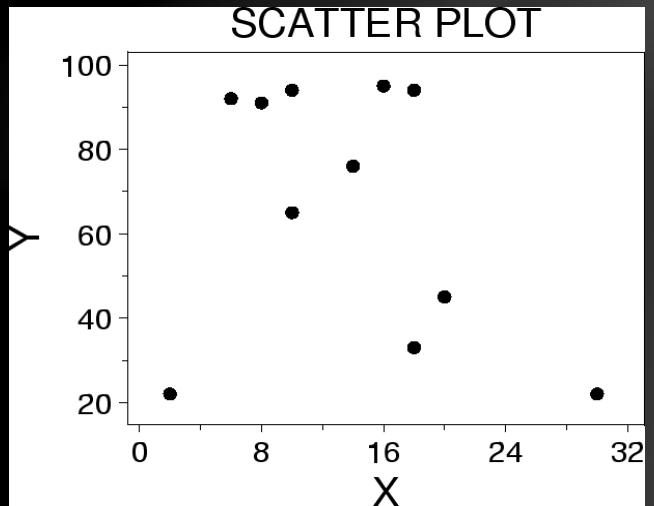
Positive & strong correlation ($r \approx 0.9$)



Negative & weak correlation ($r \approx -0{,}4$)



No correlation ($r \approx 0.1$)



Understanding the Correlation Coefficient