Project Maths
Tionscadal Mata
Development Team

Module 2

Analysing Data Numerically

Measures of Central Tendency

- Mean
- Median
- Mode

Measures of Spread

- Range
- Standard Deviation
- Inter-Quartile Range

Mean & Standard Deviation using a Calculator

Calculate the mean and standard deviation of the following 10 students heights by:

(i) using the data unsorted

(ii) creating a frequency table

Unsorted Data

	Gender	Height / cm
1.	Boy	165
2.	Girl	165
3.	Boy	150
4.	Boy	171
5.	Girl	153
6.	Boy	171
7.	Girl	153
8.	Girl	153
9.	Boy	166
10.	Boy	179

Frequency Table

Height / cm	150	153	165	166	171	179
Frequency	1	3	2	1	2	1

Mean & Standard Deviation using a Calculator

Calculate the mean and standard deviation of the following 10 students heights by:

(i) using the data unsorted

(ii) creating a frequency table

Unsorted Data

	Gender	Height / cm
1.	Boy	165
2.	Girl	165
3.	Boy	150
4.	Boy	171
5.	Girl	153
6.	Boy	171
7.	Girl	153
8.	Girl	153
9.	Boy	166
10.	Boy	179

Frequency Table

Height / cm	150	153	165	166	171	179
Frequency	1	3	2	1	2	1

Mean $= 162.6$ cm

S.D. $= 9.32$ cm

Central Tendency: The Mean

Advantages:

- Mathematical centre of a distribution
- Does not ignore any information

Disadvantages:

- Influenced by extreme scores and skewed distributions
- May not exist in the data

Central Tendency: The Mode

Advantages:

- Good with nominal data
- Easy to work out and understand
- The score exists in the data set

Disadvantages:

- Small samples may not have a mode
- More than one mode might exist

Central Tendency: The Median

Advantages:

- Not influenced by extreme scores or skewed distribution
- Good with ordinal data
- Easier to calculate than the mean
- Considered as the typical observation

Disadvantages:

- May not exist in the data
- Does not take actual data into account only its (ordered) position
- Difficult to handle theoretically

Summary: Relationship between the 3 M's

Characteristics	Mean	Median	Mode
Consider all the data in calculation			
Easily affected by extreme data			
Can be obtained from a graph			
Should be one of the data			
Need to arrange the data in ascending order			

Summary: Relationship between the 3 M's

Characteristics	Mean	Median	Mode
Consider all the data in calculation	Yes	No	No
Easily affected by extreme data	Yes	No	No
Can be obtained from a graph	No	Yes	Yes
Should be one of the data	No	No	Yes
Need to arrange the data in ascending order	No	Yes	No

Using Appropriate Averages

Example

There are 10 house in Pennylane Close.

On Monday, the numbers of letters delivered to the houses are:

0 2 5 3 34 4 0 1 0 2

Calculate the mean, mode and the median of the number of letters.

Comment on your results.

Solution

$$\text{Mean} = \frac{0+2+5+3+34+4+0+1+0+2}{10}$$

$$= 5.1$$

$$\text{Mode} = 0$$

$$\text{Median} = 2$$

In this case the **mean** (5.1) has been distorted by the large number of letters delivered to one of the houses. It is, therefore, not a good measure of a 'typical' number of letters delivered to any house in Pennylane Close.

The **mode** (0) is also not a good measure of a 'typical' number of letters delivered to a house, since 7 out of the 10 houses do acutally receive some letters.

The **median** (2) is perhaps the best measure of the 'typical' number of letters delivered to each house, since half of the houses received 2 or more letters and the other half received 2 or fewer letters.


Your turn!
2.1

2.1

Ten students submitted their Design portfolios which were marked out of 40.
The marks they obtained were

37 34 34 34 29 27 27 10 4 28

(a) For these marks find

(i) the mode **(ii)** the median **(iii)** the mean.

(b) Comment on your results.

(c) An external moderator reduced all the marks by 3.
Find the mode, median and mean of the moderated results.

Solution

2.1

Ten students submitted their Design portfolios which were marked out of 40.
The marks they obtained were

37 34 34 34 29 27 27 10 4 28

(a) For these marks find
(i) the mode **(ii)** the median **(iii)** the mean.

(b) Comment on your results.

(c) An external moderator reduced all the marks by 3.
Find the mode, median and mean of the moderated results.

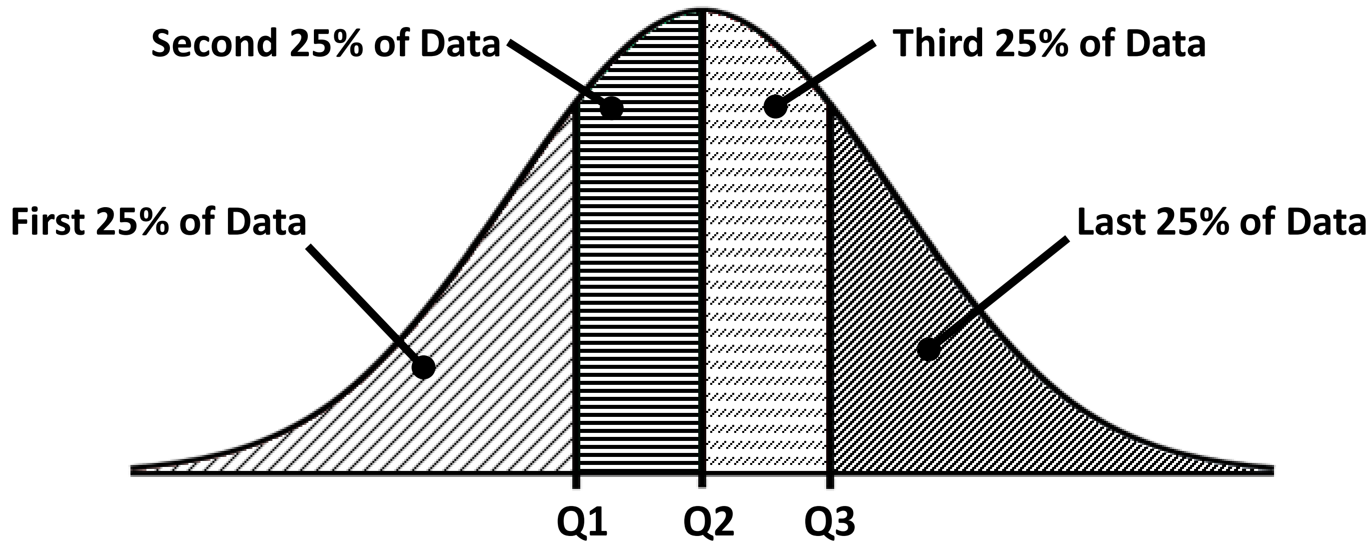
Solution

(a) **(i)** 34 **(ii)** 28.5 **(iii)** 26.4

(b) Mean is the lowest – the ‘4’ depresses its value compared to the mode and the median.

(c) 31, 25.5, 23.4

Quartiles



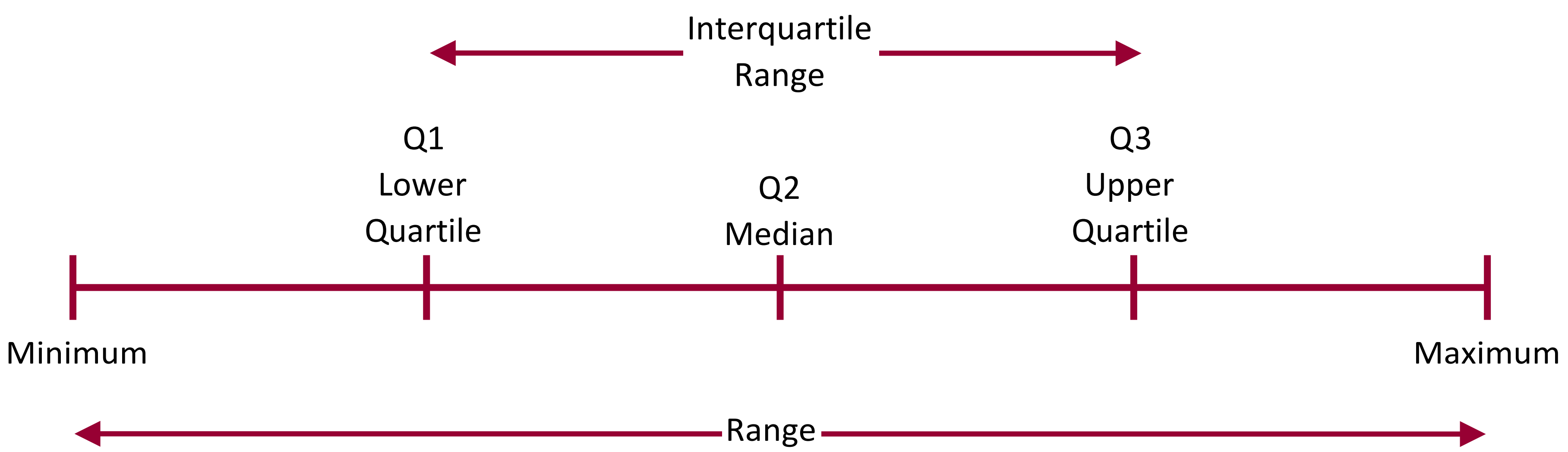
Quartiles

When we arrange the data is ascending order of magnitude and divide them into four equal parts, the values which divide the data into four equal parts are called **quartiles**.

They are usually denoted by the symbols Q_1 is the lowest quartile (or first quartile) where 25% of the data lie below it; Q_2 is the middle quartile (or second quartile or median) where 50% of the data lie below it; Q_3 is the upper quartile (or third quartile) where 75% of the data lie below it.

Interquartile Range & Range

Data in ascending order of magnitude			
First 25% of data	Second 25% of data	Third 25% of data	Last 25% of data



Standard Deviation

Example

Two machines A and B are used to measure the diameter of a washer. 50 measurements of a washer are taken by each machine. If the standard deviations of measurements taken by machine A and B are 0.4mm and 0.15mm respectively, which instrument gives more consistent measurements?

Solution

Standard deviation of A = 0.4 mm

Standard deviation of B = 0.15 mm

The smaller the standard deviation, the less widely dispersed the data is. This means that more measurements are closer to the mean. Therefore, the measurements taken by instrument B are more consistent.

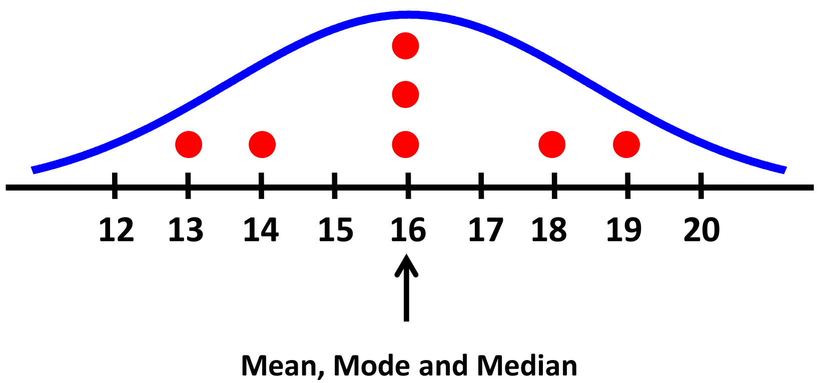
Guide to Distributions

Example 1

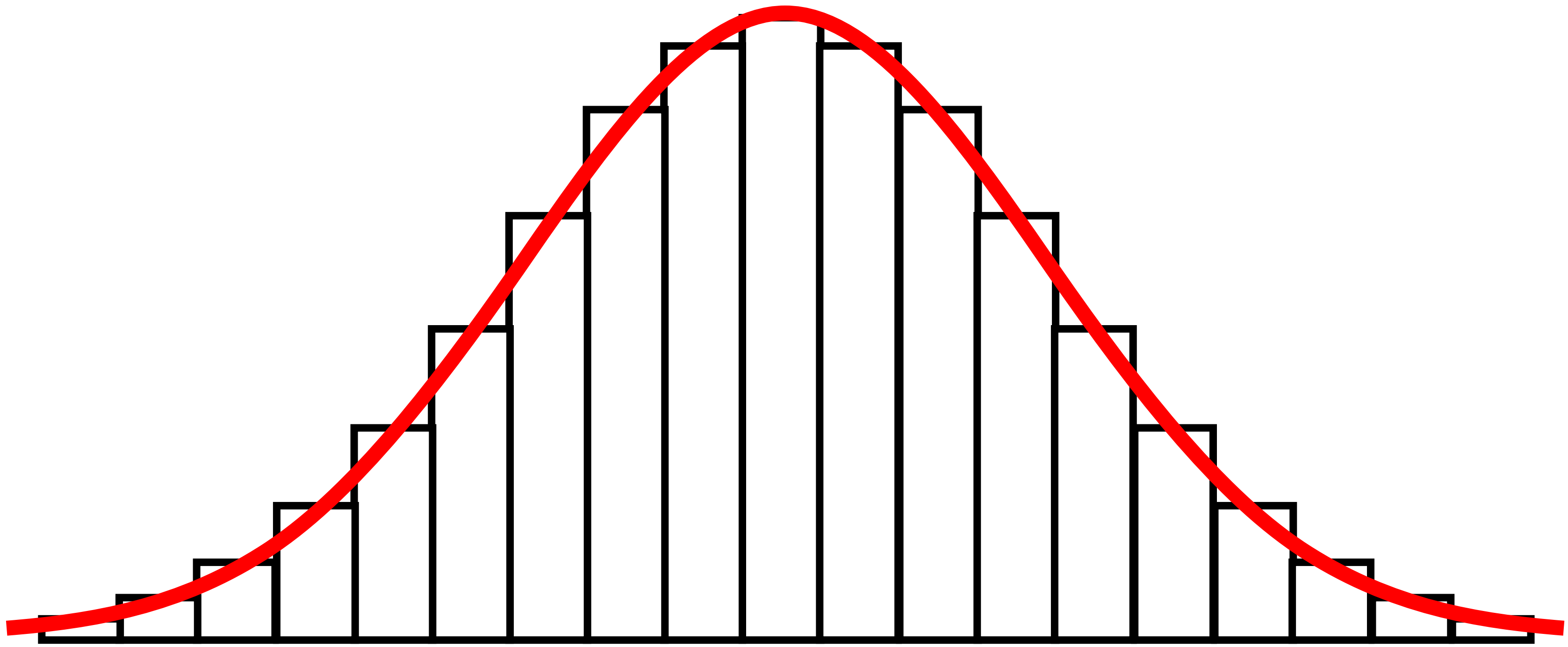
Seven teenagers at a youth club were asked their age.
They gave the following ages:

16, 14, 19, 16, 13, 18, 16

The mean, mode and the median of their ages are as follows:
Mean = 16
Mode = 16
Median = 16
If a line plot of their ages is drawn we get the following.



When the **Mean** and **Median** are the same value the plot is symmetrical (Bell Shaped).
The **Mode** affects the height of the of the Bell Shaped curve.



Example 2:

Seven people were asked how many text messages they send (on average) every week. The results were as follows:

3, 25, 30, 30, 30, 33, 40

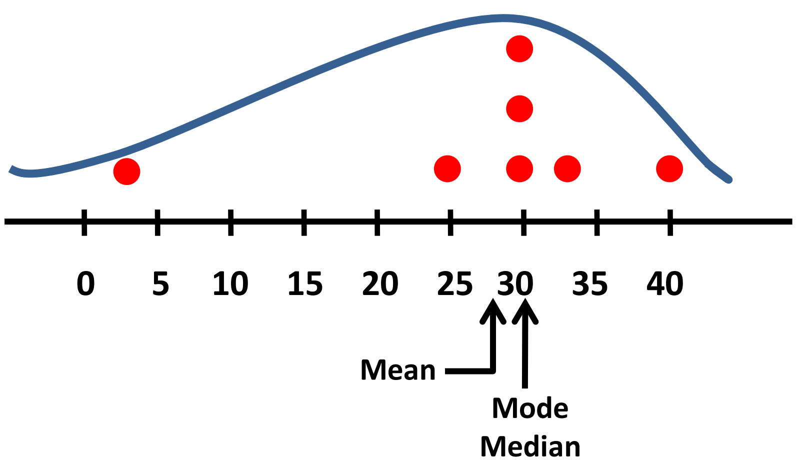
The mean mode and the median of text messages sent are as follows:

Mean = 27.29

Mode = 30

Median = 30

If a line plot of the number of texts sent is drawn we get the following:



When the **Mean** is to the left of the **Median** the data is said to be skewed to the **left** or **negatively skewed**.

The **Mode** affects the height of the curve.

Example 3:

Eight factory workers were asked to give their annual salary.

The results are as follows: (figures in thousands of Euro)

20, 22, 25, 26, 27, 27, 70

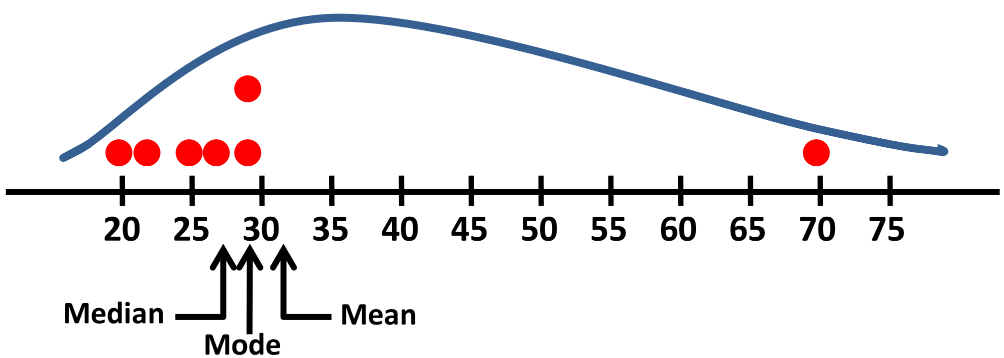
The mean mode and the median of their annual salaries are as follows:

Mean = 31

Mode = 27

Median = 26

If a line plot of their salaries is drawn we get the following.

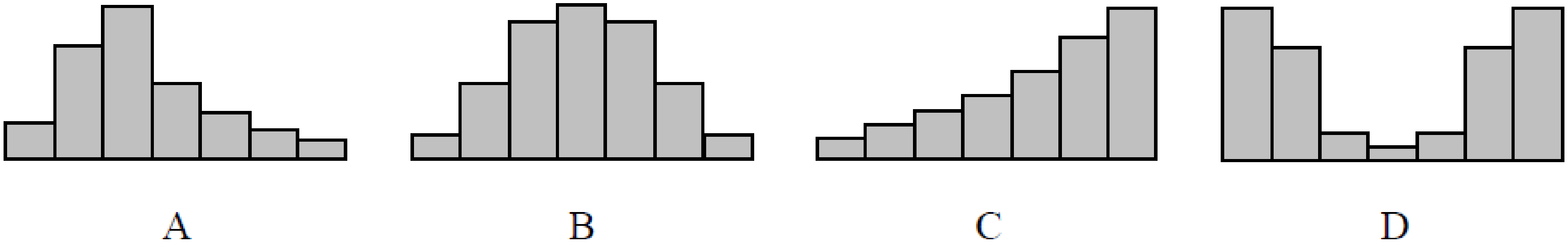


When the **Mean** is to the **right** of the **Median** the data is said to be skewed to the **right** or **positively skewed**.

The **Mode** affects the height of the curve.

Question 2 **(HL – Sample Paper)** **(25 marks)**

The shapes of the histograms of four different sets of data are shown below.

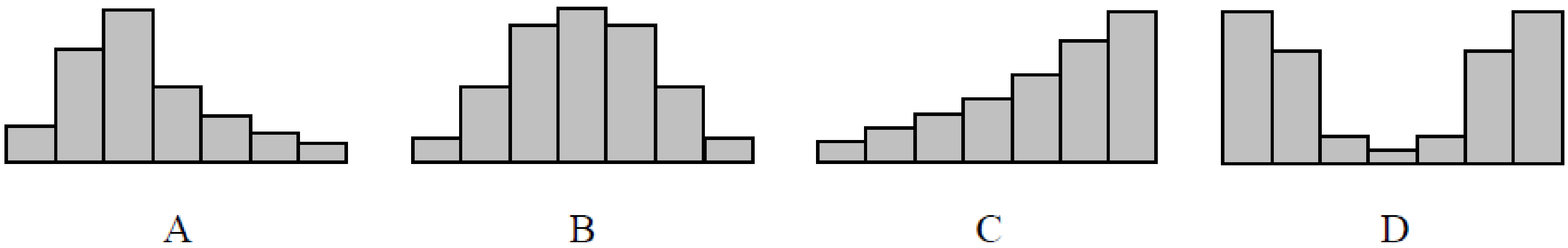


(a) Complete the table below, indicating whether the statement is correct (✓) or incorrect (×) with respect to each data set.

	A	B	C	D
The data are skewed to the left				
The data are skewed to the right				
The mean is equal to the median				
The mean is greater than the median				
There is a single mode				

Question 2 **(HL – Sample Paper)** **(25 marks)**

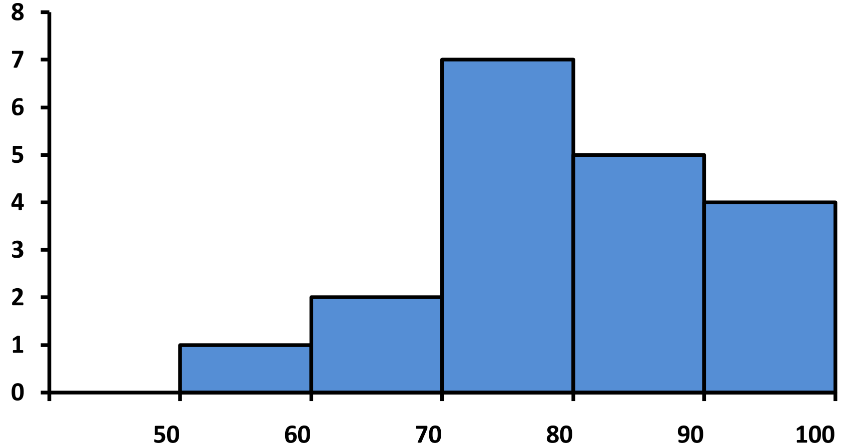
The shapes of the histograms of four different sets of data are shown below.



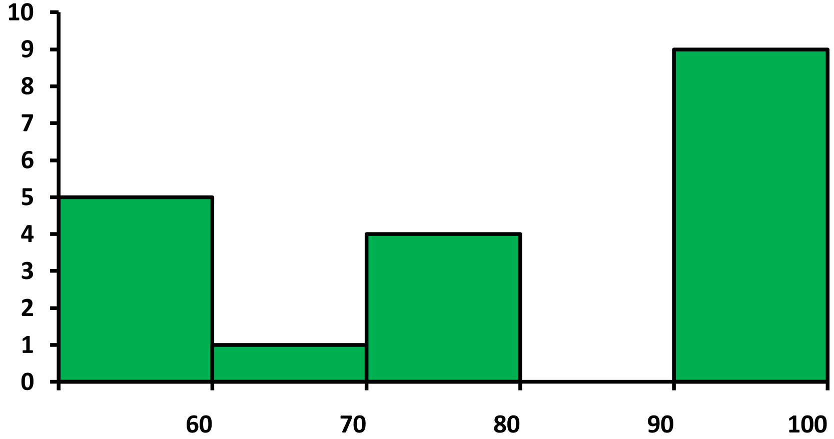
(a) Complete the table below, indicating whether the statement is correct (✓) or incorrect (✗) with respect to each data set.

	A	B	C	D
The data are skewed to the left	✗	✗	✓	✗
The data are skewed to the right	✓	✗	✗	✗
The mean is equal to the median	✗	✓	✗	✓
The mean is greater than the median	✓	✗	✗	✗
There is a single mode	✓	✓	✓	✗

Are Measures of Centre Enough?



Dataset 1

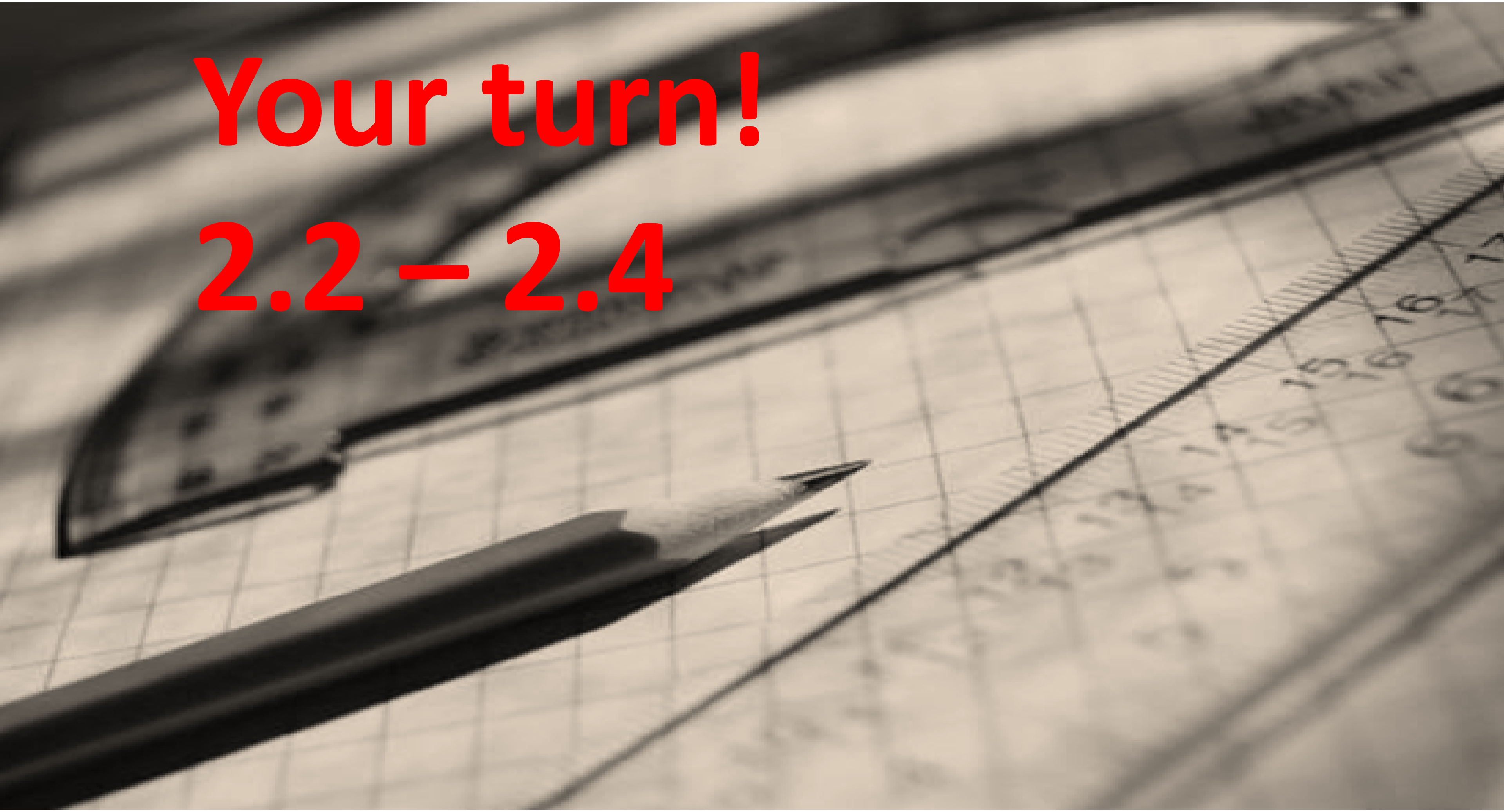


Dataset 2

	Dataset 1	Dataset 2
Median	78.0	78.0
Mean	79.1	79.1
Mode	75.0	75.0
Maximum	99	99
Minimum	58	51
Range	41	48
Standard Deviation	11.2	17.8

Your turn!

2.2 – 2.4



2.2

A clerk entering salary data into a company spreadsheet accidentally put an extra "0" in the boss's salary, listing it as €2,000,000 instead of €200,000. Explain how this error will affect these summary statistics for the company payroll:

(a) measures of centre: median and mean.

(b) measures of spread: range, IQR, and standard deviation.

Solution

2.2

A clerk entering salary data into a company spreadsheet accidentally put an extra "0" in the boss's salary, listing it as €2,000,000 instead of €200,000. Explain how this error will affect these summary statistics for the company payroll:

(a) measures of centre: median and mean.

(b) measures of spread: range, IQR, and standard deviation.

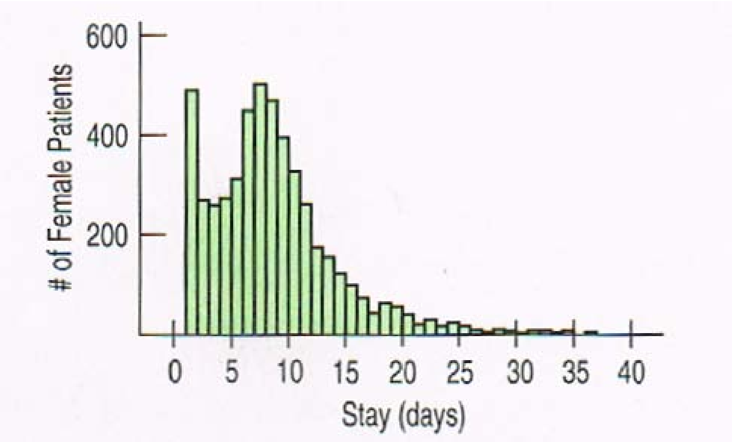
Solution

(b) As long as the boss's true salary of €200,000 is still above the median, the median will be correct.
The mean will be too large, since the total of all the salaries will decrease by €2,000,000 – €200,000=€1,800,000, once the mistake is corrected.

(b) The range will likely be too large. The boss's salary is probably the maximum, and a lower maximum would lead to a smaller range. The IQR will likely be unaffected, since the new maximum has no effect on the quartiles.
The standard deviation will be too large, because the €2,000,000 salary will have a large squared deviation from the mean.

2.4

The histogram shows the lengths of hospital stays (in days) for all the female patients admitted to hospital in New York in 1993 with a primary diagnosis of acute myocardial infarction. (heart attack)



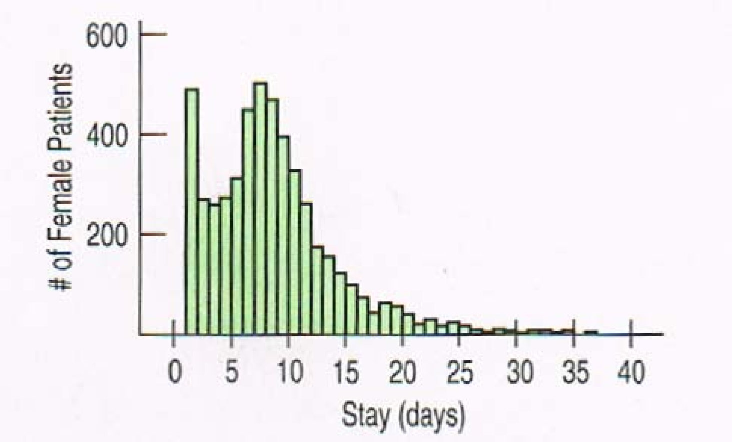
(a) From the histogram, would you expect the mean or median to be larger? Explain.

(b) Write a few sentences describing this distribution. (shape, centre, spread, unusual features).

Solution

2.4

The histogram shows the lengths of hospital stays (in days) for all the female patients admitted to hospital in New York in 1993 with a primary diagnosis of acute myocardial infarction. (heart attack)



(a) From the histogram, would you expect the mean or median to be larger? Explain.

(b) Write a few sentences describing this distribution. (shape, centre, spread, unusual features).

Solution

(a) The distribution of length of stays is skewed to the right, so the mean is larger than the median.

(b) The distribution of the length of hospital stays of female heart attack patients is skewed to the right, with stays ranging from 1 day to 36 days. The distribution is centred around 8 days, with the majority of the hospital stays lasting between 1 and 15 days. There are a relatively few hospitals stays longer than 27 days. Many patients have a stay of only one day, possibly because the patient died.

Bivariate Data

1. Involves **2 variables**
2. Deals with causes or relationships
3. The major purpose of bivariate analysis is to determine whether relationships exist

We will look at the following:

- Scatter plots
- Correlation
- Correlation coefficient
- Correlation & causality
- Line of Best Fit
- Correlation coefficient not equal to slope

Sample question:

Is there a relationship between the scores of students who study Physics and their scores in Mathematics?

Univariate Data versus Bivariate Data

Univariate data: Only one item of data is collected e.g. height

Bivariate Data: Data collected in pairs <u>to see if there is a relationship</u> between the variables e.g. height and arm span, mobile phone bill and age etc.

Examples:

Categorical paired data:	Colour of eyes and gender
Discrete paired:	Number of bars eaten per week and number of tooth fillings
Continuous paired:	Height and weight
Category and discrete paired:	Type of dwelling and number of occupants etc. [Look at C@S questionnaire]

Correlation

- Correlation: is about assessing the strength of the relationship between pairs of data. The first step in determining the relationship between 2 variables is to draw a Scatter Plot.
- After establishing if a Linear Relationship (Line of Best Fit) exists between 2 variables X and Y, the strength of the relationship can be measured and is known as the correlation coefficient (r).
- *Correlation* is a precise term describing the strength and direction of the ***linear*** relationship between quantitative variables.

Correlation Coefficient (r)

$$-1 \le r \le 1$$

r = +1

Corresponds to a perfect positive linear correlation where the points lie exactly on a straight line [The line will have positive slope]

r = 0 :

Correponds to little or no correlation i.e. as x increases there is no definite tendency for the values of y to increase or decrease in a straight line

r = –1 :

Correponds to a perfect negative linear correlation where the points lie exactly on a straight line [The line will have negative slope]

r close to + 1 :

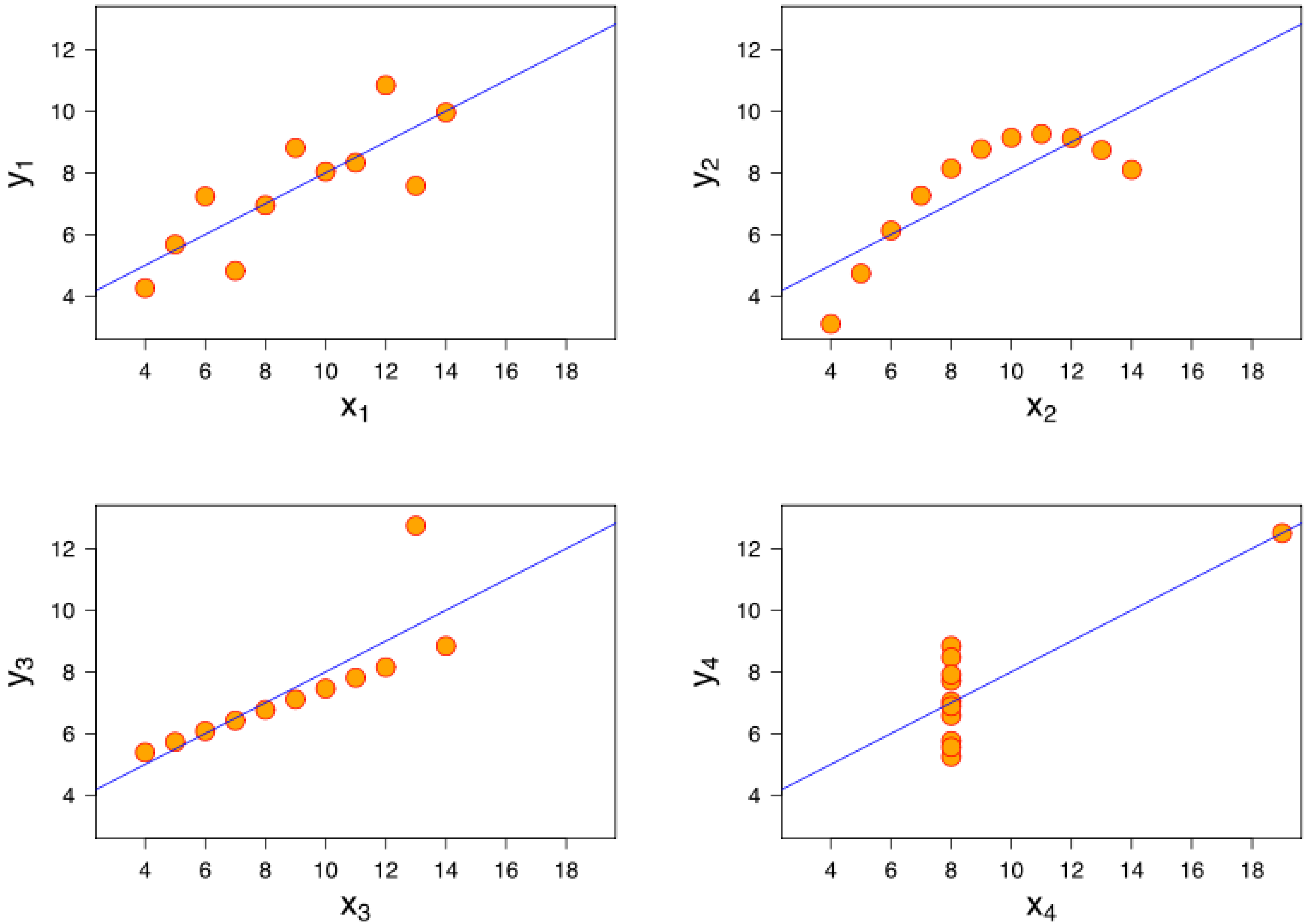
Indicates a strong positive linear correlation, i.e. y tends to increase as x increases

r close to – 1 :

Indicates a strong negative linear correlation, i.e. y tends to decrease as x increases

The correlation coefficient (r) is a numerical measure of the direction and strength of a linear association.

What's wrong? All have a Correlation Coefficient of 0.816



The four *y* variables have the same mean (7.5), standard deviation (4.12), correlation (0.816) and regression line ($y = 3 + 0.5x$). However, as can be seen on the plots, the distribution of the variables is very different.

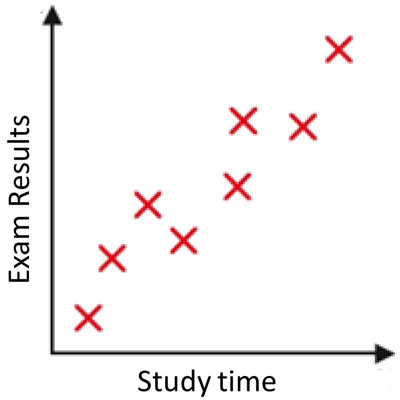
Scatter Plots

- Can show the relationship between 2 variables using ordered pairs plotted on a coordinate plane
- The data points are not joined
- The resulting pattern shows the type and strength of the relationship between the two variables
- Where a relationship exists, a line of best fit can be drawn (by eye) between the points
- Scatter plots can show positive or negative correlation, weak or strong correlation, outliers and spread of data

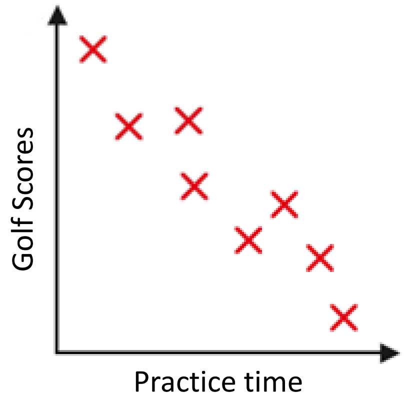
An outlier is a data point that does not fit the pattern of the rest of the data. There can be several reasons for an outlier including mistakes made in the data entry or simply an unusual value.

Describing Correlation

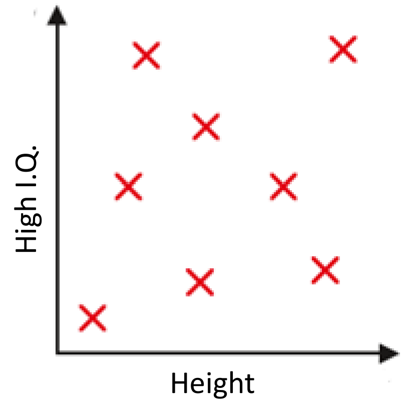
- **Form:** Straight, Curved, No pattern
- **Direction:** Positive, Negative, Neither
- **Strength:** Weak, Moderate, Strong
- **Unusual Features:** Outliers, Subgroups



Positive correlation
As one quantity increases so does the other



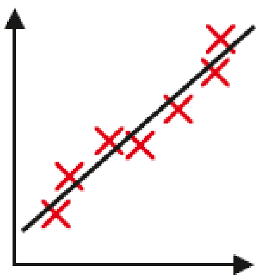
Negative correlation
As one quantity increases the other decreases.



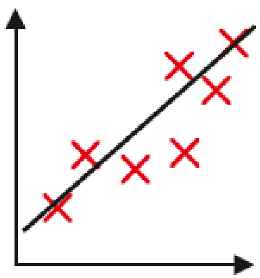
No correlation
Both quantities vary with no clear relationship

Line of Best Fit

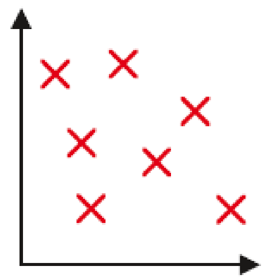
- Roughly goes through the middle of the scatter of the points
- To describe it generally: it has about as many points on one side of the line as the other, and it doesn't have to go through any of the points
- It can go through some, all or none of the points
- Strong correlation is when the scatter points lie very close to the line
- It also depends on the size of the sample from which the data was chosen



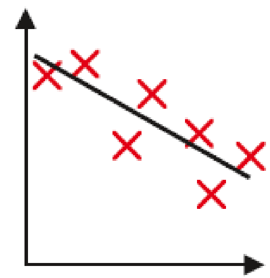
Strong positive correlation



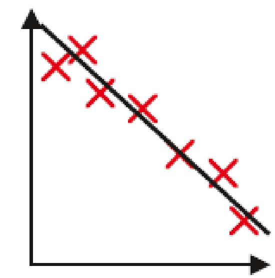
Moderate positive correlation



No correlation – no linear relationship



Moderate negative correlation



Strong negative correlation

Example

A set of students sat their mock exam in English, they sat their final exam in English at a later date. The marks obtained by the students in both examinations were as follows:

Students	A	B	C	D	E	F	G	H
Mock Results	10	15	23	31	42	46	70	75
Final Results	11	16	20	27	38	50	68	70

(a) Draw a Scatter Plot for this data and draw a line of best fit.

(b) Is there a correlation between mock results and final results?

Solution

Example

A set of students sat their mock exam in English, they sat their final exam in English at a later date. The marks obtained by the students in both examinations were as follows:

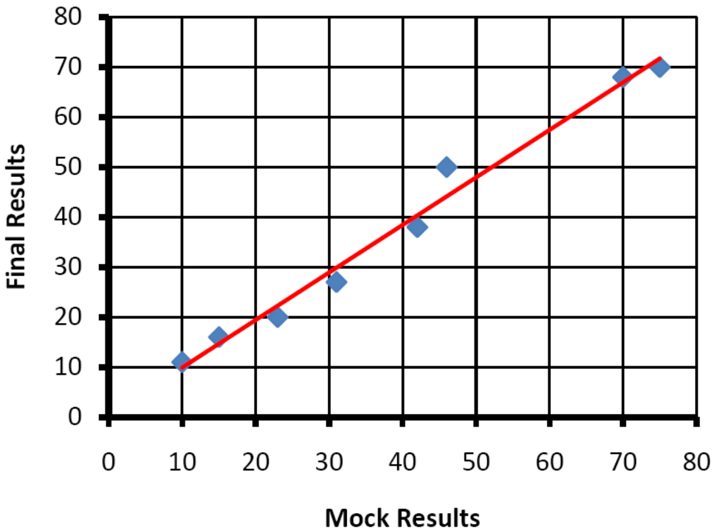
Students	A	B	C	D	E	F	G	H
Mock Results	10	15	23	31	42	46	70	75
Final Results	11	16	20	27	38	50	68	70

(a) Draw a Scatter Plot for this data and draw a line of best fit.

(b) Is there a correlation between mock results and final results?

Solution

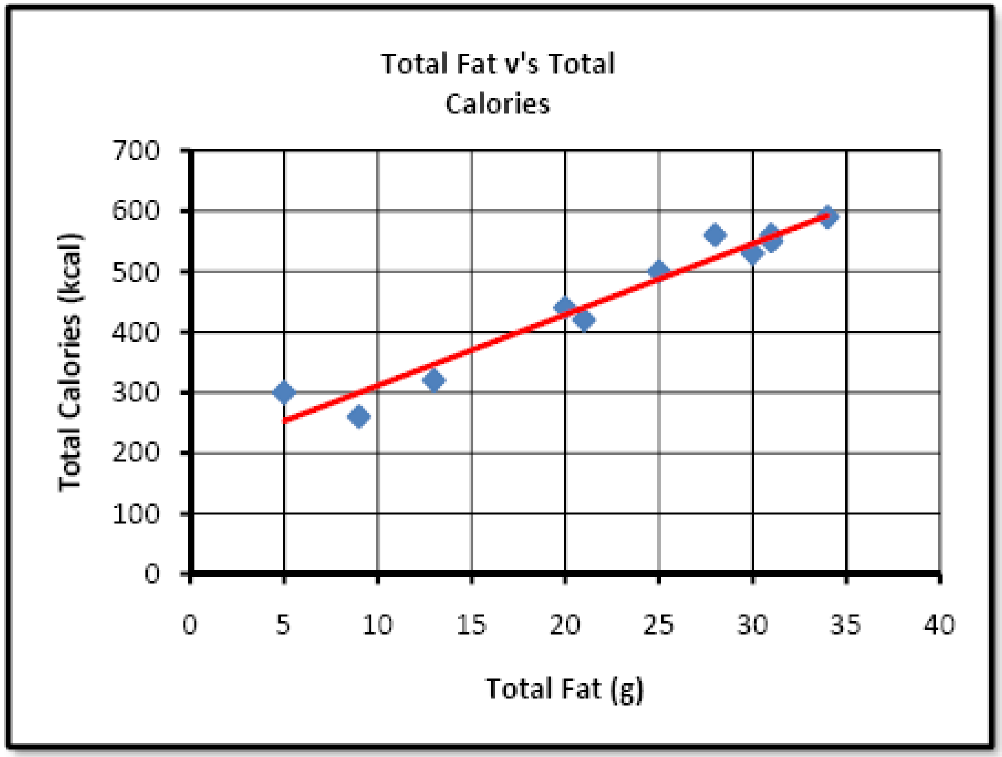
(a)



(b) There is a positive correlation between mock results and final results.

Correlation Coefficient by Calculator

Food	Total Fat (g)	Total Calories (kcal)
Hamburger	9	260
Cheeseburger	13	320
Quarter Pounder	21	420
Quarter Pounder with	30	530
Big Mac	31	560
Special	31	550
Special with Bacon	34	590
Crispy Chicken	25	500
Fish Fillet	28	560
Grilled Chicken	20	440
Grilled Chicken Light	5	300



$r = 0.9746$

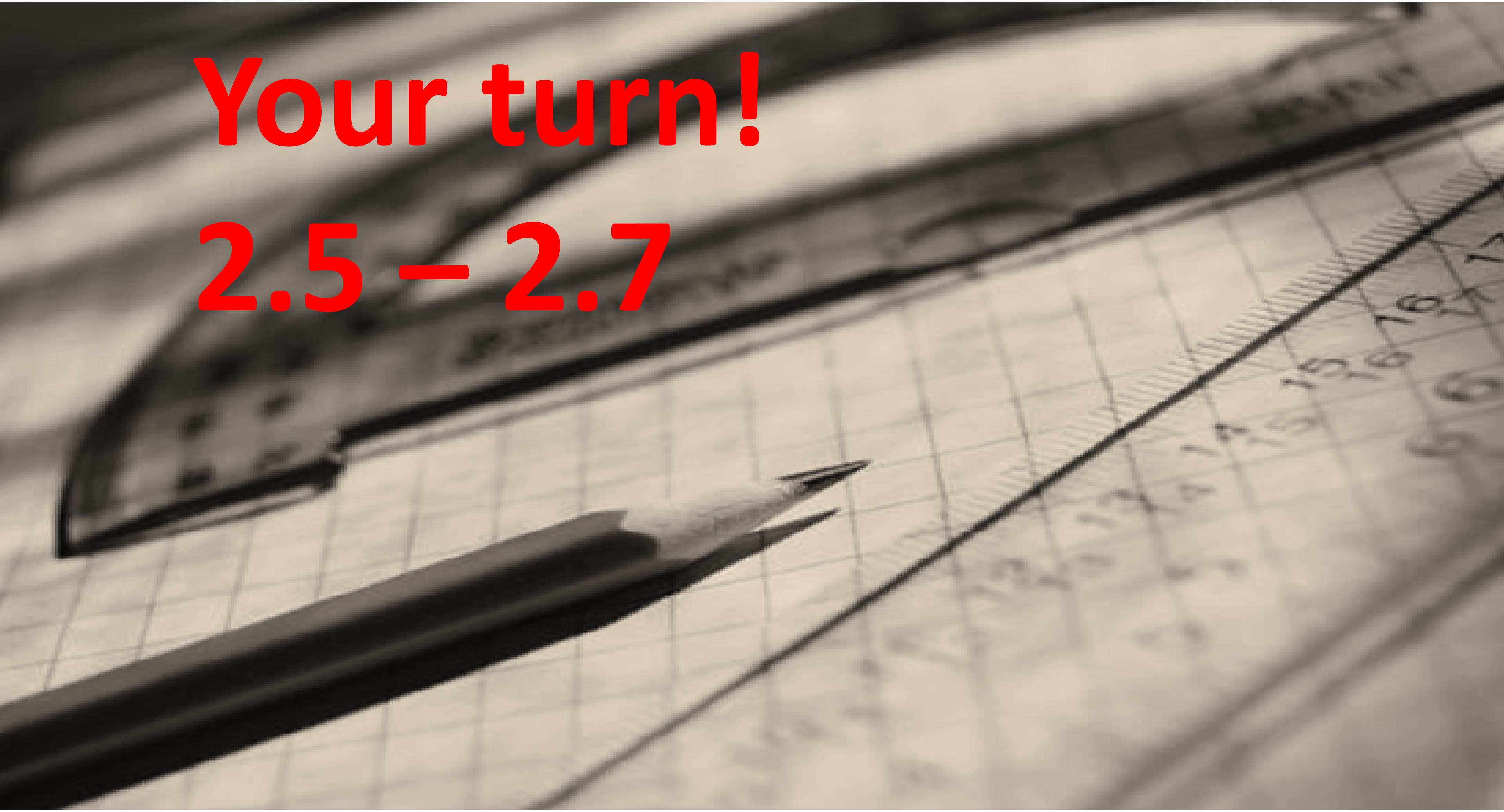
$a = 193.85$

$b = 11.73$

$y = 193.85 + 11.73x$

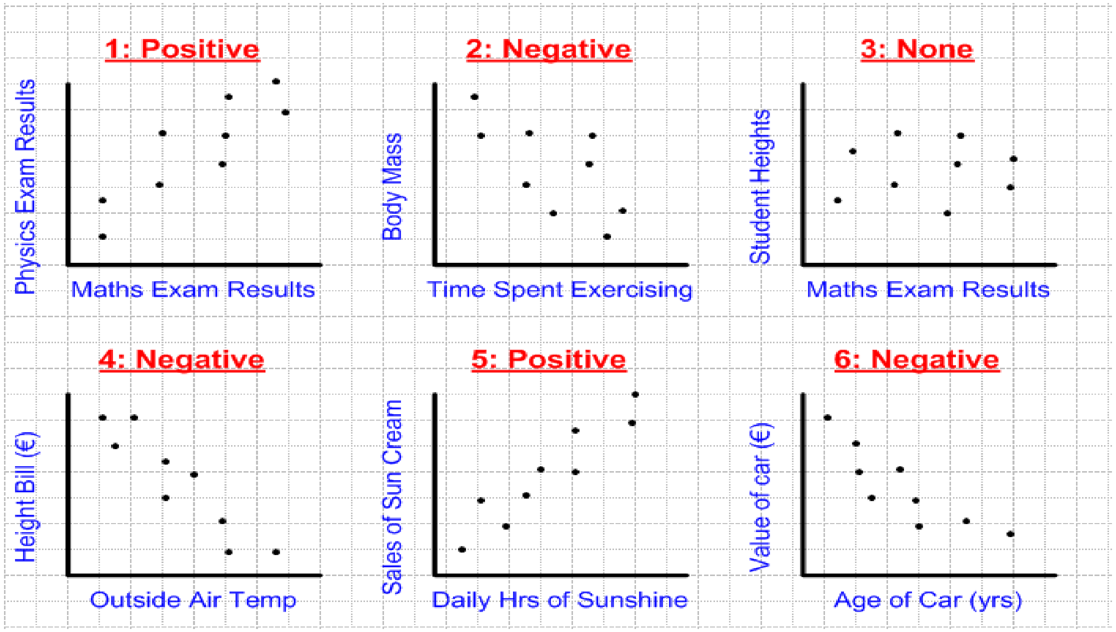
Before doing this on the calculator, the class should do a scatter plot using the data in the table. Discuss the relationship between the data (i.e. grams of fat v calories).

Your turn!
2.5 – 2.7



2.5

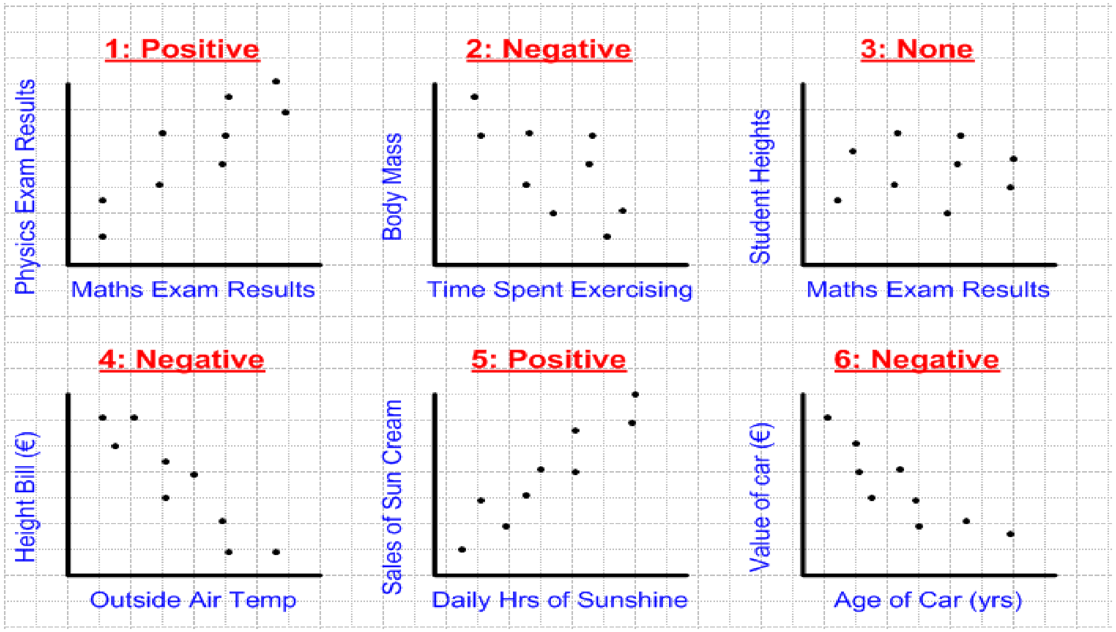
State the type of Correlation for the Scatter plots below and write a sentence describing the relationship in each case.



Solution

2.5

State the type of Correlation for the Scatter plots below and write a sentence describing the relationship in each case.



Solution

1. As maths results increases physics results tend to also increase.

2. In general the more time spend exercising, tends to lead to a decrease in body mass.

3. There is no linear relationship between Maths results and the height of students.

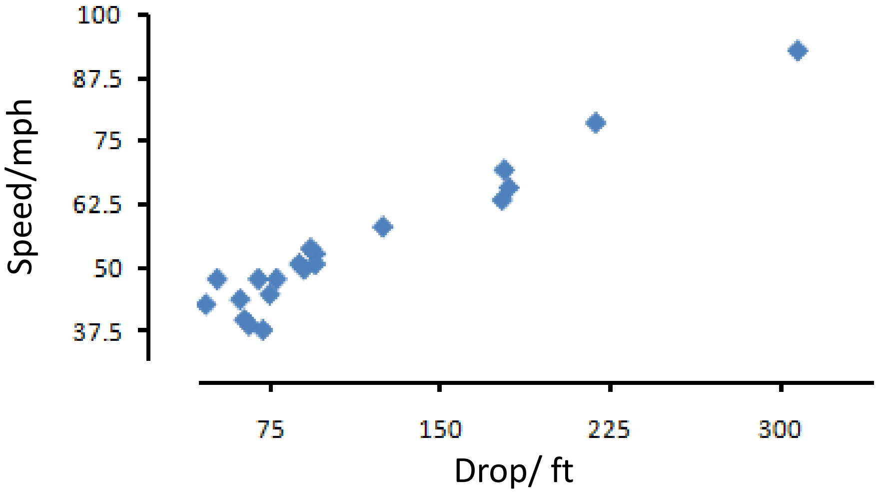
4. As the outside air temperature increases, heating bills tend to decrease.

5. As the daily hours of sunshine increases, the sale of sun cream tends to get higher.

6. In general the older the car the less its value.

2.6

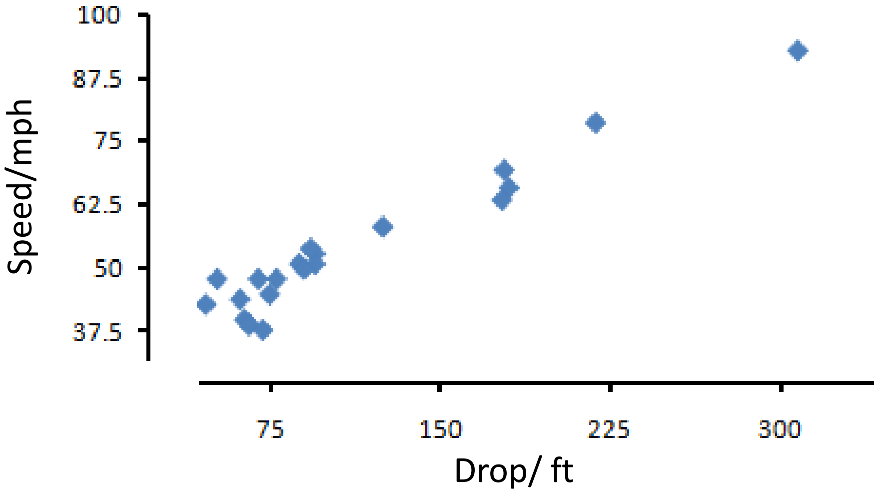
Roller coasters get all their speed by dropping down a steep initial incline,
so it makes sense that the height of that drop might be related to the speed of the coaster.
Here's a scatter plot of top Speed and largest Drop for 75 roller coasters around the world.



Solution

2.6

Roller coasters get all their speed by dropping down a steep initial incline, so it makes sense that the height of that drop might be related to the speed of the coaster. Here's a scatter plot of top Speed and largest Drop for 75 roller coasters around the world.



Solution

(a) It is appropriate to calculate correlation. Both height of the drop and speed are quantitative variables, the scatter plot shows an association that is straight enough, and there are no outliers.

(b) There is a strong, positive, linear association between drop and speed; the greater the height of the initial drop, the higher the top speed.

A candidate for office claims that “there is a correlation between television watching and crime” Criticize this statement in statistical terms.

Solution

A candidate for office claims that “there is a correlation between television watching and crime” Criticize this statement in statistical terms.

Solution

The candidate might mean that there is an association between television watching and crime. The term correlation is reserved for describing linear associations between quantitative variables. We don’t know what type of variables “television watching” and “crime” are, but they seem categorical. Even if the variables are quantitative (hours of tv watched per week, and number of crimes committed, for example), we aren’t sure that the relationship is a linear.

The politician also seems to be implying a cause-and-effect relationship between television watching and crime. Association of any kind does not imply causation.

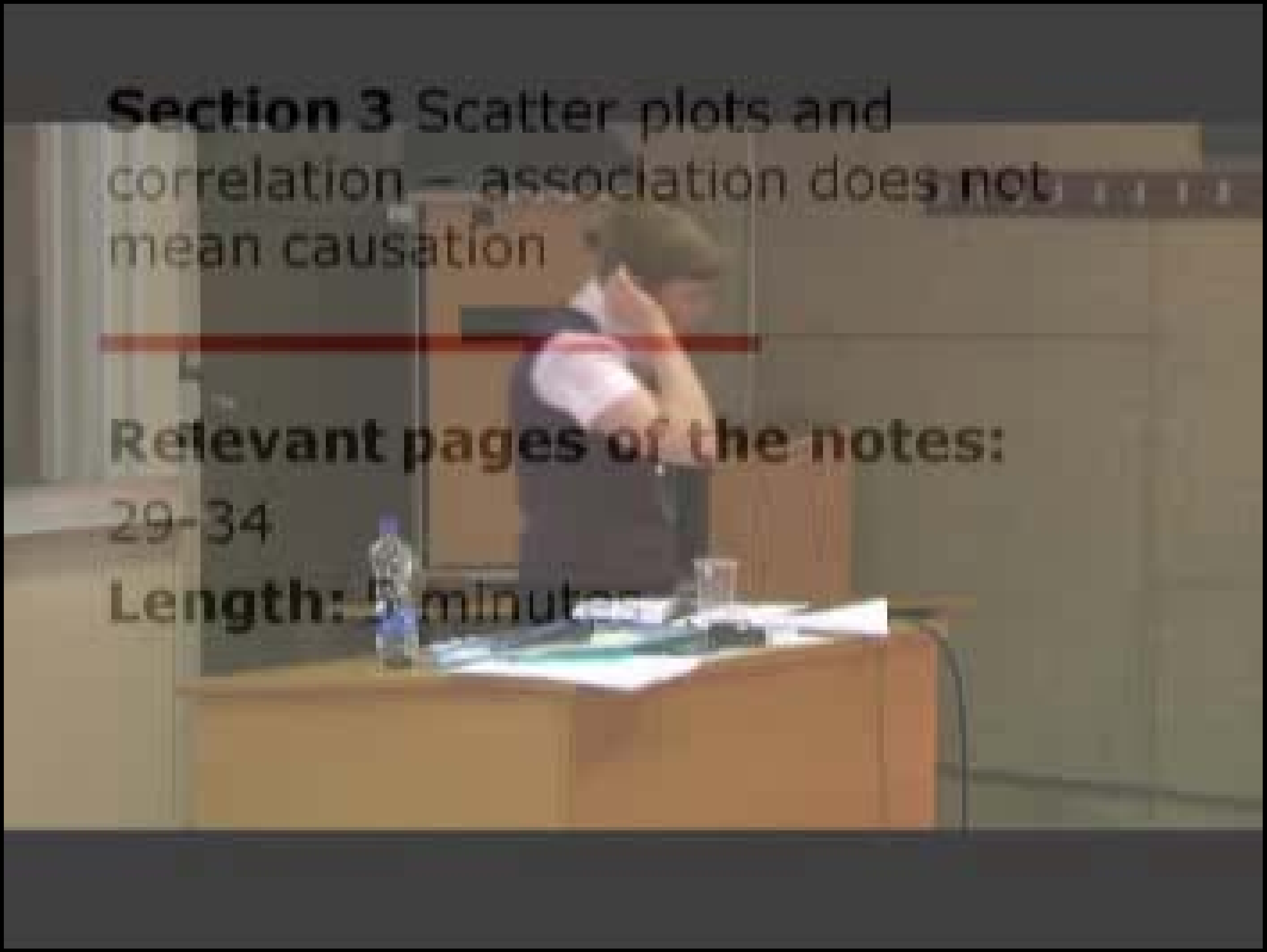
Correlation versus Causation

- Correlation is a mathematical relationship between 2 variables which are measured
- A Correlation of 0, means that there is no linear relationship between the 2 variables and knowing one does not allow prediction of the other
- Strong Correlation may be no more than a statistical association and does not imply causality
- Just because there is a strong correlation between 2 things does not mean that one causes the other. A consistently strong correlation may suggest causation but does not prove it.
- Look at these examples which show a strong correlation but do not prove causality:
- **E.g. 1** With the decrease in the number of pirates we have seen an increase in global warming over the same time period. Does this mean global warming is caused by the decrease in pirates?
- **E.g. 2** With the increase in the number of television sets sold an electrical shop has seen an increase in the number of calculators sold over the same time period. Does this mean that buying a television causes you to buy a calculator?

Criteria for Establishing Causation

- There has to be a strong consistent association found in repeated studies
- The cause has to be plausible and precede the effect in time
- Higher doses will result in stronger responses

Video Clip – Association does not mean Causation



Length: 00:02:43

Your turn!
2.8

2.8

Fast food is often considered unhealthy because much of it high in both fat and sodium. But are the two related? Here are the fat and sodium contents of several brands of burgers. Analyze the association between fat content and sodium.

Fat (g)	19	31	34	35	39	39	43
Sodium (mg)	920	1500	1310	860	1180	940	1260

Solution

2.8

Fast food is often considered unhealthy because much of it high in both fat and sodium. But are the two related? Here are the fat and sodium contents of several brands of burgers. Analyze the association between fat content and sodium.

Fat (g)	19	31	34	35	39	39	43
Sodium (mg)	920	1500	1310	860	1180	940	1260

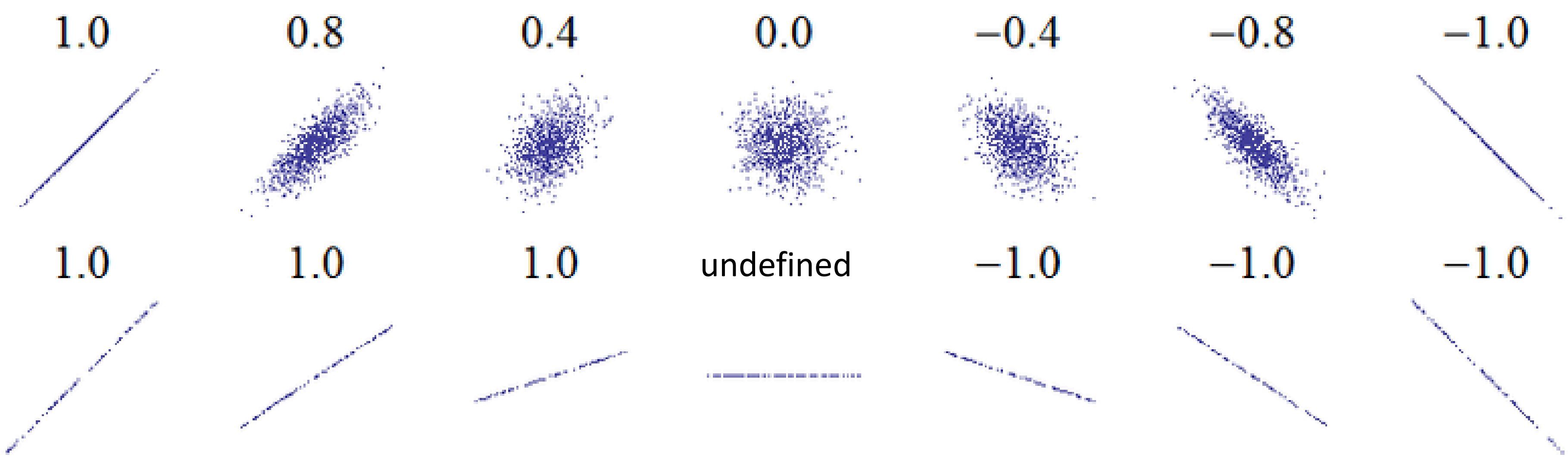
Solution

There is no apparent association between the number of grams of fat and the number of milligrams of sodium in several brands of fast food burgers.

The correlation is only $r = 0.199$, which is close to zero, an indication of no association.

One burger had a much lower fat content than the other burgers, at 19 grams of fat, with 920 milligrams of sodium. Without this (comparatively) low fat burger, the correlation would have been even lower.

Correlation & Slope



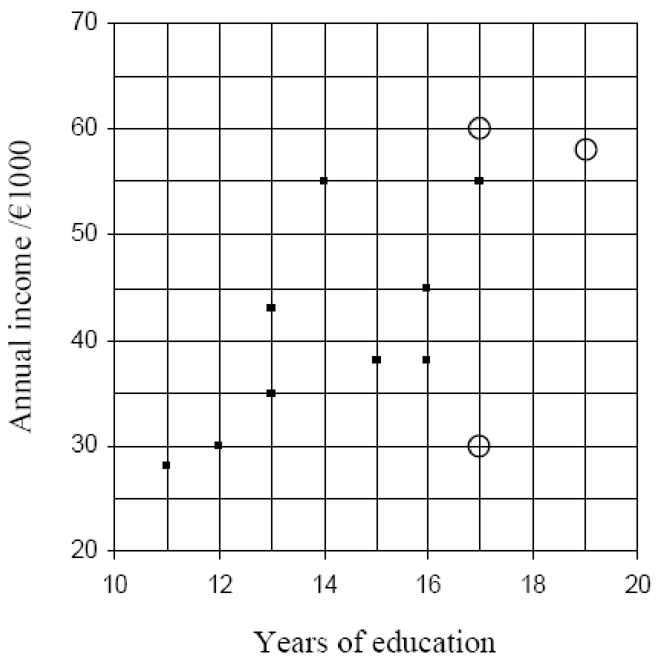
Several sets of (x, y) points, with the correlation coefficient of x and y for each set.

Note that the correlation reflects the spread and direction of a linear relationship but not the gradient (slope) of that relationship, **N.B.:** the figure in the centre of the second line has a slope of 0 but in that case the correlation coefficient is undefined because the variance of *Y* is zero.

The gradient (slope) of the line of best fit is not important when dealing with correlation, except that a vertical or horizontal line of best fit means that the variables are not connected. *[The sign of the slope of the line of best fit will be the same as that of the correlation coefficient because both will be in the same direction.]*

An economics student wants to find out whether the length of time people spend in education affects the income they earn. The student carries out a small study. Twelve adults are asked to state their annual income and the number of years they spent in full-time education. The data are given in the table below, and a partially completed scatter plot is given.

Years of education	Income /€1,000
11	28
12	30
13	35
13	43
14	55
15	38
16	45
16	38
17	55
17	60
17	30
19	58



(i) The last three rows of data have not been included on the scatter plot. Insert them now.

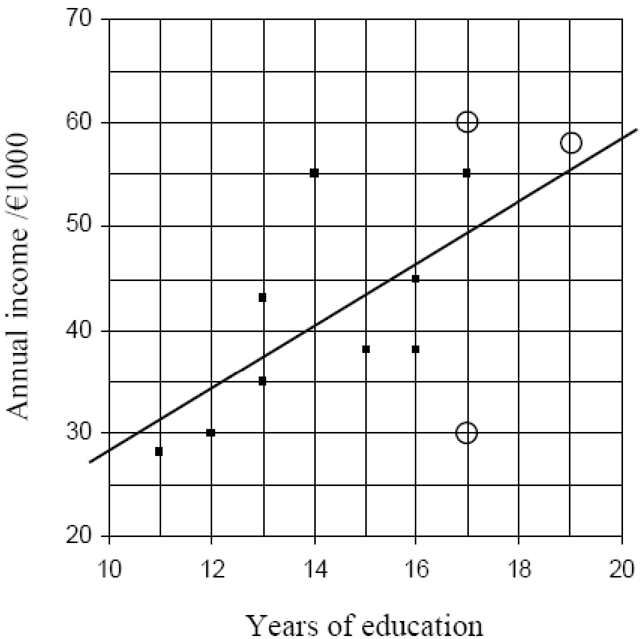
(ii) What can you conclude from the scatter plot?

It looks as though people with more education tend to have

a higher annual income. (etc.)

An economics student is interested in finding out whether the length of time people spend in education affects the income they earn. The student carries out a small study. Twelve adults are asked to state their annual income and the number of years they spent in full-time education. The data are given in the table below, and a partially completed scatter plot is given.

Years of education	Income /€1,000
11	28
12	30
13	35
13	43
14	55
15	38
16	45
16	38
17	55
17	60
17	30
19	58



(i) The last three rows of data have not been included on the scatter plot. Insert them now.

(ii) Calculate the correlation coefficient.

Answer: 0.623

(iii) What can you conclude from the scatter plot and the correlation coefficient?

There is a moderate positive correlation between the variables. That is, those with more education tend to have higher incomes

(iv) Add the line of best fit to the completed scatter plot above.

(v) Use the line of best fit to estimate the annual income of somebody who has spent 14 years in education.

Answer: €40,000

(vi) By taking suitable readings from your diagram, or otherwise, calculate the slope of the line of best fit.

line passes through (10, 28) and (20, 58).

$\text{slope} = \frac{58-28}{20-10} = \frac{30}{10} = 3$ [or €3000]

(vii) Explain how to interpret this slope in this context?

It is the expected (average) increase in income per additional year of education. That is, each additional year of education corresponds to an average increase of €3000 in annual income.

Notes