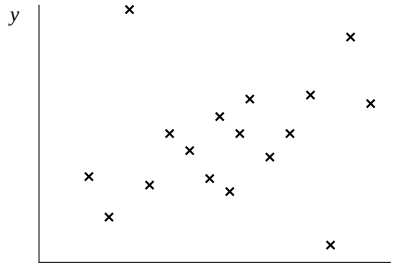
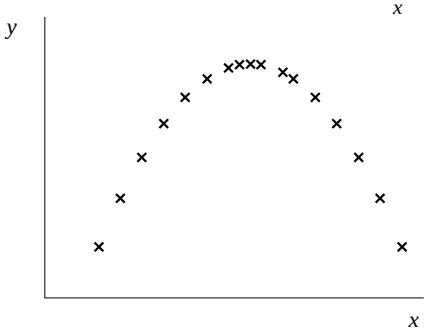
**Statistics Sector 1: Data Presentation Bivariate Data**

Aims:

- Interpret scatter graphs commenting on correlation and outliers.
- Interpret pmcc and regression lines in context
- Understand the effects on variables can have on pmcc
- When pmcc is appropriate/limitations
- Understand that cause and effect




**Scatter Diagrams**

The term correlation is used to imply that there is a linear relationship between two random variables. We can plot a scatter graph to decide whether correlation exists. We should **not** calculate the correlation coefficient for data that is not linear.

In an exam question you may be asked to comment on what the scatter diagram reveals. You must comment on whether the correlation is positive or negative and any points that do not fit the data (outliers).

Example 1

9 students were asked to measure their hand span, $x$cm, and the length of their forearm, $y$cm. The results are given in the table below.

| Pupil | A | B | C | D | E | F | G | H | J |
|---|---|---|---|---|---|---|---|---|---|
| Hand span ($x$cm) | 20.5 | 21.8 | 23.2 | 22.4 | 24.1 | 20.9 | 21.4 | 22.2 | 22.8 |
| Forearm ($y$cm) | 29.0 | 29.0 | 31.4 | 30.8 | 32.6 | 29.5 | 30.2 | 45.0 | 31.3 |

a) Use your calculator or geogebra to draw a scatter diagram for the data.
b) Give **two** distinct comments on what the scatter diagram reveals.

**Product Moment Correlation Coefficient (pmcc)**

The product moment correlation coefficient, denoted by $r$, is used to measure the strength of linear correlation between two random variables $X$ and $Y$.

**You will not be asked to calculate r in the exam, just interpret the value.**

Scientific and graphical calculators should have a correlation coefficient function ($r$) which you may use in other subjects.

**Calculating pmcc**

It is advisable to draw a scatter diagram, as above, because this gives a good visual representation of the data. The value for pmcc must be $-1 \leq r \leq 1$. Outliers identified maybe removed before calculating pmcc ($r$) as they will adversely affect results.

Example 2

9 students were asked to measure their hand span, $x$cm, and the length of their forearm, $y$cm. The results are given in the table below.

| Pupil | A | B | C | D | E | F | G | H | J |
|---|---|---|---|---|---|---|---|---|---|
| Hand span ($x$cm) | 20.5 | 21.8 | 23.2 | 22.4 | 24.1 | 20.9 | 21.4 | 22.2 | 22.8 |
| Forearm ($y$cm) | 29.0 | 29.0 | 31.4 | 30.8 | 32.6 | 29.5 | 30.2 | 45.0 | 31.3 |

The correlation coefficient is calculated $r = 0.198$.

Isla realises that one of the pieces of data has been recorded incorrectly and needs to be removed. What affect will removing the data of the pupils you identified in example 1 have on the value of $r$.

**Interpreting pmcc**

When asked to interpret the value of pmcc you must mention the strength of the correlation, whether it is positive or negative and the two variables it is between in the context of the question (you must not talk about $x$ and $y$). You must also be able to approximate the value of $r$ from scatter diagrams without calculations.

General Rules

| **Magnitude** | **Correlation** |
|---|---|
| $0.9 \leq r < 1$ | Very Strong |
| $0.7 \leq r < 0.9$ | Strong |
| $0.3 \leq r < 0.7$ | Moderate |
| $r < 0.3$ | Weak |

$r = 0$ does NOT imply no relationship just no linear relationship.

A positive correlation implies that as one variable increases the other variable also increases. Whereas negative correlation implies that as one variable increases the other decreases.

Example 3

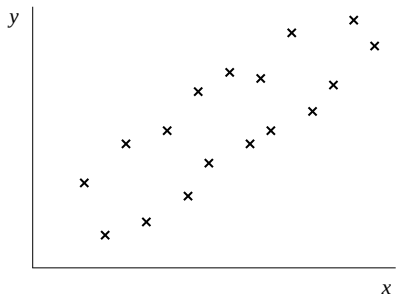
The correlation coefficient between the number of books sold, $x$, and the retail price, £$y$, is 0.768. Interpret the value of this in context.

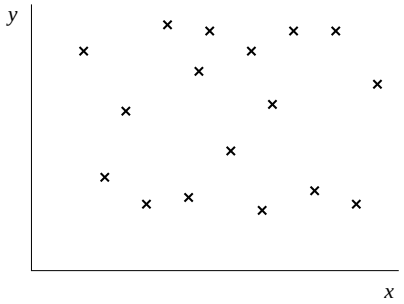
Example 4

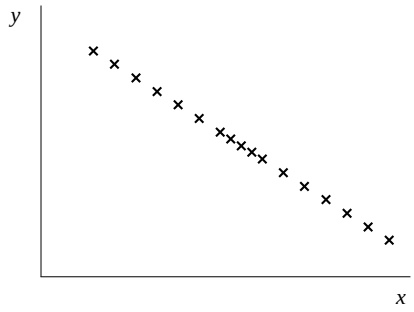
Graeme calculates the product moment correlation coefficient between the length, $x$cm , and the weight, $y$ grams, of a sample of 12 lizards. He obtains the value $-0.724$, comment, with a reason, on the likely validity of Graeme's value.

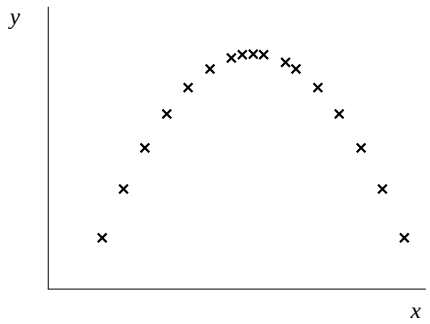
Example 5

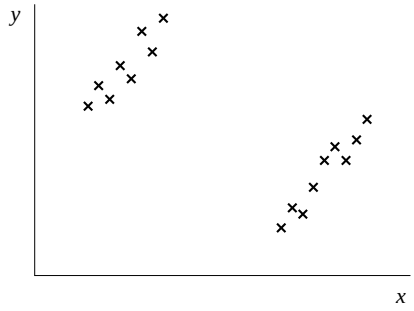
Estimate, **without undertaking any calculations**, the value of the product moment correlation coefficient between the variables $x$ and $y$ in **each** of the six scatter diagrams. Also comment, where appropriate, on any notable features revealed by the scatter diagrams.

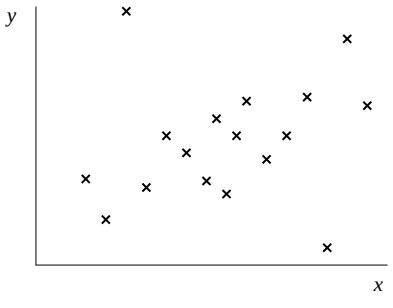


**Limitations**

It is possible to find sets of data which are highly correlated but where it would be incorrect to ascribe a connection between the two variables. It could be simply coincidence or it may be that the two variables are correlated via a third hidden variable, this is called **spurious** correlation. For example life expectancy and the number of household owning computer has increased over the last decade. Life expectancy has not increased as a result of more people owning computers but the mathematical correlation between them could be related to other factors such as technological advances and increased prosperity.

Example 6

For a project, Jamie, obtains information on a random sample of children at a local primary school. From the sample he calculates the value of the correlation coefficient:

a) between a child's age and a child's height to be -0.437
b) between a child's height and a child's weight to be 0.763
c) between a child's height and a child's average hand span to be 1.142
d) between a child's height and reading ability score to be 0.621.

Classify each of these statements as definitely correct, probably correct, probably incorrect or definitely incorrect. Give a reason for each answer.

**Regression Lines**

When bivariate data are displayed on a scatter diagram you often need to draw a line of best fit through the points. You can do this roughly by eye. However, drawing a line by eye is a somewhat haphazard process. Standard statistical software, and your graphical calculator can calculate the **least squares regression line.** Regression is used to imply the calculation of an equation for the **linear** relationship between two variables, $X$ and $Y$. The general form for the equation of the least squares regression line $y$ on $x$ is:

$$y = a + bx$$

where $b$ denotes the gradient and $a$ denotes the intercept with the $y$ axis.

**Interpreting the Regression Line**

The regression line gives important information about the relationship between the two variables. You need to be able to interpret the values of $a$ ($y$-intercept) and $b$ (gradient) in the **context** of the question.

- The intercept, $a$, is the estimated value of the dependent variable (usually $y$) when the independent variable (usually $x$) is zero. This is often extrapolation and can give an unrealistic answer.
- $b$ is the gradient. This represents the average increase/decrease in the dependent variable (usually $y$) for each one unit increase of the independent variable (usually $x$).
- The gradient of the regression line is NOT a measure of the strength of the correlation.

Example 6

The regression line $y = 25.52 + 0.64x$ gives an estimate for pulse rate, $y$ beats per minute, given a person's weight, $x$ kg.
Interpret the values of $a$ and $b$ in context.

Don't forget to include the units!

Example 7

The regression line $D = -5.18 + 1.94t$ gives an estimate of the number of cold drinks sold, $D$, dependant on the temperature, $t$°C.
Interpret the values of $a$ and $b$ in context.

**Exam Questions**

Question 1

**(a)** Three airport management trainees, Ryan, Sunil and Tim, were each instructed to select a random sample of 12 suitcases from those waiting to be loaded onto aircraft.

Each trainee also had to measure the volume, $x$, and the weight, $y$, of each of the 12 suitcases in his sample, and then calculate the value of the product moment correlation coefficient, $r$, between $x$ and $y$.

- Ryan obtained a value of $-0.843$.
- Sunil obtained a value of $+0.007$.

Explain why neither of these two values is likely to be correct. *(2 marks)*

**(b)** Peggy, a supervisor with many years' experience, measured the volume, $x$ cubic feet, and the weight, $y$ pounds, of each suitcase in a random sample of 6 suitcases, and then obtained a value of 0.612 for $r$.

- Ryan and Sunil each claimed that Peggy's value was different from their values because she had measured the volumes in cubic feet and the weights in pounds, whereas they had measured the volumes in cubic metres and the weights in kilograms.
- Tim claimed that Peggy's value was almost exactly half his calculated value because she had used a sample of size 6 whereas he had used one of size 12.

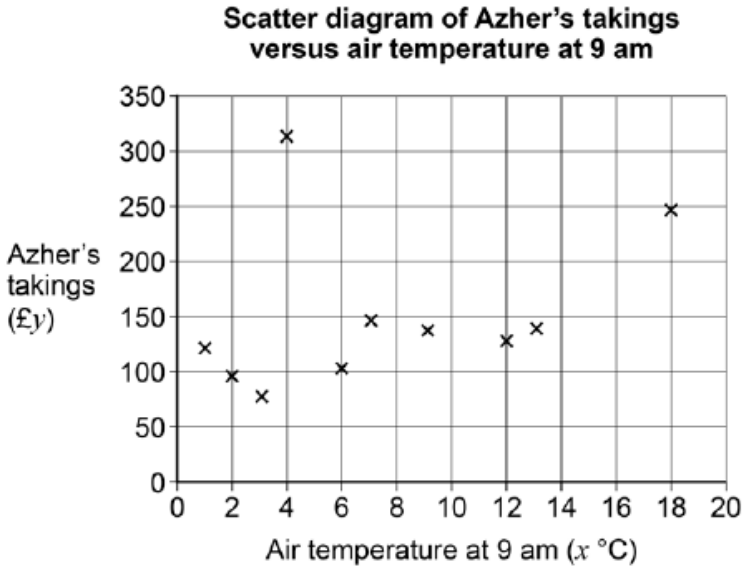
Explain why neither of these two claims is valid. *(2 marks)*

Question 2

Each Monday, Azher has a stall at a town's outdoor market. The table below shows, for each of a random sample of 10 Mondays during 2003, the air temperature, $x$°C, at 9 am and Azher's takings, £$y$.

| **Monday** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** |
|---|---|---|---|---|---|---|---|---|---|---|
| $x$ | 2 | 6 | 9 | 18 | 1 | 3 | 7 | 12 | 13 | 4 |
| $y$ | 97 | 103 | 136 | 245 | 121 | 78 | 145 | 128 | 141 | 312 |

(a) A scatter diagram of these data is shown below.




Give **two** distinct comments, in context, on what this diagram reveals.

(b) One of the Mondays is found to be Easter Monday, the busiest Monday market of the year.

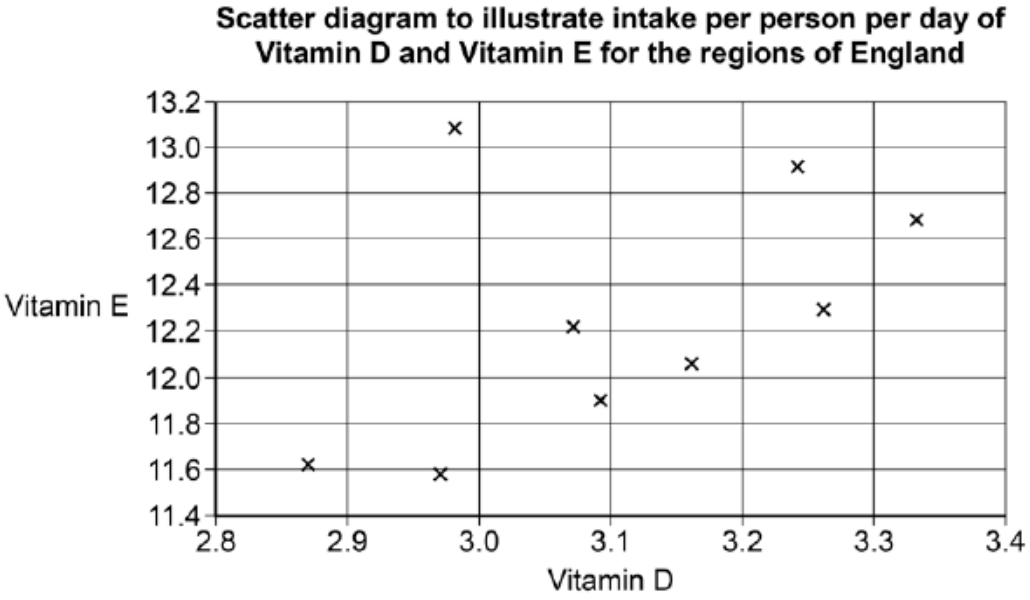
Identify which Monday this is likely to be.

Question 3

The following table gives information on the average intake, per person per day, of Vitamin D, $\mu$g and Vitamin E, mg, for nine regions in England.

| Region | Vitamin D | Vitamin E |
|---|---|---|
| North East | 2.87 | 11.63 |
| North West | 3.09 | 11.92 |
| Yorks and Humber | 2.97 | 11.59 |
| East Midlands | 3.24 | 12.95 |
| West Midlands | 3.07 | 12.24 |
| East | 3.33 | 12.72 |
| London | 2.98 | 13.12 |
| South East | 3.16 | 12.09 |
| South West | 3.26 | 12.32 |

The scatter diagram illustrates this data.




(a) Give two distinct comments on what the scatter diagram reveals.

(b) The data for the London region is removed from the table and the data point for London is removed from the scatter diagram.

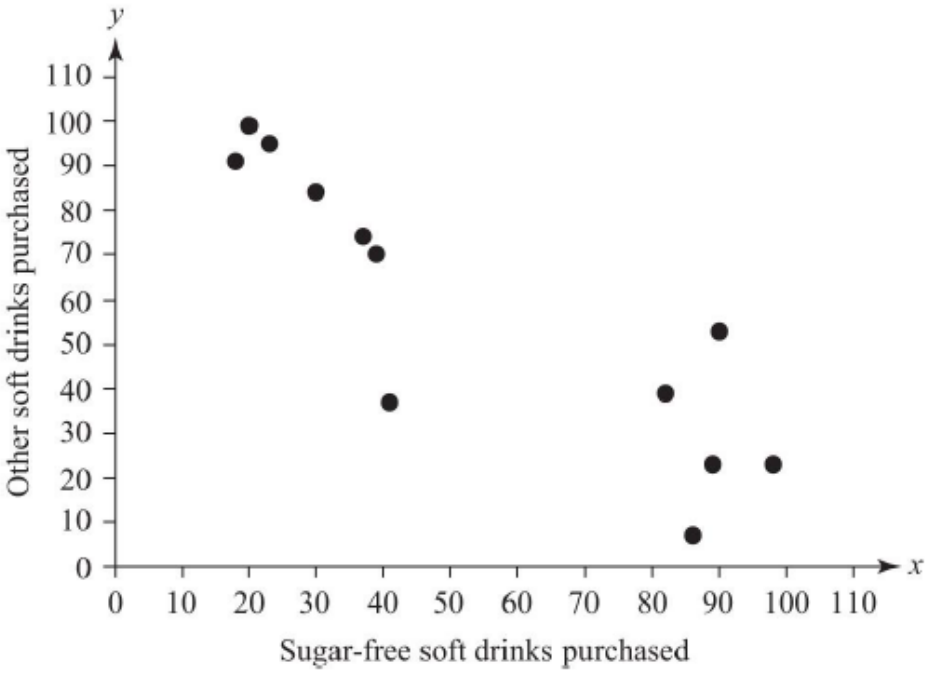
State what effect this would have on the correlation between average intake of Vitamin D and Vitamin E. Circle the correct answer.

| Correlation would be weaker and positive | Correlation would stay the same | Correlation would be negative | Correlation would be stronger and positive |
|---|---|---|---|

Question 4

For English households the scatter diagram shows the number of millilitres of sugar-free soft drink, $x$, purchased per person per week and the number of millilitres of other soft drink, $y$, purchased per person per week.

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $x$ | 23 | 37 | 30 | 39 | 86 | 98 | 90 | 82 | 89 | 41 | 18 | 20 |
| $y$ | 95 | 74 | 84 | 70 | 7 | 23 | 53 | 39 | 23 | 37 | 91 | 99 |




**a** Describe and give an interpretation of the correlation between purchases of sugar-free and other soft drinks. **[2 marks]**

The equation of the regression line of $y$ on $x$ is $y = 105.9 - 0.882x$

**b** Give an interpretation of the gradient of the regression line. **[1 mark]**