IB Mathematics SL

SMARTPREP® | Haese Mathematics

Topic 5 | Two variable statistics

Theory card T | Question cards

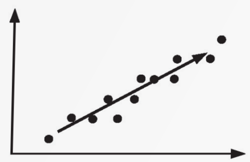
Front side

Correlation refers to the relationship or association between two variables.

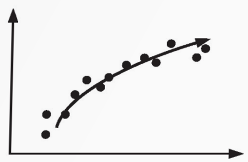
We can use a **scatter diagram** of the data to help identify **outliers** and to describe the correlation.

Linearity

These points are roughly linear.



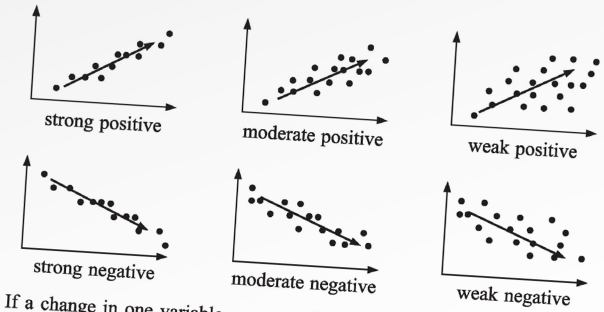
These points do not follow a linear trend.



F

Unfold

Strength and direction



If a change in one variable *causes* a change in the other variable, there is a **causal relationship** between them.

To measure the strength of the relationship between two variables, we use **Pearson's product-moment correlation coefficient** r.

- r always lies in the range $-1 \leqslant r \leqslant 1$.
- The sign of r indicates the direction of correlation. $r > 0$ indicates positive correlation. $r < 0$ indicates negative correlation.
- The size of r indicates the strength of correlation.

$\lvert r\rvert = 1$	perfect correlation
$0.95 \leqslant \lvert r\rvert < 1$	very strong correlation
$0.87 \leqslant \lvert r\rvert < 0.95$	strong correlation
$0.5 \leqslant \lvert r\rvert < 0.87$	moderate correlation
$0.1 \leqslant \lvert r\rvert < 0.5$	weak correlation
$0 < \lvert r\rvert < 0.1$	very weak correlation

The **line of best fit** drawn by eye should pass through the **mean point** $(\overline{x}, \overline{y})$.

A more accurate line of best fit is the **least squares regression line.**

When using a line of best fit to estimate values, **interpolation** is usually reliable, whereas **extrapolation** may not be.

B

IB Mathematics SL

SMARTPREP®

Haese Mathematics

Topic 5 | Two variable statistics

Theory card T | **Question cards**

Front side

114 | ? SL **Topic 5 | Two variable statistics**

The house size and number of cars owned was recorded for each household in a particular neighbourhood:

House size (m^2)	210.1	267.6	238.5	201	196.5
Number of cars	2	2	1	1	1

House size (m^2)	214.7	210	184.8	295.9	159.2
Number of cars	1	1	1	3	0

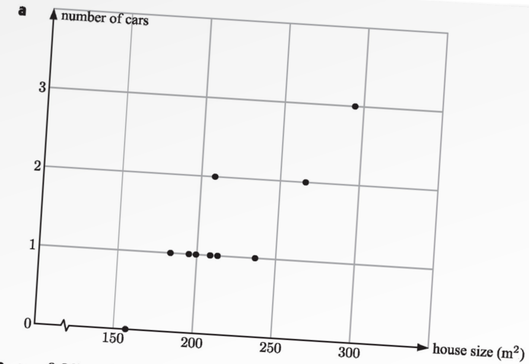
a Draw a scatter diagram for the data.

b Calculate the correlation coefficient r for the data and describe the correlation between the variables.

c Is it reasonable to conclude that there is a causal relationship between the variables? Explain your answer.

Unfold

a



b $r \approx 0.867$. There is a strong positive correlation between the variables.

c No. Despite the strong positive correlation, it is unlikely that a larger house will cause a household to buy more cars, or vice versa. Rather, an increase in both variables may be driven by the income of the household. Households with higher incomes will be more likely to afford larger houses and more cars.

IB Mathematics SL

Topic 5 | Two variable statistics

Theory card T **Question cards**

Question side

Consider the data below:

x	1	2	3	4	5
y	2.6	4.5	5.7	6.5	9.8

a Find the least squares regression line.

b Calculate $\overline{x}$ and $\overline{y}$.

c Estimate the value of y when $x = \overline{x}$.

d Hence explain why a line of best fit by eye should pass through the mean point $(\overline{x}, \overline{y})$.

Turn around

Answer side

a $y = 0.9 + 1.64x$ {using technology}

b $\overline{x} = 3$ and $\overline{y} = 5.82$ {using technology}

c When $x = \overline{x} = 3$, $y \approx 0.9 + 1.64(3)$
≈ 5.82
$= \overline{y}$

d The least squares regression line passes through the mean point $(\overline{x}, \overline{y})$ as shown in **c**. Since the least squares regression line best fits the data and a line of best fit by eye is an *estimate* of this line, any line of best fit by eye should also pass through the mean point.

IB Mathematics SL

SMARTPREP®

Haese Mathematics

Topic 5 | Two variable statistics

Theory card | Question cards

Front side

The population of an endangered bird species on an island was recorded over several years:

Years after 2008 (t years)	0	2	4	6	8
Number of birds (B)	5927	5620	5248	5011	4851

a Draw a scatter diagram for the data.

b Without calculating r, explain why B and t are negatively correlated.

c Find the equation of the least squares regression line and the r-value.

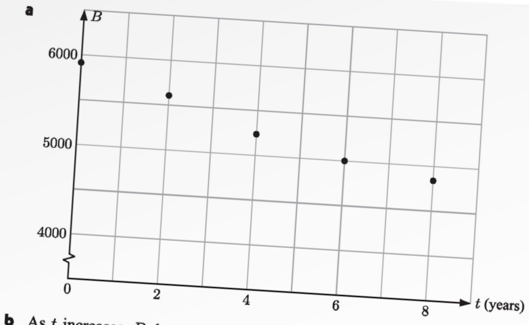
d Estimate the number of birds in:

 i 2011 **ii** 2025.

e Comment on the reliability of your answers to **d**.

Unfold

a



b As t increases, B decreases, so the variables are negatively correlated.

c $B \approx -138t + 5880$, $r \approx -0.990$ {using technology}

d **i** In 2011, $t = 2011 - 2008 = 3$

So when $t = 3$, $B \approx -138 \times 3 + 5880$

≈ 5466 birds

ii In 2025, $t = 2025 - 2008 = 17$

So when $t = 17$, $B \approx -138 \times 17 + 5880$

≈ 3534 birds

e The estimate in **d i** is an interpolation, so it is likely to be reliable.
The estimate in **d ii** is an extrapolation, so it may not be very reliable. The linear trend may not continue in the long term.

IB Mathematics SL

SMARTPREP®

Haese Mathematics

Topic 5 | Probability

Theory card T   Question cards

Front side

117 T SL **Topic 5 | Probability**

A **trial** occurs each time we perform an experiment.

An **outcome** is a possible result from one trial of an experiment. If the outcomes of an experiment have the same probability of occurring, we say they are **equally likely**.

The **sample space** U is the set of all possible outcomes.

An **event** is an outcome or set of outcomes with a particular characteristic.

Experimental probability

In many situations, we can only measure the probabilities of different outcomes by experimentation.

Experimental probability = relative frequency.

F

Unfold

Theoretical probability

For an event A containing equally likely outcomes, the probability of A occurring is $\mathrm{P}(A) = \frac{n(A)}{n(U)}$.

For any event A, $0 \leqslant \mathrm{P}(A) \leqslant 1$.

The **union** A *or* B is $A \cup B$. We take A or B to mean "A or B or both".

The **intersection** A *and* B is $A \cap B$.

$\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \cap B)$.

For **disjoint** or **mutually exclusive** events, $A \cap B = \varnothing$ and $\mathrm{P}(A \cap B) = 0$.

For any event A, A' is the event that A does not occur. A and A' are **complementary events**, and $\mathrm{P}(A) + \mathrm{P}(A') = 1$.

$A \cup A' = U$ and $A \cap A' = \varnothing$.

Two events are **independent** if the occurrence of each of them does not affect the probability that the other occurs. An example of this is sampling **with replacement**.

Two events are **dependent** if the occurrence of one of them *does* affect the probability that the other occurs. An example of this is sampling **without replacement**.

You should be able to use **Venn diagrams**, **tree diagrams**, **counting principles**, and **tables of outcomes** to calculate probabilities. You should also be able to use Venn diagrams to verify set identities.

Conditional probability

For any two events A and B, $\mathrm{P}(A \mid B) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)}$.

For independent events, $\mathrm{P}(A) = \mathrm{P}(A \mid B) = \mathrm{P}(A \mid B')$ and $\mathrm{P}(A \cap B) = \mathrm{P}(A)\mathrm{P}(B)$.

B

IB Mathematics SL

SMARTPREP®

Haese Mathematics

Topic 5 | Probability

Theory card T | **Question cards**

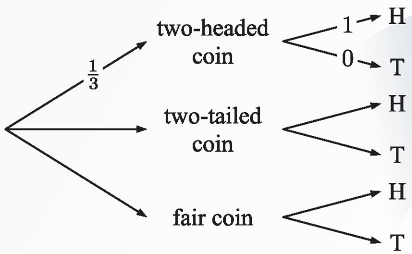
Question side

A two-headed coin, a two-tailed coin, and a fair coin with a head and a tail, are placed in a box. A coin is randomly selected from the box, and is flipped.

a Copy and complete this tree diagram.

b Find the probability that the result of the toss is:

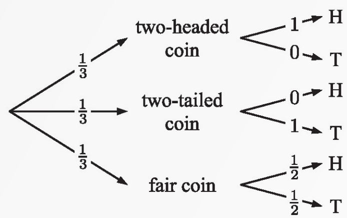
i a head **ii** a tail.



Turn around

Answer side

a



b **i** $P(H) = 1 \times \frac{1}{3} + 0 \times \frac{1}{3} + \frac{1}{2} \times \frac{1}{3}$

$= \frac{1}{3} + \frac{1}{6}$

$= \frac{1}{2}$

ii $P(T) = 1 - P(H) = \frac{1}{2}$ {using **i**}

IB Mathematics SL

SMARTPREP® | Haese Mathematics

Topic 5 | Probability

Theory card | Question cards

Question side

a Show that for events A and B that have equally likely outcomes, the probability of A given B is $\mathrm{P}(A \mid B) = \dfrac{n(A \cap B)}{n(B)}$.

b In a survey, 30 people were asked if they drive or take public transport to work. 5 people said that they only drive, 10 people said that they only take public transport, and the remaining 15 people said they do both. Given that a randomly selected person from the survey takes public transport, what is the probability that they also drive to work?

Turn around

Answer side

a Since the outcomes in A and B are equally likely,

$\mathrm{P}(A \cap B) = \dfrac{n(A \cap B)}{n(U)}$ and $\mathrm{P}(B) = \dfrac{n(B)}{n(U)}$

$$\therefore\ \mathrm{P}(A \mid B) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)} = \frac{n(A \cap B)}{n(U)} \times \frac{n(U)}{n(B)} = \frac{n(A \cap B)}{n(B)}$$

b Let D be the event that a person drives and let P be the event that a person takes public transport.

$$\therefore\ \mathrm{P}(D \mid P) = \frac{n(D \cap P)}{n(P)} \quad \{\text{from } \mathbf{a}\}$$

$$= \frac{15}{10 + 15} = \frac{15}{25} = \frac{3}{5}$$

IB Mathematics SL

SMARTPREP® | Haese Mathematics

Topic 5 | Probability

Theory card T | **Question cards**

Question side

120 | ? SL **Topic 5 | Probability**

A and B are events such that $P(A) = 0.6$ and $P(B) = 0.55$.

a Explain why A and B cannot be mutually exclusive.

b **i** Determine $P(A \cap B)$ if $P(A \cup B) = 0.8$.

ii Hence calculate $P(A \mid B)$.

Turn around

Answer side

120 | ! SL **Topic 5 | Probability**

a If A and B were mutually exclusive, then

$P(A \cup B) = P(A) + P(B)$ must be $\leqslant 1$.

But $P(A) + P(B) = 0.6 + 0.55 = 1.15 > 1$.

$\therefore$ the intersection of A and B must be non-empty.

$\therefore$ A and B cannot be mutually exclusive.

b **i** $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

$\therefore P(A \cap B) = P(A) + P(B) - P(A \cup B)$

$= 0.6 + 0.55 - 0.8$

$= 0.35$

ii $P(A \mid B) = \frac{P(A \cap B)}{P(B)}$

$= \frac{0.35}{0.55}$ {using **i**}

$= \frac{7}{11}$

IB Mathematics SL

Topic 5 | Probability

Theory card T | **Question cards**

Question side

A blood test is used to test for a rare disease. Amongst diseased individuals, the disease is detected 98% of the time. Amongst non-diseased individuals, the test result is a false positive 15% of the time. It is known that 1% of the population has the disease.

a Find the probability that a randomly selected person tests positive for the disease.

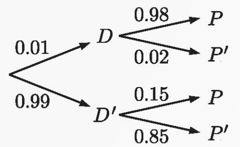
b If a randomly selected person tests positive for the disease, find the probability that they actually have it.

c Suggest a reason why doctors should not rely on a single test to determine whether a patient has the disease.

Turn around

Answer side

a The tree diagram uses D for diseased
P for positive test result.



$\therefore$ $P(P) = P(D \cap P) + P(D' \cap P)$
$= 0.01 \times 0.98 + 0.99 \times 0.15$
$= 0.1583$

b $P(D \mid P) = \dfrac{P(D \cap P)}{P(P)}$
$= \dfrac{0.01 \times 0.98}{0.1583} \approx 0.0619$

c Doctors should not rely on a single test to determine whether a patient has the disease because the probability of a false positive is so high. A false positive could result in a patient being treated unnecessarily.

IB Mathematics SL

SMARTPREP® | Haese Mathematics

Topic 5 | Probability

Theory card T | **Question cards**

Front side

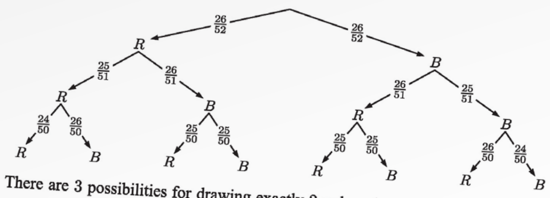
Three cards are drawn from a standard deck of 52 cards without replacement. Use a tree diagram to find the probability that:

a exactly 2 red cards are drawn

b 3 queens are drawn.

Unfold

a The tree diagram uses R for red
B for black.

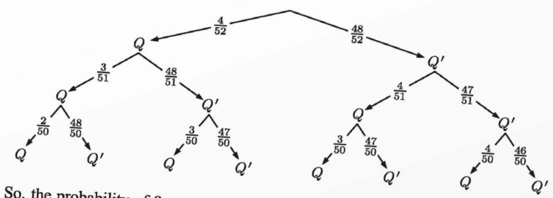


There are 3 possibilities for drawing exactly 2 red cards:

- RRB
- RBR
- BRR

So, the probability of exactly 2 red cards
$= P(RRB \text{ or } RBR \text{ or } BRR)$
$= P(RRB) + P(RBR) + P(BRR)$
$= \left(\frac{26}{52} \times \frac{25}{51} \times \frac{26}{50}\right) + \left(\frac{26}{52} \times \frac{26}{51} \times \frac{25}{50}\right) + \left(\frac{26}{52} \times \frac{26}{51} \times \frac{25}{50}\right)$
$= \frac{13}{34}$
≈ 0.382

b The tree diagram uses Q for queen.



So, the probability of 3 queens
$= P(QQQ)$
$= \frac{4}{52} \times \frac{3}{51} \times \frac{2}{50}$
$\approx 1.81 \times 10^{-4}$