REVISION SHEET – STATISTICS 1 (AQA)

THE BINOMIAL DISTRIBUTION & PROBABILITY

The main ideas are:

- Probabilities based on selecting or arranging objects.
- Probabilities based on the binomial distribution.
- The expected value of a binomial distribution.
- Expected frequencies from a series of trials.

Before the exam you should know:

- $n!$ is the number of ways of ordering a collection of n objects and ${}^nC_r$ is the number of ways of selecting a group of r objects from a total of n objects.
- when a situation can be modelled by the binomial distribution.
- the formula: $P(X = r) = {}^nC_r\, p^r q^{n-r}$ and how to use it.
- how to use the binomial distribution tables (in particular that they give cumulative probabilities).
- the mean or expected value of $X \sim B(n,p)$ is np.

Discrete Random Variables

Notation

- A discrete random variable is usually denoted by a capital letter (X, Y etc).
- Particular values of the variable are denoted by small letters (r, x etc)
- $P(X=r_1)$ means the probability that the discrete random variable X takes the value r_1
- $\Sigma P(X=r_k)$ means the sum of the probabilities for all values of r, in other words

$\Sigma P(X=r_k) = 1$

Example: A child throws two fair dice and adds the numbers on the faces. Find the probability that

(i) $P(X=4)$ (the probability that the total is 4)

(ii) $P(X<7)$ (the probability that the total is less than 7)

Answer:

(i) $P(X=4) = \frac{3}{36} = \frac{1}{12}$ (ii) $P(X<7) = \frac{15}{36} = \frac{5}{12}$

Probabilities based on selecting or arranging

- $n! = n \times (n-1) \times (n-2) \ldots \times 2 \times 1$ is the number of ways of ordering a collection of n objects.
- ${}^nC_r = \dfrac{n!}{(n-r)!r!}$ is the number of ways of selecting r objects from n.

Example

Find the number of different 4-digit numbers that can be made using each of the digits 7, 8, 9, 0 once.

Solution

This is the number of ways of ordering the digits 7, 8, 9, 0. For example 7890 and 7809 are two such orderings. This is given by $4! = 4 \times 3 \times 2 \times 1 = 24$.
This can be thought of as: "there are 4 possibilities for the 1st number, then there are 3 possibilities for the 2nd number, then there are 2 possibilities for the 3rd number, leaving only one possibility for the 4th number.

Example

Eddie is cooking a dish that requires 3 different spices and 2 different herbs, but he doesn't remember which ones. In his cupboard he has 10 different jars of spices and 5 different types of herb and he knows from past experience that the ones he needs are there.

(i) How many ways can he choose the 3 spices?

(ii) How many ways can he choose the 2 herbs?

(iii) If he chooses the herbs and spices at random what is the probability that he makes the correct selection?

Solution

(i) ${}^{10}C_3 = 120$ (ii) ${}^{5}C_2 = 10$ (You can work these out using the ${}^{n}C_r$ function on a calculator.)

(iii) $1 \div (120 \times 10) = 0.000833$

In part (iii) we multiply the results of (i) & (ii) to get 1200 different possible combinations. Only 1 of these is the correct selection so the probability of making the correct selection is $1 \div 1200$.

Probabilities based on the binomial distribution

The binomial distribution may be used to model situations in which:

1. you are conducting n trials where for each trial there are two possible outcomes, often referred to as success and failure.
2. the outcomes, success and failure, have fixed possibilities, p and q, respectively and $p + q = 1$.
3. the probability of success in any trial is independent of the outcomes of previous trials.

The binomial distribution is then written $X \sim B(n, p)$ where X is the number of successes. The probability that X is r, is given by $P(X = r) = {}^{n}C_r p^r (1-p)^{n-r}$

Example

A card is taken at random from a standard pack of 52 (13 of each suit: Spades, Hearts, Clubs, Diamonds). The suit is noted and the card is returned to the pack. This process is repeated 20 times and the number of Hearts obtained is counted.

(i) State the binomial distribution that can be used to model this situation.

(ii) What is the probability of obtaining exactly 6 Hearts?

(iii) What is the probability of obtaining 6 or less Hearts?

(iv) What is the probability of obtaining less than 4 Hearts?

(v) What is the probability of obtaining 6 or more Hearts?

Solution

(i) $X \sim B(20, 0.25)$ (ii) $P(X = 6) = {}^{20}C_6 \times 0.25^6 \times (0.75)^{20-6} = 0.1686$

(iii) $P(X \le 6) = 0.7858$ (This can be read straight from the tables as it is a "$\le$ probability").

(iv) $P(X < 4) = P(X \le 3) = 0.2252$

(v) $P(X \ge 6) = 1 - P(X \le 5) = 1 - 0.6172 = 0.3828$

You need to be very careful with $>$, $<$ or $\ge$. These must all be converted to $\le$ if you are going to use the tables. In (iv) 'less than 4' is the same as '3 or less'. In (v) the complement of '6 or more' is '5 or less'.

The expected value of a binomial distribution

The expected value (mean) of a binomial distribution $X \sim B(n,p)$ is $E[X] = np$.

Example A die is rolled 120 times. How many 3's would you expect to obtain.

Solution Here success would be defined as getting a 3, and failure not getting a 3. Therefore $n = 120$, $p = 1/6$ and $q = 5/6$. X, the number of 3s obtained is modelled by $X \sim B(120, 1/6)$ and so $E[X] = np = 120 \times (1/6) = 20$.

REVISION SHEET – STATISTICS 1 (AQA)

CORRELATION AND REGRESSION

The main ideas are:

- Scatter Diagrams and Lines of Best Fit
- Pearson's Product Moment Correlation
- The Least Squares Regression Line

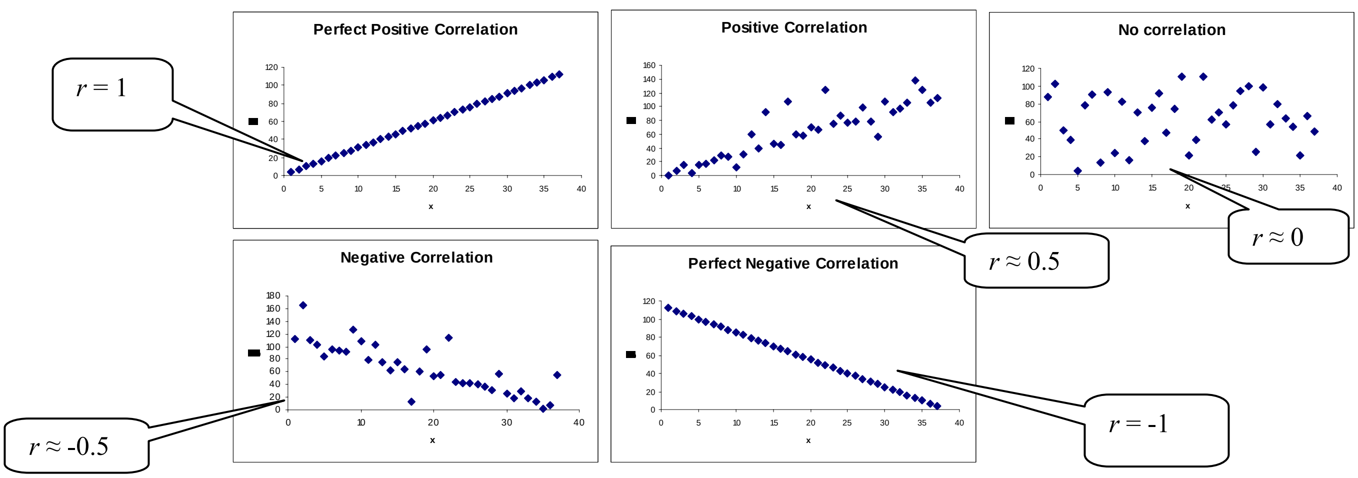
Before the exam you should know:

- Know when to use Pearson's product moment correlation coefficient
- How to use summary statistics such as $\sum x, \sum x^2, \sum y, \sum y^2, \sum xy$ to calculate S_{xx}, S_{yy}, S_{xy}.
- Know how to recognise when a 1 or 2-tail test is required.
- What is meant by a residue and the "least squares" regression line.

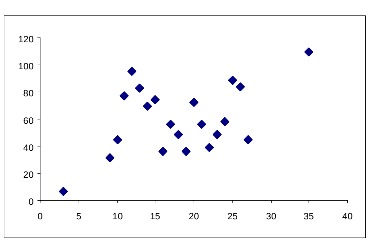
Scatter Diagrams

With Bivariate Data we are usually trying to investigate whether there is a correlation between the two underlying variable, usually called x and y.

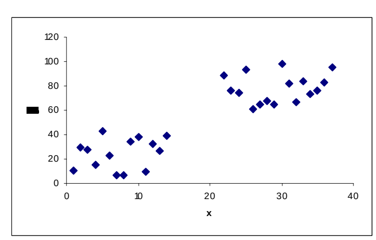
Pearson's product moment correlation coefficient, r, is a number between -1 and +1 which can be calculated as a measure of the correlation in a population of bivariate data.



Beware of diagrams which appear to indicate a linear correlation but in fact to not:



Here two outliers give the impression that there is a linear relationship where in fact there is no correlation.



Here there are 2 distinct groups, neither of which have a correlation.

Product Moment Correlation

Pearson's product Moment Correlation Coefficient:

$$S_{xy} = \sum (x-\bar{x})(y-\bar{y}) = \sum xy - n\bar{x}\bar{y}$$

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\sum (x-\bar{x})^2 \sum (y-\bar{y})^2}} \qquad \text{where : } S_{xx} = \sum (x-\bar{x})^2 = \sum x^2 - n\bar{x}^2$$

$$S_{yy} = \sum (y-\bar{y})^2 = \sum_{i=1}^{n} y^2 - n\bar{y}^2$$

A value of +1 means perfect positive correlation, a value close to 0 means no correlation and a value of -1 means perfect negative correlation. The closer the value of r is to +1 or -1, the stronger the correlation.

Example

A 'games' commentator wants to see if there is any correlation between ability at chess and at bridge.

A random sample of eight people, who play both chess and bridge, were chosen and their grades in chess and bridge were as follows:

Player	A	B	C	D	E	F	G	H
Chess grade x	160	187	129	162	149	151	189	158
Bridge grade y	75	100	75	85	80	70	95	80

Using a calculator:

$n = 8$, $\Sigma x = 1285$, $\Sigma y = 660$, $\Sigma x^2 = 209141$, $\Sigma y^2 = 55200$, $\Sigma xy = 107230$ $\bar{x} = 160.625$, $\bar{y} = 82.5$

$$r = \frac{107230 - 8\times 160.625 \times 82.5}{\sqrt{(209141 - 8\times 160.625^2)(55200 - 8\times 82.5^2)}} = 0.850 \text{ (3 s.f.)}$$

Rank Correlation

The Least Squares Regression Line

This is a line of best fit which produces the least possible value of the sum of the squares of the residuals (the vertical distance between the point and the line of best fit).

It is given by: $y - \bar{y} = \frac{S_{xy}}{S_{xx}}(x-\bar{x})$ Alternatively, $y = a + bx$ where, $b = \frac{S_{xy}}{S_{xx}}, a = \bar{y} - b\bar{x}$

Predicted values

For any pair of values (x, y), the *predicted value* of y is given by $\hat{y}$ **= *a* + *bx*.**

If the regression line is a good fit to the data, the equation may be used to predict y values for x values within the given domain, i.e. *interpolation*.

It is unwise to use the equation for predictions if the regression line is *not* a good fit for any part of the domain (set of x values) or the x value is outside the given domain, i.e. the equation is used for *extrapolation*.

The corresponding residual $= \boldsymbol{\varepsilon} = \boldsymbol{y} - \hat{y} = \boldsymbol{y - (a + bx)}$

The sum of the residuals = $\Sigma\varepsilon = 0$

The least squares regression line minimises the sum of the squares of the residuals, $\Sigma\varepsilon^2$.

Acknowledgement: Some material on these pages was originally created by Bob Francis and we acknowledge his permission to reproduce such material in this revision sheet.

REVISION SHEET – STATISTICS 1 (AQA)

EXPLORING DATA

The main ideas are:

- Types of data
- Measures of central tendency
- Measures of spread
- Confidence Intervals

Before the exam you should know:

- And be able to identify whether the data is categorical, discrete or continuous.
- How to calculate and give comments on the mean, mode, median and mid-range.
- How to calculate the range, variance and standard deviation of the data.
- How to apply the Central Limit Theorem.
- How to construct a Confidence Interval for various confidence levels.

Types of data

Categorical data or qualitative data are data that are listed by their properties e.g. colours of cars.

Numerical or quantitative data

Discrete data are data that can only take particular numerical values. e.g. shoe sizes.

Continuous data are data that can take any value. It is often gathered by measuring e.g. length, temperature.

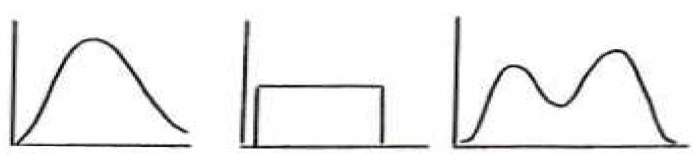
Frequency Distributions

Frequency distributions: data are presented in tables which summarise the data. This allows you to get an idea of the shape of the distribution.

Grouped discrete data can be treated as if it were continuous, e.g. distribution of marks in a test.

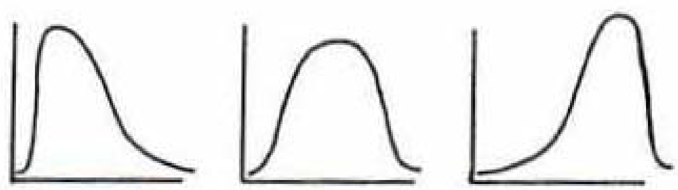
Shapes of distributions *(Not explicitly mentioned on specification)*

Symmetrical (Unimodal) **Uniform** **Bimodal** *bimodal does not mean that the peaks have to be the same height*



Skew *(not explicitly mentioned on the specification)*

Positive Skew **Symmetrical** **Negative Skew**



Central Tendency (averages)

Mean: $\bar{x} = \dfrac{\Sigma x}{n}$ (raw data) $\bar{x} = \dfrac{\Sigma xf}{\Sigma f}$ (grouped data)

Median: mid-value when the data are placed in rank order

Mode: most common item or class with the highest frequency

Mid-range: (minimum + maximum) value ÷ 2

Outliers

These are pieces of data which are at least two standard deviations from the mean

i.e. beyond $\bar{x} \pm 2s$

Dispersion (spread)

(Note. notation may vary slightly between each specification)

Range: maximum value – minimum value

Sum of squares: $S_{xx} = \Sigma(x-\bar{x})^2 \equiv \Sigma x^2 - n\bar{x}^2$ (raw data)

$S_{xx} = \Sigma(x-\bar{x})^2 f \equiv \Sigma x^2 f - n\bar{x}^2$ (frequency dist.)

Mean square deviation: $\text{msd} = \dfrac{S_{xx}}{n}$

Root mean squared deviation: $rmsd = \sqrt{\dfrac{S_{xx}}{n}}$

Variance: $s^2 = \dfrac{S_{xx}}{n-1}$ **Standard deviation**: $s = \sqrt{\dfrac{S_{xx}}{n-1}}$

Example: Heights measured to nearest cm:
159, 160, 161, 166, 166, 166, 169, 173, 173, 174, 177, 177, 177, 178, 180, 181, 182, 182, 185, 196.

Modes = 166 and 177 (i.e. data set is *bimodal*), **Midrange** = (159 +196) ÷ 2 = 177.5 , **Median** = (174 + 177) ÷ 2 = 175.5

Mean: $\bar{x} = \frac{\Sigma x}{n} = \frac{3472}{20}$ = 174.1

Range = 196 – 159 = 37

Sum of squares: $S_{xx} = \Sigma x^2 - n\bar{x}^2$ = 607886 – 20 ×174.1^2 = 1669.8

Root mean square deviation: rmsd = $\sqrt{\frac{S_{xx}}{n}} = \sqrt{\frac{1669.8}{20}}$ = 9.14 (3 s.f.) **Standard deviation:** $s = \sqrt{\frac{S_{xx}}{n-1}} = \sqrt{\frac{1669.8}{19}}$ = 9.37 (3 s.f.)

Outliers (a): 174.1 ± 2 × 9.37 = 155.36 *or* 192.84 - *the value* 196 *lies beyond these limits*, *so one outlier*

Example

A survey was carried out to find how much time it took a group of pupils to complete their homework. The results are shown in the table below. Calculate an estimate for the mean and standard deviation of the data.

Time taken (hours), t	0<t≤1	1<t≤2	2<t≤3	3<t≤4	4<t≤6
Number of pupils, f	14	17	5	1	3

Answer

Time taken (hours), t	0<t≤1	1<t≤2	2<t≤3	3<t≤4	4<t≤6
Mid interval, *x*	0.5	1.5	2.5	3.5	5
Number of pupils, f	14	17	5	1	3
fx	7	25.5	12.5	3.5	15
fx^2	88.2	38.25	31.25	12.25	75

$$\bar{x} = \frac{7+25.5+12.5+3.5+15}{14+17+5+1+3} = \frac{63}{40} = 1.575$$

S_{xx} = (88.2+38.25+31.25+12.25+75) – (40 X 1.575^2) = 2.4686

s = √(2.4686/39) = 0.252 (3dp)

Central limit theorem

The central limit theorem states that for samples of size n, drawn from a distribution with mean μ and variance σ², the distribution of the sample mean is approximately $N\left(\mu, \frac{\sigma^2}{n}\right)$ for sufficiently large n.

Confidence Intervals

A *P*% confidence interval for the mean of a population is an interval constructed from sample data such that *P*% of intervals constructed from samples of the same size will include the true population mean.
The confidence limits are found by:

$$\bar{x} \pm z\frac{\sigma}{\sqrt{n}}$$

where $\bar{x}$ is the sample mean, z is the value of the Normal variable for the confidence interval, σ is the population standard deviation, and *n* is the sample size.

Confidence level

For a confidence interval, a *P*% confidence interval means that *P*% of intervals constructed from samples of the same size will include the true population mean.

You should practice questions using the above techniques and be able to make judgments (or inferences) from their outcomes.

REVISION SHEET – STATISTICS 1 (AQA)

NORMAL DISTRIBUTION

The main ideas are:

- Properties of the Normal Distribution
- Mean, SD and Var

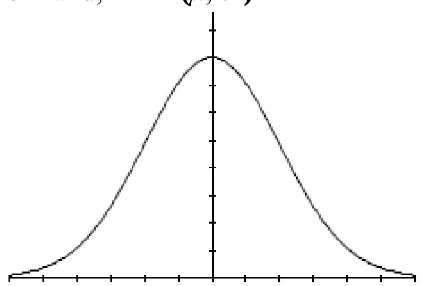
Before the exam you should know:

- All of the properties of the Normal Distribution.
- How to use the relevant tables.
- How to calculate mean, standard deviation and variance.

Definition

A continuous random variable X which is bellshaped and has mean (expectation) μ and standard deviation σ is said to follow a **Normal Distribution** with **parameters** μ and σ.

In shorthand, $\boldsymbol{X \sim N(\mu, \sigma^2)}$



This may be given in *standardised* form by using the transformation

$$z = \frac{x-\mu}{\sigma} \Rightarrow x = \sigma z + \mu, \text{ where } Z \sim \mathbf{N}(0, 1)$$

Calculating Probabilities

The area to the left of the value z, representing $P(Z \le z)$, is denoted by $\Phi(z)$ and is read from tables for $z \ge 0$.

Useful techniques for $z \ge 0$:

- $P(Z > z) = 1 - P(Z \le z)$
- $P(Z > -z) = P(Z \le z)$
- $P(Z < -z) = 1 - P(Z \le z)$

The *inverse normal tables* may be used to find $z = \Phi^{-1}(p)$ for $p \ge 0.5$. For $p < 0.5$, use symmetry properties of the Normal distribution.

99.73% of values lie within 3 s.d. of the mean

Estimating μ and/or σ

Use (simultaneous) equations of the form: $x = \sigma z + \mu$ for matching (x, z) pairs – where z is given or may be deduced from $\Phi^{-1}(p)$ for given value(s) of x.

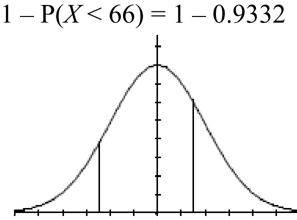
Example 1

$X \sim N(60, 16) \Rightarrow z = \frac{x-60}{4}$;

find (a) $P(X < 66)$, (b) $P(X \ge 66)$, (c) $P(55 \le X \le 63)$, (d) x_0 s.t. $P(X > x_0) = 99\%$

(a) $P(X < 66) = P(Z < 1.5) = \mathbf{0.9332}$

(b) $P(X \ge 66) = 1 - P(X < 66) = 1 - 0.9332 = \mathbf{0.0668}$



(c) $P(55 \le X \le 63) = P(-1.25 \le Z \le 0.75)$
$= P(Z \le 0.75) - P(Z < -1.25)$
$= P(Z \le 0.75) - P(Z > 1.25)$
$= P(Z \le 0.75) - [1 - P(Z \le 1.25)]$
$= 0.7734 - [1 - 0.8944] = \mathbf{0.6678}$

(d) $P(Z > -2.326) = 0.99$ from tables

Since $z = \frac{x-60}{4}$, $x = 4z + 60$

$\Rightarrow x_0 = 60 + 4\times(-2.326) = \mathbf{50.7}$ (to 3 s.f.)

Example 2

For a certain type of apple, 20% have a mass greater than 130g and 30% have a mass less than 110g.

(a) Estimate μ and σ.

(b) When 5 apples are chosen at random, find the probability that all five have a mass exceeding 115g

(a) $P(Z > 0.8416) = 0.2$ $(X = 130)$
$P(Z < -0.5244) = 0.3$ $(X = 110)$

$\Rightarrow 130 = 0.8416\,\sigma + \mu$
$110 = -0.5244\,\sigma + \mu$

Solving equations simultaneously gives: $\boldsymbol{\mu} = \mathbf{117.68}$, $\boldsymbol{\sigma} = \mathbf{14.64}$

(b) $X \sim N(117.68, 14.64^2) \Rightarrow z = \frac{x-117.68}{14.64}$;

$P(X > 115)^5 = P(Z > -0.183)^5 = 0.5726^5 = \mathbf{0.0616}$ (to 3 s.f.)

Acknowledgement: Material on this page was originally created by Bob Francis and we acknowledge his permission to reproduce it here.

REVISION SHEET – STATISTICS 1 (AQA)

PROBABILITY

The main ideas are:

- Measuring probability
- Estimating probability
- Expectation
- Combined probability
- Two trials
- Conditional probability

The experimental probability of an event is = $\dfrac{\text{number of successes}}{\text{number of trials}}$

If the experiment is repeated 100 times, then the *expectation* (expected frequency) is equal to $n \times P(A)$.

Before the exam you should know:

- The theoretical probability of an event A is given by

  $P(A) = \dfrac{n(A)}{n(\xi)}$ where A is the set of favourable outcomes and ξ is the set of all possible outcomes.
- The complement of A is written A' and is the set of possible outcomes not in set A. $P(A') = 1 - P(A)$
- For any two events A and B:
  $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
  [or P(A or B) = P(A) + P(B) – P(A and B)]
- Tree diagrams are a useful way of illustrating probabilities for both independent and dependent events.
- Conditional Probability is the probability that event B occurs if event A has already happened. It is given by

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

The sample space for an experiment illustrates the set of all possible outcomes. Any event is a sub-set of the sample space. Probabilities can be calculated from first principles.

Example: If two fair dice are thrown and their scores added the sample space is

+	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

If event *A* is "the total is 7" then $P(A) = \frac{6}{36} = \frac{1}{6}$

If event *B* is "the total > 8" then $P(B) = \frac{10}{36} = \frac{5}{18}$

If the dice are thrown 100 times, the expectation of event B is

$100 \times P(B) = 100 \times \frac{5}{18} = 27.7778$

or 28 (to nearest whole number)

More than one event

Events are **mutually exclusive** if they cannot happen at the same time so P(A **and** B) = $P(A \cap B) = 0$

Addition rule for mutually exclusive events:
P(A **or** B) = $P(A \cup B) = P(A) + P(B)$

For non-mutually exclusive events
$P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Example: An ordinary pack of cards is shuffled and a card chosen at random.

Event *A* (card chosen is a picture card): $P(A) = \frac{12}{52}$

Event *B* (card chosen is a 'heart'): $P(B) = \frac{13}{52}$

Find the probability that the card is a picture card **and** a heart.

$P(A \cap B) = \frac{12}{52} \times \frac{13}{52} = \frac{3}{52}$:

Find the probability that the card is a picture card **or** a heart.

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

$= \frac{12}{52} + \frac{13}{52} - \frac{3}{52} = \frac{22}{52} = \frac{11}{26}$

Tree Diagrams

Remember to multiply probabilities along the branches (*and*) and add probabilities at the ends of branches (*or*)

Independent events

P(*A* and *B*) = P($A \cap B$) = P(*A*) $\times$ P(*B*)

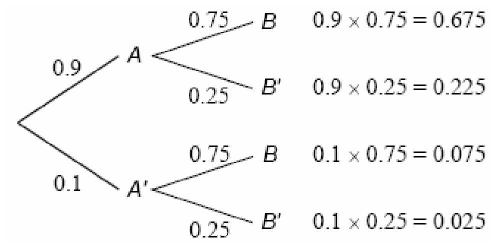
Example 1: A food manufacturer is giving away toy cars and planes in packets of cereals. The ratio of cars to planes is 9:1 and 25% of toys are red. Joe would like a car that is not red.

Construct a tree diagram and use it to calculate the probability that Joe gets what he wants.

Answer:

Event *A* (the toy is a car): P(*A*) = 0.9
Event *B* (the toy is not red): P(*B*) = 0.75

The probability of Joe getting a car that is not red is 0.675



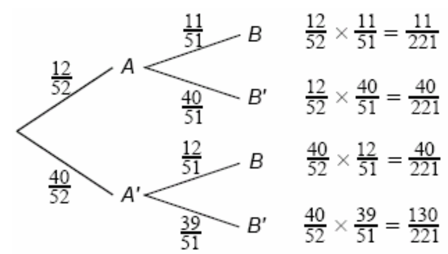
Example 2: dependent events

A pack of cards is shuffled; Liz picks two cards at random without replacement. Find the probability that both of her cards are picture cards

Answer:

Event *A* (1st card is a picture card)
Event *B* (2nd card is a picture card)

The probability of choosing two picture cards is $\frac{11}{221}$



Conditional probability

If A and B are **independent events** then the probability that event B occurs is not affected by whether or not event A has already happened. This can be seen in example 1 above. For independent events P(B/A) = P(B)

If A and B are dependent, as in example 2 above, then P(B/A) = $\dfrac{P(A \cap B)}{P(A)}$

so that probability of Liz picking a picture card on the second draw card given that she has already picked one picture card is given by P(B/A) = $\dfrac{P(A \cap B)}{P(A)} = \dfrac{11/221}{3/13} = \dfrac{11}{51}$

The multiplication law for dependent probabilities may be rearranged to give P(*A* and *B*) = P($A \cap B$) = P(*A*) $\times$ P(*B*|*A*)

Example: A survey in a particular town shows that 35% of the houses are detached, 45% are semi-detached and 20% are terraced. 30% of the detached and semi-detached properties are rented, whilst 45% of the terraced houses are rented. A property is chosen at random.

(i) Find the probability that the property is rented

(ii) Given that the property is rented, calculate the probability that it is a terraced house.

Answer

Let A be the event (the property is rented)
Let B be the event (the property is terraced)

(i) P(rented) = (0.35 X 0.3) + (0.45 X 0.3) + (0.2 X 0.45) = 0.33

The probability that a house is detached and rented	*The probability that a house is semi-detached and rented*	*The probability that a house is terraced and rented*

(ii) P(A) = 0.33 from part (i)

P(B/A) = $\dfrac{P(A \cap B)}{P(A)} = \dfrac{(0.2 \times 0.45)}{(0.33)}$ = 0.27 (2 decimal places)