

REVISION SHEET – STATISTICS 1 (OCR)

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 ${}^n C_r$ 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: $P(X = r) = {}^n C_r p^r q^{n-r}$ 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 .
- how to calculate expected frequencies when a set of trials is repeated.

Probabilities based on selecting or arranging

- $n! = n \times (n-1) \times (n-2) \dots \times 2 \times 1$ is the number of ways of ordering a collection of n objects.
- ${}^n C_r = \frac{n!}{(n-r)!r!}$ 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.

- How many ways can he choose the 3 spices?
- How many ways can he choose the 2 herbs?
- If he chooses the herbs and spices at random what is the probability that he makes the correct selection?

Solution

- ${}^{10}C_3 = 120$
- ${}^5C_2 = 10$ (You can work these out using the ${}^n C_r$ function on a calculator.)
- $1 \div (120 \times 10) = 0.000833$

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

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

$$(iii) \quad P(X \leq 6) = 0.7858 \text{ (This can be read straight from the tables as it is a “} \leq \text{ probability”)}.$$

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

$$(v) \quad P(X \geq 6) = 1 - P(X \leq 5) = 1 - 0.6172 = 0.3828$$

You need to be very careful with $>$, $<$ or \geq . These must all be converted to \leq 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 $E[X] = np = 120 \times (1/6) = 20$.

Expected frequencies from a series of trials

If a situation modelled by a binomial distribution is repeated then the expected frequency of a given number of successes is found by multiplying the probability of that number of successes by the number of times the set of trials is repeated.

Example

The probability of an individual egg being broken during packing is known to be 0.01.

- (i) What is the probability that a box of 6 eggs will have exactly 1 broken egg in it?
- (ii) In a consignment of 100 boxes how many boxes would you expect to contain exactly 1 broken egg?

Solution

$$(i) \quad \text{Using } X \sim B(6, 0.01), P(X = 1) = {}^6C_1 \times 0.01^1 \times (0.99)^{6-1} = 0.057.$$

$$(ii) \quad 0.057 \times 100 = 5.7 \text{ boxes. (This is an expected value and does not have to be an integer).}$$

REVISION SHEET – STATISTICS 1 (OCR)

CORRELATION AND REGRESSION

The main ideas are:

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

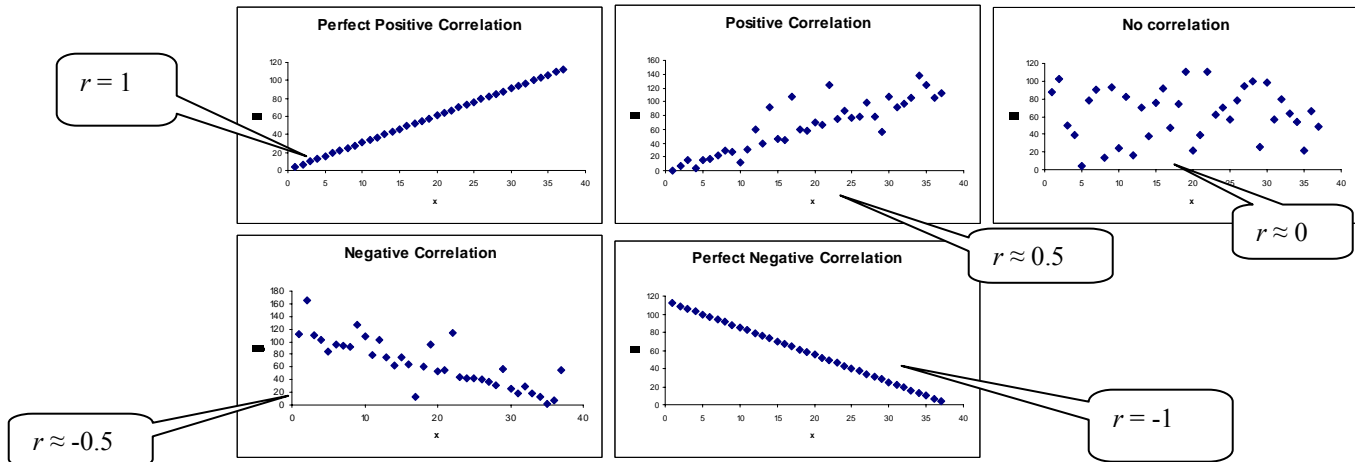
Before the exam you should know:

- 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 and apply it to the PPMCC and Spearman’s Ranking.
- What is meant by a residue and the “least squares” regression line.

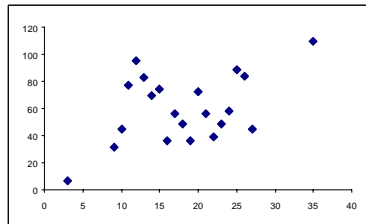
Scatter Diagrams

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

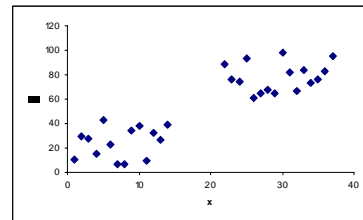
Pearson’s product moment correlation coefficient, r , is a number between -1 and +1 which can be calculated as a measure of the correlation in a population of bivariate data.



Beware of diagrams which appear to indicate a linear correlation but in fact do 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:

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

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

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:

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

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

Rank Correlation

Spearman's Rank Correlation

This is given by: $r_s = 1 - \frac{6 \sum d^2}{n(n^2 - 1)}$, where d represents the difference in ranks for each of the n pairs of

rankings. Spearman's coefficient of rank correlation can be used to investigate whether there is a general increase or decrease (i.e. non-linear correlation), which is not possible with Pearson's product moment correlation coefficient.

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 = $\epsilon = y - \hat{y} = y - (a + bx)$. The sum of the residuals = $\Sigma \epsilon = 0$

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

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 (OCR)

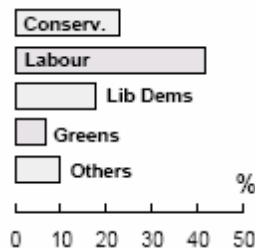
DATA PRESENTATION

The main ideas are:

- Bar charts
- Pie charts
- Histograms
- Cumulative frequency

Discrete data

Horizontal bar chart



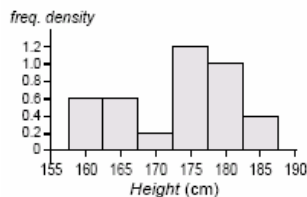
Example

A random sample of cyclists were asked how many days they had used their bicycles in the last week. The results are given in the following table.

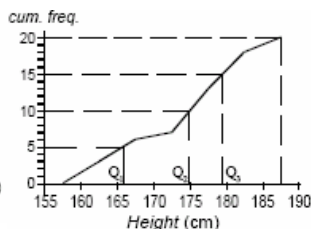
Number of days (x)	1	2	3	4	5	6	7
Frequency (f)	20	12	26	15	12	9	6

Illustrate the distribution using a suitable diagram and describe its shape.

Histogram



Cumulative frequency curve



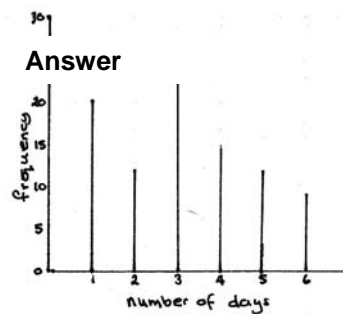
Quartiles and Percentiles

Lower (Q_1) and upper (Q_3) quartiles: values $\frac{1}{4}$ way and $\frac{3}{4}$ way through the distribution.

Percentile: The n^{th} percentile is the value $n / 100$ way through the distribution.

Before the exam you should know:

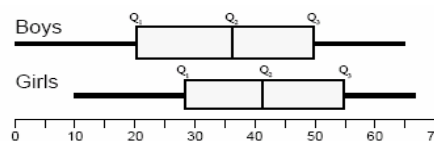
- And be able to draw, interpret and comment on:
 - Bar charts and pie charts for categorical and discrete data
- Histograms are used to illustrate grouped, continuous data. The groups can have different width and the area of each column is proportional to the frequency. The vertical axis is frequency density, which is calculated by dividing frequency by class width. There are no gaps between the columns.
- About cumulative frequency.
- Points are plotted at the upper class boundary. The curve is used to find estimates for the median, upper and lower quartiles and the inter-quartile range.
- The upper and lower quartiles can also be calculated from the data.
- A box and whisker plot is a useful way of showing the median, inter quartile range, range and any outliers



This is a good example of where to use a vertical line graph

The distribution is bimodal with a slight positive skew

Box and whisker plot



The box and whisker plot shows the range, median and quartiles. It is a good way of comparing two distributions.

Inter quartile range (IQR)

A measure of spread calculated by subtracting the lower quartile from the upper quartile: $IQR = Q_3 - Q_1$

An outlier can also be defined as a piece of data at least $1.5 \times IQR$ beyond the nearer quartile (below $Q_1 - 1.5 \times IQR$ or above $Q_3 + 1.5 \times IQR$)

184	192	175	171	195	186	178	177	182	165
183	180	186	170	196	171	187	151	186	199

- (i) Calculate the median and the inter-quartile range.
 (ii) Using the inter-quartile range, show that there is just one outlier. Find the effect of its removal on the median and the inter-quartile range.

Answer

Arrange into numerical order: 151, 165, 170, 171, 171, 175, 177, 178, 180, 182, 183, 184, 186, 186, 186, 187, 192, 195, 196, 199.

(i) Median (Q_2) = $(182+183)/2 = 182.5$

Lower quartile (Q_1) = $(171+175)/2 = 173$ Upper Quartile (Q_3) = $(186+187)/2 = 186.5$

IQR = $186.5 - 173 = 13.5$

(ii) $1.5 \times \text{IQR} = 20.25$. $Q_1 - 20.25 = 152.75$ $Q_3 + 20.25 = 206.75$

So 151 is the outlier.

Remove 151: $Q_1 = 173$ $Q_2 = 182$ $Q_3 = 189.5$ IQR = 16.5

If 151 is removed, the median drops by 0.5 but the IQR increases by 3.

Example

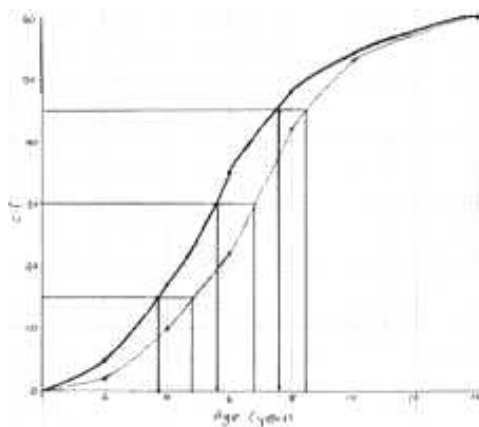
A magazine carried out a survey of the ages of 60 petrol and 60 diesel cars for sale. The results of the survey are summarised in the table.

- (i) Display the data for petrol cars on a histogram
 (ii) Draw a cumulative frequency table for each set of results. On the same axes draw the corresponding cumulative frequency graphs
 (iii) Use your curves to estimate the median and inter-quartile range for each type of car
 (iv) Comment on the differences in the two distributions

Age in years (x)	petrol cars	diesel cars
$0 \leq x < 2$	5	2
$2 \leq x < 4$	12	8
$4 \leq x < 6$	18	12
$6 \leq x < 8$	13	20
$8 \leq x < 10$	6	11
$10 \leq x < 14$	6	7

Answer

Age in years (x)	petrol	Class width	Frequency density	c.f	diesel	c.f
$0 \leq x < 2$	5	2	2.5	5	2	2
$2 \leq x < 4$	12	2	6	17	8	10
$4 \leq x < 6$	18	2	9	35	12	22
$6 \leq x < 8$	13	2	6.5	48	20	42
$8 \leq x < 10$	6	2	3	54	11	53
$10 \leq x < 14$	6	4	1.5	60	7	60



	Petrol	Diesel
Median	5.6	6.8
Q_1	3.7	4.8
Q_3	7.6	8.3
IQR	3.9	3.5

The median age of diesel cars is higher, suggesting that diesel cars are generally older than petrol cars.

The inter-quartile range for petrol cars is greater than for diesel cars.

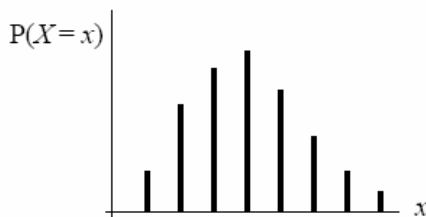
REVISION SHEET – STATISTICS 1 (OCR)

DISCRETE RANDOM VARIABLES

The main ideas are:

- Discrete random variables
- Expectation (mean) of a discrete random variable
- Variance of a discrete random variable

Discrete random variables with probabilities $p_1, p_2, p_3, p_4, \dots, p_n$ can be illustrated using a vertical line chart:



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
- $\sum P(X=r_k)$ means the sum of the probabilities for all values of r , in other words $\sum P(X=r_k) = 1$

Before the exam you should know:

- Discrete random variables are used to create mathematical models to describe and explain data you might find in the real world.
- You must understand the notation that is used.
- You must know that a discrete random variable X takes values $r_1, r_2, r_3, r_4, \dots, r_n$ with corresponding probabilities: $p_1, p_2, p_3, p_4, \dots, p_n$.
- Remember that the sum of these probabilities will be 1 so $p_1 + p_2 + p_3 + p_4, \dots + p_n = \sum P(X=r_k) = 1$.
- You should understand that the expectation (mean) of a discrete random variable is defined by

$$E(X) = \mu = \sum rP(X=r_k)$$

- You should understand that the variance of a discrete random variable is defined by:

$$\text{Var}(X) = \sigma^2 = E(X - \mu)^2 = \sum (r - \mu)^2 P(X=r)$$

$$\text{Var}(X) = \sigma^2 = E(X^2) - [E(X)]^2$$

- Geometric distribution.

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:

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

Example: X is a discrete random variable given by $P(X = r) = \frac{k}{r}$ for $r = 1, 2, 3, 4$ Find the value of k and illustrate the distribution.

Answer:

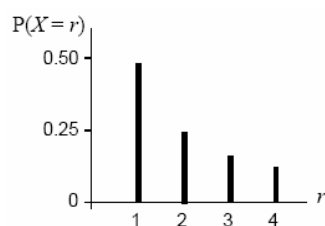
To find the value of k , use $\sum P(X = x_i) = 1$

$$\sum P(X = x_i) = \frac{k}{1} + \frac{k}{2} + \frac{k}{3} + \frac{k}{4} = 1$$

$$\Rightarrow \frac{25}{12}k = 1$$

$$\Rightarrow k = \frac{12}{25} = 0.48$$

Illustrate with a vertical line chart:



Example

Calculate the expectation and variance of the distribution

Answer:

Expectation is

$$E(X) = \mu = \sum rP(X = r) = 1 \times 0.48 + 2 \times 0.24 + 3 \times 0.16 + 4 \times 0.12 = 1.92$$

$$E(X^2) = \sum r^2 P(X = r) = 1^2 \times 0.48 + 2^2 \times 0.24 + 3^2 \times 0.16 + 4^2 \times 0.12 = 4.5$$

$$\text{Variance is } \text{Var}(X) = E(X^2) - [E(X)]^2 = 4.5 - 1.92^2 = 0.8136$$

Using tables:

For a small set of values it is often convenient to list the probabilities for each value in a table

r_i	r_1	r_2	r_3	...	r_{n-1}	r_n
$P(X = r_i)$	p_1	p_2	p_3	...	p_{n-1}	p_n

Using formulae:

Sometimes it is possible to define the probability function as a formula, as a function of r , $P(X = r) = f(r)$

Calculating probabilities:

Sometimes you need to be able to calculate the probability of some compound event, given the values from the table or function.

Explanation of probabilities:

Often you need to explain how the probability $P(X = r_k)$, for some value of k , is derived from first principles.

Example:

The discrete random variable X has the distribution shown in the table

r	0	1	2	3
$P(X = r)$	0.15	0.2	0.35	0.3

- (i) Find $E(X)$.
- (ii) Find $E(X^2)$.
- (iii) Find $\text{Var}(X)$ using (a) $E(X^2) - \mu^2$ and (b) $E(X - \mu)^2$.
- (iv) Hence calculate the standard deviation.

r	0	1	2	3	totals
$P(X = r)$	0.15	0.2	0.35	0.3	1
$rP(X = r)$	0	0.2	0.7	0.9	1.8
$r^2P(X = r)$	0	0.2	1.4	2.7	4.3
$(r - \mu)^2$	3.24	0.64	0.04	1.44	5.36
$(r - \mu)^2P(X = r)$	0.486	0.128	0.014	0.432	1.06

This is the expectation (μ)

This is $E(X^2)$

This is $\text{Var}(X) = \sum(r - \mu)^2P(X = r)$

(i) $E(X) = \mu = \sum r P(X = r) = 0 \times 0.15 + 1 \times 0.2 + 2 \times 0.35 + 3 \times 0.3$
 $= 0 + 0.2 + 0.7 + 0.9$
 $= \mathbf{1.8}$

(ii) $E(X^2) = \sum r^2 P(X = r) = 0^2 \times 0.15 + 1^2 \times 0.2 + 2^2 \times 0.35 + 3^2 \times 0.3$
 $= 0 + 0.2 + 1.4 + 2.7$
 $= \mathbf{4.3}$

(iii) (a) $\text{Var}(X) = E(X^2) - \mu^2 = 4.3 - 1.8^2 = \mathbf{1.06}$

(b) $\text{Var}(X) = E(X - \mu)^2 = 0.15(0 - 1.8)^2 + 0.2(1 - 1.8)^2 + 0.35(2 - 1.8)^2 + 0.3(3 - 1.8)^2$
 $= 0.486 + 0.128 + 0.014 + 0.432$
 $= \mathbf{1.06}$

(iv) $s = \sqrt{1.06} = 1.02956 = \mathbf{1.030}$ (3d.p.)

Notice that the two methods give the same result since the formulae are just rearrangements of each other.

standard deviation (s) is the square root of the variance

Geometric Distribution:

This describes the probability distribution of the number $X = k - 1$ of failures before the first success.

If the probability of success on each trial is p , then the probability that k trials are needed to get one success is:

$P(X = k) = (1 - p)^{k-1} p$

It is a necessary condition that each of the events is independent.

$E(X) = \frac{1}{p}$, $\text{var}(X) = \frac{q}{p^2}$

REVISION SHEET – STATISTICS 1 (OCR)

EXPLORING DATA

The main ideas are:

- Types of data
- Stem and leaf
- Measures of central tendency
- Measures of spread

Before the exam you should know:

- And be able to identify whether the data is categorical, discrete or continuous.
- How to describe the shape of a distribution, say whether it is skewed positively or negatively and be able to identify any outliers.
- And be able to draw an ordered stem and leaf and a back to back stem and leaf diagram.
- And be able to calculate and comment on the mean, mode, median and mid-range.
- And be able to calculate the range, variance and standard deviation of the data.

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.

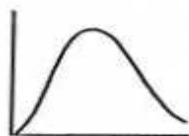
Frequency Distributions

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

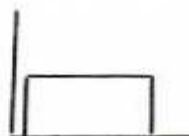
Grouped discrete data can be treated as if it were continuous, e.g. distribution of marks in a test.

Shapes of distributions

Symmetrical (Unimodal)

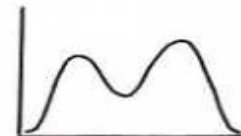


Uniform



Bimodal

bimodal does not mean that the peaks have to be the same height

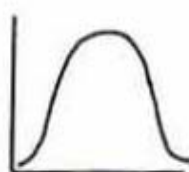


Skew (Not explicitly mentioned in Specification)

Positive Skew



Symmetrical



Negative Skew



Stem and leaf diagrams

A concise way of displaying discrete or continuous data (measured to a given accuracy) whilst retaining the original information. Data usually sorted in ascending order and can be used to find the mode, median and quartiles. You are likely to be asked to comment on the shape of the distribution.

Example

Average daily temperatures in 16 cities are recorded in January and July. The results are

January: 2, 18, 3, 6, -3, 23, -5, 17, 14, 29, 28, -1, 2, -9, 28, 19

July: 21, 2, 16, 25, 5, 25, 19, 24, 28, -1, 8, -4, 18, 13, 14, 21

Draw a back to back stem and leaf diagram and comment on the shape of the distributions.

	Jan	July	
Answer	9 5 3 1	-0 1 4	
	6 3 2 2	0 2 5 8	
	9 8 7 4	10 3 4 6 8 9	
	9 8 8 3	20 1 1 4 5 5 8	

The January data is uniform but the July data has a negative skew

Central Tendency (averages)

Mean: $\bar{x} = \frac{\sum x}{n}$ (raw data) $\bar{x} = \frac{\sum xf}{\sum f}$ (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 $\div 2$

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)

Range: maximum value – minimum value

Sum of squares:

$$S_{xx} = \sum (x - \bar{x})^2 \equiv \sum x^2 - n\bar{x}^2 \text{ (raw data)}$$

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

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

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

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

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** = $(159 + 196) \div 2 = 177.5$, **Median** = $(174 + 177) \div 2 = 175.5$

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

Range = $196 - 159 = 37$

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

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): $174.1 \pm 2 \times 9.37 = 155.36$ 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

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

$$S_{xx} = (88.2+38.25+31.25+12.25+75) - (40 \times 1.575^2) = 2.4686$$

$$s = \sqrt{(2.4686/39)} = 0.252 \text{ (3dp)}$$

REVISION SHEET – STATISTICS 1 (OCR)

PROBABILITY

The main ideas are:

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

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

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

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)

Before the exam you should know:

- The theoretical probability of an event A is given by

$$P(A) = \frac{n(A)}{n(\xi)} \text{ where } A \text{ is the set of favourable outcomes}$$

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:

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

$$[\text{or } P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)]$$

- 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

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

More than one event

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

Addition rule for mutually exclusive events:

$$P(A \text{ or } B) = P(A \cup B) = P(A) + P(B)$$

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

$$P(A \text{ and } B) = P(A \cap B) = P(A) \times P(B)$$

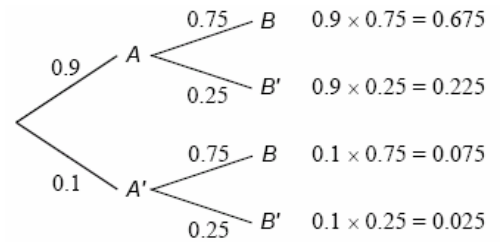
Example 1: A food manufacturer is giving away toy cars and planes in packets of cereals. The ratio of cars to planes is 9:1 and 25% of toys are red. Joe would like a car that is not red.

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



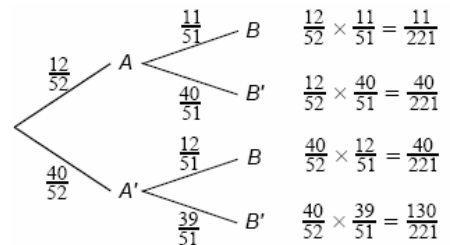
Example 2: dependent events

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

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 $P(B/A) = \frac{P(A \cap B)}{P(A)} = \frac{11/221}{3/13} = \frac{11}{51}$

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

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) $P(\text{rented}) = (0.35 \times 0.3) + (0.45 \times 0.3) + (0.2 \times 0.45) = 0.33$

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)

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