

S1 EXAM TIPS

EXAM TIPS

READ THE QUESTION CAREFULLY

ALWAYS DRAW A DIAGRAM

CHECK THAT YOUR ANSWERS MAKE
SENSE

PROBABILITIES ARE ALWAYS
BETWEEN 0 AND 1

CHECK THE VALUES THAT YOU INPUT
INTO YOUR CALCULATOR

CHECK SCATTERGRAPHS ON YOUR
CALCULATOR FOR ANOMALIES

STATISTICS

STANDARD DEVIATION

The formula that you need depends on the information you are given:

- 1) If you are given raw data
- 2)

Check if the question states that it is a sample.

Type it into your graphical calculator, click *CALC* and *1-var*, look down the list.

If it is a sample you need σ_{n-1} .

If it doesn't, you need σ_n .

2) If you are given the sums, you need to use the formula:

$$\sigma = \sqrt{\frac{\sum (x - \mu)^2}{n}} = \sqrt{\frac{1}{n} \left(\sum x^2 - \frac{(\sum x)^2}{n} \right)}$$

If you are given a sample, the n in the first formula needs replacing with $n-1$ and in the second formula the $\frac{1}{n}$ needs

replacing with $\frac{1}{n-1}$, the second n stays the same.

If you can't remember this formula, look at the formula for S_{xx} in the formula book, it is very similar.

Just multiply by $\frac{1}{n}$ or $\frac{1}{n-1}$ before square rooting.

GROUPED CONTINUOUS DATA

To find the mean:

- 1) Find the mid-point of each class interval.
- 2) Multiply each mid-point by the frequency for that class.
- 3) Find the sum of all the mid-points X frequencies
- 4) Divide by the total frequency

To find the median:

- 1) Find the overall position of the median $(n+1)/2$
- 2) Work out which class interval this falls into and the position of the median in that class interval.
- 3) Divide the class interval by the frequency and multiply by the position of the median in that class interval.
- 4) Add on the lower class bound. If the data is rounded, then remember that the lower bound is less.

To find the quartiles:

- 1) The position of the quartiles is at $(n+1)/4$ and $3(n+1)/4$
- 2) Use the same method as the median.

The interquartile range is the upper - lower quartile.

BI-VARIATE DATA

This just means comparing two sets of data. You will get a table with two rows of data and be asked to draw a scattergraph and/ or find the correlation co-efficient or regression line.

To find the correlation coefficient

1) If you are given raw data:

Type it into your graphical calculator, click GRPH, GPH1, X,
The correlation coefficient is $r=...$

(If you were too tight to buy a graphical calculator, you are going to regret it now!)

NB. Yes you get that many marks just for writing down a number from your calculator display, but double check that you typed in the data correctly!

2) If you are given the data as sums, use this formula from the formula book:

$$r = \frac{S_{xy}}{\sqrt{S_{xx} \times S_{yy}}} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\{\sum (x_i - \bar{x})^2\} \{\sum (y_i - \bar{y})^2\}}} = \frac{\sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}}{\sqrt{\left(\sum x_i^2 - \frac{(\sum x_i)^2}{n}\right) \left(\sum y_i^2 - \frac{(\sum y_i)^2}{n}\right)}}$$

Before using the Product moment correlation coefficient:

1) Look at the scattergraph

a) Is it a linear correlation?

Yes - then use PMCC,

No - PMCC is useless!

b) Are there any anomalous points?

Yes - Read the question carefully, these should be removed to give a meaningful answer or you should comment on it accordingly.

No - The PMCC is a good measure of correlation.

PMCC IS UNCHANGED IF YOU CHANGE THE UNITS OF THE VARIABLES.

INTERPRETATION IN CONTEXT

Comment on the variables as well as the correlation.

A negative PMCC means a negative correlation, as one variable increases the other decreases.

A positive PMCC means a positive correlation, as one variable increases, so does the other.

REGRESSION ANALYSIS

To find the formula for the regression line:

1) If you are given raw data:

Input it into the tables in your calculator, press GRPH, GPH1 and then X.

It gives the formula as $y=ax+b$ with the values for a and b given above.

NB. The a and b are the opposite way round from the formula book. Just make sure that you follow the way that the calculator sets it out. So long as you get the right value as the gradient (in front of the x) it will give an acceptable answer.

To draw the line on your graph, you need 2 points from the regression line and the mean point to check that it is correct. (They should all be in a straight line!)

Point 1: Use the y -intercept - b

Point 2: Use the mean point - you can get \bar{x} and \bar{y} directly from the calculator, go to your input table, press calc and then scroll through the results.

Point 3: Pick a convenient x value to the right of the mean point and put it into the regression line formula to get the y -coordinate.

2) If you are given sums:

The formula is given in the formula book.

$$\text{The regression coefficient of } y \text{ on } x \text{ is } b = \frac{S_{xy}}{S_{xx}} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

Least squares regression line of y on x is $y = a + bx$, where $a = \bar{y} - b\bar{x}$

Use the formulae for S_{xy} and S_{xx} from above the regression line formula in the formula book:

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

$$S_{xx} = \sum (x_i - \bar{x})^2 = \sum x_i^2 - \frac{(\sum x_i)^2}{n}$$

Interpreting the regression line:

- 1) The y -intercept gives the minimum possible value.
- 2) The gradient gives the rate of change, for every 1 increase in the independent variable on the x -axis, how much the dependent variable on the y -axis changes.
- 3) Remember to comment on any anomalies that would affect the accuracy of the regression line.

NB. When commenting on the values for a and b , remember to give your answer in terms of what the x and y variables represent.

Remember to give your interpretation in a way that makes sense, if the gradient is very small, it may be better to give an answer for a change in 10 or 100 on the x axis.

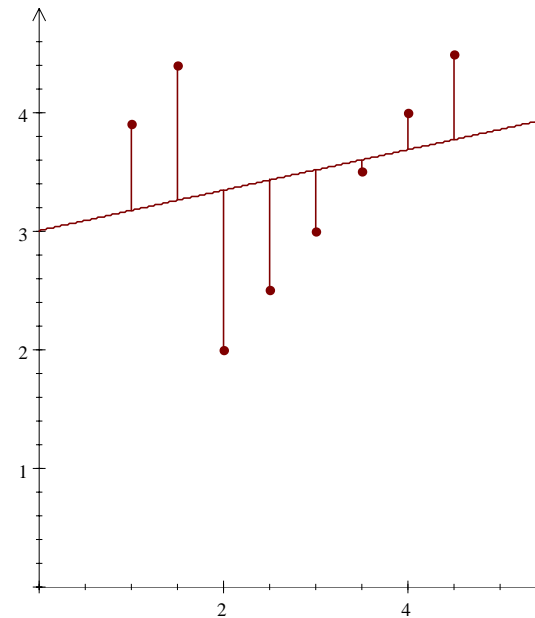
When you give readings from the regression line, use the formula, not a reading from the graph!

When asked to comment on your readings remember that:

INTERPOLATIONS (within the range of the given data) are usually reliable (but see information on residuals below)

EXTRAPOLATION (outside the range of the given data) is **NEVER** seen to be reliable.

RESIDUALS



The residuals are the vertical distance from the regression line of the points. These are shown by dotted lines on the graph.

You may be asked to find residuals or comment on an interpolation value taking the residuals into account.

1) To find residuals:

Take the x values of each of the data points, put into the regression formula and calculate the y-coordinate of the regression line.

Then find the difference between that and the y-coordinate of the actual data point.

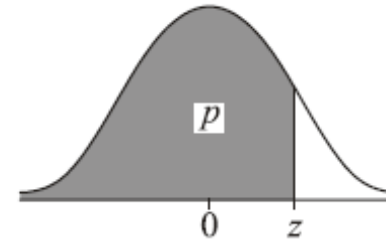
A point above the regression line has a positive residual and a point below the regression line has a negative residual.

2) To use residuals to comment on the reliability of an interpolation: You need to find the residual as a percentage of the interpolation value:

$$\frac{\text{Residual}}{\text{Interpolation}} \times 100$$

A percentage of around 10 or more is considered to be large and makes the interpolation result unreliable.

NORMAL DISTRIBUTION



The area under the curve is the probability of getting less than the z score. The total area is 1.

The tables give the probability for z-scores in the distribution $X \sim N(0,1)$, that is mean = 0, s.d. = 1.

NB. The distribution gives the mean and VARIANCE.

ALWAYS SKETCH A DIAGRAM

Read the question carefully and shade the area you want to find. If the shaded area is more than half then you can read the probability directly from the table, if it is less than half, then you need to subtract it from 1.

NB If your z-score is negative then you would look up the positive from the table. The rule for the shaded area is the same as above: more than half - read from the table, less than half subtract the reading from 1.

If you need to find the probability in between two z-scores you need to:

- 1) Draw a diagram.
- 2) Shade the required area.
- 3) Find the probability for the larger value from your table.
- 4) Find the probability for the smaller value from your table.
- 5) Subtract the two values, remembering that probabilities must be between 0 and 1.

NB. You cannot find the probability for an individual value on the normal distribution.

STANDARDISING:

If your normal distribution has a mean $\neq 0$ and s.d. $\neq 1$, then you will have to standardise to use the table.

$$z = \frac{x - \mu}{\sigma}$$

REMEMBER TO SKETCH A DIAGRAM FIRST

So, take the x value, subtract the mean and then divide by the s.d.

Look up the value from the table. If your shaded area on the diagram is more than half, then the probability is the value on the table. If it is less than half, then subtract the probability from 1.

PERCENTAGE POINTS TABLE

This is the inverse of the process above. You will be given a percentage and have to find the z-score from it. Use the small normal distribution table.

To find the central percentage of probability:

Take the probability, half what is left from 100% (or 1) and add it on before you read the value from the table, the interval is then $(-z, z)$.

Non-standard normal distributions:

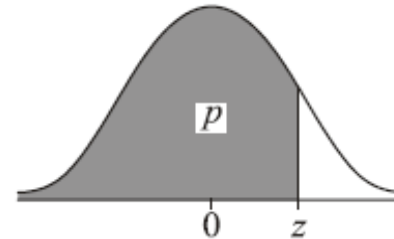
If your normal distribution has a mean $\neq 0$ and s.d. $\neq 1$, then you will have to UNstandardise after using the table.

Read the value(s) from the table and then put into the formula:

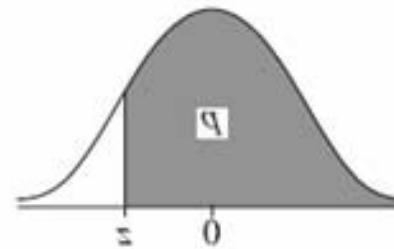
$$x = \pm z\sigma + \mu$$

NB. You will not get \pm the same number for a central interval.

INTERPRETING THE QUESTION



Shading to the left represents phrases like:
Find the probability that ... is less than...
Find the probability that ... is exceeded by z .
Find the score that exceeds...% of the population.



Shading to the right represents phrases like:
Find the probability that... is greater than...
Find the probability that ... exceeds z .
Find the score that is exceeded by...% of the population.

DISTRIBUTION OF THE SAMPLE MEAN

If you are given a sample of values from a normal distribution and asked to calculate a confidence interval, then the distribution of the mean will be normal with mean μ and s.d. $\frac{\sigma}{\sqrt{n}}$ where n is the sample size.

- 1) Calculate the mean of the sample \bar{x}
- 2) Find the z-scores for the central interval for the given level of confidence (from the % point table).
- 3) UNstandardise the z-scores, using the formula:

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

Interpreting the confidence interval:

You may be asked to comment on a claim based on the confidence interval - If the claim falls outside the confidence interval then the claim is always rejected, no matter how close it is.

If the claim is inside the confidence interval, then the claim is accepted with the observation that some values will fall outside.

THE CENTRAL LIMIT THEOREM

The central limit theorem is used when the shape of the parent population is unknown.

The sample size n , has to be over 30 for it to be valid.

The sample mean distribution then approximates a normal distribution with mean μ and s.d. $\frac{\sigma}{\sqrt{n}}$

All calculations are exactly the same as in the previous section.

NB. THE CENTRAL LIMIT THEOREM IS NOT NEEDED IF THE PARENT POPULATION IS NORMALLY DISTRIBUTED.

PROBABILITY

Non-conditional probability

Non-conditional probability means that events are **INDEPENDENT**, that is, that the probability of one event happening is not changed by the outcome of another event.

MUTUALLY EXCLUSIVE events cannot happen at the same time.

AND and OR rules for independent events:

AND (\cap) means multiply,
OR (\cup) means add.

FOR SUCCESSIVE EVENTS (ONE EVENT FOLLOWING ANOTHER), DRAW A TREE DIAGRAM.

Conditional Probability

This is where the probability of one event is changed by the outcome of the previous event.

The formulae for conditional probabilities are given in the formula book as:

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

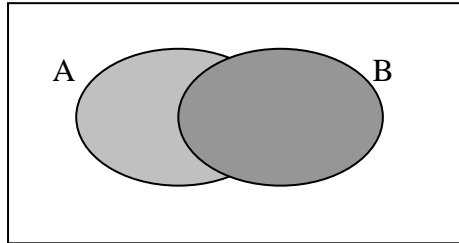
$$P(A \cap B) = P(A) \times P(B | A)$$

You may need to rearrange these and substitute different letters in depending on the question.

$P(B|A)$ means the probability that B happens given that A has already happened.

You may find a Venn diagram helpful:

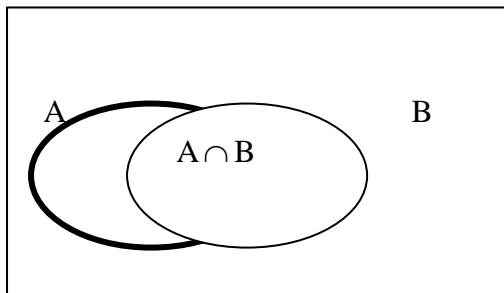
$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$



The formula means that the probability of event A OR B occurring is the probability of A plus the probability of B subtract the overlap (Probability of A AND B).

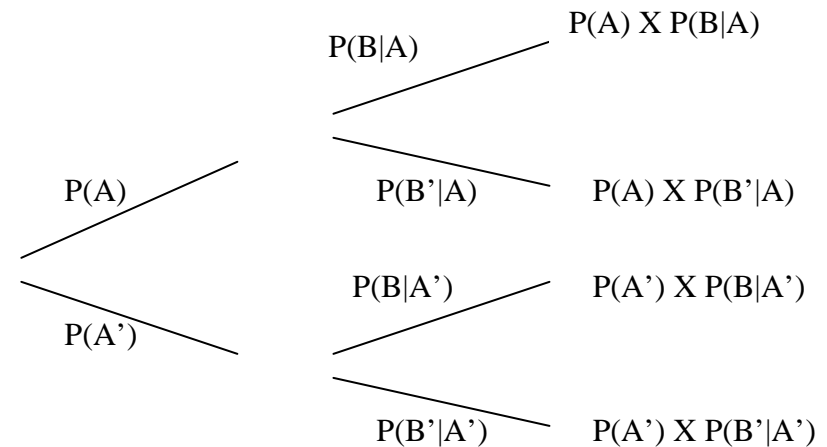
NB. For conditional probability, $P(A \cap B)$ is NOT $P(A) \times P(B)$.

$$P(A \cap B) = P(A) \times P(B | A)$$



This means that the probability of A AND B happening, is the probability of A happening AND the probability that B happens if A has already happened.

You can also use a tree diagram to find the probabilities of successive events by using a tree diagram.



Remember to ALWAYS draw a diagram to help you.

THE BINOMIAL DISTRIBUTION

The binomial distribution is a discrete distribution.

To use it you need independent probabilities (the probability doesn't change) that are mutually exclusive (cannot happen at the same time).

You should use the tables where possible.

The distribution is of the form $X \sim B(n,p)$, where n is the number of trials and p is the probability.

NB, the smaller probability is the one given for success and should be read from the table. If a probability greater than 0.5 is given in the question, subtract it from 1 and read the resulting probability from the table. As the binomial distribution is symmetrical it doesn't matter.

You need to find the correct table for n and look up the given probability at the top and the required number of outcomes x down the side.

The table gives a cumulative probability, that is that the probability is less than or equal to a given value.

If you need:

$P(X \leq x)$ Then use the value directly from the table.

$P(X < x)$ Then use the value $x-1$ from the table, as the probability doesn't include the x value.

$P(X = x)$ Then find the value for x and the value for $x-1$ and subtract them, as it doesn't include the values below x .

$P(X > x)$ Then find the value for x and subtract it from 1, as x is included in the unwanted values.

$P(X \geq x)$ Then find the value for $x-1$ and subtract it from 1, as you want the x value but want to get rid of all the values below it.