AQA STATISTICS 1 REVISION NOTES

AVERAGES AND MEASURES OF SPREAD

- **Mode :** the most common or most popular data value
  the only average that can be used for qualitative data
  not suitable if the data values are very varied

- **Mean :** important as it uses all the data values
  Disadvantage – **affected by extreme values**

If the data is grouped – use the mid-point of each group as your x

- **Median :** the middle value when the data are in order
  For n data values the median is the $\frac{n+1}{2}$th value

  Not affected by extreme values

For 10 values the median will be the 5½ th value – halfway between the 5th and the 6th values

- **Range** – biggest value – smallest value
  - greatly affected by extreme values

- **Interquartile Range** – Upper quartile – Lower quartile
  - measures the spread of the middle 50% of the data and is not affected by extreme values

3	4	5	7	8	9	9
	LQ		M		UQ	

= 5

2	3	4	4	5	7	8	9
	LQ		M		UQ		

IQR = 7.5 – 3.5 = 4

- **Standard Deviation**
  Deviation from the mean is the difference from a value from the mean value
  The standard deviation is the **average of all of these deviations**

  Formulas to work out standard deviation are given in the

- **SCALING DATA**

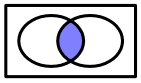
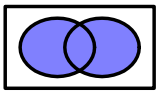
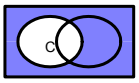
  Addition if you add **a** to each number in the list of data :
  New mean = old mean **+ a**
  New median = old median **+ a**
  New mode = old mode **+ a**
  Standard Deviation is UNCHANGED

  Multiplication - If you multiply each number in the list of data by **b** :
  New mean = old mean × **b**
  New median = old median × **b**
  New mode = old mode × **b**
  New Standard Deviation = old standard deviation × b

PROBABILITY

Outcome : each thing that can happen in an experiment
Sample Space : list of all the possible values

- NOTATION

$A \cap B$	A and B both happen	
$A \cup B$	either A or B or both happen	
C'	C doesn't happen	

  - P(C') = 1 – P(C)
  - $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Mutually Exclusive Events two or more events that cannot happen at the same time
$P(A \cup B) = P(A) + P(B)$

Independent Events the outcome of one event does not affect the outcome of another
P($A \cap B$) = P(A) P(B)

Conditional Probability : when the outcome of the first event affects the outcome of a second event, the probability of the second event depends on what has happened

  - P(B/A) means the **conditional** probability of B given A

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

- If the question states that the events are **independent** then a tree diagram might be a good idea – multiply along the branches then add the appropriate combinations together
- Probability of '**at least 1' = 1 – Probability of 'none'**
- If you are asked to find probabilities using data in a table – work out the row/column totals before you start

BINOMIAL DISTRIBUTION

- A question is binomial if:
  - Probability of an event happening is given (p)
  - Number of people/trials/objects chosen given (n)

- EQUALS or EXACTLY use the formula P(x=r)
  Make sure you write it out with the values substituted in

$$ {}_{n}C_{r}\, p^{r} (1-p)^{n-r} $$

Check that your 'powers' add to make n (number of trials)

from your calculator make sure you write down this value

- MORE/LESS THAN/AT LEAST use tables
  Remember tables give less than or equal to
  Make sure you list the numbers and identify which ones you need to include

P(X>5) 0 1 2 3 4 5 6 7 8 9 $= 1 - P(X \leq 5)$

P(X<5) 0 1 2 3 4 5 6 7 8 $= P(X \leq 4)$

- MEAN and VARIANCE
  Mean = np Variance = np(1-p) standard deviation = $\sqrt{np(1-p)}$

- ASSUMPTIONS
  Independent events with a fixed probability of success
  Randomly selected

- COMPARISONS
  You may be asked to calculate the mean and standard deviation of a binomial distribution and compare them to the mean and standard deviation of a sample (table of results) - if both the means are approximately the same AND both standard deviations (or variances) are approximately equal then you can say that binomial model appears to fit the data and that it must be independent, random observations.

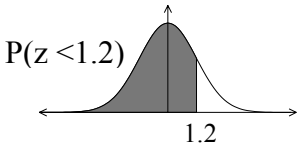
NORMAL DISTRIBUTION

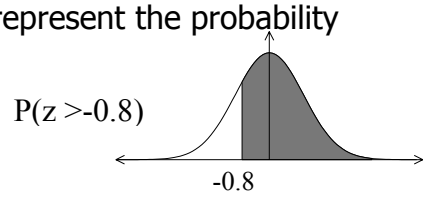
FINDING PROBABILITIES

- State the mean and variance (standard deviation)
- Standardise to find the z value

$$z = \frac{x - \text{mean}}{\text{standard deviation}}$$

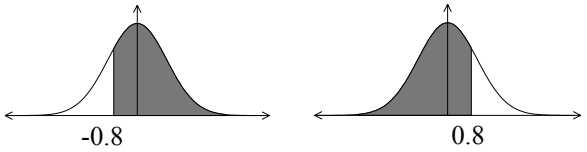
- Sketch a graph and shade in the area to represent the probability



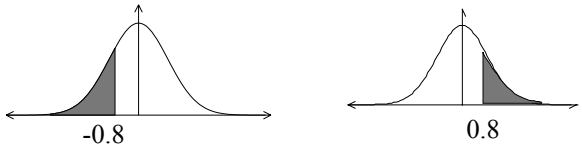


- Use the table to find the probability –
  - take care with negative z-values –

P(z > -0.8) = P(z < 0.8)



P(z < -0.8) = 1 – P(z < 0.8)

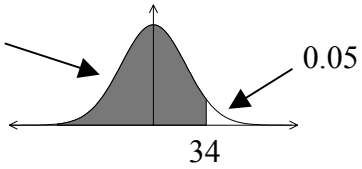


- ALWAYS CHECK YOUR ANSWER WITH YOUR GRAPH – if your shaded area is more that ½ and your answer is 0.4 (for example) – you know you have gone wrong somewhere!!

WORKING BACKWARDS – to find the mean/standard deviation – or both (simultaneous equations)

- State the probability you know and sketch a graph

P(X < 34) = 0.95



Standard deviation = 8

- Use tables to find the appropriate z value
- Write down the equation used to standardise with all of the known values substituted

$$1.6445 = \frac{34 - \text{mean}}{8}$$

- Rearrange to find the mean

SAMPLES and PROBABILITIES

If you are asked to calculate probabilities involving samples remember to divide the (population) standard deviation by the square root of the sample size when you standardise.

$$z = \frac{x - \text{mean}}{\frac{\sigma}{\sqrt{n}}} \quad \text{or} \quad z = \frac{x - \text{mean}}{\sqrt{\frac{\text{variance}}{n}}}$$

ESTIMATION

– estimating the population mean from a sample mean and finding a confidence interval

If a random sample of size n is taken from a normal population and the sample mean $\bar{x}$ is found, then the 95% confidence interval of the population mean is given by

$$\bar{x} - 1.96 \times \sqrt{\frac{\sigma^2}{n}} \quad , \quad \bar{x} + 1.96 \times \sqrt{\frac{\sigma^2}{n}}$$

(Z value)

where σ^2 is either the population variance (if given) or an unbiased estimate of the population variance (found from the sample see below)

- CASE 1 : **Standard deviation or Variance** of the population is **stated** in the question
  - from the data you only need to calculate the mean
  - use your tables to find the appropriate 'z value'
  - **write out** the above expressions for the confidence intervals with all your values substituted in
  - calculate the two values for your confidence intervals and state clearly (3 sf)

  CHECK your answer – add the two answers together and divide by 2
  – this should be the sample mean!!!

- CASE 2 : **Standard deviation or Variance** of the population is **UNKNOWN**
  - you will need to use the data to calculate the variance of the sample and then an **unbiased estimate of the population variance**
  - your calculator will give you value of the standard deviation for the sample you have entered

  σ_{xn} - square this to calculate the **sample variance**

  An **unbiased estimate of the population variance** is $\frac{n}{n-1} \times$ Sample variance

  Use this as the value of σ^2 in your confidence interval when substituting the values in

- **INTERPRETATION – a 95% confidence interval tells us that if we took the same size sample 100 times then 95 of the confidence intervals we would calculate should contain the TRUE population mean.**

CENTRAL LIMIT THEOREM

– you only need to use the central limit theorem if you are **not told** that the sample is selected from a population which is Normally Distributed

The theorem concerns the distribution of the sample means and as long as the sample size is large enough (greater than 30) then the sample means will be normally distributed – and so we can calculate confidence intervals

REGRESSION – finding the equation of the line of best fit – least squares

$y = a + bx$

Gradient – the change in y for each unit change in x

e.g for every 1 degree rise in temperature sales increase by £b

Intercept - the value of 'y' when 'x' is zero.
e.g. When the temperature is 0°C ice cream sales are £10

- If you are asked to interpret the values of a and b, make sure you discuss it in the context of the question NOT in terms of x and y (see examples above)

- If you are give a table of values – use your calculator to find a and b
  In your workings state a = …. b =…..
  and show your values substituted into $y = a + bx$

- If you are calculating a and b using the formulae – make sure you use the formula book – showing how you substituted the values in
  Always work out 'b' first
  Use 'b' and the means of x and y to work out a a = (mean of y) - b(mean of x)

- TO PLOT THE REGRESSION LINE – choose 2 different values of x – use your equation y=a + bx to work out the predicted y-values - plot the two points and join with a straight line

- RESIDUAL

  = OBSERVED*(actual value)* – PREDICTED*(using equation y= a + bx)*

  - the smaller the residuals the greater the accuracy of the line of best fit in predicting values
  - sometimes an 'average' residual can be used to make predictions using the line of best fit –*e.g if an individual has an average residual of 5 – then to predict for this particular person using the line, 5 should be added to the value predicted using the equation.*

- **RELIABILITY OF PREDICTIONS**
  - **Interpolation**
    predicting using an x-value within the range of x-values used to calculate the a and b considered to be a reliable prediction

  - **Extrapolation**
    predicting using an x-value outside of the range of x-values used to calculate a and b UNRELIABLE estimate
    - because you are assuming that the linear trend continues indefinitely – use your common sense to explain why this may be incorrect
    - watch out for NEGATIVE (or unrealistic) y values which may result for the x values suggested – again use your common sense to explain why this is unrealistic

- SCALING either the x or the y values will change the equation

  e.g $y = 0.5 – 1.2x$

  If the x values are doubled then the equation becomes $y = 0.5 – 1.2(2x)$

  $y = 0.5 – 2.4x$

  If 5 is added on to each of the y values then the equation becomes

  $y + 5 = 0.5 – 1.2x$

  $y = - 4.5 – 1.2 x$ *(always rearrange to get y = 'a' + 'b'x)*

CORRELATION

The Product Moment Correlation Coefficient : is a numerical measures of the strength and type of correlation – denoted by **r** and will lie in the range $-1 \leq r \leq 1$

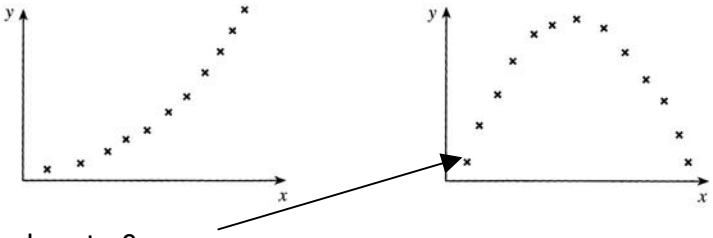
- indicates how well the data, when plotted in a scatter graph, fits a **straight line** pattern
- NOT APPROPRIATE if the data does not follow a **linear pattern** when plotted (straight line) – so scatter graph is needed to check this
- If you are give a table of values – use your calculator to find r
  (If you have time it's a good idea to check that you have entered your values correctly)
- If you are calculating using summary values -make sure you use the formula book – showing how you substituted the values into the formula

INTERPRETING r – make sure you do this in the context of the question (not just positive correlation)
e.g. There appears to be a fairly strong relationship between temperature and ice cream sales, higher temperatures appear to correspond to higher values of ice-cream sales and vice versa.

- Scaling data – a linear transformation or scaling of one or both of the variables will not affect the correlation coefficient – all of the points will stay in the same position RELATIVE to each other

TAKE CARE

- Not all correlation will be linear



For this data, the correlation coefficient is close to 0
This does not mean that there is no correlation but simply mean that there is no **linear correlation** (pattern appears to be quadratic)

- **Spurious Correlation**
  A strong correlation between 2 variables does not mean that one thing **causes** the other –high marks in a maths exam do not necessarily cause high marks in a Statistics exam, they are likely to both be dependent on a common third variable : the students mathematical ability
- **Outliers**
  One or two outliers can have a dramatic effect on a correlation coefficient