Statistics GCSE Higher Revision Sheet

This document attempts to sum up the contents of the Higher Tier Statistics GCSE.

There is one exam, two hours long. A calculator is allowed. It is worth 75% of the whole GCSE; the other 25% is coursework.

Before you go into the exam make sure you are fully equipped with two pens, two pencils, a calculator, a ruler, a protractor and a pair of compasses. Also be sure not to panic; it is not uncommon to get stuck on a question (I've been there!). Just continue with what you can do and return at the end to the question(s) you have found hard. If you have time check all your work, especially the first question you attempted... always an area prone to error.

I am always available on `jonathan.m.stone@gmail.com` to answer any questions you may have. Please do not hesitate.

J M S

Summarising Data

- For any set of data we first tend to look for a *measure of central tendency* and a *measure of spread*. The former tells us on average where a piece of data is located and the measure of spread tells us how spread out it is.
- The *mode* is the value that occurs most often. A set of data can have two (or more) modes or no mode at all.
- The *mean* is the value when you add them all up and divide by how many there are.
- The *median* is the middle value when they are written out in order. If there is a gap in the middle then you take the average of the two numbers either side. You should think $\left(\frac{n+1}{2}\right)^{\text{th}}$ when determining which data point it is.
- Given a frequency distribution you create a new column (the product of the first two; xf) and sum it to help you work out the *mean*. For example

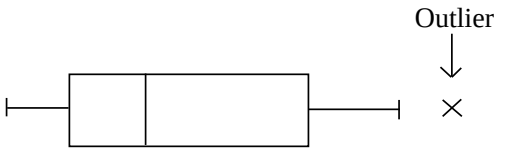
x	f					x	f	xf
0	1	$\Rightarrow$	Create	$\Rightarrow$		0	1	0
1	3	$\Rightarrow$	Extra	$\Rightarrow$		1	3	3
2	5	$\Rightarrow$	Column	$\Rightarrow$		2	5	10
3	4	$\Rightarrow$	&	$\Rightarrow$		3	4	12
4	2	$\Rightarrow$	Sum	$\Rightarrow$		4	2	8
	15						15	33

  So the mean is $\overline{x} = \frac{\sum xf}{\sum f} = 33/15$.
- If you have grouped data ($0 \leqslant x < 5$, $5 \leqslant x < 10$, etc.) then the best you can do is *estimate the mean* by using the mid-point (2.5, 7.5, etc.) and carry out the same process as above.
- The *modal class* will be the class with the highest frequency.

- For grouped data you can only state the class interval that contains the median (e.g. it lies in the class $15 \leqslant x < 20$). You can also do slightly better than this by drawing a cumulative frequency curve and then reading across and down from half the maximum of the vertical axis (P115).

- An outlier is any value which is more than 1.5 times the interquartile range below the lower quartile or 1.5 times the interquartile range above the upper quartile.

- When drawing a box and whisker diagram you must include any outliers you have found as 'x's. For example



- The variance of a set of data is defined to be $\text{Var}(x) = \frac{\sum(x_i - \overline{x})^2}{n}$. In practice it is easier to use the formula

$$\text{Var}(x) = \frac{\sum x^2}{n} - \overline{x}^2.$$

- The standard deviation is just the square root of the variance. Example; calculate the standard deviation of the data set $\{3, 4, 6, 10, 12, 13\}$. Firstly we note that $\overline{x} = 8$ and $n = 6$. $\sum x^2 = 3^2 + 4^2 + 6^2 + 10^2 + 12^2 + 13^2 = 474$. Therefore

$$\text{sd} = \sqrt{\frac{\sum x^2}{n} - \overline{x}^2} = \sqrt{\frac{474}{6} - 8^2} = 3.873 \text{ (3dp)}.$$

- For a frequency distribution the formula for the standard deviation is slightly different.

$$\text{sd} = \sqrt{\frac{\sum x^2 f}{\sum f} - \overline{x}^2}.$$

We can add columns to any frequency distribution to help us calculate the elements of the calculation. Two columns have been added to the original two columns below.

x	f	xf	$x^2 f$
1	4	4	4
2	5	10	20
3	7	21	63
4	5	20	80
5	4	20	100
	$n = \sum f = 25$	$\sum(xf) = 75$	$\sum(x^2 f) = 267.$

So $\overline{x} = \dfrac{\sum(xf)}{n} = \dfrac{75}{25} = 3$ and $s = \sqrt{\dfrac{\sum(x^2 f) - n\overline{x}^2}{n-1}} = \sqrt{\dfrac{267 - 25 \times 3^2}{24}} = 1.3228\ldots$

- Standardised scores can be used to compare two different sets of data. They are defined

$$\text{Standardised score}(z) = \frac{\text{score} - \text{mean}}{\text{standard deviation}}.$$

A standardised score of zero is the mean. A standardised score of 2.2 is a score just over two standard deviations above the mean (a very good score!).

Scatter Diagrams and Correlation

- Spearman's rank correlation coefficient (r_s) is bound to come up. You will be given a table and you will need to (in the next 2 columns) rank the data. If two data points are tied then you (eg the 2nd and 3rd are tied) then you rank them both 2.5.

%	IQ	Rank %	Rank IQ	d	d^2
89	143	2.5	1	1.5	2.25
55	89	7	8	-1	1
72	102	5	6	1	1
91	136	1	2	-1	1
89	126	2.5	3	0.5	0.25
30	60	9	9	0	0
71	115	6	4	-2	4
53	100	8	7	-1	1
78	103	4	5	1	1

Now $\boxed{r_s = 1 - \dfrac{6\Sigma d^2}{n(n^2-1)}}$. Σd^2 is just the sum of the d^2 column in the table and n is the number of pairs of data; here $n = 9$. We therefore find $r_s = 1 - \frac{6\times 11.5}{9(81-1)} = 0.904\dot{1}\dot{6}$. Therefore we see a strong degree of positive linear correlation.

- If r_s is close to -1 then strong negative linear correlation. If close to zero then no linear correlation.

Time Series

- A *time series* is a set of observations of a variable taken over a period of time. It can be drawn in a *line graph*.

- A *trend line* is a line that shows the overall trend trend of the data. It can remove any *seasonal variation* (think ice cream sales through a year; peaks and troughs).

- A *moving average* is a way of smoothing out these seasonal variations. For example to calculate the 3 point moving average of a set of data you take the first three pieces of data and work out their mean. Then you take the 2nd, 3rd and 4th and work out the mean. Etcetera until you reach the last three. For example the data $\{3, 6, 5, 7, 4, 7, 9, 8\}$ would give $\{4\frac{2}{3}, 6, 5\frac{1}{3}, 6, 6\frac{2}{3}, 8\}$ as its 3 point moving average.

- You can plot these MA points on a line graph. See precisely where to plot these on the graph in the example on P196. These points then help you to draw a trend line.

Probability

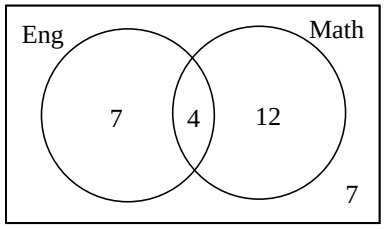
- Conditional probability is found difficult. This is because it can give counter-intuitive answers. For example the probability of a person chosen at random in the world being female is approximately 0.5. However the probability of a person chosen at random being female, *given that they are a student at Eton* is zero.

- Conditional probabilities can be read from tables and Venn diagrams with ease. The following table gives food eaten at a restaurant.

	Lamb	Beef	Total
Sorbet	4	7	11
Crumble	8	4	12
Total	12	11	23

If asked for P(person chosen has eaten beef) then the answer is clearly $\frac{11}{23}$. However if asked for P(eaten beef given that had crumble), this is a different question. We must imagine we are in the row for crumble only. We can therefore see the answer is $\frac{4}{12}$. You should read P(A|B) as P(A given that B has happened).

- Consider the following students surveyed as to whether they studied Maths of English for A Level.



If asked P(Maths|English) then we must imagine we live in the English circle only. We can see the answer is therefore $\frac{4}{11}$. Similarly P(English|Maths)=$\frac{4}{16}$.

Probability Distributions

- For any possible experiment, the probabilities *must* sum to one. So for example if the outcomes of an experiment were a, b, c, and d and we saw

x	a	b	c	d
$P(x)$	0.2	0.5	k	0.1

Then the missing probability must be $k = 1 - (0.2 + 0.5 + 0.1) = 0.3$. In the harder example

x	a	b	c	d
$P(x)$	k	$2k$	$3k$	0.4

We still have that $k + 2k + 3k + 0.4 = 1$ so we discover $k = 0.1$. Therefore

x	a	b	c	d
$P(x)$	0.1	0.2	0.3	0.4

- The *discrete uniform distribution* is where all the probabilities of each outcome are equally likely. The perfect example is the dice. $P(1) = P(2) = P(3) = P(4) = P(5) = P(6) = \frac{1}{6}$.
- The *binomial distribution* is applicable where an experiment is repeated and each time the probabilities remain constant (we say that the events are *independent*). In such an experiment we view one outcome as a '*success*' and one a '*failure*'. We denote $P(\text{success}) = p$ and $P(\text{failure}) = q$. Obviously $p + q = 1$ so we can also say $P(\text{failure}) = 1 - p$.

- At A Level we discover

$$P(r \text{ successes from } n \text{ trials}) = \frac{n!}{r!(n-r)!} \times p^r \times q^{n-r} = {}^nC_r \times p^r \times q^{n-r}.$$

The $\frac{n!}{r!(n-r)!}$ bit is the number of ways r objects can be chosen from n and can be accessed on your calculator by the 'nCr' button.

- For example; "The probability I am on time for work is 0.9. In a 13 day period what is the probability I will be on time exactly 11 times?"

Well $p = 0.9$, $q = 0.1$ so the answer is ${}^{13}C_{11} \times 0.9^{11} \times 0.1^2 = 0.245$ (3dp).

- Your textbook (and the exam) has tried to 'help' you by giving you information like

$$(p+q)^4 = p^4 + 4p^3q + 6p^2q^2 + 4pq^3 + q^4.$$

This is a way of telling you that the ${}^nC_r$ coefficients in this case are {1,4,6,4,1}.

So for example; "The probability that a tulip bulb will flower is 0.3. If I plant 4 in my garden what is the probability that more than one will flower?"

Well $p = 0.3$, $q = 0.7$ and more than one means two, three or four. So the answer is $(6 \times 0.3^2 \times 0.7^2) + (4 \times 0.3^3 \times 0.7^1) + (1 \times 0.3^4 \times 0.7^0) = 0.3483$.

- Notation. In general we write $X \sim B(n, p)$. We read this as "The random variable X is distributed binomially with n trials and a probability of p of success".

In the example above about being on time for work we write $X \sim B(13, 0.9)$. X is the number of times I am on time. 13 trials. 0.9 chance of success. X can take the values $\{0, 1, \ldots 12, 13\}$.

- The *normal distribution* is a distribution often found in nature for continuous variables. For example you can expect people's height or weight to be normally distributed. It is sometimes described as the 'bell curve' or the 'Gaussian distribution'.

It is fully described by the mean, μ and the standard deviation σ. The mean is the line of symmetry in the middle and the standard deviation describes how spread out the data is. The mean, median and mode are the same in a normal distribution.

- 95% of the data lie within 2 standard deviations of the mean and 99.8% of the data lie within 3. Beware that the 5% and the 0.2% not contained within these bounds are at *both ends of the distribution* and care should be taken with exam questions. A sketch should always be drawn of the situation!