OCR Statistics 1 Module Revision Sheet

The S1 exam is 1 hour 30 minutes long. You are allowed a graphics calculator.

Before you go into the exam make sure you are fully aware of the contents of the formula booklet you receive. 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.

J.M.S.

"Without data, all you are is just another person with an opinion."

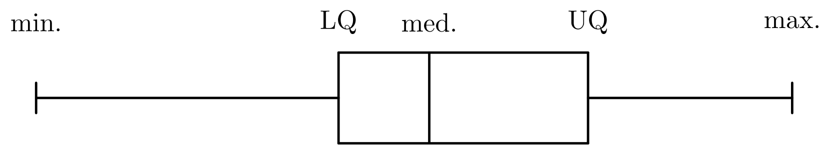
Representation Of Data

- You must be happy constructing unordered, back-to-back and ordered stem and leaf diagrams. They show the overall distribution of the data and back-to-back diagrams allow you to compare two sets of data.
- Cumulative frequency graphs. The cumulative frequency is a "running total" of the frequencies as you go up the values. For example

x	f				x (upper limit of)	cum. freq
$0 \leqslant x < 5$	8	$\Rightarrow$	Create	$\Rightarrow$	5	8
$5 \leqslant x < 10$	13	$\Rightarrow$	Cumulative	$\Rightarrow$	10	21
$10 \leqslant x < 15$	17	$\Rightarrow$	Frequency	$\Rightarrow$	15	38
$15 \leqslant x < 20$	10				20	48

  Plot the second of these tables and join it with a smooth curve to form the *cumulative frequency curve*. From this the median and the two quartiles can be found.

- Once these values are found we can draw a *box and whisker diagram*. The box and whisker diagram uses five values: the minimum, the maximum, the lower quartile, the upper quartile and the median. It is good for showing spread and comparing two quantities.



- Histograms are usually drawn for continuous data in classes. If the classes have equal widths, then you merely plot amount against frequency.
- If the classes do *not* have equal widths then we need to create a new column for *frequency density*. Frequency density is defined by f.d. $= \frac{\text{frequency}}{\text{class width}}$. The *area* of the bars are what represents the frequency, *not* the height.
- Frequency polygons are made by joining together the mid-points of the bars of a histogram with a ruler.

Measures Of Location

- The *mean* (arithmetic mean) of a set of data $\{x_1, x_2, x_3 \dots x_n\}$ is given by

$$\overline{x} = \frac{\text{sum of all values}}{\text{the number of values}} = \frac{\sum x}{n}.$$

When finding the mean[1] of a frequency distribution the mean is given by

$$\frac{\sum(xf)}{\sum f} = \frac{\sum(xf)}{n}.$$

- If a set of numbers is arranged in ascending (or descending) order the *median* is the number which lies half way along the series. It is the number that lies at the $\left(\frac{n+1}{2}\right)^{\text{th}}$ position. Thus the median of $\{13, 14, 15, 15\}$ lies at the $2\frac{1}{2}$ position $\Rightarrow$ average of 14 and 15 $\Rightarrow$ median = 14.5.

- The *mode* of a set of numbers is the number which occurs the most frequently. Sometimes no mode exists; for example with the set $\{2, 4, 7, 8, 9, 11\}$. The set $\{2, 3, 3, 3, 4, 5, 6, 6, 6, 7\}$ has two modes 3 and 6 because each occurs three times. One mode $\Rightarrow$ "unimodal". Two modes $\Rightarrow$ "bimodal". More than two modes $\Rightarrow$ "multimodal".

	Advantages	Disadvantages
Mean	⋆ The best known average. ⋆ Can be calculated exactly. ⋆ Makes use of all the data. ⋆ Can be used in further statistical work.	⋆ Greatly affected by extreme values. ⋆ Can't be obtained graphically. ⋆ When the data are discrete can give an impossible figure (2.34 children).
Median	⋆ Can represent an actual value in the data. ⋆ Can be obtained even if some of the values in a distribution are unknown. ⋆ Unaffected by irregular class widths and unaffected by open-ended classes. ⋆ Not influenced by extreme values.	⋆ For grouped distributions its value can only be estimated from an ogive. ⋆ When only a few items available or when distribution is irregular the median may not be characteristic of the group. ⋆ Can't be used in further statistical calculations.
Mode	⋆ Unaffected by extreme values. ⋆ Easy to calculate. ⋆ Easy to obtain from a histogram.	⋆ May exist more than one mode. ⋆ Can't be used for further statistical work. ⋆ When the data are grouped its value cannot be determined exactly.

Measures Of Spread

- The simplest measure of spread is the *range*. Range $= x_{\text{max}} - x_{\text{min}}$.

- The interquartile range is simply the upper quartile take away the lower quartile. Both of these values are usually found from a cumulative frequency graph (above).

- The *sum of squares from the mean* is called the *sum of squares* and is denoted

$$S_{xx} = \sum(x - \overline{x})^2 = \sum x^2 - n\overline{x}^2.$$

[1]Statistics argues that the average person has one testicle and that 99.999% of people have more than the average number of arms...

For example given the data set $\{3, 6, 7, 8\}$ the mean is 6; $\sum x^2 = 9 + 36 + 49 + 64 = 158$; so $S_{xx} = \sum x^2 - n\overline{x}^2 = 158 - 4 \times 6^2 = 14$.[2]

- The *standard deviation* (σ) is defined: $\sigma = \sqrt{\text{variance}} = \sqrt{\frac{S_{xx}}{n}} = \sqrt{\frac{\sum x^2}{n} - \overline{x}^2}$.

- *Example*: Given the set of data $\{5, 7, 8, 9, 10, 10, 14\}$ calculate the standard deviation. Firstly we note that $\overline{x} = 9$.

$$\sigma = \sqrt{\frac{\sum x^2}{n} - \overline{x}^2} = \sqrt{\frac{(5^2 + \cdots + 14^2)}{7} - 9^2}$$
$$= \sqrt{\frac{615}{7} - 81} = 2.6186\ldots$$

- When dealing with frequency distributions such as

x	1	2	3	4	5
f	4	5	7	5	4

, we *could* calculate σ by writing out the data[3] and carrying out the calculations as above, but this is clearly slow and inefficient. To our rescue comes a formula for σ that allows direct calculation from the table. This is

$$\sigma = \sqrt{\frac{\sum(x^2 f)}{n} - \overline{x}^2}.$$

- *Example*: Calculate mean and sd for the above frequency distribution. For easy calculation we need to add certain columns to the usual x and f columns thus;

x	f	xf	x^2f
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 $\overline{x} = \frac{\sum(xf)}{n} = \frac{75}{25} = 3$ and $\sigma = \sqrt{\frac{\sum(x^2 f)}{n} - \overline{x}^2} = \sqrt{\frac{267}{25} - 3^2} = 1.2961\ldots$

- *Linear Coding*. Given the set of data $\{2, 3, 4, 5, 6\}$ we can see that $\overline{x} = 4$ and it can be calculated that $\sigma = 1.414$ (3dp). If we add 20 to all the data points we can see that the mean becomes 24 and the standard deviation will be unchanged. If the data set is multiplied by 3 we can see that the mean becomes 12 and the standard deviation would become three times as large (4.743 (3dp)).

- If, instead of being given $\sum x$ and $\sum x^2$, you were given $\sum(x - a)$ and $\sum(x - a)^2$ for some constant a, you just use the substitution $u = x - a$ and use $\sum u$ and $\sum u^2$ to work out the mean of u and the standard deviation of u. Then, using the above paragraph, we know $\bar{x} = \bar{u} + a$ and $\sigma_x = \sigma_u$.

[2] Or we could have done $S_{xx} = \Sigma(x - \overline{x})^2 = (3-6)^2 + (6-6)^2 + (7-6)^2 + (8-6)^2 = 14$.
[3] $\{1, 1, 1, 1, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 3, 4, 4, 4, 4, 4, 5, 5, 5, 5\}$!!!

Probability

- An *independent event* is one which has no effect on subsequent events. The events of spinning a coin and then cutting a pack of cards are independent because the way in which the coin lands has no effect on the cut. For two *independent* events A & B

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

  For example a fair coin is tossed and a card is then drawn from a pack of 52 playing cards. Find the probability that a head and an ace will result.

  $$\mathbb{P}(\text{head}) = \tfrac{1}{2}, \qquad \mathbb{P}(\text{ace}) = \tfrac{4}{52} = \tfrac{1}{13}, \qquad \text{so } \mathbb{P}(\text{head and ace}) = \tfrac{1}{2} \times \tfrac{1}{13} = \tfrac{1}{26}.$$

- *Mutually Exclusive Events.* Two events which cannot occur at the same time are called mutually exclusive. The events of throwing a 3 or a 4 in a single roll of a fair die are mutually exclusive. For any two mutually exclusive events

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

  For example a fair die with faces of 1 to 6 is rolled once. What is the probability of obtaining either a 5 or a 6?

  $$\mathbb{P}(5) = \tfrac{1}{6}, \qquad \mathbb{P}(6) = \tfrac{1}{6}, \qquad \text{so } \mathbb{P}(5 \text{ or } 6) = \tfrac{1}{6} + \tfrac{1}{6} = \tfrac{1}{3}.$$

- *Non-Mutually Exclusive Events.* When two events can both happen they are called non-mutually exclusive events. For example studying English and studying Maths at A Level are non-mutually exclusive. By considering a Venn diagram of two events A & B we find

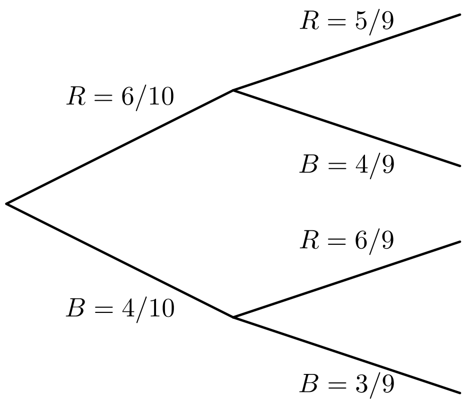
  $$\mathbb{P}(A \text{ or } B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \text{ and } B),$$
  $$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B).$$

- *Tree Diagrams.* These may be used to help solve probability problems when more than one event is being considered. The probabilities on any branch section must sum to one. You multiply along the branches to discover the probability of that branch occurring.

  For example a box contains 4 black and 6 red pens. A pen is drawn from the box and it is not replaced. A second pen is then drawn. Find the probability of

  (i) two red pens being obtained.
  (ii) two black pens being obtained.
  (iii) one pen of each colour being obtained.
  (iv) two red pens *given* that they are the same colour.

  Draw tree diagram to discover:



  (i) $\mathbb{P}(\text{two red pens}) = \frac{6}{10} \times \frac{5}{9} = \frac{30}{90} = \frac{1}{3}$.
  (ii) $\mathbb{P}(\text{two black pens}) = \frac{4}{10} \times \frac{3}{9} = \frac{12}{90} = \frac{2}{15}$.
  (iii) $\mathbb{P}(\text{one of each colour}) = 1 - \frac{30}{90} - \frac{12}{90} = \frac{8}{15}$.
  (iv) $\mathbb{P}(\text{two reds} \mid \text{same colour}) = \frac{1/3}{1/3+2/15} = \frac{5}{7}$.

- *Conditional Probability.* In the above example we see that the probability of two red pens is $\frac{1}{3}$, but the probability of two red pens *given that both pens are the same colour* is $\frac{5}{7}$. This is known as conditional probability. $\mathbb{P}(A \mid B)$ mean the probability of A *given* that B has happened. It is governed by

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \text{ and } B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}.$$

For example if there are 120 students in a year and 60 study Maths, 40 study English and 10 study both then

$$\mathbb{P}(\text{study English} \mid \text{study Maths}) = \frac{\mathbb{P}(\text{study Maths \& English})}{\mathbb{P}(\text{study Maths})} = \frac{10/120}{60/120} = \frac{1}{6}.$$

- A is independent of B if $\mathbb{P}(A) = \mathbb{P}(A \mid B) = \mathbb{P}(A \mid B')$. (i.e. whatever happens in B the probability of A remains unchanged.) For example flicking a coin and then cutting a deck of cards to try and find an ace are independent because

$$\mathbb{P}(\text{cutting ace}) = \mathbb{P}(\text{cutting ace} \mid \text{flick head}) = \mathbb{P}(\text{cutting ace} \mid \text{flick tail}) = \tfrac{1}{13}.$$

Permutations And Combinations

- Factorials are defined $n! \equiv n \times (n-1) \times (n-2) \times \cdots \times 2 \times 1$. Many expressions involving factorials simplify with a bit of thought. For example $\frac{n!}{(n-2)!} = n(n-1)$. Also there is a convention that $0! = 1$.

- The number of ways of arranging n different objects in a line is $n!$ For example how many different arrangements are there if 4 different books are to be placed on a bookshelf? There are 4 ways in which to select the first book, 3 ways in which to choose the second book, 2 ways to pick the third book and 1 way left for the final book. The total number of different ways is $4 \times 3 \times 2 \times 1 = 4!$

- Permutations. The number of ways of selecting r objects from n when *the order of the selection matters* is ${}^nP_r$. It can be calculated by

$${}^nP_r = \frac{n!}{(n-r)!}.$$

For example in how many ways can the gold, silver and bronze medals be awarded in a race of ten people? The order in which the medals are awarded matters, so the number of ways is given by ${}^{10}P_3 = 720$.

In another example how many words of four letters can be made from the word CONSIDER? This is an arrangement of four out of eight different objects where the order matters so there are ${}^8P_4 = \frac{8!}{4!} = 1680$ different words.

- Combinations. The number of ways of selecting r objects from n when *the order of the selection does not matter* is ${}^nC_r$. It can be calculated by

$${}^nC_r \equiv \binom{n}{r} = \frac{n!}{r!\,(n-r)!}.$$

For example in how many ways can a committee of 5 people be chosen from 8 applicants? Solution is given by ${}^8C_5 = \frac{8!}{5! \times 3!} = 56$.

In another example how many ways are there of selecting your lottery numbers (where one selects 6 numbers from 49)? It does not matter which order you choose your numbers, so there are ${}^{49}C_6 = 13\,983\,816$ possible selections.

- If letters are repeated in a 'word', then you just divide through by the factorials of each repeat. Therefore there are $\frac{11!}{4!\times 4!\times 2!}$ arrangements of the word 'MISSISSIPPI'.[4]
- You must be good at 'choosing committee' questions [be on the lookout, they can be in disguise]. For example how many ways are there of choosing a committee of 3 women and 4 men from a group containing 10 women and 5 men? There are $\binom{10}{3}$ ways of choosing the women (the order doesn't matter) and $\binom{5}{4}$ ways of choosing the men. Therefore overall there are $\binom{5}{4} \times \binom{10}{3}$ ways of choosing the committee.
- Example: If I deal six cards from a standard deck of cards, in how many ways can I get exactly four clubs? Well there are $\binom{13}{4}$ ways of getting the clubs, and $\binom{39}{2}$ ways of getting the non-clubs, so therefore the answer to the original question is $\binom{13}{4} \times \binom{39}{2}$.
- When considering lining things up in a line we start from the principle that there are $n!$ ways of arranging n objects. In the harder examples you need to be a cunning.

  For example three siblings join a queue with 5 other people making 8 in total.

  1. How many way are there of arranging the 8 in a queue? Easy; 8!
  2. How many ways are there of arranging the 8, such that the siblings are together? We, we imaging the three siblings tied together. There are therefore 6! ways of arranging the 5 and the bundle of siblings and then there are 3! ways of arranging the siblings in the bundle. Therefore the answer is $6! \times 3!$
  3. How many ways are there of arranging the siblings so they are not together? There are 5! ways of arranging the five without the siblings. There are then 6 places for the first sibling to go, 5 for the second, and 4 for the third. Therefore $5! \times 6 \times 5 \times 4$.
- To calculate probabilities we go back to first principles and remember that probability is calculated from the number of ways of getting what we want over the total number of possible outcomes. So in the above example, if the 8 are arranged randomly in a line, what is the probability of the siblings being together? $\mathbb{P}(\text{together}) = \frac{6!\times 3!}{8!}$.

  Going back to the four club question if it asked for the probability of getting exactly four clubs if I dealt exactly six cards from the pack, the answer would be $\frac{\binom{13}{4}\times\binom{39}{2}}{\binom{52}{6}}$. The $\binom{52}{6}$ represents the total number of ways I can deal six cards from the 52.

Probability Distributions

- A random variable is a quantity whose value depends on chance. The outcome of a random variable is usually denoted by a capital letter (e.g. X). We read $\mathbb{P}(X = 2)$ as the probability that the random variable takes the value 2. For a fair die, $\mathbb{P}(X = 5) = \frac{1}{6}$.
- For discrete random variables they are usually presented in a table. For example for a fair die:

x	1	2	3	4	5	6
$\mathbb{P}(X = x)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

- In general, for any event, the probability distribution is of the form

x	x_1	x_2	x_3	x_4	x_5	x_6	$\cdots$
$\mathbb{P}(X = x)$	p_1	p_2	p_3	p_4	p_5	p_6	$\cdots$

[4] As an non-mathematical aside, find two fruits that are anagrams of each other.

- As before, it is crucial that we remember the probabilities sum to one. This can be useful at the start or problems where a constant must be evaluated. For example in:

x	1	2	3	4
$\mathbb{P}(X=x)$	k	k	$2k$	$4k$

  We discover $k+k+2k+4k=1$, so $k=\frac{1}{8}$.

Binomial And Geometric Distributions

- The **binomial distribution** is applicable when you have fixed number of *repeated, independent* 'trials' such that each trial can be viewed as 'success' (p) or 'fail' ($q=1-p$). In justifying a binomial distribution you must not just quote the previous sentence; you must apply it to the situation in the question. For example: "Binomial is applicable because the probability of each tulip flowering is independent of each other tulip and the probability of flowering is a constant".

- For example if I throw darts at a dart board and my chance of hitting a double is 0.1 and I throw 12 darts at the board and my chance of hitting a double is independent of all the other throws then a binomial distribution will be applicable. We let X be the number of doubles I hit. X can therefore take the values $\{0,1,2\ldots,11,12\}$; i.e. there are 13 possible outcomes. $p=0.1$; the probability of success and $q=1-p=0.9$; the probability of failure. We write $X\sim B(n,p)$ which here is $X\sim B(12,0.1)$.

- I would always advise you to visualise the tree diagram. From this we can 'see' that $\mathbb{P}(X=12)=0.1^{12}$ and $\mathbb{P}(X=0)=0.9^{12}$. In general

$$\mathbb{P}(X=x)=\binom{n}{x}\times p^x\times q^{n-x}.$$

  So in the example, the probability I hit exactly 7 doubles is $\mathbb{P}(X=7)=\binom{12}{7}\times 0.1^7\times 0.9^5$.

- For questions such as $\mathbb{P}(X\leqslant 5)$ or $\mathbb{P}(X\geqslant 8)$ you must be able to use the tables in the formula book. The tables always give $\mathbb{P}(X\leqslant \text{something})$. You must be able to convert probabilities to this form and then read off from the table. For $X\sim B(10,0.35)$.

$$\begin{aligned}
&\mathbb{P}(X\leqslant 7)=0.9952,\\
&\mathbb{P}(X<5)=\mathbb{P}(X\leqslant 4)=0.7515,\\
&\mathbb{P}(X\geqslant 7)=1-\mathbb{P}(X\leqslant 6)=1-0.9740=0.0260,\\
&\mathbb{P}(X>3)=1-\mathbb{P}(X\leqslant 3)=1-0.5138=0.4862.
\end{aligned}$$

- The **geometric distribution** is applicable when you are looking for how long you wait until an event has occurred. The events must be *repeated, independent and success/fail.* Potentially you could wait forever until a success occurs; something to look for if you are unsure what distribution to apply. Similar to the binomial you must justify *in the context of the question.*

- Going back to the darts example, we could rephrase it as how long must I wait until I hit a double? Let X be the number of throws until I hit a double. We write $X\sim\mathrm{Geo}(0.1)$. X can take the values $\{1,2,3,\ldots\}$.

- Obviously $\mathbb{P}(X=1)=0.1$. Less obviously $\mathbb{P}(X=4)=0.9^3\times 0.1$ (I must have three failures and *then* my success). In general

$$\mathbb{P}(X=x)=q^{x-1}\times p.$$

- There are no tables for the geometric distribution because there does not need to be. To calculate $\mathbb{P}(X \geqslant 5)$ we must have had 4 failures. Therefore $\mathbb{P}(X \geqslant 5) = q^4 = (1-p)^4$. Also to calculate $\mathbb{P}(X \leqslant 6)$ we use the fact that $\mathbb{P}(X \leqslant 6) = 1-\mathbb{P}(X \geqslant 7) = 1-q^6 = 1-(1-p)^6$. In general

$$\mathbb{P}(X \geqslant x) = (1-p)^{x-1} \qquad \text{and} \qquad \mathbb{P}(X \leqslant x) = 1-(1-p)^x.$$

Expectation And Variance Of A Random Variable

- The expected value of the event is denoted $\mathbb{E}(X)$ or μ. It is defined

$$\mathbb{E}(X) = \mu = \boxed{\sum x\mathbb{P}(X=x)}.$$

For example for a fair die with

x	1	2	3	4	5	6
$\mathbb{P}(X=x)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

we find:

$$\begin{aligned}\mathbb{E}(X) &= \left(1\times\tfrac{1}{6}\right) + \left(2\times\tfrac{1}{6}\right) + \left(3\times\tfrac{1}{6}\right) + \left(4\times\tfrac{1}{6}\right) + \left(5\times\tfrac{1}{6}\right) + \left(6\times\tfrac{1}{6}\right)\\ &= 3\tfrac{1}{2}.\end{aligned}$$

- The variance of an event is denoted $\mathrm{Var}(X)$ or σ^2 and is defined

$$\mathrm{Var}(X) = \sigma^2 = \mathbb{E}(X^2) - (\mathbb{E}(X))^2 = \mathbb{E}(X^2) - \mu^2 = \boxed{\sum x^2\mathbb{P}(X=x) - \mu^2}.$$

So for the *biased* die with distribution

x	1	2	3	4	5	6
$\mathbb{P}(X=x)$	$\frac{1}{3}$	$\frac{1}{6}$	0	0	$\frac{1}{6}$	$\frac{1}{3}$

we find that

$$\mathbb{E}(X) = \left(1\times\tfrac{1}{3}\right) + \left(2\times\tfrac{1}{6}\right) + (3\times 0) + (4\times 0) + \left(5\times\tfrac{1}{6}\right) + \left(6\times\tfrac{1}{3}\right) = 3\tfrac{1}{2}$$

and

$$\begin{aligned}\mathrm{Var}(X) &= \sum x^2\mathbb{P}(X=x) - \mu^2\\ &= \left(1^2\times\tfrac{1}{3}\right) + \left(2^2\times\tfrac{1}{6}\right) + \left(3^2\times 0\right) + \left(4^2\times 0\right) + \left(5^2\times\tfrac{1}{6}\right) + \left(6^2\times\tfrac{1}{3}\right) - 3\tfrac{1}{2}^2\\ &= 17\tfrac{1}{6} - 3\tfrac{1}{2}^2 = 4\tfrac{11}{12}.\end{aligned}$$

- The expectation of a binomial distribution $B(n,p)$ is np. The variance of $B(n,p)$ is npq.
- The expectation of a geometric distribution $\mathrm{Geo}(p)$ is $\frac{1}{p}$.

Correlation

- The Product Moment Correlation Coefficient is a number (r) calculated on a set of bivariate data that tells us how correlated two data sets are.
- The value of r is such that $-1 < r < 1$. If $r = 1$ you have perfect positive linear correlation. If $r = -1$ you have perfect negative linear correlation. If $r = 0$ then there exists no correlation between the data sets.
- It is defined
$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$
where we define the individual components as
$$\begin{aligned} S_{xx} &= \sum x^2 - \tfrac{1}{n}\left(\sum x\right)^2, \\ S_{yy} &= \sum y^2 - \tfrac{1}{n}\left(\sum y\right)^2, \\ S_{xy} &= \sum xy - \tfrac{1}{n}\sum x \sum y. \end{aligned}$$
- So to calculate r for the data set

x	14	12	16	18	21	13	15	17
y	1	2	4	5	2	8	5	6

we write the data in columns and add extra ones. We then sum the columns and calculate from these sums. Note that in the above example $n = 8$ (i.e. the number of pairs, not the number of individual data pieces).

x	y	x^2	y^2	xy
14	1	196	1	14
12	2	144	4	24
16	4	256	16	64
18	5	324	25	90
21	2	441	4	42
13	8	169	64	104
15	5	225	25	75
17	6	289	36	102
126	**33**	**2044**	**175**	**515**

Therefore
$$\begin{aligned} S_{xx} &= \sum x^2 - \tfrac{1}{n}\left(\sum x\right)^2 = 2044 - \tfrac{126^2}{8} = 59.5, \\ S_{yy} &= \sum y^2 - \tfrac{1}{n}\left(\sum y\right)^2 = 175 - \tfrac{33^2}{8} = 38.875, \\ S_{xy} &= \sum xy - \tfrac{1}{n}\sum x \sum y = 515 - \tfrac{126 \times 33}{8} = -4.75. \end{aligned}$$
Therefore
$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{-4.75}{\sqrt{59.5 \times 38.875}} = -0.09876\ldots$$
Therefore the data has very, very weak negative correlation. Basically it has no *meaningful* correlation.
- It can be shown that if one (or both) of the variables are transformed in a linear fashion i.e. if we replace the x values by, say, $\frac{x-4}{3}$ (or any transformation formed by $+$, $-$, $\div$ or $\times$ with constants) then the value of r will be unchanged.

- You need to be able to calculate Spearman's rank correlation coefficient (r_s). 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 (e.g. 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 $\boxed{r_s = 1 - \frac{6\sum d^2}{n(n^2-1)}}$. $\sum d^2$ 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 $r_s = 1 - \frac{6\times 11.5}{9(81-1)} = 0.9041\dot{6}$. Therefore we see a strong degree of positive association.

- If r_s is close to -1 then strong negative association. If close to zero then no meaningful association/agreement.

Regression

- For any set of bivariate data (x_i, y_i) there exist two possible regression lines; 'y on x' and 'x on y'.
- If neither is controlled (see below) then if you want to predict y from a given value of x, you use the 'y on x' line. If you want to predict x from a given value of y, you use the 'x on y' line.
- The 'y on x' line is defined

$$y = a + bx \quad \text{where} \quad b = \frac{S_{xy}}{S_{xx}} \quad \text{and} \quad a = \bar{y} - b\bar{x}.$$

- The 'x on y' line is defined

$$x = a' + b'y \quad \text{where} \quad b' = \frac{S_{xy}}{S_{yy}} \quad \text{and} \quad a' = \bar{x} - b'\bar{y}.$$

- Both regression lines pass through the average point $(\bar{x}, \bar{y})$.
- In the example in the book (P180) the height of the tree is the dependent variable and the circumference of the tree is the independent variable. This is because the experiment has been constructed to see how the height of the tree depends on its circumference.
- If one variable is being controlled by the experimenter (e.g. x), it is called a controlled variable. If x is controlled you would never use the 'x on y' regression line. Only use the 'y on x' line. You would use this to predict y from x (expected) *and* x from y (not-expected)