Statistics 1 Module Revision Sheet

The S1 exam is 1 hour 30 minutes long and is in two sections.

Section A (36 marks) 5 questions worth no more than 8 marks each.

Section B (36 marks) 2 questions worth about 18 marks each.

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.

JMS

1. Exploring Data

Measures of Central Tendency

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

$$\overline{x} = \frac{\text{sum of all values}}{\text{the number of values}} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{\sum x}{n}.$$

When finding the mean 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 $\frac{n+1}{2}$ position. Thus the median of $\{13, 14, 15, 15\}$ lies at the $2\frac{1}{2}$ position ⇒ average of 14 and 15 ⇒ 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 ⇒ "unimodal". Two modes ⇒ "bimodal". More than two modes ⇒ "multimodal".

- The *mid-range* is given by the average of the minimum and maximum values. Mid-range = $(x_{\text{max}} + x_{\text{min}})/2$.

	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 *mean absolute deviation from the mean* is given by $\frac{1}{n}\sum|x-\overline{x}|$. For example in the data set $\{4, 5, 7, 8\}$ the mean is 6, so the absolute deviations are $2, 1, 1, 2$ so the mean absolute deviation is $\frac{1}{4}(2 + 1 + 1 + 2) = 1.5$.
- The *sum of squares from the mean* is called the *sum of squares* and is denoted
$$S_{xx} = \sum(x-\overline{x})^2 = \sum x^2 - n\overline{x}^2.$$
For example given the data set $\{3, 6, 7, 8\}$ the mean is 6; $\sum x^2 = 9 + 36 + 49 + 64 = 158$; so $S_{xx} = \sum x^2 - n\overline{x}^2 = 158 - 4 \times 6^2 = 14$.[1]
- The *mean square deviation* is defined: msd $= \dfrac{S_{xx}}{n} = \dfrac{\sum x^2 - n\overline{x}^2}{n}$.
- The *root mean square deviation* is defined: rmsd $= \sqrt{\text{msd}} = \sqrt{\dfrac{S_{xx}}{n}} = \sqrt{\dfrac{\sum x^2 - n\overline{x}^2}{n}}$.
- The *variance* is defined: variance $= \dfrac{S_{xx}}{n-1} = \dfrac{\sum x^2 - n\overline{x}^2}{n-1}$.
- The *standard deviation* (s) is defined: $s = \sqrt{\text{variance}} = \sqrt{\dfrac{S_{xx}}{n-1}} = \sqrt{\dfrac{\sum x^2 - n\overline{x}^2}{n-1}}$.
- On graphical calculators from Casio the rmsd is given by '$x\sigma_n$' and the sd by '$x\sigma_{n-1}$'.
- *Example*: Given the set of data $\{5, 7, 8, 9, 10, 10, 14\}$ calculate the standard deviation. Firstly we note that $\overline{x} = 9$.
$$s = \sqrt{\frac{S_{xx}}{n-1}} = \sqrt{\frac{\sum x^2 - n\overline{x}^2}{n-1}} = \sqrt{\frac{(5^2 + \cdots + 14^2) - 7 \times 9^2}{6}}$$
$$= \sqrt{\frac{615 - 567}{6}} = \sqrt{8} = 2.8284\ldots$$

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

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

- When dealing with frequency distributions such as the table above, we *could* calculate the rmsd or the sd by writing out the data[2] and carrying out the calculations as above, but this is clearly slow and inefficient.[3] To our rescue come formulae for rmsd and sd that allow direct calculation from the table. They are

$$\text{rmsd} = \sqrt{\frac{\sum(x^2f) - n\overline{x}^2}{n}} \qquad \text{sd} = \sqrt{\frac{\sum(x^2f) - n\overline{x}^2}{n-1}}.$$

- *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
	$n = \sum f = 25$	$\sum(xf) = 75$	$\sum(x^2f) = 267.$

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

- An item of data is an *outlier* if it is more than two standard deviations from the mean (i.e. outlier if $|x - \overline{x}| > 2s$). It means that some more investigation is needed to see if it needs to be discarded. 95% of the data lie within two standard deviations and 99.75% lie within three standard deviations (assuming normally distributed population).

- *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 $s = 1.581$ (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)).

  Combining the above ideas we find that given a data set x_i and we transform it to create a new data set $y_i = ax_i + b$ then the new mean will be $\overline{y} = a\overline{x} + b$ and the new standard deviation will be $s_y = as_x$. This can be used to make certain calculations easier. For example;

x	f				y	f
164	2	⇒		⇒	1	2
168	3	⇒	Convert $y = \frac{x-160}{4}$,	⇒	2	3
172	4	⇒	therefore $x = 4y + 160$	⇒	3	4
176	4	⇒		⇒	4	4

  Once we find $\overline{y}$ and s_y we find that $\overline{x} = 4\overline{y} + 160$ and $s_x = 4s_y$.

[2] $\{1, 1, 1, 1, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 3, 4, 4, 4, 4, 4, 5, 5, 5, 5\}$!!!

[3] Indeed, it would be nearly impossible if the frequencies were in the thousands.

3. 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

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

  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.

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

- *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

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

  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?

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

- *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

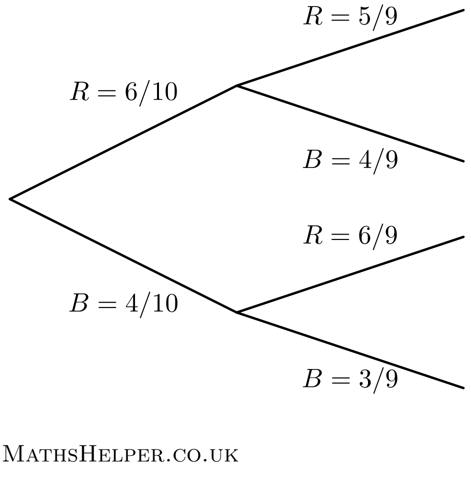
  $$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B),$$
  $$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

- *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) $P(\text{two red pens}) = \frac{6}{10} \times \frac{5}{9} = \frac{30}{90} = \frac{1}{3}.$

(ii) $P(\text{two black pens}) = \frac{4}{10} \times \frac{3}{9} = \frac{12}{90} = \frac{2}{15}.$

(iii) $P(\text{one of each colour}) = 1 - \frac{30}{90} - \frac{12}{90} = \frac{8}{15}.$

(iv) $P(\text{two reds} \mid \text{same colour}) = \frac{1/3}{1/3+2/15} = \frac{5}{7}.$

- *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. $P(A \mid B)$ mean the probability of A *given* that B has happened. It is governed by

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

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

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

- A is independent of B if $P(A) = P(A \mid B) = P(A \mid B')$. (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

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

4. Discrete Random Variables

- The table below shows the probability distribution for the outcome (X) of a die.

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

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

r	r_1	r_2	r_3	r_4	r_5	r_6	$\cdots$
$P(X=r)$	p_1	p_2	p_3	p_4	p_5	p_6	$\cdots$

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

$$E(X) = \mu = \boxed{\sum rP(X=r)}.$$

For example for a fair die

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

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

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

So for the *biased* die with distribution

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

we find that

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

and

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

- The other way of calculating these quantities is by using a table. We will consider the example of the bias die above.

r	$P(X = r)$	$rP(X = r)$	$r^2P(X = r)$
1	$\frac{1}{3}$	$1 \times \frac{1}{3} = \frac{1}{3}$	$1^2 \times \frac{1}{3} = \frac{1}{3}$
2	$\frac{1}{6}$	$2 \times \frac{1}{6} = \frac{1}{3}$	$2^2 \times \frac{1}{6} = \frac{2}{3}$
3	0	$3 \times 0 = 0$	$3^2 \times 0 = 0$
4	0	$4 \times 0 = 0$	$4^2 \times 0 = 0$
5	$\frac{1}{6}$	$5 \times \frac{1}{6} = \frac{5}{6}$	$5^2 \times \frac{1}{6} = 4\frac{1}{6}$
6	$\frac{1}{3}$	$6 \times \frac{1}{3} = 2$	$6^2 \times \frac{1}{3} = 12$
		$\sum rP(X = r) = 3\frac{1}{2}$	$\sum r^2P(X = r) = 17\frac{1}{6}$

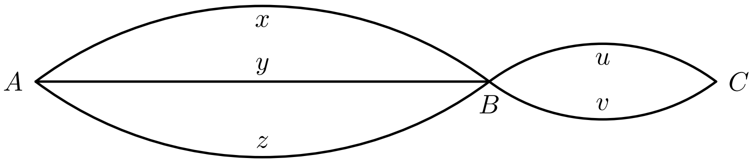
So, as before $E(X) = 3\frac{1}{2}$ and $\mathrm{Var}(X) = 17\frac{1}{6} - 3\frac{1}{2}^2 = 4\frac{11}{12}$.

5. Further Probability

- Factorials are defined $n! = n \times (n-1) \times (n-2) \times \cdots \times 2 \times 1$. Many expressions involving factorials simplify with a bit of thought. For example $n!/(n-2)! = n(n-1)$. 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 $4 \times 3 \times 2 \times 1 = 4!$

- Several events. If there are 3 roads from A to B and 2 roads from B to C. How many routes are there from A to C?



The solution to our problem is $3 \times 2 = 6$ because the set of possible routes is

$$x \to u \qquad y \to u \qquad z \to u \qquad x \to v \qquad y \to v \qquad z \to v.$$

In general if there are a ways for trial A to result, b ways for trial B to result and c ways for trial C to result then there are $a \times b \times c$ different possible outcomes.

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

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

  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 ${}^{8}P_{4} = 8!/4! = 1680$ different words.

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

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

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

  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.