OCR Further Statistics Revision Sheet

The OCR further maths A level is examined with four 90 minute exams. Each paper carries equal weight (25%) and each paper is marked out of 75 marks. Two of the papers are compulsory:
Pure Core 1.
Pure Core 2.

Then you (or your school) selects two out of:
Statistics.
Mechanics.
Discrete Mathematics.
Additional Pure Mathematics.

This revision sheet *should* cover all of statistics you need for the optional statistics paper. It represents 25% of your further maths A level. *Please* get in contact if you spot anything missing.

I hope you find this revision sheet useful and wish you the very best of luck with your studies.

J.M.S.

Representation Of Data

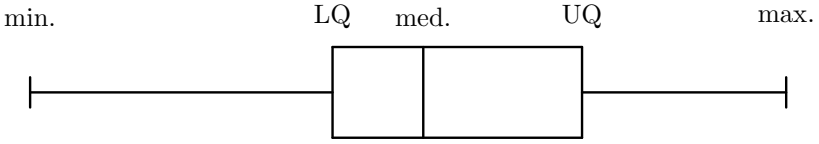
- You must be happy constructing unordered, back-to-back and ordered stem and leaf diagrams. They show the overall distribution of the data and back-to-back diagrams allow you to compare two sets of data.
- Cumulative frequency graphs. The cumulative frequency is a "running total" of the frequencies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

x	f
$0 \leqslant x < 5$	8
$5 \leqslant x < 10$	13
$10 \leqslant x < 15$	17
$15 \leqslant x < 20$	10

$\Rightarrow$ Create $\Rightarrow$ Cumulative $\Rightarrow$ Frequency $\Rightarrow$

x (upper limit of)	cum. freq
5	8
10	21
15	38
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 $\{x_1, x_2, x_3 \ldots x_n\}$ 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 $= x_{\max} - x_{\min}$.

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

- 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

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

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: $\sigma = \sqrt{\text{variance}} = \sqrt{\dfrac{S_{xx}}{n}} = \sqrt{\dfrac{\sum x^2}{n} - \overline{x}^2}$.
- *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	$x^2 f$
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 $\sigma = \sqrt{\dfrac{\sum (x^2 f)}{n} - \overline{x}^2} = \sqrt{\dfrac{267}{25} - 3^2} = 1.2961\ldots$

- *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

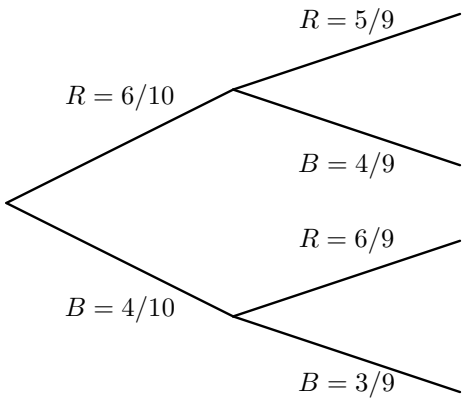
$$\mathbb{P}(A \text{ or } B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \text{ and } B),$$
$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B).$$

- *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 arranging objects 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 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

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

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	$\cdots$

- 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 $X\sim \text{Geo}(0.1)$. 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

$$\begin{aligned} S_{xx} &= \sum x^2 - \tfrac{1}{n}(\sum x)^2, \\ S_{yy} &= \sum y^2 - \tfrac{1}{n}(\sum y)^2, \\ S_{xy} &= \sum xy - \tfrac{1}{n}\sum x \sum y. \end{aligned}$$

- 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

$$\begin{aligned} S_{xx} &= \sum x^2 - \tfrac{1}{n}(\sum x)^2 = 2044 - \tfrac{126^2}{8} = 59.5, \\ S_{yy} &= \sum y^2 - \tfrac{1}{n}(\sum y)^2 = 175 - \tfrac{33^2}{8} = 38.875, \\ S_{xy} &= \sum xy - \tfrac{1}{n}\sum x \sum y = 515 - \tfrac{126\times 33}{8} = -4.75. \end{aligned}$$

  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 $\boxed{r_s = 1 - \dfrac{6\sum d^2}{n(n^2-1)}}$. $\sum d^2$ 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 $r_s = 1 - \frac{6\times 11.5}{9(81-1)} = 0.90416\dot{}$. 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)

Continuous Random Variables

- A continuous random variable (crv) is usually described by means of a probability density function (pdf) which is defined for all real x. It must satisfy

$$\int_{-\infty}^{\infty} f(x)\,dx = 1 \quad \text{and} \quad f(x) \geqslant 0 \text{ for all } x.$$

- Probabilities are represented by areas under the pdf. For example the probability that X lies between a and b is

$$\mathbb{P}(a < X < b) = \int_a^b f(x)\,dx.$$

  It is worth noting that for any specific value of X, $\mathbb{P}(X = \text{value}) = 0$ because the area of a single value is zero.

- The median is the value m such that

$$\int_{-\infty}^{m} f(x)\,dx = \frac{1}{2}.$$

  That is; the area under the curve is cut in half at the value of the median. Similarly the lower quartile (Q_1) and upper quartile (Q_3) are defined

$$\int_{-\infty}^{Q_1} f(x)\,dx = \frac{1}{4} \quad \text{and} \quad \int_{-\infty}^{Q_3} f(x)\,dx = \frac{3}{4}.$$

- The expectation of X is defined

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x f(x)\,dx.$$

  Compare this to the discrete definition of $\sum x\mathbb{P}(X = x)$. Always be on the lookout for symmetry in the distribution before carrying out a long integral; it could save you a lot of time. You should therefore always sketch the distribution if you can.

- The variance of X is defined

$$\text{Var}(X) = \int_{-\infty}^{\infty} x^2 f(x)\,dx - \mu^2.$$

  Again, compare this to the discrete definition of $\sum x^2\mathbb{P}(X = x) - \mu^2$. Don't forget to subtract μ^2 at the end; someone always does!

- The main use for this chapter is to give you the basics you may need for the normal distribution. The normal distribution is by far the most common crv.

The Normal Distribution

- The normal distribution (also known as the Gaussian distribution[5]) is the most common crv. It is found often in nature; for example daffodil heights, human IQs and pig weights can all be modelled by the normal curve. A normal distribution can be summed up by two parameters; its mean (μ) and its variance (σ^2). For a random variable X we say $X \sim N(\mu, \sigma^2)$.

[5] I *do* wish we would call it the Gaussian distribution. Carl Friedrich Gauss. Arguably the greatest mathematician ever. German…

- As with all crvs probabilities are given by areas; i.e. $\mathbb{P}(a < X < b) = \int_a^b f(x)\,dx$. However the $f(x)$ for a normal distribution is complicated and impossible to integrate exactly. We therefore need to use tables to help us. Since there are an infinite number of $N(\mu, \sigma^2)$ distributions we use a special one called the standard normal distribution. This is $Z \sim N(0, 1^2)$.

- The tables given to you work out the areas to the left of a value. The notation used is $\Phi(z) = \int_{-\infty}^{z} f(z)\,dz$. So $\Phi(0.2)$ is the area to the left of 0.2 in the standard normal distribution. The tables do not give Φ(negative value) so there are some tricks of the trade you must be comfortable with. These and they are always helped by a sketch and remembering that the area under the whole curve is one. For example

$$\Phi(z) = 1 - \Phi(-z)$$
$$\mathbb{P}(Z > z) = 1 - \Phi(z)$$

- Real normal distributions are related to the standard distribution by

$$Z = \frac{X - \mu}{\sigma} \qquad (\dagger).$$

  So if $X \sim N(30, 16)$ and we want to answer $\mathbb{P}(X > 24)$ we convert $X = 24$ to $Z = (24-30)/4 = -1.5$ and answer $\mathbb{P}(Z > -1.5) = \mathbb{P}(Z < 1.5) = 0.9332$.

- Another example; If $Y \sim N(100, 5^2)$ and we wish to calculate $\mathbb{P}(90 < Y < 105)$. Converting to $\mathbb{P}(-2 < Z < 1)$ using †. Then finish off with

$$\mathbb{P}(-2 < Z < 1) = \Phi(1) - \Phi(-2) = \Phi(1) - (1 - \Phi(2)) = 0.8413 - (1 - 0.9772) = 0.8185.$$

- You must also be able to do a 'reverse' lookup from the table. Here you don't look up an area from a z value, but look up a z value from an area.

  For example find a such that $\mathbb{P}(Z < a) = 0.65$. Draw a sketch as to what this means; to the left of some value a the area is 0.65. Therefore, reverse looking up we discover $a = 0.385$.

- Harder example; Find b such that $\mathbb{P}(Z > b) = 0.9$. Again a sketch shows us that the area to the right of b must be 0.9, so b must be negative. Considering the sketch carefully, we discover $\mathbb{P}(Z < -b) = 0.9$, so reverse look up tells us $-b = 1.282$, so $b = -1.282$.

- Reverse look up is then combined with † in questions like this. For $X \sim N(\mu, 5^2)$ it is known $\mathbb{P}(X < 20) = 0.8$; find μ. Here you will find it easier if you draw both a sketch for the X and also for Z and marking on the important points. The z value by reverse look up is found to be 0.842. Therefore by † we obtain, $0.842 = (20 - \mu)/5$, so $\mu = 15.79$.

- Harder example; $Y \sim (\mu, \sigma^2)$ you know $\mathbb{P}(Y < 20) = 0.25$ and $\mathbb{P}(Y > 30) = 0.4$. You should obtain two † equations;

$$-0.674 = \frac{20 - \mu}{\sigma} \quad \text{and} \quad 0.253 = \frac{30 - \mu}{\sigma} \quad \Rightarrow \quad \mu = 27.27 \text{ and } \sigma = 10.79.$$

- The binomial distribution can sometimes be approximated by the normal distribution. If $X \sim B(n, p)$ and $np > 5$ and $nq > 5$ then we can use $V \sim N(np, npq)$ as an approximation. Because we are going from a discrete distribution to a continuous, a continuity correction must be used.

- For example if $X \sim B(90, \frac{1}{3})$ we can see $np = 30 > 5$ and $nq = 60 > 5$ so we can use $V \sim N(30, 20)$. Some examples of the conversions:

$$\mathbb{P}(X = 29) \approx \mathbb{P}(28.5 < V < 29.5),$$
$$\mathbb{P}(X > 25) \approx \mathbb{P}(V > 25.5),$$
$$\mathbb{P}(5 \leqslant X < 40) \approx \mathbb{P}(4\tfrac{1}{2} < V < 39\tfrac{1}{2}).$$

The Poisson Distribution

- The Poisson distribution is a discrete random variable (like the binomial or geometric distribution). It is defined
$$\mathbb{P}(X = x) = e^{-\lambda}\frac{\lambda^x}{x!}.$$
X can take the values $0, 1, 2, \ldots$ and the probabilities depend on only one parameter, λ. Therefore we find

x	0	1	2	3	…
$\mathbb{P}(X = x)$	$e^{-\lambda}\frac{\lambda^0}{0!}$	$e^{-\lambda}\frac{\lambda^1}{1!}$	$e^{-\lambda}\frac{\lambda^2}{2!}$	$e^{-\lambda}\frac{\lambda^3}{3!}$	…

- For a Poisson distribution $\mathbb{E}(X) = \mathrm{Var}(X) = \lambda$. We write $X \sim \mathrm{Po}(\lambda)$.
- As for the binomial we use tables to help us and they are given (for various different λs) in the form $\mathbb{P}(X \leqslant x)$. So if $\lambda = 5$ and we wish to discover $\mathbb{P}(X < 8)$ we do $\mathbb{P}(X < 8) = \mathbb{P}(X \leqslant 7) = 0.8666$. Also note that if we want $\mathbb{P}(X \geqslant 4)$ we would use the fact that probabilities sum to one, so $\mathbb{P}(X \geqslant 4) = 1 - \mathbb{P}(X \leqslant 3) = 1 - 0.2650 = 0.7350$.
- The Poisson distribution can be used as an approximation to the binomial distribution provided $n > 50$ and $np < 5$. If these conditions are met and $X \sim B(n, p)$ we use $W \sim \mathrm{Po}(np)$. [No continuity correction required since we are approximating a discrete by a discrete.]
- For example with $X \sim B(60, \frac{1}{30})$ both conditions are met and we use $W \sim \mathrm{Po}(2)$. Therefore some example of some calculations:
$$\mathbb{P}(X \leqslant 3) \approx \mathbb{P}(W \leqslant 3) = 0.8571 \text{ (from tables)}$$
$$\mathbb{P}(3 < X \leqslant 7) \approx \mathbb{P}(3 < W \leqslant 7) = \mathbb{P}(W \leqslant 7) - \mathbb{P}(W \leqslant 3) = 0.9989 - 0.8571 = 0.1418.$$
- The normal distribution can be used as an approximation to the to the Poisson distribution if $\lambda > 15$. So if $X \sim \mathrm{Po}(\lambda)$ we use $Y \sim N(\lambda, \lambda)$. However, here we *are* approximating a discrete by a continuous, so a continuity correction must be applied.
- For example if $X \sim \mathrm{Po}(50)$ we can use $Y \sim N(50, 50)$ since $\lambda > 15$. To calculate $\mathbb{P}(X = 49)$ we would calculate (using $Z = (X - \mu)/\sigma$)
$$\begin{aligned}\mathbb{P}(X = 49) \approx \mathbb{P}(48.5 < Y < 49.5) &= \mathbb{P}(-0.212 < Z < -0.071)\\ &= \mathbb{P}(0.071 < Z < 0.212)\\ &= \Phi(0.212) - \Phi(0.071)\\ &= 0.5840 - 0.5283 = 0.0557.\end{aligned}$$
Similarly
$$\begin{aligned}\mathbb{P}(X < 55) &\approx \mathbb{P}(Y < 54.5)\\ &= \mathbb{P}\left(Z < \frac{54.5 - 50}{\sqrt{50}}\right)\\ &= \mathbb{P}(Z < 0.6364)\\ &= 0.738.\end{aligned}$$

Sampling

- If a sample is taken from an underlying population you can view the mean of this sample as a random variable in its own right. This is a subtle point and you should dwell on it! If you can't get to sleep sometime, you should lie awake thinking about it. (I had to.)
- If the underlying population has $\mathbb{E}(X) = \mu$ and $\text{Var}(X) = \sigma^2$, then the distribution of the mean of the sample, $\bar{X}$, is

$$\mathbb{E}(\bar{X}) = \mu \text{ (the same as the underlying) and } \text{Var}(\bar{X}) = \frac{\sigma^2}{n}.$$

  This means that the larger your sample, the less likely it is that the mean of this sample is a long way from the population mean. So if you are taking a sample, make it as big as you can!
- If your sample is sufficiently large (roughly > 30) the central limit theorem (CLT) states that the distribution of the sample mean is approximated by

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

  no matter what the underlying distribution is.
- If the underlying population is discrete you need to include a $\frac{1}{2n}$ correction factor when using the CLT. For example $\mathbb{P}(\bar{X} > 3.4)$ for a discrete underlying with a sample size of 45 would mean you calculate $\mathbb{P}(\bar{X} > 3.4 + \frac{1}{90})$.
- If the underlying population is a normal distribution then no matter how large the sample is (e.g. just 4) we can say

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

- If you have the whole population data available to you then to calculate the mean you use $\mu = \frac{\sum x}{n}$ and to calculate the variance you use

$$\sigma^2 = \frac{\sum x^2}{n} - \bar{x}^2 = \frac{\sum x^2 - n\bar{x}^2}{n}.$$

  However you do not usually have all the data. It is more likely that you merely have a sample from the population. From this sample you may want to estimate the population mean and variance. As you would expect your best estimate of the population mean is the mean of the sample $\frac{\sum x}{n}$. However the best estimate of the population variance is not the variance of the sample. You must calculate s^2 where

$$s^2 = \frac{\sum x^2 - n\bar{x}^2}{n-1} = \frac{n}{n-1}\left(\frac{\sum x^2 - n\bar{x}^2}{n}\right) = \frac{n}{n-1}\left(\frac{\sum x^2}{n} - \bar{x}^2\right).$$

  Some textbooks use $\hat{\sigma}$ to mean s; they both mean 'the unbiased estimator of the population σ'. So

$$\text{(Estimate of population variance)} = \frac{n}{n-1} \times \text{(Sample variance)}.$$

- You could be given raw data ($\{x_1, x_2, \ldots x_n\}$) in which you just do a direct calculation. Or summary data ($\sum x^2$, $\sum x$ and n). Or you could be given the sample variance and n. From all of these you should be able to calculate s^2. It should be clear from the above section how to do this.

Continuous Hypothesis Testing

- In *any* hypothesis test you will be testing a 'null' hypothesis H_0 against an 'alternative' hypothesis H_1. In S2, your H_0 will only *ever* be one of these three:

  $H_0 : p$ =something

  $H_0 : \lambda$ =something

  $H_0 : \mu$ =something

  Don't deviate from this and you can't go wrong. Notice that it does *not* say $H_0 = p$ =something.

- The book gives three approaches to continuous hypothesis testing, but they are all essentially the same. You always compare the probability of what you have seen (under H_0) and anything more extreme, and compare this probability to the significance level. If it is less than the significance level, then you reject H_0 and if it is greater, then you accept H_0.

- Remember we connect the real (X) world to the standard (Z) world using $Z = \frac{X-\mu}{\sigma}$.

- You can do this by:

  1. Calculating the probability of the observed value and anything more extreme and comparing to the significance level.
  2. Finding the critical Z-values for the test and finding the Z-value for the observed event and comparing. (e.g. critical Z-values of 1.96 and -1.96; if observed Z is 1.90 we accept H_0; if observed is -2.11 the reject H_0.)
  3. Finding the critical values for $\bar{X}$. For example critical values might be 17 and 20. If X lies between them then accept H_0; else reject H_0.

- Example: P111 Que 8. Using method 3 from above.

  Let X be the amount of magnesium in a bottle. We are told $X \sim N(\mu, 0.18^2)$. We are taking a sample of size 10, so $\bar{X} \sim N(\mu, \frac{0.18^2}{10})$. Clearly

  $$\begin{aligned} H_0 &: \mu = 6.8 \\ H_1 &: \mu \neq 6.8. \end{aligned}$$

  We proceed assuming H_0 is correct. Under H_0, $\bar{X} \sim N(6.8, \frac{0.18^2}{10})$. This is a 5% two-tailed test, so we need $2\frac{1}{2}\%$ at each end of our normal distribution. The critical Z values are (by reverse lookup) $Z_{\text{crit}} = \pm 1.960$. To find how these relate to $\bar{X}_{\text{crit}}$ we convert thus

  $$\begin{aligned} Z_{\text{crit}} &= \frac{\bar{X}_{\text{crit}} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \\ 1.960 &= \frac{\bar{X}_{\text{crit}} - 6.8}{\sqrt{\frac{0.18^2}{10}}} \\ \text{and} \quad -1.960 &= \frac{\bar{X}_{\text{crit}} - 6.8}{\sqrt{\frac{0.18^2}{10}}} \end{aligned}$$

  These solve to $\bar{X}_{\text{crit}} = 6.912$ and $\bar{X}_{\text{crit}} = 6.688$. The observed $\bar{X}$ is 6.92 which lies just outside the acceptance region. We therefore reject H_0 and conclude that the amount of magnesium per bottle is probably *different* to 6.8. [The book is in error in claiming that we conclude it is bigger than 6.8.]

Discrete Hypothesis Testing

- For any test with discrete variables, it is usually best to find the critical value(s) for the test you have set and hence the critical region. The critical value is the first value *at which you would reject* the null hypothesis.
- For example if testing $X \sim B(16, p)$ we may test (at the 5% level)

$$\begin{aligned} \mathrm{H}_0 &: p = \tfrac{5}{6} \\ \mathrm{H}_1 &: p < \tfrac{5}{6}. \end{aligned}$$

  We are looking for the value at the lower end of the distribution (remember the "<" acts as an arrow telling us where to look in the distribution). We find $\mathbb{P}(X \leqslant 11) = 0.1134$ and $\mathbb{P}(X \leqslant 10) = 0.0378$. Therefore the critical value is 10. Thus the critical region is $\{0, 1, 2 \ldots 9, 10\}$. So when the result for the experiment is announced, if it lies in the critical region, we reject H_0, else accept H_0.
- Another example: If testing $X \sim B(20, p)$ at the 10% level with

$$\begin{aligned} \mathrm{H}_0 &: p = \tfrac{1}{6} \\ \mathrm{H}_1 &: p \neq \tfrac{1}{6}. \end{aligned}$$

  Here we have a two tailed test with 5% at either end of the distribution. At the lower end we find $\mathbb{P}(X = 0) = 0.0261$ and $\mathbb{P}(X \leqslant 1) = 0.1304$ so the critical value is 0 at the lower end. At the upper end we find $\mathbb{P}(X \leqslant 5) = 0.8982$ and $\mathbb{P}(X \leqslant 6) = 0.9629$. Therefore

$$\begin{aligned} \mathbb{P}(X \geqslant 6) &= 1 - \mathbb{P}(X \leqslant 5) = 1 - 0.8982 = 0.1018 \\ \mathbb{P}(X \geqslant 7) &= 1 - \mathbb{P}(X \leqslant 6) = 1 - 0.9629 = 0.0371 \end{aligned}$$

  So at the upper end we find $X = 7$ to be the critical value. [Remember that at the upper end, the critical value is always one more than the upper of the two values where the gap occurs; here the gap was between 5 and 6 in the tables, so 7 is the critical value.] The critical region is therefore $\{0, 7, 8 \ldots 20\}$.
- There is a Poisson example in the 'Errors in hypothesis testing' section.

Errors In Hypothesis Testing

- A Type I error is made when a true null hypothesis is rejected.
- A Type II error is made when a false null hypothesis is accepted.
- For continuous hypothesis tests, the $\mathbb{P}$(Type I error) is just the significance level of the test. [This fact should be obvious; if not think about it harder!]
- For a Type II error, you must consider something like the example on page 140/1 which is superbly explained. From the original test, you will have discovered the acceptance and the rejection region(s). When you are told the real mean of the distribution and asked to calculate the $\mathbb{P}$(Type II error), you must use the new, real mean and the old standard deviation (with a new normal distribution; e.g. $N(\mu_{\text{new}}, \sigma^2_{\text{old}}/n)$) and work out the probability that the value lies within the old acceptance region. [Again, the book is *very* good on this and my explanation is poor.]

- For discrete hypothesis tests, the $\mathbb{P}$(Type I error) is not merely the stated significance level of the test. The stated value (e.g. 5%) is merely the 'notional' value of the test. The true significance level of the test (and, therefore, the $\mathbb{P}$(Type I error)) is the probability of all the values in the rejection region, given the truth of the null hypothesis.

  For example in a binomial hypothesis test we might have discovered the rejection region was $X \leqslant 3$ and $X \geqslant 16$. If the null hypothesis was "H_0: $p = 0.3$", then the true significance level of the test would be $\mathbb{P}(X \leqslant 3 \text{ or } X \geqslant 16 \mid p = 0.3)$.

- To calculate $\mathbb{P}$(Type II error) you would, given the true value for p (or λ for Poisson), calculate the probability of the *complementary* event. So in the above example, if the true value of p was shown to be 0.4, you would calculate $\mathbb{P}(3 < X < 16 \mid p = 0.4)$.

- Worked example for Poisson: A hypothesis is carried out to test the following:

$$\begin{aligned} H_0 &: \lambda = 7 \\ H_1 &: \lambda \neq 7 \\ \alpha &= 10\% \\ &\text{Two tailed test.} \end{aligned}$$

  Under H_0, $X \sim \text{Po}(7)$. We discover the critical values are $X = 2$ and $X = 13$. The critical region is therefore $X \leqslant 2$ and $X \geqslant 13$.

  Therefore $\mathbb{P}$(Type I error) and the true value of the test is therefore

$$\begin{aligned} \mathbb{P}(X \leqslant 2 \text{ or } X \geqslant 13 \mid \lambda = 7) &= \mathbb{P}(X \leqslant 2) + \mathbb{P}(X \geqslant 13) \\ &= \mathbb{P}(X \leqslant 2) + 1 - \mathbb{P}(X \leqslant 12) \\ &= 0.0296 + 1 - 0.9730 \\ &= 0.0566 = 5.66\%. \end{aligned}$$

  Given that the true value of λ was shown to be 10, then $\mathbb{P}$(Type II error) would be

$$\begin{aligned} \mathbb{P}(2 < X < 13 \mid \lambda = 10) &= \mathbb{P}(X \leqslant 12) - \mathbb{P}(X \leqslant 2) \\ &= 0.7916 - 0.0028 \\ &= 0.7888 = 78.88\%. \end{aligned}$$

Preliminaries

- In S1 when calculating the variance you will mostly have used $\sigma^2 = \dfrac{\sum x^2}{n} - \bar{x}^2$. This was for ease of calculation. However in S3 the equivalent formula $\sigma^2 = \dfrac{\sum (x - \bar{x})^2}{n}$ appears to make a storming comeback. You will often be given $\sum (x - \bar{x})^2$ summary data and you must know how to handle it.

- The unbiased estimator of variance from a sample (s^2) simplifies to

$$s^2 \equiv \frac{n}{n-1}\left(\frac{\sum x^2}{n} - \bar{x}^2\right) = \frac{n}{n-1}\left(\frac{\sum (x-\bar{x})^2}{n}\right) = \frac{\sum (x-\bar{x})^2}{n-1}.$$

- Because of this, if you need to calculate (for a two sample t-test) $s_p^2 = \frac{(n_x-1)s_x^2 + (n_y-1)s_y^2}{n_x+n_y-2}$ and are

given $\sum(x-\bar{x})^2$ and $\sum(y-\bar{y})^2$ then s_p^2 simplifies thus

$$\begin{aligned} s_p^2 &= \frac{(n_x-1)s_x^2+(n_y-1)s_y^2}{n_x+n_y-2} \\ &= \frac{(n_x-1)\left(\frac{\sum(x-\bar{x})^2}{n_x-1}\right)+(n_y-1)\left(\frac{\sum(y-\bar{y})^2}{n_y-1}\right)}{n_x+n_y-2} \\ &= \frac{\sum(x-\bar{x})^2+\sum(y-\bar{y})^2}{n_x+n_y-2}. \end{aligned}$$

Continuous Random Variables

- In S2 you met probability density functions (pdf) $f(x)$. They measured where events were more likely to occur than others. To find $\mathbb{P}(a < X < b)$ we needed to calculate the area between $x = a$ and $x = b$; i.e. $\int_a^b f(x)\,dx$. In S3 we have cumulative distribution functions (cdf) $F(x)$ which are defined $F(x) \equiv \mathbb{P}(X \leqslant x)$. We can think of $F(x)$ as the area to the left of x in the pdf. So $F(4)$ is the area to the left of 4 and $F(3)$ is the area to the left of 3. Therefore $\mathbb{P}(3 < X < 4) = F(4) - F(3)$. This is an example of:

$$\mathbb{P}(a < X < b) = F(b) - F(a).$$

- Cdfs make calculating the median (M) very easy. You just solve $F(M) = \frac{1}{2}$. Likewise the upper (Q_3) and lower (Q_1) quartiles are very easy to calculate; $F(Q_1) = \frac{1}{4}$ and $F(Q_3) = \frac{3}{4}$.

  You must understand the concept of percentiles and how to get them from a cdf. The 85th percentile (say) is such that 85% of the data lies to the left of that point. Therefore $F(P_{85}) = \frac{85}{100}$.

- You cannot write

$$\int_1^x x^2\,dx = \left[\frac{x^3}{3}\right]_1^x = \frac{x^3}{3} - \frac{1}{3}.$$

  You must use a dummy variable thus:

$$\int_1^x t^2\,dt = \left[\frac{t^3}{3}\right]_1^x = \frac{x^3}{3} - \frac{1}{3}.$$

  Basically whenever you find yourself putting an x on the upper limit of an integral, change all future x's to t's.

- To calculate $f(x)$ from $F(x)$ is easy; just differentiate $F(x)$. For example given

$$F(x) = \begin{cases} 0 & x < 0 \\ \frac{x^3}{27} & 0 \leqslant x \leqslant 3 \\ 1 & x > 3. \end{cases}$$

  When we differentiate the constants 0 and 1 they become 0. The $\frac{x^3}{27}$ becomes $\frac{x^2}{9}$ so the pdf is

$$f(x) = \begin{cases} \frac{x^2}{9} & 0 \leqslant x \leqslant 3 \\ 0 & \text{otherwise.} \end{cases}$$

- To calculate $F(x)$ from $f(x)$ is a little trickier. You must remember that $F(x)$ is the *entire* area to the left of a point. Therefore given

$$f(x) = \begin{cases} k & 0 \leqslant x < 2 \\ k(x-1) & 2 \leqslant x \leqslant 3 \\ 0 & \text{otherwise.} \end{cases}$$

Firstly we calculate[6] $k = \frac{2}{7}$. For the section $0 \leqslant x < 2$ we do the expected $\int_0^x \frac{2}{7}\,dt = \left[\frac{2}{7}t\right]_0^x = \frac{2}{7}x$. However, for the next region we *do not* just do $\int_2^x \frac{2}{7}(x-1)\,dt$. We need to *add in* the contribution from the first part (i.e. the value of $F(2)$ from the first result; $\frac{4}{7}$ in this case). So we do $\frac{4}{7} + \int_2^x \frac{2}{7}(t-1)\,dt = \frac{1}{7}(x^2 - 2x + 4)$. Therefore

$$F(x) = \begin{cases} 0 & x < 0 \\ \frac{2}{7}x & 0 \leqslant x < 2 \\ \frac{1}{7}(x^2 - 2x + 4) & 2 \leqslant x \leqslant 3 \\ 1 & x > 3. \end{cases}$$

- Once you have calculated your $F(x)$ a nice check to see whether your cdf is correct is to see if your $F(x)$ is continuous[7] *which it must be*. For example let's say you discovered that

$$F(x) = \begin{cases} 0 & x < 0 \\ \frac{1}{3}x & 0 \leqslant x < 1 \\ x^2 - \frac{5}{2}x + 2 & 1 \leqslant x \leqslant 2 \\ 1 & x > 2. \end{cases}$$

  You then check the 'boundary' values where the functions are being joined; here they are $x = 0$, $x = 1$ and $x = 2$. In this case there is no problem for $x = 0$ nor $x = 2$, but when we look at $x = 1$ there is a problem. $\frac{1}{3}x$ gives $\frac{1}{3}$ but $x^2 - \frac{5}{2}x + 2$ gives $\frac{1}{2}$. Therefore we must have made a mistake which must be fixed.

- Given a cdf $F(x)$ you can find a related cdf $F(y)$ where X and Y are related; i.e $Y = g(X)$. The idea here is that $F(x) \equiv \mathbb{P}(X \leqslant x)$. Start with the original cdf. Then write $F(y) = \mathbb{P}(Y \leqslant y) = \mathbb{P}(g(X) \leqslant y) = \mathbb{P}(X \leqslant g^{-1}(y))$. Then replace *every* x in the original cdf by $g^{-1}(y)$ (even the ones in the limits).

  For example given

$$F(x) = \begin{cases} 0 & x < 2 \\ \frac{1}{8}(x^2 - 2x) & 2 \leqslant x \leqslant 4 \\ 1 & x > 4. \end{cases}$$

  and $Y = 4X^2$ find $F(y)$.

  So, $F(y) = \mathbb{P}(Y \leqslant y) = \mathbb{P}(4X^2 \leqslant y) = \mathbb{P}(X \leqslant \frac{\sqrt{y}}{2})$ (we don't have to worry about $\pm$ when square rooting because the cdf is only defined for positive x). Therefore

$$F(y) = \begin{cases} 0 & \frac{\sqrt{y}}{2} < 2 \\ \frac{1}{8}((\frac{\sqrt{y}}{2})^2 - 2(\frac{\sqrt{y}}{2})) & 2 \leqslant \frac{\sqrt{y}}{2} \leqslant 4 \\ 1 & \frac{\sqrt{y}}{2} > 4. \end{cases}$$

  And so

$$F(y) = \begin{cases} 0 & y < 16 \\ \frac{1}{32}(y - 4\sqrt{y}) & 16 \leqslant y \leqslant 64 \\ 1 & y > 64. \end{cases}$$

- You must be a little careful if (say) $Y = \frac{1}{X}$. You start $F(y) = \mathbb{P}(Y \leqslant y) = \mathbb{P}(\frac{1}{X} \leqslant y) = \mathbb{P}(X \geqslant \frac{1}{y}) = 1 - \mathbb{P}(X \leqslant \frac{1}{y})$. Notice this reversal of the inequality sign; this is because if $\frac{a}{b} > \frac{c}{d}$ then $\frac{b}{a} < \frac{d}{c}$.

[6] By remembering $\int_{-\infty}^{\infty} f(x)\,dx = 1$.

[7] A function is continuous if you can draw it without taking your pen off the paper…basically.

- In S2 expectation for a pdf $f(x)$ is $\mathbb{E}(X) = \int_{-\infty}^{\infty} x f(x)\,dx$. In S3 you can find the expectation of any function $g(X)$ of the pdf $f(x)$ by the formula

$$\mathbb{E}(g(X)) = \int_{-\infty}^{\infty} g(x) f(x)\,dx.$$

For example find $\mathbb{E}(X^2 + 1)$ of

$$f(x) = \begin{cases} \frac{e^{x-1}}{e-1} & 1 \leqslant x \leqslant 2 \\ 0 & \text{otherwise.} \end{cases}$$

So,

$$\begin{aligned} \mathbb{E}(X^2 + 1) &= \int_1^2 (x^2 + 1)\frac{e^{x-1}}{e-1}\,dx \\ &= \frac{1}{e-1}\int_1^2 x^2 e^{x-1} + e^{x-1}\,dx \\ &= \text{int by parts twice on first bit...good exercise for you to do...} \\ &= \frac{3e-2}{e-1}. \end{aligned}$$

Linear Combinations Random Variables

- Any random variable X can be transformed to become a new random variable $Y = aX + b$ where a and b are constants. It can be shown that

$$\mathbb{E}(Y) = \mathbb{E}(aX + b) = a\mathbb{E}(X) + b.$$

It can also be shown that

$$\text{Var}(Y) = \text{Var}(aX + b) = a^2\text{Var}(X).$$

The b 'disappears' because it only has the effect of moving X up or down the number line and does not therefore alter the spread (i.e. variance). Note also that the a gets squared when one 'pulls it out' of the variance. Therefore $\text{Var}(-2X) = (-2)^2\text{Var}(X) = 4\text{Var}(X)$. It also makes sense with $\text{Var}(-X) = (-1)^2\text{Var}(X) = \text{Var}(X)$ because if one makes all the values of X negative from where they were they are just as spread out.

- Take any two random variables X and Y. If they are combined in a linear fashion $aX + bY$ for constant a and b then it is ***always true*** (even when X and Y are not independent) that

$$\mathbb{E}(aX + bY) = a\mathbb{E}(X) + b\mathbb{E}(Y).$$

If X and Y are *independent* then

$$\text{Var}(aX + bY) = a^2\text{Var}(X) + b^2\text{Var}(Y).$$

It is particularly useful to note that $\text{Var}(X - Y) = \text{Var}(X) + \text{Var}(Y)$. These results extend (rather obviously) to more than two variables

$$\begin{aligned} \mathbb{E}(a_1X_1 + a_2X_2 + \cdots + a_nX_n) &= a_1\mathbb{E}(X_1) + a_2\mathbb{E}(X_2) + + \cdots + a_n\mathbb{E}(X_n), \\ \text{Var}(a_1X_1 + a_2X_2 + \cdots + a_nX_n) &= a_1^2\text{Var}(X_1) + a_2^2\text{Var}(X_2) + \cdots + a_n^2\text{Var}(X_n). \end{aligned}$$

The second (of course) true if all independent.

- If X and Y are *independent* and normally distributed then $aX + bY$ is also normally distributed. Because $\mathbb{E}(aX + bY) = a\mathbb{E}(X) + b\mathbb{E}(Y)$ and $\text{Var}(aX + bY) = a^2\text{Var}(X) + b^2\text{Var}(Y)$ we find

$$X \sim N(\mu_1, \sigma_1^2) \quad \text{and} \quad Y \sim N(\mu_2, \sigma_2^2) \qquad \Rightarrow \qquad aX + bY \sim N(a\mu_1 + b\mu_2, a^2\sigma_1^2 + b^2\sigma_2^2).$$

  For example when Jon throws a shot put his distance is $J \sim N(11, 4)$. When Ali throws a shot his distance is $A \sim N(12, 9)$. Find the probability on one throw that Jon beats Ali. So we need $J - A \sim N(11 - 12, 4 + 9)$ which gives $J - A \sim N(-1, 13)$. Notice the variances have been added and that the expected value is negative (on average Jon will lose to Ali). Now

$$\begin{aligned}\mathbb{P}(J - A > 0) &= \mathbb{P}\left(Z > \frac{0 - (-1)}{\sqrt{13}}\right) \\ &= \mathbb{P}(Z > 0.277) \\ &= 1 - \mathbb{P}(Z < 0.277) = 0.3909\end{aligned}$$

- Given a random variable X you must fully appreciate the difference between two *independent* samplings of this random variable (X_1 and X_2) and two times this random variable ($2X$). For example given a random variable X such that

x	1	2
$\mathbb{P}(X = x)$	$\frac{1}{2}$	$\frac{1}{2}$

  The random variable $2X$ is doubling the outcome of *one* sampling of X, but $X_1 + X_2$ is adding *two* independent samplings of X. Thus $2X$ can *only* take values 2 and 4 with probabilities $\frac{1}{2}$ each. But $X_1 + X_2$ can take values 2, 3 and 4 with probabilities $\frac{1}{4}$, $\frac{1}{2}$ and $\frac{1}{4}$ respectively. Note that the expected values for $2X$ and $X_1 + X_2$ are the same (because $\mathbb{E}(2X) = 2\mathbb{E}(X)$ and $\mathbb{E}(X_1 + X_2) = \mathbb{E}(X_1) + \mathbb{E}(X_2) = 2\mathbb{E}(X)$), but that the variances are *not* the same; i.e. $\text{Var}(2X) \neq \text{Var}(X_1 + X_2)$. This is because $\text{Var}(2X) = 4\text{Var}(X)$ and $\text{Var}(X_1 + X_2) = \text{Var}(X_1) + \text{Var}(X_2) = 2\text{Var}(X)$.

  For example given the above shot put example $J \sim N(11, 4)$. If Jon was to throw the shot put three times (independently) and the total of all three throws recorded we would need $J_1 + J_2 + J_3 \sim N(33, 3 \times 4)$ and **not** $3J \sim N(33, 9 \times 4)$.

- Given Poisson distributed X and Y it is even simpler. Here $aX + bY$ is not distributed Poisson[8]. However the special case of $X + Y$ *is* distributed Poisson.

$$X \sim Po(\lambda_1) \quad \text{and} \quad Y \sim Po(\lambda_2) \qquad \Rightarrow \qquad X + Y \sim Po(\lambda_1 + \lambda_2).$$

  For example if Candy makes on average 3 typing errors per hour and Tiffany makes 4 typing errors per hour find the probability of fewer than 12 errors in total in a two hour period. Here we have $Po(14)$ so $\mathbb{P}(X < 12) = \mathbb{P}(X \leqslant 11) = 0.2600$ (tables).

Student's *t*-Distribution

- In S2 you learnt that if you take a sample from a normal population of *known variance* σ^2 then no matter what the sample size $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ exactly.

  The test statistic for $H_0 : \mu = c$ is $Z = \dfrac{\bar{X} - c}{\sqrt{\frac{\sigma^2}{n}}}$.

[8] Because with the Poisson we require the expectation and the variance to be the same and given $X \sim Po(\lambda_1)$ and $Y \sim Po(\lambda_2)$ we have $\mathbb{E}(aX + bY) = a\mathbb{E}(X) + b\mathbb{E}(Y) = a\lambda_1 + b\lambda_2$ and $\text{Var}(aX + bY) = a^2\text{Var}(X) + b^2\text{Var}(Y) = a^2\lambda_1 + b^2\lambda_2$ and the only time $aX + bY = a^2X + b^2Y$ is when $a = b = 1$.

- You also learnt that if you take a sample of size $n > 30$ from *any* population distribution where you know σ^2 then (by CLT) $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ approximately.

  The test statistic for $H_0 : \mu = c$ is $Z = \dfrac{\bar{X} - c}{\sqrt{\frac{\sigma^2}{n}}}$.

- You also learnt that if you take a sample of size $n > 30$ from *any* population distribution with unknown σ^2 then you estimate σ^2 by calculating s^2 and (by CLT) $\bar{X} \sim N\left(\mu, \frac{s^2}{n}\right)$ approximately.

  The test statistic for $H_0 : \mu = c$ is $Z = \dfrac{\bar{X} - c}{\sqrt{\frac{s^2}{n}}}$.

- You would therefore think that if you were drawing from a normal population with unknown σ^2 then you would estimate σ^2 by calculating s^2 and $\bar{X} \sim N\left(\mu, \frac{s^2}{n}\right)$. But ***this is not the case!!!*** In fact $\bar{X}$ is exactly described by Student's t-distribution[9].

  The test statistic for $H_0 : \mu = c$ is $T = \dfrac{\bar{X} - c}{\sqrt{\frac{s^2}{n}}}$.

- (You will notice the apparent contradiction between the last two bullet points. If a large sample ($n > 30$) is taken from a normal population with unknown variance then how can $\bar{X}$ be distributed *both* normally and as a t-distribution? Well, as the sample size gets larger, the t-distribution converges to the normal distribution. Just remember that *technically* if you have a normal population with unknown variance then $\bar{X}$ is *exactly* a t-distribution, but if $n > 30$ then CLT lets us *approximate* $\bar{X}$ as a normal. In practice the t-distribution is used only with small sample sizes.)

- There is the new concept of the degree of freedom (denoted ν) of the t-distribution. As ν gets larger the t-distribution tends towards the standard normal distribution. However if ν is small enough, then the difference between t and z becomes quite marked (as you can see yourself from the tables).

- We can do hypothesis tests here just like we did in S2, only instead of using the normal tables we use the t tables (with correct degrees of freedom ν) to find t_{crit} and compare the test statistic $\dfrac{\bar{X} - c}{\sqrt{\frac{s^2}{n}}}$ against t_{crit}. Here $\nu = n - 1$.

- For example a machine is producing circular disks whose radius is normally distributed. Their radius historically has been 5cm. The factory foreman believes that the machine is now producing disks that are too small. A sample of 9 disks are taken and their radii are

  $$4.8, \quad 4.9, \quad 4.5, \quad 5.2, \quad 4.9, \quad 4.8, \quad 5.0, \quad 4.8, \quad 5.0$$

  Test at the 10% level whether the foreman has a case.

  Let μ = the population mean radii of the disks.

  $H_0 : \mu = 5$,

  $H_1 : \mu < 5$.

[9] Named after W.S.Gosset who wrote under the pen name 'Student'. Gosset devised the t-test as a way to cheaply monitor the quality of stout. Good bloke.

$n = 9$, so $\nu = 9 - 1 = 8$.

$\alpha = 10\%$. Therefore in t_8 we lookup 90% (because one tailed) and discover 1.397. But because it is a "<" test t_{crit} must be negative to $t_{\text{crit}} = -1.397$.

$\bar{x} = \frac{\sum x}{n} = \frac{43.9}{9} = 4.8\dot{7}$.

$s^2 = \frac{n}{n-1}\left(\frac{\sum x^2}{n} - \bar{x}^2\right) = \frac{9}{8}\left(\frac{214.43}{9} - 4.8\dot{7}^2\right) = 0.03694$.

$t_{\text{obs}} = \frac{\bar{x}-c}{\sqrt{\frac{s^2}{n}}} = \frac{4.8\dot{7}-5}{\sqrt{\frac{0.03694}{9}}} = -1.908$.

$-1.908 < -1.397$. This value lies in the rejection region of the test and therefore at the 10% level we have sufficient evidence to reject H_0 and conclude that the machine is probably not working fine.

Testing For Difference Between Means

- The central pillar in this section is that if $\bar{X} \sim N\left(\mu_x, \frac{\sigma_x^2}{n_x}\right)$ (which is either exactly true if X is itself normal, or approximately true if $n_x > 30$ from CLT) and $\bar{Y} \sim N\left(\mu_y, \frac{\sigma_y^2}{n_y}\right)$ then (provided X and Y are independent)

$$\bar{X} - \bar{Y} \sim N\left(\mu_x - \mu_y, \frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}\right).$$

- If X and Y are *normally* distributed with *known* variances (σ_x^2 and σ_y^2) and we are testing H_0 : $\mu_x - \mu_y = c$ the test statistic is

$$Z = \frac{\bar{X} - \bar{Y} - c}{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}}.$$

For example[10] it is known that French people's heights (in cm) are normally distributed $N(\mu_f, 25)$. It is also known that German people's heights are normally distributed $N(\mu_g, 20)$. It is wished to test whether or not German people are taller than French people (at the $2\frac{1}{2}\%$ level). A random sample of 10 French people's heights are and their mean height recorded ($\bar{f}$). Similarly 8 German people's heights are taken and their mean recorded ($\bar{g}$).

1. State appropriate null and alternative hypotheses.
2. Find the set of values for $\bar{g} - \bar{f}$ for which we would reject the null hypothesis.
3. If in fact Germans are 7cm taller on average then find the probability of a Type II error.

1. $\text{H}_0 : \mu_g - \mu_f = 0$,
   $\text{H}_1 : \mu_g - \mu_f > 0$.
2. Given $Z = \frac{\bar{X}-\bar{Y}-c}{\sqrt{\frac{\sigma_x^2}{n_x}+\frac{\sigma_y^2}{n_y}}}$ we obtain

$$Z_{\text{crit}} = 1.960 = \frac{(\bar{G} - \bar{F})_{\text{crit}}}{\sqrt{\frac{25}{10} + \frac{20}{8}}}.$$

Therefore critical value is $(\bar{g} - \bar{f})_{\text{crit}} = 4.383$. We therefore reject the null hypothesis if $\bar{g} - \bar{f} \geqslant 4.383$.

[10] It's well worth thinking very hard about this example. It stumped me the first time I saw a similar question.

3. For a Type II error we must lie in the *acceptance region* of the original test given the new information. Here we require $\mathbb{P}(\bar{g} - \bar{f} < 4.383 \mid \mu_g - \mu_f = 7)$, so

$$\begin{aligned}\mathbb{P}(\bar{g} - \bar{f} < 4.383 \mid \mu_g - \mu_f = 7) &= P\left(Z < \frac{4.383 - 7}{\sqrt{\frac{25}{10} + \frac{20}{8}}}\right) \\ &= \mathbb{P}(Z < -1.170) \\ &= 1 - \mathbb{P}(Z < 1.170) \\ &= 1 - 0.8790 = 0.121\end{aligned}$$

- If X and Y are *not* normally distributed we need the samples to be *large* (then CLT applies). If the variances are *known* then the above is still correct. However if the population variances are unknown we replace the σ_x and σ_y by their estimators s_x and s_y.

  For example, Dr. Evil believes that people's attention spans are different in Japan and America. He samples 80 Japanese people and finds their attention spans are described (in minutes) $\sum j = 800$ and $\sum j^2 = 12000$. He samples 100 people in America and finds $\sum a = 850$ and $\sum a^2 = 11200$. Test at the 5% level whether Dr Evil is justified in his claim. So

  $\mathrm{H}_0 : \mu_j - \mu_a = 0.$

  $\mathrm{H}_1 : \mu_j - \mu_a \neq 0.$

  $\alpha = 5\%.$

  $\bar{j} = 10$, $\bar{a} = 8.5$.

  $s_j^2 = \frac{80}{79}\left(\frac{12000}{80} - 10^2\right) = 50.63.$

  $s_a^2 = \frac{100}{99}\left(\frac{11200}{100} - 8.5^2\right) = 40.15.$

  $Z_{\text{obs}} = \frac{\bar{X} - \bar{Y} - c}{\sqrt{\frac{s_j^2}{n_j} + \frac{s_a^2}{n_a}}} = \frac{10 - 8.5}{\sqrt{\frac{50.63}{80} + \frac{40.15}{100}}} = 1.475.$

  $Z_{\text{crit}} = \pm 1.960$. Therefore we reject if $|Z_{\text{obs}}| > 1.960$.

  $1.475 < 1.960$, so at the 5% level we have no reason to reject H_0 and conclude that Dr Evil is probably mistaken in his claim that the two countries have different attention spans.

- If X and Y are *normally* distributed with an *unknown,* ***common*** variance and we are testing $H_0 : \mu_x - \mu_y = c$ we use a two-sample t-test. The test statistic here is

$$T = \frac{\bar{X} - \bar{Y} - c}{\sqrt{s_p^2\left(\frac{1}{n_x} + \frac{1}{n_y}\right)}}.$$

  Here s_p^2 is the unbiased pooled estimate of the *common* variance, defined

$$s_p^2 = \frac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{n_x + n_y - 2}.$$

  Also $\nu = n_x + n_y - 2$. For example a scientist wishes to test whether new heart medication reduces blood pressure. 10 patients with high blood pressure were given the medication and their summary data is $\sum x = 1271$ and $\sum (x - \bar{x})^2 = 640.9$. 8 patients with high blood pressure were given a placebo and their summary data is $\sum y = 1036$ and $\sum (y - \bar{y})^2 = 222$. Carry out a hypothesis test at the 10% level to see if the medication is working.

  $\mathrm{H}_0 : \mu_x - \mu_y = 0.$

$H_1 : \mu_x - \mu_y < 0.$

$\alpha = 10\%.$

$\bar{x} = 127.1,\ \bar{y} = 129.5.$

$s_x^2 = \frac{10}{9}\left(\frac{640.9}{10}\right) = 71.2\dot{1}.$

$s_y^2 = \frac{8}{7}\left(\frac{222}{8}\right) = 31.71.$

$s_p^2 = \frac{9\times 71.2\dot{1}+7\times 31.71}{16} = 53.93.$

$T_{\text{obs}} = \frac{\bar{X}-\bar{Y}-c}{\sqrt{s_p^2\left(\frac{1}{n_x}+\frac{1}{n_y}\right)}} = \frac{127.1-129.5}{\sqrt{53.93\left(\frac{1}{10}+\frac{1}{8}\right)}} = -0.689.$

$\nu = 16$ so $T_{\text{crit}} = -1.337.$

$-0.689 > -1.337$, so at the 10% level we have no reason to reject H_0 and conclude that the medication is probably not lowering blood pressure.

- Also look for 'paired' data. This can only happen if $n_x = n_y$ *and* if every piece of data in x is somehow linked to a piece of data in y. Ask yourself 'would it matter if you changed the ordering of the x_i but not the y_i?' If yes, then paired. If the data is paired then you create a new set of data $d_i = x_i - y_i$.

  1. If the *population of differences* is distributed normally (or assumed to be distributed normally) then the test statistic for $H_0 : \mu_d = c$ is

$$T = \frac{\bar{D} - c}{\sqrt{\frac{s_d^2}{n}}} \qquad \text{with } \nu = n - 1.$$

     For example, Dwayne believes that his mystical crystals can boost IQs. He takes 10 students and records their IQs before and after they have been 'blessed' by the crystals. The results are

Victim	1	2	3	4	5	6	7	8	9	10
IQ Before	107	124	161	89	96	120	109	98	147	89
IQ After	108	124	159	100	101	119	110	101	146	94

     Test at the 5% level Dwayne's claim. The data is clearly paired and thus we create $d_i = IQ_{\text{after}} - IQ_{\text{before}}$ giving

$$1,\quad 0,\quad -2,\quad 11,\quad 5,\quad -1,\quad 1,\quad 3,\quad -1,\quad 5.$$

     $H_0 : \mu_d = 0,$

     $H_1 : \mu_d > 0.$

     $\alpha = 5\%$

     $\nu = 10 - 1 = 9.$

     $\bar{d} = \frac{22}{10} = 2.2$

     $s_d^2 = \frac{n}{n-1}\left(\frac{\sum d^2}{n} - \bar{d}^2\right) = \frac{10}{9}\left(\frac{188}{10} - 2.2^2\right) = 15.5\dot{1}.$

     $T_{\text{obs}} = \frac{\bar{D}-c}{\sqrt{\frac{s_d^2}{n}}} = 1.766.$

     $T_{\text{crit}} = 1.833$ (tables)

     $1.766 < 1.833$ therefore at the 5% level no reason to reject H_0 and conclude that the crystals probably don't significantly increase IQ.

2. If the *population of differences* is not distributed normally, but the sample size is large, then CLT applies and the test statistic for $H_0 : \mu_d = c$ is

$$Z = \frac{\bar{D} - c}{\sqrt{\frac{s_d^2}{n}}}.$$

- If testing for differences in population *proportions* there are two cases, each requiring *independent, large* samples (CLT).

1. For $H_0 : p_x = p_y$ (i.e. no difference in population proportions) the test statistic is

$$Z = \frac{P_{sx} - P_{sy}}{\sqrt{pq\left(\frac{1}{n_x} + \frac{1}{n_y}\right)}}.$$

Here p is the value of the *common* population proportion $p = \dfrac{x + y}{n_x + n_y}$. Also $p_{sx} = \frac{x}{n_x}$ and $p_{sy} = \frac{y}{n_y}$.

2. For $H_0 : p_x - p_y = c$ the test statistic is

$$Z = \frac{P_{sx} - P_{sy} - c}{\sqrt{\frac{P_{sx}Q_{sx}}{n_x} + \frac{P_{sy}Q_{sy}}{n_y}}}.$$

Here $q_{sy} = 1 - p_{sx}$ and $q_{sy} = 1 - p_{sy}$.

Confidence Intervals

- It has been described to me by someone I respect that a confidence interval is like an 'egg-cup' of a certain width that we throw down onto the number-line. Of all possible 'egg-cups' we want 90% (or some other percentage) of those egg cups to contain the true mean μ. This does not mean that a confidence interval has a 90% chance of containing the mean; it either contains the mean or it doesn't.

- A confidence interval is denoted $[a, b]$ which means $a < x < b$. In S3 we only consider symmetric confidence intervals about the sample mean (because $\bar{x}$ is an unbiased estimate of μ). They basically represent the acceptance region of a hypothesis test where $H_0 : \mu = \bar{x}$.

- To find the required z or t values in all of the following confidence intervals is easy. If you want (say) a 90% confidence interval then you (sort of) want to contain 90% of the data, so you must have 10% not contained which means that there must be 5% at each end of the distribution. Therefore you look up, either in the little table *beneath* the big normal table or in the correct line of the t table, 95%. This then gives you the z or t value to the left of which 95% of the data lies.

- This is fine for certain special values (90%, 95%, 99% etc.) and for the t-distribution this is all you can do. However for z values we can also do a 'reverse look-up' in the main normal tables to find more 'exotic' values. For example if I wanted a 78% confidence interval with z, then 11% would be in each end. Therefore I would reverse look-up 0.8900 *within* the main body of the table to find $z = 1.226$.

- If you are drawing from a normal *of known variance* σ^2 then the confidence interval will be

$$\left[\bar{x} - z\frac{\sigma}{\sqrt{n}}, \bar{x} + z\frac{\sigma}{\sqrt{n}}\right].$$

This result is true even for small sample sizes.

For example, an α% confidence interval is calculated from a normal population whose variance is known to be 9. The sample size is 16 and the confidence interval is $[19.68675, 22.31325]$. Find α. The midpoint of the interval is 21. Therefore the confidence interval is $[21 - z\frac{3}{\sqrt{16}}, 21 + z\frac{3}{\sqrt{16}}]$. We can then solve $21 + z\frac{3}{\sqrt{16}} = 22.31325$ to find $z = 1.751$. A forward lookup in the table reveals 0.96. Therefore there exists 4% at either end, so $\alpha = 8$; i.e. it is an 92% confidence interval.

- If you are drawing from a normal *of unknown variance* then the confidence interval will be

$$\left[\bar{x} - t\frac{s}{\sqrt{n}}, \bar{x} + t\frac{s}{\sqrt{n}}\right].$$

  The degrees of freedom here will be $\nu = n - 1$.

- If you are drawing from an unknown distribution then (provided $n > 30$ to invoke the CLT) then the confidence interval will be

$$\left[\bar{x} - z\frac{s}{\sqrt{n}}, \bar{x} + z\frac{s}{\sqrt{n}}\right].$$

- If, instead of means, we are taking a sample proportion then the confidence interval will be

$$\left[p_s - z\sqrt{\frac{p_s q_s}{n}}, p_s + z\sqrt{\frac{p_s q_s}{n}}\right].$$

- If instead of single samples we are looking for a confidence interval for the difference between two populations we use the following, depending on the situation.

  1. Difference in means being zero from two normals of *known* variances

$$\left[\bar{x} - \bar{y} - z\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}, \bar{x} - \bar{y} + z\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}\right].$$

     Or for difference in means $\bar{x} - \bar{y}$ being c,

$$\left[\bar{x} - \bar{y} - c - z\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}, \bar{x} - \bar{y} - c + z\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}\right].$$

     This can also be used for non-normal populations of known variance if the samples are *large* (CLT).

  2. The above can be altered if the samples are *large* (CLT) and the variances are not known to

$$\left[\bar{x} - \bar{y} - z\sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}, \bar{x} - \bar{y} + z\sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}\right].$$

  3. Difference in means being zero from two normals of *the same, unknown* variance

$$\left[\bar{x} - \bar{y} - ts_p\sqrt{\frac{1}{n_x} + \frac{1}{n_y}}, \bar{x} - \bar{y} + ts_p\sqrt{\frac{1}{n_x} + \frac{1}{n_y}}\right].$$

Or for difference in means $\bar{x} - \bar{y}$ being c,

$$\left[\bar{x} - \bar{y} - c - ts_p\sqrt{\frac{1}{n_x} + \frac{1}{n_y}}, \bar{x} - \bar{y} - c + ts_p\sqrt{\frac{1}{n_x} + \frac{1}{n_y}}\right].$$

Here s_p is the unbiased pooled estimate of the *common* variance $s_p^2 = \frac{(n_x-1)s_x^2+(n_y-1)s_y^2}{n_x+n_y-2}$. The degrees of freedom is $\nu = n_x + n_y - 2$.

4. If dealing with difference in population proportions we use

$$\left[p_{sx} - p_{sy} - z\sqrt{\frac{p_{sx}q_{sx}}{n_x} + \frac{p_{sy}q_{sy}}{n_y}}, p_{sx} - p_{sy} + z\sqrt{\frac{p_{sx}q_{sx}}{n_x} + \frac{p_{sy}q_{sy}}{n_y}}\right].$$

χ^2-Tests

- χ^2 tests measure how good data fits a given distribution. The test statistic here is

$$X^2 = \sum \frac{(O-E)^2}{E}.$$

Here O is the observed frequency and E the expected frequency. The larger X^2 becomes the more likely it is that the observed data does *not* come from the expected values that we have calculated.

- As with the t-distribution, the χ^2 distribution has a degree of freedom associated with it still denoted ν. This is calculated

$$\nu = \text{number of classes} - \text{number of constraints}.$$

- Given observed frequencies you need to calculate expected frequencies from theoretical probabilities. Expected frequencies are the expected probability times the total number of trials. The convention is that if an expected value is less than 5, then you combine with a larger expected value such that all values end up greater than 5. For example if you had

OBS	22	38	24	18	9	2	1	0
EXP	23.4	35.1	27.2	16.1	7.2	3.1	0.9	0.2

you would combine the final four columns to get

OBS	22	38	24	18	12
EXP	23.4	35.1	27.2	16.1	11.4

Because of this combining the total number of classes would be 5 and *not* 8.

- Fitting a Distribution
  - As with any hypothesis tests, the expected values are computed supposing that H_0 is correct. For example given the data

Outcome	0	1	2	3	4	5
Obs Frequency	22	37	23	10	6	2

test at the 5% level the hypotheses

H_0 : The data is well modelled by $B(5, \frac{1}{4})$,

H_1 : The data is not well modelled by $B(5, \frac{1}{4})$.

So, under H_0 we have $B(5, \frac{1}{4})$. We calculate the probabilities of the six outcomes from S1:

x	0	1	2	3	4	5
$\mathbb{P}(X = x)$	$\frac{243}{1024}$	$\frac{405}{1024}$	$\frac{135}{512}$	$\frac{45}{512}$	$\frac{15}{1024}$	$\frac{1}{1024}$

Then we note that the total number in the observed data is 100, so we multiply the expected probabilities by 100 to obtain expected frequencies (to 1dp).

Outcome	0	1	2	3	4	5
Exp Frequency	23.7	39.6	26.3	8.8	1.5	0.1

We see that the expected frequencies have dropped below five, so we combine the last 3 columns to obtain:

OBS	22	37	23	18
EXP	23.7	39.6	26.3	10.4

So $X^2 = \frac{2.89}{23.7} + \frac{6.76}{39.6} + \frac{10.89}{26.3} + \frac{57.76}{10.4} = 6.26$.

Now the only constraint here is the total observed frequencies of 100, so $\nu = 4 - 1 = 3$. In the tables we observe $\mathbb{P}(\chi^2_3 \leqslant 7.815) = 0.95$. Therefore the critical X^2 value is 7.815. So $6.26 < 7.815$ and we therefore have no reason to reject H_0 and conclude that $B(5, \frac{1}{4})$ is probably a good model for the data.

– Parameter Estimation. It is important to note that there is a difference in ν in the following situations:

  ⋆ H_0 : The data can be modelled by a Poisson distribution with $\lambda = 3.1$.

  ⋆ H_0 : The data can be modelled by a Poisson distribution.

The second has an extra constraint because you will need to estimate the value of λ from your observed data. In general just remember that if you estimate a parameter from observed data then this provides another constraint.

  ⋆ If you need to estimate p from a frequency table for testing the goodness of fit of a binomial distribution you calculate $\bar{x}$ from the data in the usual way and equate this with np because that is the expectation of a binomial. For example, estimate p from the following observed data:

x	0	1	2	3	4
Obs frequency	12	16	6	2	1

  So $np = \bar{x} = \frac{0\times12+1\times16+2\times6+3\times2+4\times1}{37} = \frac{38}{37}$. Therefore $p = \frac{38}{37\times4} = 0.257$ (to 3dp).

  ⋆ If you need to estimate λ from a frequency table for testing the goodness of fit of a Poisson distribution you calculate $\bar{x}$ from the data in the usual way and equate this with λ. The only potential difficulty lies in the fact that the Poisson distribution has an infinite number of outcomes $\{0, 1, 2, 3 \dots\}$. However, the examiners will take pity and give you a scenario such as

x	0	1	2	3	4 or more
Obs frequency	5	11	10	3	0

  where the "4 or more" frequency will be zero. Therefore $\lambda = \frac{0\times5+1\times11+2\times10+3\times3}{29} = 1.38$ (to 2dp).

  ⋆ Likewise the geometric distribution takes an infinite number of possible outcomes $\{1, 2, 3, 4 \dots\}$, and $\mathbb{E}(X) = \frac{1}{p}$, so to estimate p we calculate $\frac{1}{\mathbb{E}(X)}$. For example given

x	1	2	3	4	5 or more
Obs frequency	26	20	13	6	0

So, $\bar{x} = \frac{1\times26+2\times20+3\times13+4\times6}{65} = \frac{129}{65}$. Therefore $p = \frac{65}{129}$.

- For example for the following, test at the 1% level the following hypotheses:
  H_0 : The data is well modelled by a Poisson,
  H_1 : The data is not well modelled by a Poisson.

x	0	1	2	3	4	5or more
Obs frequency	14	23	14	7	2	0

So we estimate from the data (as above) $\lambda = \frac{4}{3}$. Now we calculate the first five expected values using total $\times\, \mathbb{P}(X = x) = e^{-\lambda}\frac{\lambda^x}{x!}$. The final total we calculate by 60 subtract the other five totals.

x	0	1	2	3	4	5or more
Exp frequency	15.8	21.1	14.1	6.2	2.1	0.7

.

So combining columns so that the expected values equal at least five we obtain.

OBS	14	23	14	9
EXP	15.8	21.1	14.1	9.0

Now $X^2 = 0.377$. $\nu = 4 - 2 = 2$ (2 constraints because of 60 total and estimation of λ).

From tables $\mathbb{P}(\chi_2^2 < 9.210) = 0.99$. $0.377 < 9.210$ and therefore at the 1% level we have no reason to reject H_0 and conclude that the data is probably well described by a Poisson.

- Contingency Tables
  - we are looking for *independence* (or, equivalently, dependence) between two variables. Remember that two events (A and B) are independent if $\mathbb{P}(A|B) = \mathbb{P}(A|B') = \mathbb{P}(A)$. Coupling this with the formula $\mathbb{P}(A|B) = \frac{\mathbb{P}(A\cap B)}{\mathbb{P}(B)}$ (which drops out easily from a Venn diagram with A and B overlapping) we discover that independence *implies* $\mathbb{P}(A)\times\mathbb{P}(B) = \mathbb{P}(A\cap B)$. Therefore given any contingency table showing observed values we wish to calculate the values that would be expected *if they were* independent. Then carry out the analysis as before.
  - For example 81 children are asked which of football, rugby or netball is their favourite.

OBS	Football	Rugby	Netball	Total
Boy	17	25	3	45
Girl	9	3	24	36
Total	26	28	27	81

Now, *if* the sex and choice of favourite were independent then $\mathbb{P}$(rugby and girl) = $\mathbb{P}$(rugby) $\times\, \mathbb{P}$(girl) = $\frac{28}{81}\times\frac{36}{81}$. Therefore the number of girls who like rugby best should be $81\times\frac{28}{81}\times\frac{36}{81}$. The 81 cancels to give an expected number of $\frac{28\times36}{81}$. This is an example of the general result

$$\text{expected number} = \frac{\text{column total}\times\text{row total}}{\text{grand total}}.$$

Therefore in our example we have

EXP	Football	Rugby	Netball	Total
Boy	$\frac{26\times45}{81} = 14\frac{4}{9}$	$\frac{28\times45}{81} = 15\frac{5}{9}$	$\frac{27\times45}{81} = 15$	45
Girl	$\frac{26\times36}{81} = 11\frac{5}{9}$	$\frac{28\times36}{81} = 12\frac{4}{9}$	$\frac{27\times36}{81} = 12$	36
Total	26	28	27	81

None of the expected values are less than 5, so no need to combine columns. Therefore $X^2 = \sum \frac{(O-E)^2}{E} = 35.52$ (to 2 dp). Make sure you can get my answer. A table often helps you build up to the answer. Use columns O, E, $(O-E)^2$, $\frac{(O-E)^2}{E}$.

– In an $m \times n$ contingency table the degrees of freedom is

$$\nu = (m-1)(n-1).$$

So in the above example $\nu = (3-1)\times(2-1) = 2$. So if we were to carry out a hypothesis test (at the 5% level) of

H_0 : The variables 'sex' and 'favourite sport' are independent;

H_1 : The variables 'sex' and 'favourite sport' are not independent.

We would use the correct row in the χ^2 tables to discover that $\mathbb{P}(\chi_2^2 > 5.991) = 0.05$. Now $35.52 > 5.991$ so we reject H_0 and conclude that 'sex' and 'favourite sport' are not independent.

– If you have a 2×2 contingency table you must apply Yates's correction. Here you reduce each value of $|O-E|$ by $\frac{1}{2}$. Again a table helps you build up to the answer. Use columns O, E, $|O-E|$, $\left(|O-E|-\frac{1}{2}\right)^2$, $\frac{(|O-E|-\frac{1}{2})^2}{E}$.

For example carry out a hypothesis test to see if hair colour and attractiveness are independent.

OBS	Blonde	Not blonde	Total
Fit	24	16	40
Minging	14	46	60
Total	38	62	100

Expected values are calculated as before.

EXP	Blonde	Not blonde	Total
Fit	$\frac{38\times40}{100} = 15.2$	$\frac{62\times40}{100} = 24.8$	40
Minging	$\frac{38\times60}{100} = 22.8$	$\frac{62\times60}{100} = 37.2$	60
Total	38	62	100

Therefore the table would be

O	E	$\lvert O-E\rvert$	$\left(\lvert O-E\rvert-\frac{1}{2}\right)^2$	$\frac{(\lvert O-E\rvert-\frac{1}{2})^2}{E}$
24	15.2	8.8	68.89	4.532
16	24.8	8.8	68.89	2.778
14	22.8	8.8	68.89	3.021
46	37.2	8.8	68.89	1.852
				12.183

$X^2 = 12.183$ and $\nu = 1$ and you use these values in any subsequent hypothesis test. (Note that X^2 is pretty high here and for any significance level in the tables we would reject the hypothesis that hair colour and fitness were independent. Blondes are hot.)

Preliminaries

- Your pure maths needs to be far stronger for S4 than in any other Statistics module.
- You must be strong on general binomial expansion from C4.

$$(1+x)^n = 1 + nx + \frac{n(n-1)}{2}x^2 + \frac{n(n-1)(n-2)}{3!}x^3 + \frac{n(n-1)(n-2)(n-3)}{2}x^4 + \cdots$$

This is valid only for $|x| < 1$. This is important for probability/moment generating functions.

- In particular you must be good at 'plucking out' specific coefficients (which may represent probabilities). For example find the x^8 coefficient in $\frac{x(3+x^2)}{\sqrt{4+2x}}$.

$$\begin{aligned}\frac{x(3+x^2)}{\sqrt{4+2x}} &= (3x+x^3)(4+2x)^{-\frac{1}{2}} \\ &= (3x+x^3)\left(4(1+\frac{x}{2})\right)^{-\frac{1}{2}} \\ &= \frac{1}{2}(3x+x^3)\left(1+\frac{x}{2}\right)^{-\frac{1}{2}} \\ &= \frac{1}{2}(3x+x^3)\left(1-\frac{x}{4}+\cdots-\frac{63}{8192}x^5+\cdots-\frac{429}{262,144}x^7+\ldots\right)\end{aligned}$$

  So the x^8 coefficient will be $-\frac{1}{2}(3\times\frac{429}{262,144}+1\times\frac{63}{8192}) = -\frac{3303}{524,288}$. It helps hugely to be thinking ahead about what coefficients you are going to need.

- Recall from S3 that $\mathbb{E}(g(X)) = \sum g(x_i)p_i$ for discrete random variables and $\mathbb{E}(g(X)) = \int_{-\infty}^{\infty} g(x)f(x)\,dx$ for continuous random variables.

- Recall also that $\text{Var}(X) \equiv \mathbb{E}(X^2) - (\mathbb{E}(X))^2$.

Probability

- There are three very useful ways of representing information in probability questions. Venn diagrams, tree diagrams and two-way tables. You must think hard about which approach is going to be most helpful in the question you are to answer. Read the whole question before you start!

- Set theory is very important in probability. Know the following
  - '$A \cap B$' is the intersection of the sets A and B. The overlap between the two sets. "AND"
  - '$A \cup B$' is the union of the sets A and B. Anything that lies in either A or B (or both). "OR"
  - A' means 'not A'. Everything outside A.
  - $\{\}$ (or $\emptyset$) denotes the empty set. For example $A \cap A' = \{\}$

- Events A and B are *mutually exclusive* if both A and B cannot both happen. Represented by a Venn diagram of non-overlapping circles. Here

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

- However in the general case where A and B are not mutually exclusive we have

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

  This is because we are overcounting the overlap. It is called the *addition law*.

  For three events the addition law becomes A, B and C we have (in general)

$$\mathbb{P}(A \cup B \cup C) = \mathbb{P}(A) + \mathbb{P}(B) + \mathbb{P}(C) - \mathbb{P}(A \cap B) - \mathbb{P}(A \cap C) - \mathbb{P}(B \cap C) + \mathbb{P}(A \cap B \cap C).$$

  Again this drops out easily from a Venn diagram.

- Events $A_1, A_2, \ldots$ are said to be exhaustive if $\mathbb{P}(A_1 \cup A_2 \cup \ldots) = 1$. In other words the events $A_1, A_2, \ldots$ contain all the possibilities.

- If A and B are *independent* events then

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

- We read $\mathbb{P}(A|B)$ as the probability of A given that B has occured. It is defined

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}.$$

  However this formula is not always easy to apply, so Mr Stone's patented 'collapsing universes' approach from a Venn or tree diagram is often more intuitive.

- Using $\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$ and $\mathbb{P}(B|A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)}$ we discover

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B|A) = \mathbb{P}(B)\mathbb{P}(A|B).$$

  This is called the *multiplication law* of probability and is incredibly useful in converting $\mathbb{P}(A|B)$ into $\mathbb{P}(B|A)$ and vice versa. The multiplication law drops out readily from a tree diagram.

- Bayes' Theorem[11] states

$$\mathbb{P}(A_j|B) = \frac{\mathbb{P}(A_j)\mathbb{P}(B|A_j)}{\mathbb{P}(B)} = \frac{\mathbb{P}(A_j)\mathbb{P}(B|A_j)}{\sum_{\text{all } i} \mathbb{P}(A_i)\mathbb{P}(B|A_i)}.$$

  This looks scary, but drops out from a tree diagram. The formal statement is not required for S4, but is very important.

Non-Parametric Tests

- All of the hypothesis tests studied in Stats 2 & 3 required knowledge (or at the very least an assumption) of some kind of underlying distribution for you to carry out the test. However sometimes you have no knowledge about the underlying population. Statisticians therefore developed a series of *non-parametric* tests for situations where you have no knowledge of the underlying population.

- The *sign test* is a test about the *median* (i.e. the point at which you have an equal number of data points either side). If H_0 : median = 10, say, then under H_0, whether a random piece of data lies above or below 10 has probability $\frac{1}{2}$. For n pieces of data we therefore have a binomial $B(n, \frac{1}{2})$. Rather than work out critical values, the best approach is probably to calculate (under H_0) the probability of what you have observed and anything more extreme. For example test at the 5% level whether the median of the data

$$1,\ 1,\ 2,\ 3,\ 6,\ 7,\ 8,\ 9,\ 9,\ 9,\ 10,\ 10,\ 11,\ 13$$

  is 5. Note that there are four pieces of data less than 5.

  H_0 : The median of the data is 5.

  H_1 : The median of the data is not 5.

  $\alpha = 5\%$. Two tailed test.

  Under H_0, $X \sim B(14, \frac{1}{2})$.

  $\mathbb{P}(X \leqslant 4) = 0.0898 > 0.025$, so at the 5% level there is insufficient evidence to reject H_0 and we conclude that the median of the data is probably 5. [You could have also gone through the rigmarole of demonstrating that the critical value is 2 (or 12) but my way is quicker and life's short.]

[11] Reverend Thomas Bayes from my home town of Tunbridge Wells. Wrote a document defending Newton's calculus hence a rather good bloke.

- Although there is no example in your textbook I see no reason why they couldn't ask a question where you had a large enough sample to require the normal approximation to $B(n, \frac{1}{2})$…don't forget your continuity correction.
- The sign test is a very crude test because it takes absolutely no account of how far away the data lies on either side of the median. If you want to take account of the magnitude of the deviations you need to use…
- …the *Wilcoxon signed-rank test*. Here it is assumed that the data is *symmetric*; therefore it is a test about both the median or the mean because for symmetric data the median and mean are the same.

  You calculate the deviations from the median/mean, rank the size of the deviations and then sum the positive ranks to get P and sum the negative ranks to get Q. The test statistic is T, where T is the smaller of P or Q. For example test at the 5% level whether the mean of

  $$1.3,\ 2.1,\ 7.3,\ 4.9,\ 3.2,\ 1.6,\ 5.6,\ 5.7$$

  is 3.

  The data sort of looks symmetric, so OK to proceed with Wilcoxon.

  H_0 : The mean of the data is 3.

  H_1 : The mean of the data is greater than 3.

  $\alpha = 5\%$. One tailed test.

Data	1.3	2.1	7.3	4.9	3.2	1.6	5.6	5.7
Deviation	−1.7	−0.9	+4.3	+1.9	+0.2	−1.4	+2.6	+2.7
Rank	4	2	8	5	1	3	6	7
Signed Rank	−4	−2	+8	+5	+1	−3	+6	+7

  So $P = 27$, $Q = 9$, so $T_{\text{obs}} = 9$. The lower T is, the worse it is for H_0 and the tables give the *largest* value at which you would reject H_0. $T_{\text{crit}} = 5$. $9 > 5$, so at the 5% level we have insufficient evidence to reject H_0 and conclude that the mean is probably 3.
- For large samples (i.e. when the tables don't give the values you want; running out of values) a normal approximation can be used where

  $$Z = \frac{T + 0.5 - \frac{1}{4}n(n+1)}{\sqrt{\frac{1}{24}n(n+1)(2n+1)}}.$$

  Note that because T is the smaller of P and Q that Z will always be negative (both Z_{crit} and Z_{obs}). For example if you had 100 pieces of data and you were testing at the 1% level whether the mean was some value (against H_1 of the mean not being some value) and $P = 2000$ and $Q = 3050$ then $T = 2000$. So

  $$\begin{aligned} Z_{\text{obs}} &= \frac{T_{\text{obs}} + 0.5 - \frac{1}{4}n(n+1)}{\sqrt{\frac{1}{24}n(n+1)(2n+1)}} \\ &= \frac{2000 + 0.5 - \frac{1}{4} \times 100 \times 101}{\sqrt{\frac{1}{24} \times 100 \times 101 \times 201}} \\ &= -1.803 \end{aligned}$$

  Because it is a two-tailed 1% test we reverse look-up 0.995 to obtain $Z_{\text{crit}} = -2.576$. Finally $-1.803 > -2.576$, so at the 1% level there is insufficient evidence to reject H_0 and conclude that the mean is probably whatever we thought it was under H_0.

- The *Wilcoxon rank-sum test* is the non-parametric equivalent of the two-sample t-test from S3. It tests whether two different sets of data are drawn from identical populations. The central idea for the theory is that if X and Y are drawn from identical distributions, then $P(X < Y) = \frac{1}{2}$. The tables are then constructed from tedious consideration of all the possible arrangements of the ranks (called the 'sampling distribution').

  Given two sets of data, let m be the number of pieces of data from the smaller data set and n be the number of pieces of data from the larger data set (if they are both the same size it's up to you which is m and which n). Then rank *all* the data and sum the ranks of the 'm' population; call this total R_m. Also calculate $m(n+m+1) - R_m$ and let the test statistic W be the smaller of R_m and $m(n+m+1) - R_m$. The smaller W is, the more likely we are to reject H_0 and the tables give the largest W at which we reject H_0.

  For example test at the 5% level whether the following are drawn from identical populations.

A	23	14	42	12	30	40
B	16	21	9	35		

  H_0 : Data drawn from identical distributions.

  H_1 : Data not drawn from identical distributions.

  $\alpha = 5\%$. Two tailed test.

Data	9	12	14	16	21	23	30	35	40	42
Rank	1	2	3	4	5	6	7	8	9	10

  So $m = 4$, $n = 6$, $R_m = 18$, $m(n+m+1) - R_m = 26$, $W_{\text{obs}} = 18$. Looking at the tables we see $W_{\text{crit}} = 12$, and $18 > 12$, so at the 5% level there is insufficient evidence to reject H_0 and we conclude that the data is probably drawn from identical distributions.

- For large samples (i.e. when the tables don't give the values you want; running out of values) a normal approximation can be used where

$$Z = \frac{W + 0.5 - \frac{1}{2}m(m+n+1)}{\sqrt{\frac{1}{12}mn(m+n+1)}}.$$

Probability Generating Functions

- In Stats 1 & 2 you met discrete random variables (DRVs) such that each outcome had a probability attached. Sometimes there were rules which related the probability to the outcome (binomial, geometric, Poisson). However, in general we had:

x	x_1	x_2	x_3	x_4	$\dots$
$\mathbb{P}(X = x)$	p_1	p_2	p_3	p_4	$\dots$

  Recall that $\sum p_i = 1$ because the sum of all the probabilities must total 1 and that $\mathbb{E}(X) = \sum p_i x_i$. Also $\mathbb{E}(f(X)) = \sum p_i f(x_i)$ from Stats 3 and $\text{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2 = \sum p_i x_i^2 - (\sum p_i x_i)^2$ from Stats 2.

- At some point some bright spark decided to consider the properties of

$$G_X(t) = \mathbb{E}(t^X) = \sum p_i t^{x_i} = p_1 t^{x_1} + p_2 t^{x_2} + p_3 t^{x_3} + p_4 t^{x_4} + \cdots$$

where t is a 'dummy variable' unrelated to x. You can see that this will create either a finite or infinite series. This is called the probability generating function of X. It is a single function that contains within it all of the (potentially infinite) probabilities of X.

For example given

x	-2	-1	0	1	2
$\mathbb{P}(X=x)$	$\frac{1}{6}$	$\frac{1}{4}$	$\frac{1}{3}$	$\frac{1}{8}$	$\frac{1}{8}$

the generating function is $G_X(t) = p_1t^{x_1} + p_2t^{x_2} + p_3t^{x_3} + p_4t^{x_4} + \cdots = \frac{1}{6}t^{-2} + \frac{1}{4}t^{-1} + \frac{1}{3} + \frac{1}{8}t + \frac{1}{8}t^2$. We can therefore see that if (say) we saw a term $\frac{5}{24}t^6$, then we can see that $\mathbb{P}(X=6) = \frac{5}{24}$. Note that if you see a constant term then that tells you $\mathbb{P}(X=0)$ because $t^0 = 1$.

- An important property is that $G_X(1) = 1$ because $G_X(1)$ is just the sum of all the probabilities of X, i.e. $\sum p_i$.
- Another useful thing to do is consider the derivative $G'_X(t)$ with respect to t;

$$G'_X(t) = \sum p_i x_i t^{x_i - 1} = p_1x_1t^{x_1-1} + p_2x_2t^{x_2-1} + p_3x_3t^{x_3-1} + \cdots$$

Again, if we consider $G'(1)$ we obtain

$$G'(1) = \sum p_i x_i = p_1x_1 + p_2x_2 + p_2x_2 + p_3x_3 + \cdots = \mathbb{E}(X).$$

- Variances can also be calculated by

$$\text{Var}(X) = G''_X(1) + G'_X(1) - (G'_X(1))^2.$$

- Some standard pgfs are given in the formula book:

Distribution	$B(n,p)$	$Po(\lambda)$	$Geo(p)$
pgf	$(1-p+pt)^n$	$e^{\lambda(t-1)}$	$\frac{pt}{1-(1-p)t}$

Any good candidate should be able to derive these…

- For two *independent* random variables X and Y (with pgfs $G_X(t)$ and $G_Y(t)$ respectively) the pgf of $X+Y$ is $G_{X+Y}(t) = G_X(t) \times G_Y(t)$. This extends to three or more *independent* random variables.

Moment Generating Functions

- You will recall from FP2 that the Maclaurin expansion for e^x is

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \frac{x^5}{5!} + \cdots.$$

This is valid for all values of x (and you should know why from your Pure teachings). An alternative notation used is $e^x \equiv \exp(x)$.

- The nth moment of a distribution is $\mathbb{E}(X^n)$. So the first moment is just $\mathbb{E}(X)$. The second moment is $\mathbb{E}(X^2)$, which is useful in calculating variances. The zeroth moment is $\mathbb{E}(X^0) = \mathbb{E}(1) = 1$.

- The moment generating function (mgf) is defined for

x	x_1	x_2	x_3	x_4	$\ldots$
$\mathbb{P}(X=x)$	p_1	p_2	p_3	p_4	$\ldots$

by

$$\begin{aligned}
M_X(t) = \mathbb{E}(e^{tX}) = \sum p_i e^{x_i t} &= p_1 e^{x_1 t} + p_2 e^{x_2 t} + p_3 e^{x_3 t} + p_4 e^{x_4 t} + \cdots \\
&= \quad p_1 + p_1 x_1 t + p_1 \frac{x_1^2 t^2}{2!} + p_1 \frac{x_1^2 t^3}{3!} + \cdots \\
&\quad + p_2 + p_2 x_2 t + p_2 \frac{x_2^2 t^2}{2!} + p_2 \frac{x_2^2 t^3}{3!} + \cdots \\
&\quad + p_3 + p_3 x_3 t + p_3 \frac{x_3^2 t^2}{2!} + p_3 \frac{x_3^2 t^3}{3!} + \cdots \\
&\quad + p_4 + p_4 x_2 t + p_4 \frac{x_4^2 t^2}{2!} + p_4 \frac{x_4^2 t^3}{3!} + \cdots \\
&= \quad (p_1 + p_2 + p_3 + p_4 + \cdots) \\
&\quad + (p_1 x_1 + p_2 x_2 + p_3 x_3 + p_4 x_4 + \cdots)\, t \\
&\quad + \left(p_1 x_1^2 + p_2 x_2^2 + p_3 x_3^2 + p_4 x_4^2 + \cdots\right) \frac{t^2}{2!} \\
&\quad + \left(p_1 x_1^3 + p_2 x_2^3 + p_3 x_3^3 + p_4 x_4^3 + \cdots\right) \frac{t^3}{3!} \\
&\quad + \cdots \\
&= \mathbb{E}(1) + \mathbb{E}(X)t + \mathbb{E}\left(X^2\right) \frac{t^2}{2!} + \mathbb{E}\left(X^3\right) \frac{t^3}{3!} + E\left(X^4\right) \frac{t^4}{4!} + \cdots
\end{aligned}$$

So you can see that the constant term of $M_X(t)$ should always be $\mathbb{E}(1) = 1$ because it represents the sum of the probabilities. The coefficient of t will be $\mathbb{E}(X)$ and the coefficient of $\frac{t^2}{2!}$ (*not* just the coefficient of t^2) will be $\mathbb{E}(X^2)$. In general the coefficient of $\frac{t^n}{n!}$ will be $\mathbb{E}(X^n)$, that is, the nth moment.

- As with pgfs, differentiating mgfs (with respect to t) is a 'good thing'. However, instead of letting $t = 1$ we let $t = 0$ (because $a^0 = 1$). So differentiating $M_X(t)$ we find:

$$\begin{aligned}
M_X(t) &= p_1 e^{x_1 t} + p_2 e^{x_2 t} + p_3 e^{x_3 t} + p_4 e^{x_4 t} + \cdots \\
M_X'(t) &= p_1 x_1 e^{x_1 t} + p_2 x_2 e^{x_2 t} + p_3 x_3 e^{x_3 t} + p_4 x_4 e^{x_4 t} + \cdots \\
M_X'(0) &= p_1 x_1 + p_2 x_2 + p_3 x_3 + p_4 x_4 + \cdots \\
&= \sum x_i p_i = \mathbb{E}(X).
\end{aligned}$$

So $M_X'(0) = \mathbb{E}(X)$.

Differentiating again we find:

$$\begin{aligned}
M_X'(t) &= p_1 x_1 e^{x_1 t} + p_2 x_2 e^{x_2 t} + p_3 x_3 e^{x_3 t} + p_4 x_4 e^{x_4 t} + \cdots \\
M_X''(t) &= p_1 x_1^2 e^{x_1 t} + p_2 x_2^2 e^{x_2 t} + p_3 x_3^2 e^{x_3 t} + p_4 x_4^2 e^{x_4 t} + \cdots \\
M_X''(0) &= p_1 x_1^2 + p_2 x_2^2 + p_3 x_3^2 + p_4 x_4^2 + \cdots \\
&= \sum x_i^2 p_i = \mathbb{E}(X^2).
\end{aligned}$$

So using $\mathrm{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2$ we find $\mathrm{Var}(X) = M_X''(0) - (M_X'(0))^2$.

- Notice that with mgfs there are two ways to obtain the expectation and variance of your random variable. All things being equal I would choose the differentiation method, but you must ensure that your mgf is defined for $t = 0$. Also read the question carefully to see what they are wanting.

- Moment generating functions can also be defined for continuous random variables:

$$M_X(t) = \int_{-\infty}^{\infty} f(x)e^{tx}\,dx.$$

As before $M_X(0) = 1$, $M_X'(0) = \mathbb{E}(X)$, $M_X''(0) = \mathbb{E}(X^2)$. Convergence issues can arise EXAMPLE!!!!!!

- Some standard mgfs are given in the formula book:

Distribution	Uniform on $[a, b]$	Exponential	$N(\mu, \sigma^2)$
mgf	$\frac{e^{bt}-e^{at}}{(b-a)t}$	$\frac{\lambda}{\lambda-t}$	$e^{\mu t+\frac{1}{2}\sigma^2 t^2}$

Any good candidate should be able to derive these too…

- As with pgfs, for two *independent* random variables X and Y (with mgfs $G_X(t)$ and $G_Y(t)$ respectively) the mgf of $X + Y$ is $M_{X+Y}(t) = M_X(t) \times M_Y(t)$. This extends to three or more *independent* random variables.

Estimators

- It is vital to recall here that $\mathbb{E}(X) = \mu$ and $\mathrm{Var}(X) = \sigma^2$ (by definition).

- Given a population there may be many parameters that we may wish to know. For example we might like to know the mean μ, the variance σ^2, the median M, the maximum or minimum, the IQR, etc. In general we shall call this parameter θ.

Usually we will never know θ because we won't have the whole population. But we will be able to take a random sample from the population. From this sample we can calculate a quantity U which we shall use to estimate θ. We call U an estimator of θ.

- U is said to be an *unbiased estimator* of θ if

$$\mathbb{E}(U) = \theta.$$

i.e. if we take an average of *all possible* U (remember that U is a random variable) we will get the desired θ. If $\mathbb{E}(U) \neq \theta$ then the estimator is said to be biased (not giving the desired result on average).

For example to show that $K = \frac{X_1+2X_2+5X_3}{8}$ is an unbiased estimator of μ we merely consider $\mathbb{E}(K)$ and keep whittling down as far as we can go (using S3 expectation and variance algebra)

$$\begin{aligned}\mathbb{E}(K) &= \mathbb{E}\left(\frac{X_1 + 2X_2 + 5X_3}{8}\right)\\ &= \mathbb{E}\left(\frac{X_1}{8} + \frac{X_2}{4} + \frac{5X_3}{8}\right)\\ &= \frac{1}{8}\mathbb{E}(X_1) + \frac{1}{4}\mathbb{E}(X_2) + \frac{5}{8}\mathbb{E}(X_3)\\ &= \frac{1}{8}\mu + \frac{1}{4}\mu + \frac{5}{8}\mu = \mu.\end{aligned}$$

- For continuous random variables just remember that $\mathbb{E}(X) = \int_{-\infty}^{\infty} x f(x)\,dx$. For example find the value of k which makes $L = k(X_1 + X_2)$ an unbiased estimator of θ for

$$f(x) = \begin{cases} \frac{2}{\theta}\left(1 - \frac{x}{\theta}\right) & 0 \leqslant x \leqslant \theta \\ 0 & \text{otherwise} \end{cases}$$

First calculate $\mathbb{E}(X)$ *à la* S2:

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x f(x)\,dx = \int_0^{\theta} x\,\frac{2}{\theta}\left(1-\frac{x}{\theta}\right)\,dx = \left[\frac{x^2}{\theta} - \frac{2x^3}{3\theta^2}\right]_0^{\theta} = \frac{\theta}{3}.$$

So for L to be unbiased we need $\mathbb{E}(L) = \theta$, so

$$\begin{aligned}
\mathbb{E}(L) &= \theta \\
\mathbb{E}(k(X_1 + X_2) &= \theta \\
k(\mathbb{E}(X_1) + \mathbb{E}(X_2)) &= \theta \\
k(2 \times \mathbb{E}(X)) &= \theta \\
k\left(\frac{2\theta}{3}\right) &= \theta \\
k &= \frac{3}{2}.
\end{aligned}$$

- Given two *unbiased* estimators the *most efficient* estimator (of the two) is the one where $\mathrm{Var}(U)$ is smaller. A smaller variance is a 'good thing'.
- Sometimes you may need calculus to work out the most efficient estimator from an infinite family. For example X_1, X_2 and X_3 are three independent measurements of X.

$$S = \frac{aX_1 + 2X_2 + 4X_3}{a+6} \quad (\text{with } a \neq -6)$$

is suggested as an estimator for μ. Prove that S is unbiased whatever the value of a and find the value of a which makes S most efficient. So

$$\begin{aligned}
\mathbb{E}(S) &= \mathbb{E}\left(\frac{aX_1 + 2X_2 + 4X_3}{a+6}\right) \\
&= \frac{1}{a+6}\mathbb{E}(aX_1 + 2X_2 + 4X_3) \\
&= \frac{1}{a+6}[a\mathbb{E}(X_1) + 2\mathbb{E}(X_2) + 4\mathbb{E}(X_3)] \\
&= \frac{1}{a+6}[a\mu + 2\mu + 4\mu] \\
&= \frac{\mu}{a+6}(a+6) = \mu.
\end{aligned}$$

So S is unbiased for all values of a. Now consider

$$\begin{aligned}
\mathrm{Var}(S) &= \mathrm{Var}\left(\frac{aX_1 + 2X_2 + 4X_3}{a+6}\right) \\
&= \frac{1}{(a+6)^2}\mathrm{Var}(aX_1 + 2X_2 + 4X_3) \\
&= \frac{1}{(a+6)^2}[a^2\mathrm{Var}(X_1) + 4\mathrm{Var}(X_2) + 16\mathrm{Var}(X_3)] \\
&= \frac{a^2+20}{(a+6)^2}\sigma^2.
\end{aligned}$$

To minimise $\mathrm{Var}(S)$ we need $\frac{d}{da}\mathrm{Var}(S) = 0$. So

$$\begin{aligned}
0 &= \frac{d}{da}\left(\frac{a^2+20}{(a+6)^2}\sigma^2\right) = \frac{2a(a+6)^2 - 2(a+6)(a^2+20)}{(a+6)^4}\sigma^2 \\
\text{So } 0 &= 2a(a+6)^2 - 2(a+6)(a^2+20) \\
0 &= 2(a+6)[a(a+6) - (a^2+20)] \\
0 &= (a+6)(6a-20).
\end{aligned}$$

So $a = -6$ or $a = \frac{10}{3}$, but $a \neq -6$ so $a = \frac{10}{3}$ is the value of a that makes S most efficient[12].

- Here is a tough type of problem that caught me out the first two (or three (or four (…))) times I saw it. Slot away the method just in case. For example consider

$$f(x) = \begin{cases} \frac{2x}{\theta^2} & 0 \leqslant x \leqslant \theta \\ 0 & \text{otherwise} \end{cases}$$

An estimate of θ is required and a suggestion is made to calculate $\frac{5L}{4}$ where L is the maximum of two independent observations of X (X_1 and X_2). Show that this estimator is unbiased.

The thing to remember is that for L to be the maximum of X_1 and X_2, then X_1 *and* X_2 must *both* be less than or equal to L; i.e. we are going to calculate a cdf. So

$$\mathbb{P}(L \leqslant l) = \mathbb{P}(X_1 \leqslant l) \times \mathbb{P}(X_2 \leqslant l).$$

(This can be extended to three or more independent samplings of X.)

By sketching $f(x)$ we can see that the probability that one observation is less than or equal to l is given by a triangle in this case of area $\frac{l^2}{\theta^2}$ (or by the integral $\int_0^l f(x)\,dx$ for a more general $f(x)$). So $\mathbb{P}(L \leqslant l) = \mathbb{P}(X_1 \leqslant l) \times \mathbb{P}(X_2 \leqslant l) = \frac{l^2}{\theta^2} \times \frac{l^2}{\theta^2} = \frac{l^4}{\theta^4}$. Differentiating wrt to l we find the pdf of l to be

$$f(l) = \begin{cases} \frac{4l^3}{\theta^4} & 0 \leqslant l \leqslant \theta \\ 0 & \text{otherwise} \end{cases}$$

Therefore we calculate $\mathbb{E}\left(\frac{5L}{4}\right)$ as follows:

$$\begin{aligned} \mathbb{E}\left(\frac{5L}{4}\right) &= \frac{5}{4}\mathbb{E}(L) \\ &= \frac{5}{4}\int_0^\theta l \times \frac{4l^3}{\theta^4}\,dl \\ &= \frac{5}{4}\left[\frac{4l^5}{5\theta^4}\right]_0^\theta = \theta. \end{aligned}$$

Therefore $\frac{5L}{4}$ is an unbiased estimator of θ. I will leave it as an exercise for the reader to demonstrate that $\text{Var}\left(\frac{5L}{4}\right) = \frac{\theta^2}{24}$.

Discrete Bivariate Distributions

- The discrete random variables you have met thus far have been in one variable only. For example

x	2	3	5	7
$\mathbb{P}(X = x)$	$\frac{1}{2}$	$\frac{1}{9}$	$\frac{1}{3}$	$\frac{1}{18}$

However we can have discrete *bivariate* distributions. For example

		X		
		2	3	5
Y	4	0	$\frac{1}{2}$	$\frac{1}{10}$
	5	$\frac{1}{5}$	$\frac{3}{20}$	$\frac{1}{20}$

[12] I suppose we should consider the second derivative to show that this value of a minimises rather than maximises the variance, but life's too short…

From this we can see, say, $\mathbb{P}(X = 3, Y = 5) = \frac{3}{20}$.

- The marginal distribution is what one obtains if one of the variables is 'ignored'. In the above example the marginal distribution of X can be written

x	2	3	5
$\mathbb{P}(X = x)$	$\frac{1}{5}$	$\frac{13}{20}$	$\frac{3}{20}$

This can be added to the bivariate distribution thus:

		X		
		2	3	5
Y	4	0	$\frac{1}{2}$	$\frac{1}{10}$
	5	$\frac{1}{5}$	$\frac{3}{20}$	$\frac{1}{20}$
		$\frac{1}{5}$	$\frac{13}{20}$	$\frac{3}{20}$

$\mathbb{E}(X)$ and $\mathrm{Var}(X)$ can be calculated in the usual way obtaining $\mathbb{E}(X) = \frac{31}{10}$ and $\mathrm{Var}(X) = \frac{79}{100}$ (do it!). Similarly you can work out the marginal distribution of Y if you are so inclined.

- The *conditional* distribution of a bivariate distribution can be calculated *given that* one of the variables (X or Y) has taken a specific value. For the above example the "distribution of X conditional on $Y = 4$" is calculated by rewriting the 4 row with all the values divided by $\mathbb{P}(Y = 4) = \frac{3}{5}$.

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

This is all from our friend $\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$.

- A way to check whether X and Y are *independent* of each other in a bivariate distribution is to check whether every entry in the distribution is the product of the two relevant marginal probabilities. For example

		X			
		1	2	3	
Y	1	$\frac{1}{3}$	$\frac{2}{9}$	$\frac{1}{9}$	$\frac{2}{3}$
	2	$\frac{1}{6}$	$\frac{1}{9}$	$\frac{1}{18}$	$\frac{1}{3}$
		$\frac{1}{2}$	$\frac{1}{3}$	$\frac{1}{6}$	

Here we see $\mathbb{P}(X = 2, Y = 1) = \frac{2}{9}$ is the same as $\mathbb{P}(X = 2) \times \mathbb{P}(Y = 1) = \frac{1}{3} \times \frac{2}{3} = \frac{2}{9}$. The same is true for *every* entry in the table, so X and Y are independent. It only takes one entry not to satisfy this to ensure X and Y are *not* independent.

- The *covariance* of a discrete bivariate distribution is defined

$$\mathrm{Cov}(X, Y) \equiv \mathbb{E}((X - \mu_X)(Y - \mu_Y)).$$

However this tends to be cumbersome to calculate so we use the equivalent formula

$$\mathrm{Cov}(X, Y) = \mathbb{E}(XY) - \mu_X \mu_Y.$$

The covariance can be thought of as the correlation coefficient (r from Stats 1) for two probability distributions (sort of). The covariance can be both positive or negative (like the correlation coefficient).

- To calculate the covariance, first create the marginal distributions:

		X		
		1	3	4
Y	2	$\frac{1}{3}$	$\frac{1}{4}$	0
	5	$\frac{1}{6}$	$\frac{1}{8}$	$\frac{1}{8}$

$\Rightarrow$

		X			
		1	3	4	
Y	2	$\frac{1}{3}$	$\frac{1}{4}$	0	$\frac{7}{12}$
	5	$\frac{1}{6}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{5}{12}$
		$\frac{1}{2}$	$\frac{3}{8}$	$\frac{1}{8}$	

Then use the marginal distributions to calculate μ_X and μ_Y.

$\mu_X = \mathbb{E}(X) = \sum xp = 1 \times \frac{1}{2} + 3 \times \frac{3}{8} + 4 \times \frac{1}{8} = \frac{17}{8}$.

$\mu_Y = \mathbb{E}(Y) = \sum yp = 2 \times \frac{7}{12} + 5 \times \frac{5}{12} = \frac{13}{4}$.

Now we use this to calculate the covariance thus:

$$\begin{aligned}\text{Cov}(X,Y) &= \mathbb{E}(XY) - \mu_X\mu_Y \\ &= (1\times 2\times\tfrac{1}{3}) + (1\times 5\times\tfrac{1}{6}) + (3\times 2\times\tfrac{1}{4}) + (3\times 5\times\tfrac{1}{8}) + (4\times 2\times 0) + (4\times 5\times\tfrac{1}{8}) - \tfrac{17}{8}\times\tfrac{13}{4} \\ &= \frac{15}{32}.\end{aligned}$$

- If X and Y are independent then $\text{Cov}(X,Y) = 0$. However, if $\text{Cov}(X,Y) = 0$ this does not *necessarily* mean that X and Y are independent. But if $\text{Cov}(X,Y) \neq 0$ then X and Y cannot be independent.

- With an understanding of covariance we can write the relationship for $\text{Var}(aX \pm bY)$ when X and Y are *not* independent:

$$\text{Var}(aX \pm bY) = a^2\text{Var}(X) + b^2\text{Var}(Y) \pm 2ab\text{Cov}(X,Y).$$

Notice the extra term at the end of the formula we are used to from S3 for *independent* X and Y.