OCR Statistics 3 Module Revision Sheet

The S3 exam is 1 hour 30 minutes long. You are allowed a graphics calculator.

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

$\mathscr{J}.\mathscr{M}.\mathscr{S}.$

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[1] $k = \frac{2}{7}$. 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[2] *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.

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

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

- 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 $\frac{b}{a} < \frac{d}{c}$.

- In S2 expectation for a pdf $f(x)$ is $\mathbb{E}(X) = \int_{-\infty}^{\infty} xf(x)\,dx$. 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,

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

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[3]. 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 $\text{H}_0 : \mu = c$ 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 where you know σ^2 then (by CLT) $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ approximately.

  The test statistic for $\text{H}_0 : \mu = c$ 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 $\text{H}_0 : \mu = c$ 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[4].

  The test statistic for $\text{H}_0 : \mu = c$ is $T = \dfrac{\bar{X} - c}{\sqrt{\frac{s^2}{n}}}$.

[3] 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 $a = b = 1$.

[4] 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.

- (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 $\mu =$ the population mean radii of the disks.

  $\text{H}_0 : \mu = 5$,

  $\text{H}_1 : \mu < 5$.

  $n = 9$, 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 $\mathrm{H}_0 : \mu_x - \mu_y = c$ 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[5] 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. $\mathrm{H}_0 : \mu_g - \mu_f = 0$,
   $\mathrm{H}_1 : \mu_g - \mu_f > 0$.
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 $\bar{g} - \bar{f} \geqslant 4.383$.
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

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

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

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

$\alpha = 5\%$.

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

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

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

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

$Z_{\text{crit}} = \pm 1.960$. 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.

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

  $\text{H}_1 : \mu_x - \mu_y < 0$.

  $\alpha = 10\%$.

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

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

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

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

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

  $\nu = 16$ 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

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

$\mathrm{H}_0 : \mu_d = 0,$
$\mathrm{H}_1 : \mu_d > 0.$
$\alpha = 5\%$
$\nu = 10 - 1 = 9.$
$\bar{d} = \frac{22}{10} = 2.2$
$s_d^2 = \frac{n}{n-1}\left(\frac{\sum d^2}{n} - \bar{d}^2\right) = \frac{10}{9}\left(\frac{188}{10} - 2.2^2\right) = 15.5\dot{1}.$
$T_{\text{obs}} = \frac{\bar{D}-c}{\sqrt{\frac{s_d^2}{n}}} = 1.766.$
$T_{\text{crit}} = 1.833$ (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 $\mathbb{P}(\text{rugby and girl}) = \mathbb{P}(\text{rugby}) \times \mathbb{P}(\text{girl}) = \frac{28}{81} \times \frac{36}{81}$. 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 $\frac{28\times 36}{81}$. 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	$\frac{26\times 45}{81} = 14\frac{4}{9}$	$\frac{28\times 45}{81} = 15\frac{5}{9}$	$\frac{27\times 45}{81} = 15$	45
Girl	$\frac{26\times 36}{81} = 11\frac{5}{9}$	$\frac{28\times 36}{81} = 12\frac{4}{9}$	$\frac{27\times 36}{81} = 12$	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.)