OCR Statistics 2 Module Revision Sheet

The S2 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.

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

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[1]) 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 $X \sim N(\mu, \sigma^2)$.

- 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 $\Phi(\text{negative value})$ 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 $\dagger$. 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 $\dagger$ 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 $\dagger$ we obtain, $0.842 = (20 - \mu)/5$, so $\mu = 15.79$.

---

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

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

- For a Poisson distribution $\mathbb{E}(X) = \text{Var}(X) = \lambda$. We write $X \sim \text{Po}(\lambda)$.

- 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 $W \sim \text{Po}(np)$. [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 $W \sim \text{Po}(2)$. 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 $X \sim \text{Po}(\lambda)$ 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 $X \sim \text{Po}(50)$ 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

  $$H_0 \;:\; \mu = 6.8$$
  $$H_1 \;:\; \mu \neq 6.8.$$

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

$$Z_{\text{crit}} = \frac{\bar{X}_{\text{crit}} - \mu}{\sqrt{\frac{\sigma^2}{n}}}$$

$$1.960 = \frac{\bar{X}_{\text{crit}} - 6.8}{\sqrt{\frac{0.18^2}{10}}}$$

$$\text{and} \quad -1.960 = \frac{\bar{X}_{\text{crit}} - 6.8}{\sqrt{\frac{0.18^2}{10}}}$$

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)

$$\begin{aligned} H_0 &: p = \tfrac{5}{6} \\ H_1 &: p < \tfrac{5}{6}. \end{aligned}$$

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

$$\begin{aligned} H_0 &: p = \tfrac{1}{6} \\ H_1 &: p \neq \tfrac{1}{6}. \end{aligned}$$

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

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

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:

$$
\begin{aligned}
\text{H}_0 &: \lambda = 7 \\
\text{H}_1 &: \lambda \neq 7 \\
\alpha &= 10\% \\
&\text{Two tailed test.}
\end{aligned}
$$

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

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