Quality of Tests Cheat Sheet

Hypothesis tests are not perfect. They can lead to incorrect conclusions which can have serious consequences, depending on the context of the hypothesis test. The aim of this chapter is to be able to quantify the reliability of a hypothesis test. We will look at the different errors that can occur, learn to calculate the probability with which they occur and understand what is meant by the size and power of a test.

Type I and Type II errors

You need to understand what Type I and Type II errors are, and how to find the probability with which they occur in a given scenario.

- A Type I error occurs when you incorrectly reject the null hypothesis. (i.e. the null hypothesis was correct, but you rejected it)
  - ⇒ The probability of a Type I error is the same as the actual significance level of the test.
  - ⇒ When dealing with a continuous distribution, the probability of a Type I error is the same as the significance level of the test.
- A Type II error occurs when you incorrectly accept the null hypothesis. (i.e. the null hypothesis was wrong, but you accepted it)

Example 1: Accidents occurred on a stretch of motorway at an average rate of 6 per month. Many of the accidents that occurred involved vehicles skidding into the back of other vehicles. By way of a trial, a new type of road surface that is said to reduce the risk of vehicles skidding is laid on this stretch of road, and during the first month of operation 4 accidents occurred.

(a) Test this result to see if it gives evidence that there has been an improvement at the 5% level of significance.
(b) Calculate $P(Type\ I\ error)$ for this test.
(c) If the true average rate of accidents occurring with the new type of road surface was 3.5, calculate the probability of a Type II error.

(a) We are dealing with a Poisson distribution. We start by defining the distribution:	Let X represent the number of accidents occurring in a given month, then $X \sim Po(6)$.
Now we state our hypotheses. We are testing for a reduction in the risk of vehicles skidding, so the alternative hypothesis is $\lambda < 6$.	$H_0: \lambda = 6$ $H_1: \lambda < 6$
Finding $P(X \le 4)$ and comparing it to the significance level:	$P(X \le 4) = 0.2851 > 0.05$ Insufficient evidence to reject H_0. The average number of accidents per month has not been reduced.
(b) To find $P(Type\ I\ error)$, we take the null hypothesis to be true and find the probability that our variable falls in the critical region.	$P(Type\ I\ error) = P(X\ falls\ in\ the\ critical\ region)$
Finding the critical region: You can use the statistical tables or your calculator to do this.	$P(X \le 2) = 0.0620 > 0.05$ $P(X \le 1) = 0.0174 < 0.05$ $\therefore critical\ region\ is\ X \le 1.$
So the required probability is:	$\therefore P(Type\ I\ error) = P(X \le 1) = 0.0174$
(c) To find $P(Type\ II\ error)$, we assume that the mean is in fact 3.5 and find the probability that our variable doesn't fall in the critical region.	We need to find the probability that X does not lie in the critical region given that the mean λ is in fact 3.5. Now we have that $X \sim Po(3.5)$.
So the required probability is:	$P(Type\ II\ error) = P(X \ge 2) = 1 - P(X \le 1)$ $= 1 - 0.1359 = 0.8641$

- To find $P(Type\ I\ error)$, we take the null hypothesis to be true and find the probability that our variable falls in the critical region.
- To find $P(Type\ II\ error)$, we take the null hypothesis to be incorrect and find the probability that our variable doesn't fall in the critical region. You will be given the true value of the population parameter being tested if you are asked to find $P(Type\ II\ error)$.

Example 2: The random variable X is geometrically distributed, and it is desired to test $H_0: p = 0.2$ against $H_1: p < 0.2$, using a 5% level of significance.

(a) Calculate the critical region for this test.
(b) State the probability of a Type I error for this test and, given that the true probability was found to be $p = 0.05$, calculate the probability of a Type II error.

(a) We start by defining the distribution we are dealing with:	$X \sim Geo(0.2)$
We are testing for $p < 0.2$, so we need to find the lowest value of c such that $P(X \ge c) < 0.05$	$P(X \ge c) = (1 - 0.2)^{c-1} < 0.05$ $0.8^{c-1} < 0.05$
Taking logs of both sides then making use of the power rule for logs:	$\log(0.8^{c-1}) < \log(0.05)$ $(c-1)\log 0.8 < \log 0.05$
Dividing through by $\log 0.8$: Since $\log 0.8 < 0$, the inequality flips.	$c - 1 > \frac{\log 0.05}{\log 0.8}$ $\therefore c > 14.425$
All integers greater than 14.425 form the critical region.	So the critical region is $X \ge 15$.
(b) $P(Type\ I\ error) = P(X\ is\ in\ the\ critical\ region)$	$\therefore P(Type\ I\ error) = P(X \ge 15) = (1 - 0.2)^{15-1} = 0.8^{14}$ $= 0.0440\ (4\ d.p.)$
To find $P(Type\ II\ error)$, we take $p = 0.05$ and find the probability X is not in the critical region.	So we have $X \sim Geo(0.05)$. $P(Type\ II\ error) = P(X \le 14)$ $= 1 - (1 - 0.05)^{14} = 0.5123\ (4\ d.p.)$

Finding Type I and Type II errors with normal distributions

You also need to be able to find Type I and Type II errors using the normal distribution. The principle is the same as before.

Example 3: The weight of jam in a jar, measured in grams, is distributed normally with a mean of 150g and a standard deviation of 6g. The production process occasionally leads to a change in the mean weight of jam per jar but the standard deviation remains unaltered.

The manager monitors the production process and for every new batch takes a random sample of 25 jars and weighs their contents to see if there has been any reduction in the mean weight of jam per jar.

Find the critical values for the test statistic $\bar{X}$, the mean weight of jam in a sample of 25 jars, using:
(a) a 5% level of significance
(b) a 1% level of significance.
Given that the true value of μ for the new batch is in fact 147,
(c) Find the probability of a Type II error for each of the above critical regions.

(a) We start by stating our hypotheses. We are testing for a reduction in the mean, so we have a one-tail test.	$H_0: \mu = 150$ $H_1: \mu < 150$
Use the central limit theorem to denote the distribution of the sample mean. The sample size is 25.	$\bar{X} \sim N\left(150, \frac{6^2}{25}\right)$
Find the critical value at the 5% level of significance.	For a 5% significance level, $Z < -1.6449$ So $\frac{\bar{X}-150}{\frac{6}{\sqrt{25}}} \le -1.6449$
Rearrange the expression and round the solution to 2 d.p.	$\bar{X} \le 148.03\ (2\ d.p.)$
(b) Find the critical value at the 1% level of significance.	For a 1% significance level, $Z < -2.3263$ So $\frac{\bar{X}-150}{\frac{6}{\sqrt{25}}} \le -2.3263$
Rearrange the expression and round the solution to 2 d.p.	$\bar{X} \le 147.21\ (2\ d.p.)$
c) Find the probability that $\bar{X}$ is outside the critical region we found in part a given that the mean is actually 147.	Now we have $\bar{X} \sim N\left(147, \frac{6^2}{25}\right)$ For part a, $P(Type\ II\ error) = P(\bar{X} > 148.03)$ $= 0.1963\ (4\ d.p.)$
Find the probability that $\bar{X}$ is outside the critical region we found in part b given that the mean is actually 147.	For part b, $P(Type\ II\ error) = P(\bar{X} > 147.21)$ $= 0.4311\ (4\ d.p.)$

The relationship between Type I and Type II errors

The above example gives us an insight into the relationship between Type I and Type II errors. From part **a** to part **b** we lowered the significance level from 5% to 1%. In other words, we reduced the $P(Type\ I\ error)$ from 5% to 1%. As we did this however, the $P(Type\ II\ error)$ increased. In part c you found that $P(Type\ II\ error)$ went from 0.1963 to 0.4311. This highlights the inverse relationship between Type I and Type II errors; as you reduce the Type I error, the Type II error will increase.

This is why a smaller significance level isn't necessarily optimal. A smaller significance level will increase the probability of a Type II error, which could potentially have serious consequences depending on the context of the hypothesis test. For example, if you are carrying out a test to determine whether a testing procedure for a disease functions as required, then a Type II error would cause you to conclude that the procedure works well even though it in fact doesn't. This could lead to patients being told they don't have the disease even when they do, which as you can imagine is highly serious since diseases would potentially go untreated.

A Type I error on the other hand would cause the procedure is concluded inadequate even though it does in fact work fine, which isn't such a bad outcome since we would then naturally resort to other testing procedures. In this case, a Type II error is much more serious than a Type I error and thus we would aim to minimise the Type II error, possibly by avoiding a low significance level such as 1%.

Calculating the size and power of a test

You need to be able to understand what is meant by the size and power of a test and be able to calculate them.

- The size of a test is equal to the probability of a Type I error. This is the probability of incorrectly rejecting the null hypothesis.
- The probability of rejecting the null hypothesis when it is false is known as the power of the test.
- $Power = 1 - P(Type\ II\ error)$

Simply put, the power is a measure of how good the test performs when the null hypothesis is in fact incorrect. Since the power is a probability, it can only take values in the range $0 \le power \le 1$.

Example 4: The random variable X has a Poisson distribution. A sample is taken, and it is desired to test $H_0: \lambda = 4.5$ against $H_1: \lambda < 4.5$ using a 5% significance level.
(a) Find the size of this test.
(b) Given that $\lambda = 4.1$, find the power of this test.

(a) We begin by defining the distribution:	$X \sim Po(8)$
The size is equal to $P(Type\ I\ error)$. This is the probability that X falls in the critical region. So finding the critical region:	$P(X \le 1) = 0.0611 > 0.05$ $P(X = 0) = 0.0111 < 0.05$ So the critical region is $X = 0$
We already saw that $P(X = 0) = 0.0111$	$\therefore P(Type\ I\ error) = P(X = 0) = 0.0111$
(b) Our distribution becomes:	$X \sim Po(4.1)$
We simply need to find $P(X = 0)$ using $\lambda = 4.1$.	$P(rejecting\ H_0) = P(X = 0) = \frac{e^{-4.1}(4.1)^0}{0!} = 0.0166$

The power function

In part **b** of the above example, we were told the actual value of the population parameter λ, which allowed us to find the power. In practice however, population parameters are often unknown which means we cannot determine the power of the test in the same way we did above. In such situations, what we can do instead is find a function which describes the power of the test in terms of the relevant population parameter. This allows us to see how the power varies for different values of the parameter.

- The power function of a test is the function of the parameter θ which gives the probability that the test statistic will fall in the critical region of the test if θ is the true value of the parameter.
- You can use the power function to plot a graph of power against θ.
- When comparing two tests of similar size, you should recommend the test with the higher power within the likely range of the parameter.

Example 5: In a binomial experiment consisting of 12 trials, the random variable X representsd the number of success and p the probability of success. In a test of $H_0: p = 0.45$ against $H_1: p < 0.45$ the null hypothesis is rejected if the number of successes is 2 or less. Show that the power function for this test is given by $(1-p)^{12} + 12(p)(1-p)^{11} + 66(p)^2(1-p)^{10}$.

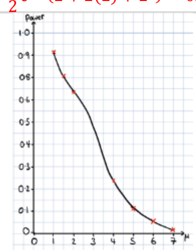
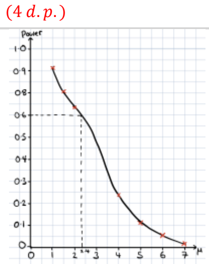
We begin by defining the distribution:	$X \sim B(12, 0.45)$
$Power = P(X\ in\ C.R.) = P(X \le 2)$	$Power = P(X \le 2) = P(X = 0) + P(X = 1) + P(X = 2)$
Working out each probability individually using the probability mass function for a binomial random variable:	$P(X = 0) = (1-p)^{12}$, $P(X = 1) = 12(1-p)(p)^{11}$ $P(X = 2) = 66(1-p)^2(p)^8$
Adding the probabilities together:	$\therefore Power = (1-p)^{12} + 12(p)(1-p)^{11} + 66(p)^2(1-p)^{10}$

Example 6: A single observation, x, is to be taken form a Poisson distribution with parameter μ. This observation is to be used to test $H_0: \mu = 6$ against $H_1: \mu < 6$. The critical region is chosen to be $x \le 2$.

(a) Show that the power function is given by $\frac{1}{2}e^{-\mu}(2 + 2\mu + \mu^2)$.
The table gives the values of the power function to 2 decimal places.

μ	1.0	1.5	2.0	4.0	5.0	6.0	7.0
Power	0.92	0.81	s	0.24	0.12	0.06	0.03

(b) Calculate the value of s.
(c) Draw a graph of the power function.
(d) Estimate the range of values of μ for which the power of this test is greater than 0.6.

(a) First denoting our distribution:	$X \sim Po(6)$
The power is equal to the probability that X falls within the critical region. This is $P(X \le 2)$.	$Power = P(X \le 2) = P(X = 0) + P(X = 1) + P(X = 2)$ $P(X = 0) = \frac{e^{-\mu}(\lambda)^0}{0!} = e^{-\mu}$ $P(X = 1) = \frac{e^{-\mu}(\lambda)^1}{1!} = \mu e^{-\mu}$ $P(X = 2) = \frac{e^{-\mu}(\mu)^2}{2!} = \frac{1}{2}\mu^2 e^{-\mu}$
	$\therefore Power = e^{-\mu} + \mu e^{-\mu} + \frac{1}{2}\mu^2 e^{-\mu}$ $= \frac{1}{2}e^{-\mu}(2 + 2\mu + \mu^2)$ as required.
(b) Substituting $\mu = 2$ into the power function:	$s = \frac{1}{2}e^{-2}(2 + 2(2) + 2^2) = 0.6767\ (4\ d.p.)$
(c) Sketching the graph: (d) We need to first estimate the value of μ for which the power is 0.6. To do this, we find where the curve intersects the line $Power = 0.8$ and look at the corresponding value of μ.	(c)  (d) 
The curve is above $Power = 0.6$ for $\mu < 2.4$.	The line $Power = 0.6$ intersects the curve at approximately $\mu = 2.4$. So the power of the test is greater than 0.6 for $\mu < 2.4$.