7 Hypothesis testing

Very often we want to use data to check a statement about the population. Is the gender ratio 0.5? Is the vote share in elections going to be less than 50%? Are wages higher than last year? Such a statement about the nature of the population is called a statistical *hypothesis* and is often in terms of a population parameter like the population mean. For example, the statement about the gender ratio is actually asking if the mean of the random variable denoting gender in the population is 0.5. What we want to do is to take a sample from the population and be able to say whether the data supports the hypothesis or not. In this lecture we will understand the procedure for this called hypothesis testing.

Let us take an example. Suppose we want to know if the occurrence of flu in the APU campus is the same as in the general population. We know that in the population, the chance of catching a flu in any given year is 10%. We know that out of 225 students and residents at APU, 35 have caught the flu over the past year (all numbers are made up here!), giving the probability of 15.6%. The sample variance will be 0.132, (how do you know this just from the sample mean?) and hence the sample standard deviation will be $\sqrt{0.132} = 0.363$ Now we want to test whether this higher number is just because of chance or is the probability of catching flu actually different at APU.

First we state the *null hypothesis*. The null hypothesis is the one we would like to check. The test may either *reject* the null hypothesis, or it may *fail to reject* it. In other words, the null hypothesis is the one that the test is trying to disprove. Here, the null hypothesis is that the population mean of getting flu at APU is 10%. We write it as follows.

$$H_0 : \mu = 0.1$$

If the test disproves the null hypothesis, then what we believe to be true is called the alternative hypothesis. In this case, it is that the probability of getting the flu at APU is different from the rest of the population. (Note that we are not

talking about whether it is higher or lower, we will come back to that a bit later). We write it as follows.

$$H_1 : \mu \neq 0.1$$

Now we will construct a *test statistic* using the sample and the null. Since the test statistic is constructed using the sample, it will be a random variable. We will consider the distribution of this random variable assuming that the null is true and find the probability of observing a value as far away as the one we observe in our sample.

For hypothesis about the sample mean, the test statistic we use is called the *t-statistic*. It just tells us how many standard deviations (of the sample mean distribution) away the sample mean is from the population mean.

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

This sample statistic will be distributed as the t-distribution, since we are using the sample variance to calculate it. We assume the null to be true so the population mean, which is also the expected value of the sample mean, is $\mu = 0.1$. The observed value of the sample mean is $\bar{x} = 0.156$. And the estimate of the standard deviation of the sample mean distribution is $s/\sqrt{n} = 0.363/\sqrt{225} = 0.024$. Therefore, the value of the t-statistic for this sample is

$$t = \frac{0.156 - 0.1}{0.024} = 2.3$$

Now we need to devise a test that will tell us that given this value of the t-statistic, if the null hypothesis should be rejected or not. We will do this by defining regions of the distribution of the where the probability of finding the observed value of the t-statistic is very low if the null is true. This is shown in Figure 1.

The area under the curve for the values of the t-statistic where we will reject the null, represent the probability that the test will reject the null even when it

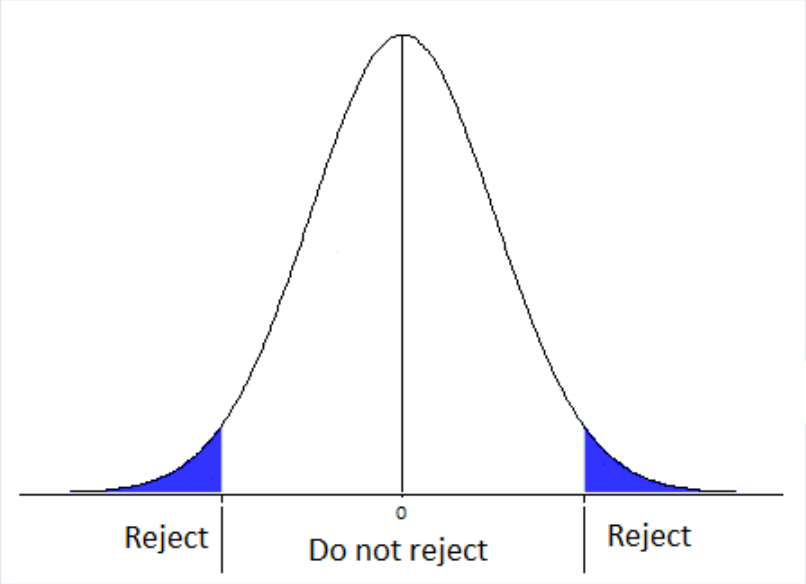

Figure 1: Areas of the distribution of the test statistic where the null will be rejected

is true. This is called the significance level of the test and is typically represented by α. Typically it is 5% or 1%. A 5% significance level will imply that the area of the blue region in Figure 1 is 0.05. This means that even if the population mean is actually what the null says, if we conduct this test an infinite number of times, the null will be rejected by the test 5% of the time. It also means that if the null is true, the chances of observing a value of the test statistic in this region is 5%.

Coming back to our test, let us say we want to conduct a test with a 5% significance level. Then the value of the t-statistic above which we will reject the null is the value where the *cdf* will be 0.95+0.025=0.975. This value is called the *critical value* of the test statistic. If we look up the table for the t-distribution, with degrees of freedom 224, then the critical value is 1.97. Since $t > 1.97$, we can *reject the null* at 5% significance. What if we conduct the test at 1% significance level. The critical value is 2.598. Since $t < 2.598$, we *fail to reject the null* at 1% significance. How do we decide what significance level to assume? We cannot assume really low significance levels, as the test will be so strict that even if the

null is incorrect, it will fail to reject it. We consider this trade-off between over-rejection and under-rejection in section 7.2.

Note: In this entire example there is one basic problem because of which all results are wrong. What is it?

7.1 P-values

A p-value for a given value of a test statistic is that level of significance where the null will be just rejected. In other words, p-value is the probability, given the null is true, of observing a value of the test statistic at least as extreme as the actual value observed. In our previous (incorrect) example, the p-value for $t = 2.3$, would be 0.022. This means that if the null is true then the probability of the value of the test statistic being at least as far away from 0 as 2.3 is 2.2%. From our test earlier, we knew that this value would be between 1% and 5%. Another way of doing the hypothesis test is to just compare the p-value of the test statistic to the desired significance level. If p-value $<$ significance level, then the null is rejected.

Sometimes the p-value is misinterpreted as the probability that the null is true. Why is this interpretation incorrect?

7.2 Power of a test

Now there are four possibilities, outlined in Table 1. If in reality the null is true and the test doesn't reject it, or if the null is false and the test rejects it then the test result is correct. But the test result can be wrong in two ways. First, if the null is in fact true but the test rejects it. This is called a Type I error. The second is that the null is false but the test fails to reject it. This is called a Type II error.

We can try to understand this using Figure 2. The null hypothesis is that the

		Test result	
		Null rejected	*Null not rejected*
Reality	*Null true*	Type I error	Test is correct
	Null false	Test is correct	Type II error

Table 1: Two types of error in hypothesis testing

population mean is a, i.e. $H_0 : \mu = a$. Hence, the solid line curve is the distribution of the sample mean is the null hypothesis is true. For a given significance level, say 5%, the rejection area will be like the area shaded grey in the figure. It should be evident that the significance level α is the probability of making a Type I error, i.e. the probability of rejecting the null even when it is true. Now, if the reality is that the null is not true and the real population mean is b, then the actual distribution of the sample mean is the curve drawn with a dashed line. The probability of making a Type II error, i.e. the probability of not rejecting the null even when it is not true, is given by the area shaded in blue. It is the area under the curve of the dashed distribution that falls in the acceptance area of the null distribution.The probability of *not* making a type II error is called the *power* of the test (it is given by the area of the dashed curve *not* shaded in blue). It is easy to see that there is a trade off between the power and the significance level. If we want to reduce the probability of making a Type I error, we will reduce the significance level from 5% to ,say, 1%. This would result in an increase in the probability of making a Type II error, thus reducing the power of the test.

The power of the test can only be determined if we assume a particular value for the alternative hpothesis, i.e. what particular value will the population mean take if the null is not true. If this value is very different from the null, i.e. b is very far from a, then the area of overlap will be very small and the test will have high power, regardless of the significance level we choose. But given a particular value of the population mean that we have in mind, how can we increase power without changing the significance level? This can be done by increasing the sample size.

A larger sample size would reduce the standard deviation of the sample mean and would hence reduce the area of overlap and increase the power of the test.

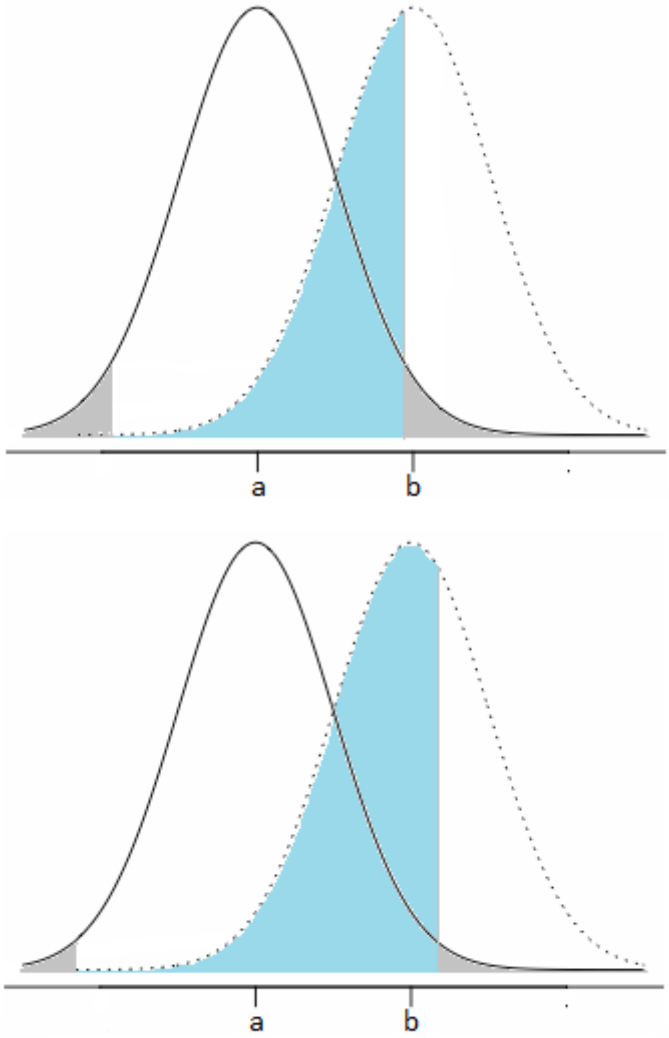

Figure 2: Probabilities of Type I (shaded in grey) and Type II (blue) errors

7.3 One-sided test

In the hypothesis tests that we have see so far, the null hypothesis is that the population mean takes a particular value, and we reject the null if the sample

mean is far enough from it on either side, i.e. samller or larger than the null. This is what is called a two-sided or a two-tailed test. A one sided test is when the null hypothesis is that the population mean can take a maximum or minimum value. For example, in the APU flu example, it may be the case the medical journals say that the probability of catching a flu can be at most 10%. In that case, the null and alternative hypotheses would be the following.

$$H_0 : \mu \leq 0.1$$

$$H_1 : \mu > 0.1$$

We will proceed with the hypothesis test as earlier, assuming that the population mean is 0.1 and constructing a distribution of sample means around it. The difference will be that the rejection region will be on only one side of the distribution. We will not reject the null even if the sample mean is much smaller than 0.1, but we will reject it if the sample mean is much larger. Figure 3 shows the rejection area for a two sided test compared to that for a one-sided test for the same significance level.

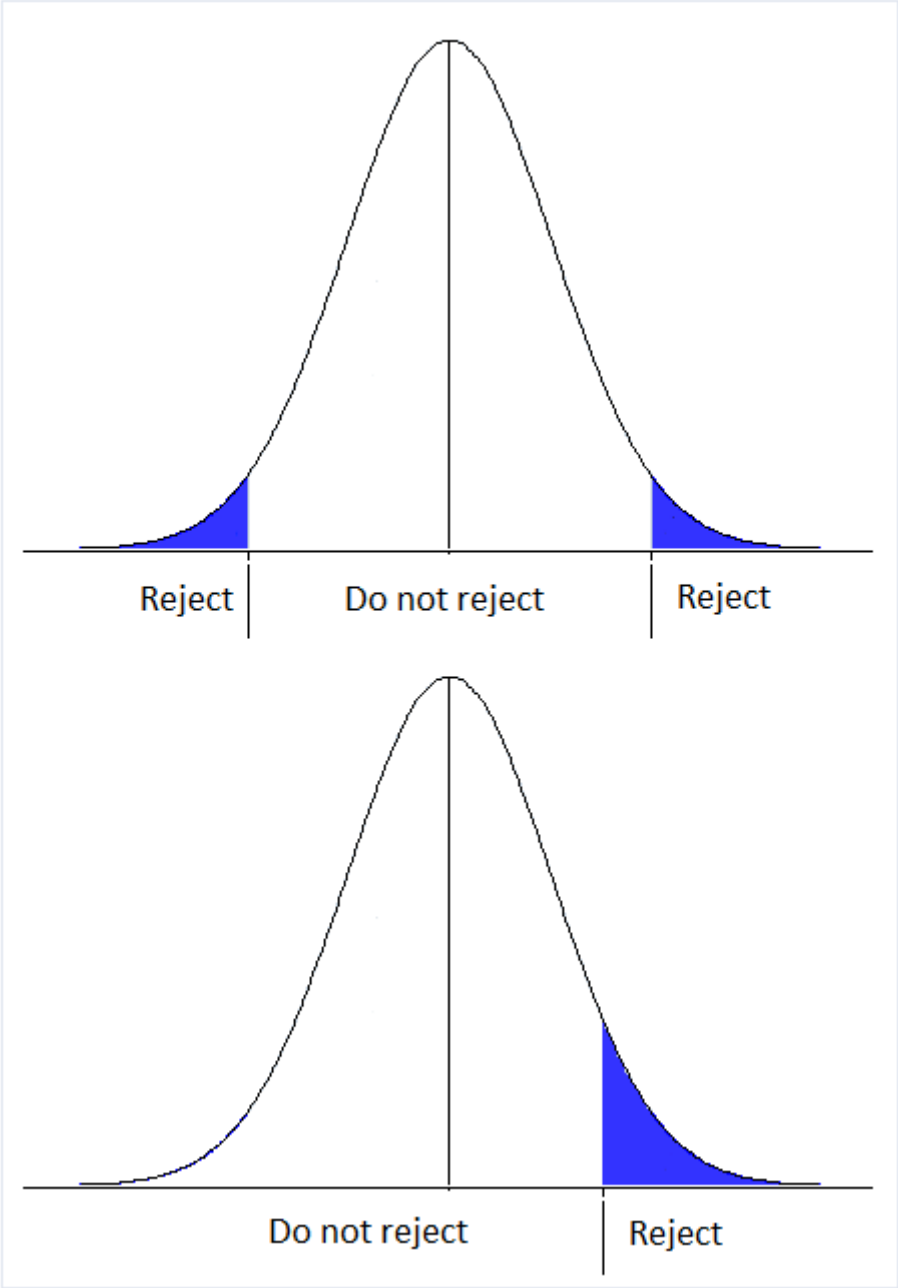

Figure 3: The rejection areas for a two-sided test (top) and for a one-sided test (bottom) for the same significance level