

## Lecture 7

# The Z Test

### 7.1 Introduction

In Lecture 1 we saw that statistics has a crucial role in the scientific process and that we need a good understanding of statistics in order to avoid reaching invalid conclusions concerning the experiments that we do. In Lectures 2 and 3 we saw how the use of statistics necessitates an understanding of probability. This led us to study how to calculate and manipulate probabilities using a variety of probability rules. In Lectures 4, 5 and 6 we consider three specific probability distributions that turn out to be very useful in practical situations. Effectively, all of these previous lectures have provided us with the basic tools we need to use statistics in practical situations. In this lecture we consider the general framework used to test a specific hypothesis by examining some basic examples that utilise our knowledge of the Normal distribution.

### 7.2 The logic of significance tests

#### Example 7.1: Baby-boom hypothesis test

Consider the following hypothetical situation: Suppose we think that UK newborns are heavier than Australian newborns. We know from large-scale studies that UK newborns average 3426g, with an SD of 538g. (See, for example, [NNGT02].) The weights are approximately normally distributed. We think that maybe

babies in Australia have a mean birth weight smaller than 3426g and we would like to test this hypothesis.

Intuitively we know how to go about testing our hypothesis. We need to take a sample of babies from Australia, measure their birth weights and see if the sample mean is *significantly smaller* than 3426g. Now, we have a sample of 44 Australian newborns, presented in Table 1.2, and with histogram presented in Figure 7.1. (Ignore for the moment that these are not really a sample of all Australian babies...)

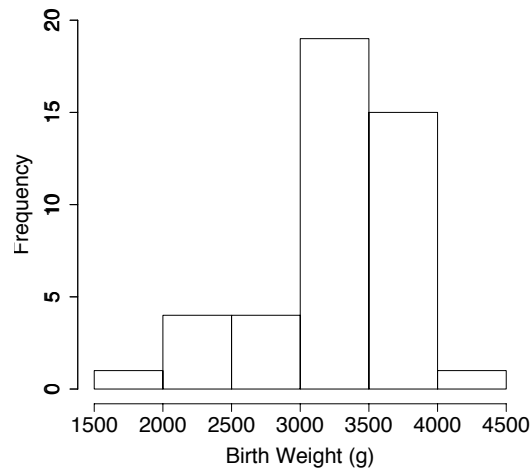


Figure 7.1: A Histogram showing the birth weight distribution in the Baby-boom dataset.

We observe that the sample mean of these 44 weights is 3276g. So we might just say that we're done. The average of these weights is smaller than 3426g, which is what we wanted to show.

“But wait!” a skeptic might say. “You might have just happened by chance to get a particularly heavy group of newborns. After all, even in England lots of newborns are lighter than 3276g. And 44 isn't such a big sample.”

How do we answer the skeptic? Is 44 big enough to conclude that there is a real difference in weights between the Australian

sample and the known English average? We need to distinguish between

The research hypothesis: “Australian newborns have a mean weight greater than 3426g;”

and

The null hypothesis: “There’s no difference in mean weight; the apparent difference is purely due to chance.”

How do we decide which is true? We put the null hypothesis to the test. It says that the 44 observed weights are just like what you might observe if you picked 44 at random from the UK newborn population; that is, from a normal distribution with mean 3426g and SD 538g.

Let  $X_1, \dots, X_{44}$  be 44 weights picked at random from a  $\mathcal{N}(3426, 538^2)$  distribution, and let  $X = \frac{1}{44}(X_1 + \dots + X_{44})$  be their mean. How likely is it that  $X$  is as small as 3276? Of course, it’s never impossible, but we want to know how plausible it is.

We know from section 6.6 that

$$X_1 + \dots + X_{44} \sim \mathcal{N}(3426 \times 44, 538^2 \times 44), \text{ and}$$

$$X = \frac{1}{44}(X_1 + \dots + X_{44}) \sim \mathcal{N}(3426, 538^2/44) = \mathcal{N}(3000, 81^2).$$

Thus,

$$P(X \leq 3276) = P\left(Z \leq \frac{3276 - 3426}{81}\right) = P(Z \leq -1.81) = 1 - P(Z < 1.81),$$

where  $Z = (X - \mu)/\sigma = (X - 3426)/81$  has standard normal distribution. Looking this up on the standard normal table, we see that the probability is about 0.0351. ■

The probability 0.0351 that we compute at the end of Example ?? is called the **p-value** of the test. It tells us how likely it is that we would observe such an extreme result if the null hypothesis were true. The lower the p-value, the stronger the evidence *against* the null hypothesis. We are faced with the alternative: Either the null hypothesis is false, or we have by chance happened to get a result that would only happen about one time in 30. This seems unlikely, but not impossible.

Pay attention to the double negative that we commonly use for significance tests: We have a research hypothesis, which we think would be interesting if it were true. We don't test it directly, but rather we use the data to challenge a less interesting **null hypothesis**, which says that the apparently interesting differences that we've observed in the data are simply the result of chance variation. We find out whether the data support the research hypothesis by showing that the null hypothesis is false (or unlikely). If the null hypothesis passes the test, then we know only that this particular challenge was inadequate. We haven't proven the null hypothesis. After all, we may just not have found the right challenger; a different experiment might show up the weaknesses of the null. (The potential strength of the challenge is called the "power" of the test, and we'll learn about that in Lecture ??.)

What if the challenge succeeds? We can then conclude with confidence (how much confidence depends on the p-value) that the null was wrong. But in a sense, this is shadow boxing: We don't exactly know who the challenger is. We have to think carefully about what the plausible alternatives are. (See, for instance, Example 7.2.)

### 7.2.1 Outline of significance tests

The basic steps carried out in Example 7.1 are common to most significance tests:

- (i). Begin with a **research (alternative) hypothesis**.
- (ii). Set up the **null hypothesis**.
- (iii). Collect a sample of data.
- (iv). Calculate a **test statistic** from the sample of data.
- (v). Compare the test statistic to its **sampling distribution** under the null hypothesis and calculate the **p-value**. The strength of the evidence is larger, the smaller the p-value.

### 7.2.2 Significance tests or hypothesis tests? Breaking the .05 barrier

We use p-values to weigh scientific evidence. What if we need to make a decision?

One common situation is that the null hypothesis is being compared to an alternative that implies a definite course of action. For instance, we may be testing whether daily doses of vitamin C prevent colds: We take 100 subjects, and give them vitamin C supplements every day for a year, and no vitamin C supplement for another year, and compare the numbers of colds. At the end, we have two alternatives: either make a recommendation for vitamin C, or not.

The standard approach is to start by saying: The neutral decision is to make no recommendation, and we associate that with the null hypothesis, which says that any difference observed may be due to chance. In this system, the key goal is to control the likelihood of falsely making a positive recommendation (because we have rejected the null hypothesis). This situation, where we incorrectly reject the null hypothesis is called a **Type I Error**. The opposite situation, where we retain the null hypothesis although it is false, is called a **Type II Error**.

By definition, if the null hypothesis is true, the probability that the p-value is less than a given number  $\alpha$  is exactly  $\alpha$ . Thus, we begin our hypothesis test by fixing  $\alpha$ , the probability of a Type I error, to be some tolerably low number. We call this  $\alpha$  the **significance level** of the test. (A common choice is  $\alpha = 0.05$ , but the significance level can be anything you choose. If the consequences of a Type I Error would be extremely serious — for instance, if we are testing a new and very expensive cancer drug, with the expectation that h

In our current example, the p-value is about  $10^{-6}$  which is lower than 0.05. In this case, we would conclude that

“there is significant evidence against the null hypothesis at the 5% level”

Another way of saying this is that

“we reject the null hypothesis at the 5% level”

If the p-value for the test were much larger, say 0.23, then we would conclude that

“the evidence against the null hypothesis is not significant at the 5% level”

Another way of saying this is that

“we cannot reject the null hypothesis at the 5% level”

Note that the conclusion of a hypothesis test, strictly speaking, is binary: We either reject or retain the null hypothesis. There are no gradations, no

Decision \ Truth	$H_0$ True	$H_0$ False
Retain $H_0$	Correct (Prob. $1 - \alpha$ )	Type II Error (Prob.= $\beta$ )
Reject $H_0$	Type I Error (Prob.=level= $\alpha$ )	Correct (Prob.=Power= $1 - \beta$ )

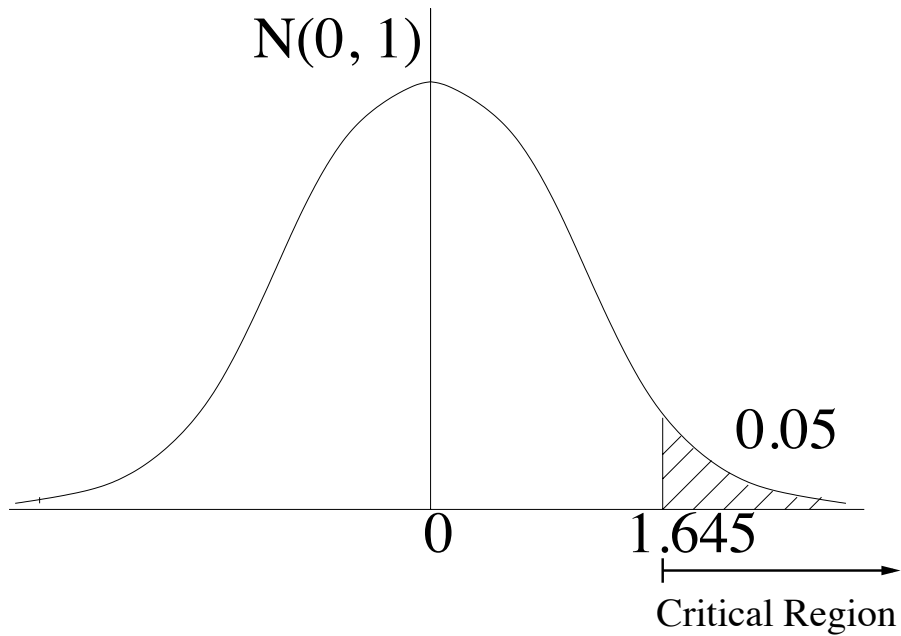
Table 7.1: Types of errors

strong rejection or borderline rejection or barely retained. The fact that our p-value was  $10^{-6}$  ought not to be taken, retrospectively, as stronger evidence against the null hypothesis than a p-value of 0.04 would have been.

By the strict logic imposed the data are completely used up in the test: If we are testing at the 0.05 level and the p-value is 0.06, we cannot then collect more data to see if we can get a lower p-value. We would have to throw away the data, and start a new experiment. Needless to say, this is not what scientists really do, which makes even the apparently clear-cut yes/no decision set-up of the hypothesis test in reality rather difficult to interpret.

It is also quite common to confuse this situation with using significance tests to judge scientific evidence. So common, in fact, that many scientific journals impose the 0.05 significance threshold to decide whether results are worth publishing. An experiment that resulted in a statistical test with a p-value of 0.10 is considered to have failed, even if it may very well be providing reasonable evidence of something important; if it resulted in a statistical test with a p-value of 0.05 then it is a success, even if the effect size is minuscule, and even though 1 out of 20 true null hypotheses will fail the test at significance level 0.05.

Another way of thinking about hypothesis tests is that there is some **critical region** of values such that if the test statistic lies in this region then we will reject  $H_0$ . If the test statistic lies outside this region we will not reject  $H_0$ . In our example, using a 5% level of significance this set of values will be the most extreme 5% of values in the right hand tail of the distribution. Using our tables backwards we can calculate that the boundary of this region, called the **critical value**, will be 1.645. The value of our test statistic is 3.66 which lies in the critical region so we reject the null hypothesis at the 5% level.



### 7.2.3 Overview of Hypothesis Testing

Hypothesis tests are identical to significance tests, except for the choice of a *significance level* at the beginning, and the nature of the conclusions we draw at the end:

- (i). Begin with a **research (alternative) hypothesis** and decide upon a **level of significance** for the test.
- (ii). Set up the **null hypothesis**.
- (iii). Collect a sample of data.
- (iv). Calculate a **test statistic** from the sample of data.
- (v). Compare the test statistic to its **sampling distribution** under the null hypothesis and calculate the **p-value**,

*or equivalently,*

Calculate the **critical region** for the test.

(vi). Reject the null hypothesis if

the p-value is less than the **level of significance**,

*or equivalently,*

the test statistic lies in the **critical region**.

Otherwise, retain the null hypothesis.

### 7.3 The one-sample Z test

A common situation in which we use hypothesis tests is when we have multiple independent observations from a distribution with unknown mean, and we can make a test statistic that is normally distributed. The null hypothesis should then tell us what the mean and standard error are, so that we can normalise the test statistic. The normalised test statistic is then commonly called  $Z$ . We always define  $Z$  by

$$Z = \frac{\text{observation} - \text{expectation}}{\text{standard error}}. \quad (7.2)$$

The expectation and standard error are the mean and the standard deviation of the *sampling distribution*: that is, the mean and standard deviation that the observation has when seen as a random variable, whose distribution is given by the null hypothesis. Thus,  $Z$  has been *standardised*: its distribution is standard normal, and the p-value comes from looking up the observed value of  $Z$  on the standard normal table.

We call this a “one-sample” test because we are interested in testing the mean of samples from a single distribution. This is as opposed to the “two-sample test” (discussed in section ??), in which we are testing the difference in means between two populations.

#### 7.3.1 Test for a population mean $\mu$

We know from Lecture 6 that if

$$X_1 \sim \mathcal{N}(\mu, \sigma^2) \quad X_2 \sim \mathcal{N}(\mu, \sigma^2)$$

then

$$\begin{aligned} \bar{X} = \frac{1}{2}X_1 + \frac{1}{2}X_2 &\sim \mathcal{N}\left(\frac{1}{2}\mu + \frac{1}{2}\mu, \left(\frac{1}{2}\right)^2\sigma^2 + \left(\frac{1}{2}\right)^2\sigma^2\right) \\ \Rightarrow \bar{X} &\sim \mathcal{N}\left(\mu, \frac{\sigma^2}{2}\right) \end{aligned}$$

In general,

If  $X_1, X_2, \dots, X_n$  are  $n$  independent and identically distributed random variables from a  $\mathcal{N}(\mu, \sigma^2)$  distribution then

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

Thus, if we are testing the null hypothesis

$$H_0 : \text{The } X_i \text{ have } \mathcal{N}(\mu, \sigma) \text{ distribution,}$$

the expectation is  $\mu$ , and the standard error is  $\sigma/\sqrt{n}$ . Thus,

When testing the sample mean of  $n$  normal samples, with known SD  $\sigma$ , for the null hypothesis mean =  $\mu$ , the test statistic is

$$Z = \frac{\text{sample mean} - \mu}{\sigma/\sqrt{n}}.$$

Thus, under the assumption of the null hypothesis the sample mean of 44 values from a  $\mathcal{N}(3426, 538^2)$  distribution is

$$\bar{X} \sim \mathcal{N}\left(3426, \frac{538^2}{44}\right) = \mathcal{N}(3426, 81^2)$$

### 7.3.2 Test for a sum

Under some circumstances it may seem more intuitive to work with the sum of observations rather than the mean. If  $S = X_1 + \dots + X_n$ , where the  $X_i$  are independent with  $\mathcal{N}(\mu, \sigma^2)$  distribution, then  $S \sim \mathcal{N}(n\mu, n\sigma^2)$ . That is, the expectation is  $n\mu$  and the standard error is  $\sigma\sqrt{n}$ .

When testing the sum of  $n$  normal samples, with known SD  $\sigma$ , for the null hypothesis mean =  $\mu$ , the test statistic is

$$Z = \frac{\text{observed sum of samples} - n\mu}{\sigma \times \sqrt{n}}.$$

### 7.3.3 Test for a total number of successes

Suppose we are observing independent trials, each of which has unknown probability of success  $p$ . We observe  $X$  successes. We have the estimate  $\hat{p} = X/n$ . Suppose we have some possible value  $p_0$  of interest, and we wish to test the null hypothesis

$$H_0 : p = p_0$$

against the alternative

$$H_1 : p > p_0.$$

We already observed in section 6.8 that the random variable  $X$  has distribution very close to normal, with mean  $pn$  and standard error  $\sqrt{np(1-p)}$ , as long as  $n$  is reasonably large. We have then the test statistic

When testing the number of successes in  $n$  trials, for the null hypothesis  $P(\text{success}) = p_0$ , the test statistic is

$$Z = \frac{\text{observed number of successes} - np_0}{\sqrt{np_0(1-p_0)}}.$$

#### Example 7.2: The Aquarius Machine, continued

We repeat the computation of Example 6.9. The null hypothesis, corresponding to “no extrasensory powers”, is

$$H_0 : p = p_0 = 0.25;$$

the alternative hypothesis, Tart’s research hypothesis, is

$$H_1 : p > 0.25.$$

With  $n = 7500$ , the expected number of successes under the null hypothesis is  $7500 \times \frac{1}{4} = 1875$ , and the standard error is  $\sqrt{7500 \times \frac{1}{4} \times \frac{3}{4}} = 37.5$ . We compute the test statistic

$$\begin{aligned} Z &= \frac{\text{observed number of successes} - \text{expected number of successes}}{\text{standard error}} \\ &= \frac{2006 - 1875}{37.5} \\ &= 3.49. \end{aligned}$$

(This is slightly different from the earlier computation ( $z = 3.48$ ) because we conventionally ignore the continuity correction when computing test statistics.) Thus we obtain from the standard normal table a p-value of 0.0002.

So it is extraordinary unlikely that we would get a result this extreme purely by chance, if the null hypothesis holds. If  $p_0 = 1/4$ , then Tart happened to obtain a result that one would expect to see just one time in 5000. Must we then conclude that  $p_0 > 1/4$ ? And must we then allow that at least some subjects had precognitive powers? Actually, in this case we know what happened to produce this result. It seems that there were defects in the random number generator, making the same light less likely to come up twice in a row. Subjects presumably cued in to this pattern after a while — they were told after each guess whether they'd been right — and made use of it for their later guesses. Thus, the binomial distribution did not hold — the outcomes of different tries were not independent, and did not all have probability  $1/4$  — but not in the way that Tart supposed. Thus, one needs always to keep in mind: Statistical tests tell us that the come from our chance model, but it doesn't necessarily follow that our favourite alternative is true.

Some people use the term **Type III error** to refer to the mistake of correctly rejecting the null hypothesis, but for the wrong reason. Thus, to infer that the subjects had extrasensory powers from these data would have been a Type III error. ■

### 7.3.4 Test for a proportion

When testing for probability of success in independent trials, it often seems natural to consider the *proportion* of successes rather than the number of

successes as the fundamental object. Under the null hypothesis

$$H_0 : p = p_0$$

the expected proportion of successes  $X/n$  is  $p_0$ , and the standard error is  $\sqrt{p_0(1-p_0)/n}$ .

When testing the proportion of successes in  $n$  trials, for the null hypothesis  $P(\text{success}) = p_0$ , the test statistic is

$$Z = \frac{\text{proportion of successes} - p_0}{\sqrt{p_0(1-p_0)/n}}.$$

$Z$  has standard normal distribution.

The test statistic will come out exactly the same, regardless of whether we work with numbers of successes or proportions.

### Example 7.3: The Aquarius machine, again

We repeat the computations of Example 7.2, treating the proportion of correct guesses as the basic object. The observed proportion of successes is  $\hat{p} = k/n = 2006/7500 = 0.26747$ . The standard error for the proportion is

$$SE_{\hat{p}} = \sqrt{p_0(1-p_0)/n} = \sqrt{\frac{1}{4} \cdot \frac{3}{4} / 7500} = 0.005.$$

Thus, the test statistic is

$$Z = \frac{0.26747 - 0.25}{0.005} = 3.49,$$

which is exactly the same as what we computed before. ■

### Example 7.4: GBS and swine flu vaccine

In Example 5.7 we fit a Poisson distribution to the number of GBS cases by week after vaccination. We noted that the fit, given in Table 5.3, didn't look very good, and concluded that

GBS cases were not independent of the time of vaccination. But we did not test this goodness of fit formally.

In the current formalism, the null hypothesis formalises the notion that GBS is independent of vaccination, so that numbers of GBS cases are Poisson distributed, with parameter  $\lambda = 1.33$ . We test this by looking at the number of weeks with 0 GBS cases, which was observed to be 16. The formal null hypothesis is

$$H_0 : P(0 \text{ cases in a week}) = e^{-1.33} = 0.2645.$$

The alternative hypothesis is

$$H_1 : P(0 \text{ cases in a week}) \neq 0.2645.$$

The observed proportion is  $16/40 = 0.4$ . The standard error is

$$SE = \sqrt{0.2645 \times 0.7355/40} = 0.0697.$$

Thus, we may compute

$$Z = \frac{0.2645 - 0.4}{0.0697} = -1.94$$

Looking this up on the table, we see that  $P(Z < -1.94) = 1 - P(Z < 1.94) = 0.026$ . Since we have a two-sided alternative, the p-value is twice this, or 0.052.

If we were doing a significance test at the 0.05 level (or any lower level), we would simply report that the result was not significant at the 0.05 level, and retain the null hypothesis. Otherwise, we simply report the p-value and let the reader make his or her own judgement. ■

### 7.3.5 General principles: The square-root law

The fundamental fact which makes statistics work is the fact that when we add up  $n$  independent observations, the expected value increases by a factor of  $n$ , while the standard error increases only by a factor of  $\sqrt{n}$ . Thus, when we divide by  $n$  to obtain a mean (or a proportion), the standard error ends up shrinking by a factor of  $\sqrt{n}$ . This corresponds to our intuition that averaging many independent samples will tend to be closer to the true value than any single measurement. If the standard deviation of the population is  $\sigma$ , the standard error of the sample mean is  $\sigma/\sqrt{n}$ . Intuitively, the standard error tells us about how far off the sample mean will be from the true population mean (or true probability of success): we will almost never be off by more than 3 SEs.

## 7.4 One and two-tailed tests

In Example 7.1 we wanted to test the research hypothesis that mean birth weight of Australian babies was less than 3426g. This suggests that we had some prior information that the mean birth weight of Australian babies was definitely not higher than 3426g, and that the interesting question was whether the weight was lower. If this were not the case then our research hypothesis would be that the mean birth weight of Australian babies was different from 3426g. This allows for the possibility that the mean birth weight could be less than or greater than 3426g.

In this case we would write our hypotheses as

$$H_0 : \mu = 3426\text{g}$$

$$H_1 : \mu \neq 3426\text{g}$$

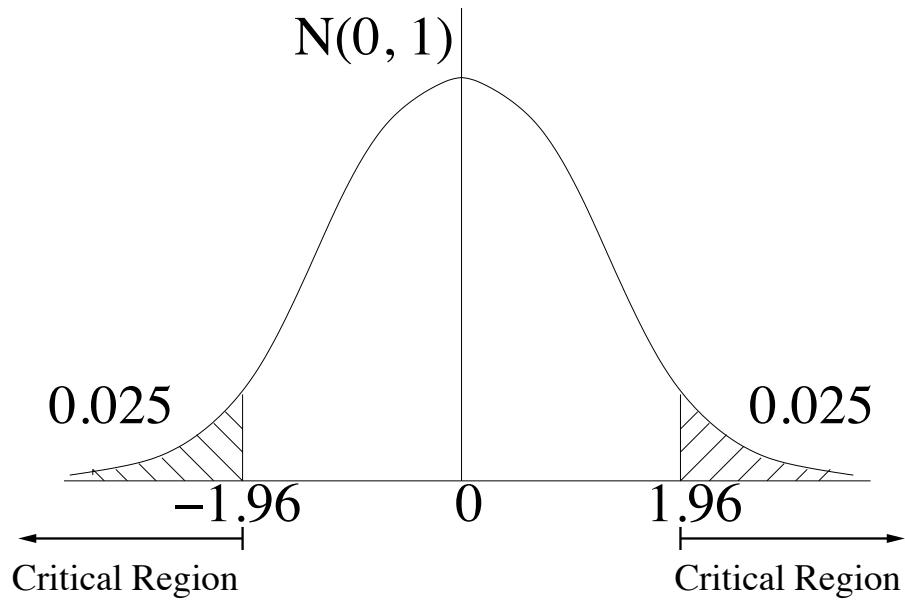
As before we would calculate our test statistic as  $-1.81$ . The p-value is different, though. We are not looking at the probability that  $Z$  is only less than  $-1.81$  (in the positive direction), but that  $Z$  is at least this big *in either direction*; so  $P(|Z| > 3.66)$ , or  $P(Z > 3.66) + P(Z < -3.66) = 2P(Z > 3.66)$ . Because of symmetry,

For a Z test, the two-tailed p-value is always twice as big as the one-tailed p-value.

In this case we allow for the possibility that the mean value is greater than 3426g by setting our critical region to be lowest 2.5% and highest 2.5% of the distribution. In this way the total area of the critical region remains 0.05 and so the level of significance of our test remains 5%. In this example, the critical values are  $-1.96$  and  $1.96$ . Thus if our test statistic is less than  $-1.96$  or greater than  $1.96$  we would reject the null hypothesis. In this example, the value of test statistic does lie in the critical region so we reject the null hypothesis at the 5% level.

This is an example of a **two-sided test** as opposed to the previous example which was a **one-sided test**. The prior information we have in a specific situation dictates what we use as our alternative hypothesis which in turn dictates the type of test that we use.

Fundamentally, though, the distinction between one-tailed and two-tailed tests is important only because we set arbitrary p-values such as 0.05 as hard



cutoffs. We should be cautious about “significant” results that depend for their significance on the choice of a one-tailed test, where a two-tailed test would have produced an “insignificant” result.