

# Lecture 7 : Hypothesis Testing

Jonathan Marchini

November 19, 2004

## Introduction

In Lecture 2 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 Lecture 3 we saw how the use of statistics necessitates an understanding of probability. This lead 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 utilize our knowledge of the Normal distribution.

## Single sample test for a population mean $\mu$

Consider the following hypothetical situation: From previous experience we know that the birth weights of babies in England are Normally distributed with a mean of 3000g and a standard deviation of 500g. We think that maybe babies in Australia have a mean birth weight greater than 3000g 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 larger* than 3000g. We use probability and statistics to decide when the sample mean is *significantly larger*.

More formally, we start by wring down our two competing hypotheses.

The main hypothesis that we are most interested in is the **research hypothesis**, denoted  $H_1$ , that the mean birth weight of Australian babies is greater than 3000g.

The other hypothesis is the **null hypothesis**, denoted  $H_0$ , that the mean birth weight is equal to 3000g.

We can write this compactly as

$$H_0 : \mu = 3000g$$
$$H_1 : \mu > 3000g$$

The null hypothesis is written first followed by the research hypothesis. The research hypothesis is often called the **alternative hypothesis** even though it is often the first hypothesis we think of.

Normally, we start with the research hypothesis and “set up” the null hypothesis to be directly counter to what we hope to show. We then try to show that, in the light of our collected data, that the null hypothesis is false. We do this by calculating the probability of the data if the null hypothesis is true. If this probability is very small it suggests that the null hypothesis is false.

Another way of thinking about this is that we set up a race between the hypotheses and then we try to disqualify the null hypothesis leaving just one hypothesis in the race.

Once we have set up our null and alternative hypothesis we can collect a sample of data. For example, we can imagine we collected the birth weights of the 44 babies in the Babyboom dataset. A histogram of the 44 birth weights is shown in Figure 1.

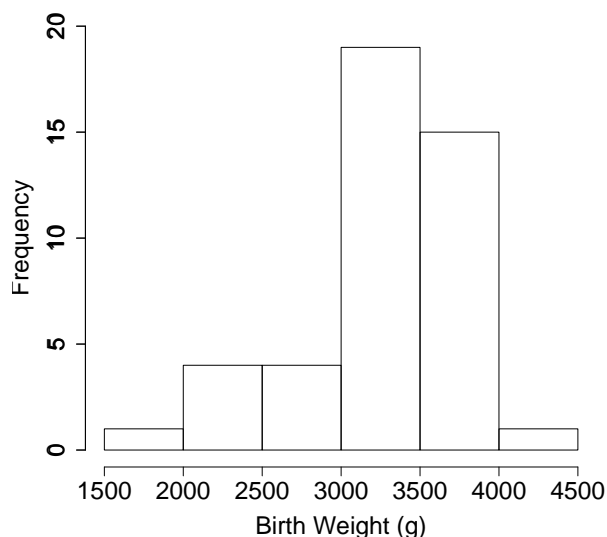


Figure 1: A Histogram showing the birth weight distribution in the Baby-boom dataset.

The sample mean of the dataset is

$$\bar{x} = 3275.955$$

We now want to calculate the probability of obtaining a sample with a mean as large as 3275.955 under the assumption of the null hypothesis  $H_0$ . To do this we need to calculate the distribution of the mean of 44 values from a  $N(3000, 500^2)$  distribution.

We know from Lecture 6 that if

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

then

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

In general,

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

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

**Note** This distribution is known as the **sampling distribution** of the mean value. The standard deviation of the distribution ( $\sigma/\sqrt{n}$ ) is known as the **standard error** of the mean.

Thus, under the assumption of the null hypothesis the sample mean of 44 values from a  $N(3000, 500^2)$  distribution is

$$\bar{X} \sim N\left(3000, \frac{500^2}{44}\right) = N(3000, 5681.818)$$

Now we can calculate the probability of obtaining a sample with a mean as large as 3275.955 using standardization.

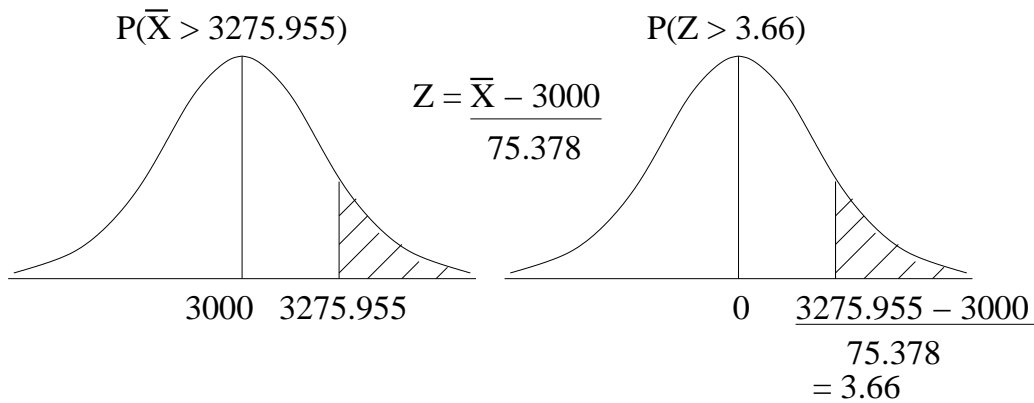
$$\begin{aligned} P(\bar{X} > 3275.955) &= P\left(\frac{\bar{X} - 3000}{75.378} > \frac{3275.955 - 3000}{75.378}\right) \\ &= P(Z > 3.66) \\ &= 1 - 0.99985 \\ &= 0.00015 \end{aligned}$$

Effectively, standardization involves calculating the **test statistic**  $Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$  which has a standard Normal distribution under the null hypothesis.

The probability we calculate is called the **p-value** of the test. In this case the

$$\bar{X} \sim N(3000, 5681.818)$$

$$Z \sim N(0, 1)$$



p-value is very low. This says that the probability of the data is very low if we assume the null hypothesis is true.

But how low does this probability have to be before we can conclude that the null hypothesis is false. The convention within statistics is to choose a **level of significance** before the experiment that dictates how low the p-value should be before we reject the null hypothesis. In practice, many people use a significance level of 5% and conclude that there is significant evidence against the null hypothesis if the p-value is less than or equal to 0.05. A more conservative approach uses a 1% significance level and conclude that there is significant evidence against the null hypothesis if the p-value is less than 0.01.

In our current example, the p-value is 0.00015 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 much larger, say 0.23, then we would conclude that

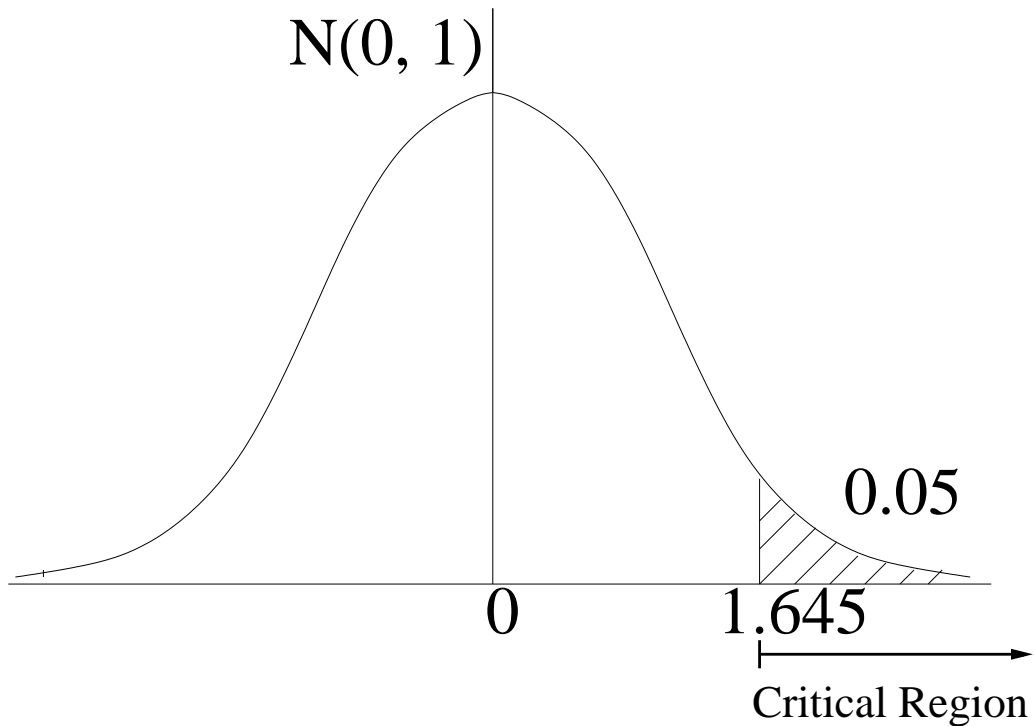
“the evidence against the null hypothesis is not significant at the 5% level”

Another way of saying this is that

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

## Calculating a critical region

Another way of thinking about this test 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.



# Overview of Hypothesis Testing

The previous example involved the following general steps that are common to most, if not all, statistical tests

1. Begin with a **research (alternative) hypothesis** and decide upon a **level of significance** for the test.
2. Set up the **null hypothesis**.
3. Collect a sample of data.
4. Calculate a **test statistic** from the sample of data.
5. 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.

6. 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.

## One and two-tailed tests

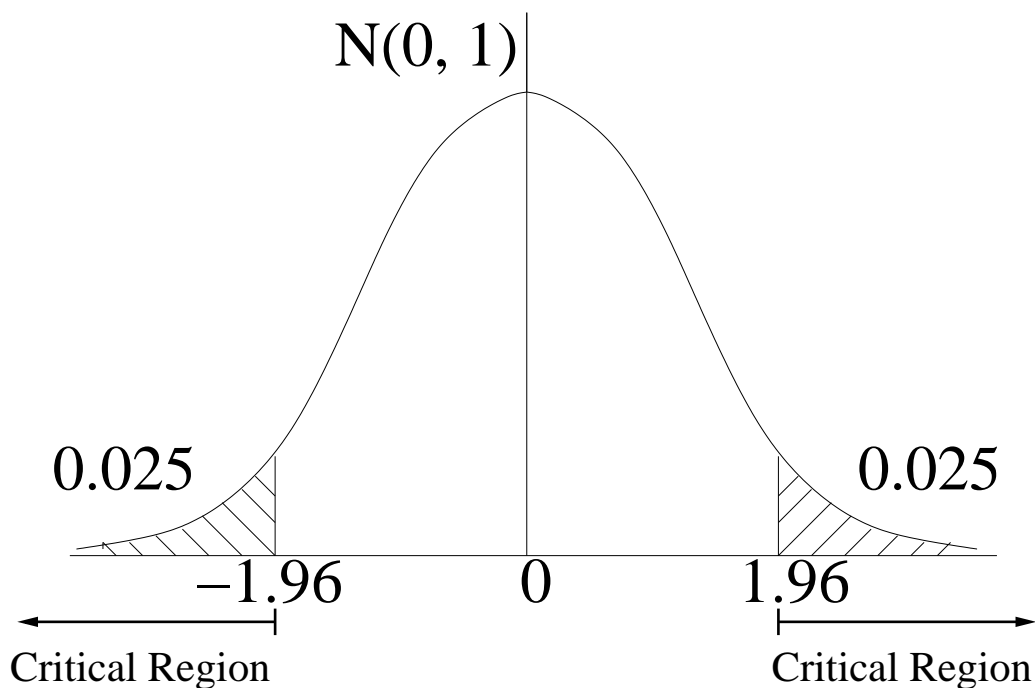
In the previous example we wanted to test the research hypothesis that mean birth weight of Australian babies was greater than 3000g. This suggests that we had some prior information that the mean birth weight of Australian babies was definitely not lower than 3000g. If this were not the case then our research hypothesis would be that the mean birth weight of Australian babies was different from 3000g. This allows for the possibility that the mean birth weight could be less than or greater than 3000g.

In this case we would write our hypotheses as

$$\begin{aligned}H_0 &: \mu = 3000\text{g} \\H_1 &: \mu \neq 3000\text{g}\end{aligned}$$

As before we would calculate our test statistic as 3.66. In this case we allow for the possibility that the mean value is less than 3000g 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.



## Two sample test for a difference between two means ( $\sigma_1$ and $\sigma_2$ known)

Suppose our research hypothesis is that the mean birth weight of boys is greater than mean birth weight of girls. Suppose we know that the standard deviation of boys weights is 500g and the standard deviation of girls weights is 400g. We want to test our research hypothesis using a significance level of 5%

We can test our hypothesis using the steps laid out in Section .

**Step 1** Our research/alternative hypothesis can be written as

$$H_1 : \mu_{boys} > \mu_{girls}$$

and we set our level of significance to be 5%. This dictates that we will carry out a one-tailed test.

**Step 2** We set up our null hypothesis to be directly counter to our research hypothesis

$$H_0 : \mu_{boys} = \mu_{girls}$$

**Step 3** In this example we will assume that we collected the Babyboom dataset. That is we have  $n_{boys} = 26$  boys and  $n_{girls} = 18$  girls.

**Step 4** We base our test statistic on the difference between the sample means of the boys and girls. Under the null hypothesis we know that

$$\begin{aligned}\bar{X}_{boys} &\sim N\left(\mu, \frac{500^2}{26}\right) \\ \bar{X}_{girls} &\sim N\left(\mu, \frac{400^2}{18}\right) \\ \Rightarrow \bar{X}_{boys} - \bar{X}_{girls} &\sim N\left(0, \frac{500^2}{26} + \frac{400^2}{18}\right)\end{aligned}$$

Thus, we can construct a test statistic as

$$Z = \frac{\bar{X}_{boys} - \bar{X}_{girls}}{\sqrt{\frac{500^2}{26} + \frac{400^2}{18}}} \sim N(0, 1)$$

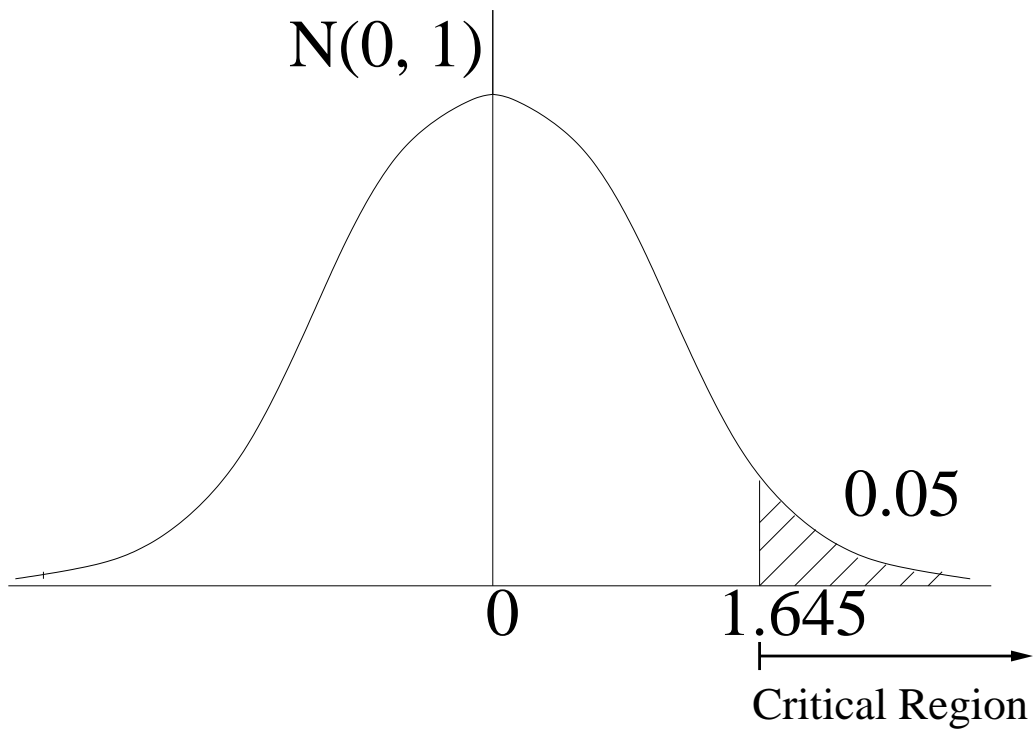
For our dataset we have  $\bar{x}_{boys} = 3375.308$  and  $\bar{x}_{girls} = 3132.444$  so we can calculate the test statistic as

$$Z = \frac{3375.308 - 3132.444}{\sqrt{\frac{500^2}{26} + \frac{400^2}{18}}} = 1.785$$

In general, to test for a difference between two means (with  $\sigma_1$  and  $\sigma_2$  known) from  $n_1$  and  $n_2$  observations from the two groups we use the test statistic

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0, 1)$$

**Step 5** The critical region of the test at the 5% level is  $Z > 1.645$



**Step 6** The test statistic lies in the critical region so we conclude that there is significant evidence against the null hypothesis at the 5% level of significance.

## One sample test for a proportion $p$

Suppose that a university claims to admit equal numbers of state and public school students. We have a research hypothesis that the university tends to admit more public school students so we collect interview 500 first year students and discover that 267 came from public schools. We want to test our hypothesis at the 5% level

First we write down our null and alternative hypotheses regarding the population proportion  $p$  of public school students

$$H_0 : p = 0.5$$

$$H_1 : p > 0.5$$

Using our sample of data we can obtain an estimate of  $p$  as

$$\hat{p} = \frac{267}{500} = 0.534$$

In this situation the test statistic used is

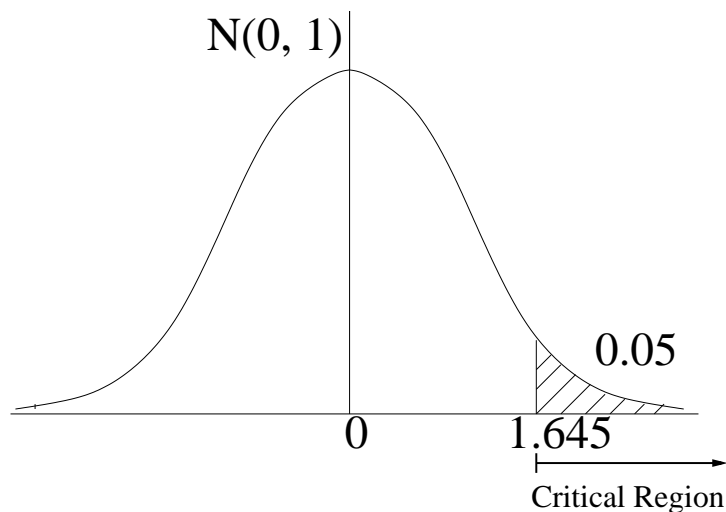
$$Z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}} \sim N(0, 1)$$

where  $p$  is the proportion dictated by the null hypothesis  $H_0$  and  $n$  is the size of our sample.

In our example, the value of the test statistic is

$$Z = \frac{0.534 - 0.5}{\sqrt{\frac{0.5 \times 0.5}{500}}} = 1.52$$

Again the critical region of the test is  $Z > 1.645$



In this example the test statistic does not lie in the critical region so we conclude that the evidence against the null hypothesis is not significant at the 5% level.

## Two sample test for a difference between proportions

Suppose that a newspaper carries out a poll in two cities A and B to ascertain the proportion of people who think the Prime Minister is doing a good job. In city A 336 out of 600 people support the PM. In city B 656 out of 1000 people support the PM. The newspaper would like to test the hypothesis that the proportion of people who think the PM is doing a good job differs between the cities.

First we write down our null and alternative hypotheses regarding the population proportions  $p_1$  and  $p_2$  of PM support in cities A and B

$$H_0 : p_1 = p_2 (= p)$$

$$H_1 : p_1 \neq p_2$$

Using our sample of data we can obtain an estimate of  $p_1$  and  $p_2$  as

$$\hat{p}_1 = \frac{336}{600} = 0.56$$

$$\hat{p}_2 = \frac{656}{1000} = 0.656$$

This is an example of a two-sided test.

In this situation the test statistic used is

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0, 1)$$

where  $n_1$  and  $n_2$  are the two sample sizes and  $p$  is either

(a) known from prior knowledge

(b) estimated from the data as  $\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$

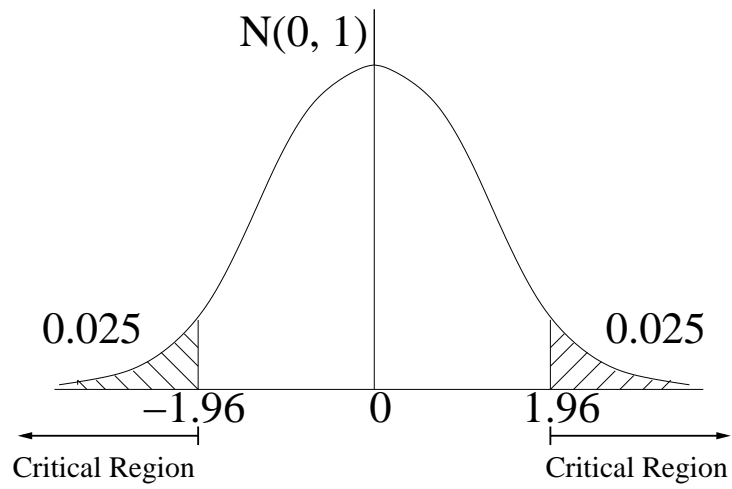
In our example, the value of  $p$  is estimated by

$$\hat{p} = \frac{600 \times 0.56 + 1000 \times 0.656}{600 + 1000} = 0.62$$

The test statistic is

$$Z = \frac{0.56 - 0.656}{\sqrt{0.62 \times (1 - 0.62)\left(\frac{1}{600} + \frac{1}{1000}\right)}} = -3.83$$

The critical region of the test is  $Z < -1.96$  or  $Z > 1.96$



In this example the test statistic lies in the critical region so we conclude that there is significant evidence against the null hypothesis at the 5% level.

## Summary of Z tests

All of the test statistics that we have used in this lecture had a standard Normal distribution. Tests of this type are called Z tests. This section provides a brief summary of the tests we have learnt. This information is repeated in the formula book.

One sample test for a population mean  $\mu$  ( $\sigma$  known)

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

Two sample test for a difference between two means (with  $\sigma_1$  and  $\sigma_2$  known)

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

One sample test for a proportion  $p$

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}}$$

Two sample test for the difference of two proportions

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$p$  is either

(a) known from prior knowledge

(b) estimated from the data as  $\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$