

Introduction to Hypothesis Testing

This deals with an issue highly similar to what we did in the previous chapter. In that chapter we used sample information to make inferences about the range of possibilities for the population statistic. That is, the confidence interval says that “we are pretty sure” that the population parameter is within this range somewhere. In this chapter we use sample information to either refute or fail to refute a given conjecture or hypothesis.

I. Basics of Hypothesis Testing

Typically we begin with some theory or claim about a particular parameter. For example, suppose a hospital administrator claims it requires 20 minutes to perform a given procedure. This claim by the administrator is referred to as the **null hypothesis**. The null hypothesis is typically given the symbol: H_0

$$H_0: \mu = 20$$

That is, we assert that the population parameter is equal to the null. If the null turns out to be false, then some alternative must be specified. The **alternative hypothesis** is specified as:

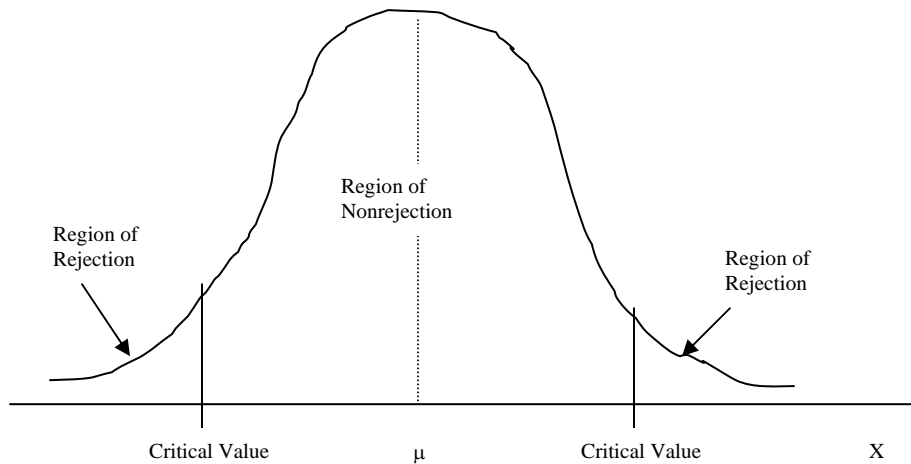
$$H_1: \mu \neq 20$$

This represents the conclusion we would reach if we reject the null. At this point it is possible to reject the null as being too low or too high. This is what is known as a **two-tailed test**, later we will see **one-tailed tests**, where the alternative is in one direction or the other.

Generally the null always refers to some specific value (it has an equal sign in it), whereas the alternative is not specific (it is the opposite of the null)

The way the null hypothesis is tested is to take a sample from the population and compare the sample results to the null. If the sample statistic is “far away” from the null then we conclude that the null must not be true, whereas if the sample statistic is “close” to the null then we conclude that there is no reason to reject the null.

NOTE: we never accept a null hypothesis, we only fail to reject. This is a big deal! If we say we accept the null, this implies that we have proven the null is true. We cannot do this, all we can do is refute the null or fail to refute. Failing to refute is not the same as accepting. Be careful with your wording.



The idea is that the sampling distribution of the test statistic is normally distributed (either because the population is normally distributed or because of the Central Limit Theorem), which means that it looks something like the above. The idea is if we have a population with mean μ then most of our sample means ought to be in the region of nonrejection. If we take a sample and get a mean outside this region then either it is a really odd sample or (more likely) the mean really is not μ .

So to make this decision we first need to determine the **critical value** of the test statistic. This divides the regions.

Type I and Type II errors

There are some risks involved in doing this. That is, there are two kinds of mistakes that we can make when conducting hypothesis testing.

Type I Error – this occurs if the null hypothesis is rejected when in fact it is true. The probability of a type I error is α

α is referred to as the level of significance. Typically this is what is chosen by the researcher. This is generally selected to be .01, .05, or .1. A critical value of .05 is equivalent to a 95% confidence interval.

Type II Error – this occurs if the null hypothesis is not rejected when in fact it is false and should be rejected. The probability of a Type II error is denoted as β .

The probability of a Type II error is dependent on the difference between the hypothesized and actual values of the population parameter.

The **Power of a Test** is denoted as $1-\beta$. This is the complement of a Type II error – it is the probability of rejecting the null hypothesis when, in fact, it is false. Generally statisticians attempt to maximize the power of a test, but as the power of the test increases so also does α . More on this below.

Decision	H ₀ True	H ₀ False
Do Not Reject H ₀	Correct Decision Confidence = $1-\alpha$	Type II Error Prob = β
Reject H ₀	Type I Error Prob = α	Correct Decision Power = $1-\beta$

II. Z tests of hypothesis for the mean -- σ known.

Going back to our example suppose we take a sample of 36 occasions on which the procedure was performed with a mean of 15.2 minutes. If the population standard deviation is 12 minutes, can the administrator’s claim be substantiated?

So:

$$H_0: \mu=20$$

$$H_1: \mu \neq 20$$

1. Using the Critical Value

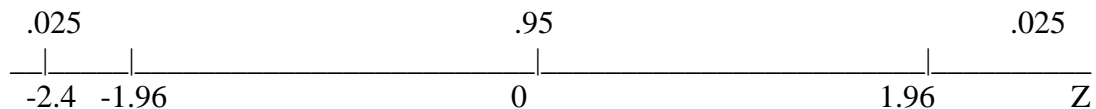
Given that the population standard deviation is known, the Z test statistic is:

$$Z = (X-\mu)/(\sigma/\sqrt{n})$$

For our example: $z = (15.2-20)/(12/6) = -2.4$

Or 15.2 is 2.4 standard errors away from the mean. Is this too unlikely?

It depends on our choice of critical value. If we choose a level of significance of .05 then the size of the rejection region is .05. Since this is a two-tailed test we divide this into parts, .025 each.



using **=normsinv(.025)** we get the Z value of -1.96 . That is there is a .025 probability of getting a value that is 1.96 standard deviations below the mean. Likewise there is a 2.5 percent chance of getting a z value 1.96 standard deviations above the mean. Thus our critical value for Z is 1.96. That is if our Z-statistic is more than 1.96 standard deviations away from the mean (in absolute value) then we say that is too unlikely and we reject our null hypothesis.

In our example, our Z-statistic is -2.4 which is more than 1.96 so we reject our null hypothesis. There is evidence from our sample that the population mean is NOT 20.

Suppose, however, our sample of 36 procedures revealed a mean of 18, how would this change the answer?

Now $Z = 18 - 20/2 = -1$. So now our z-statistic falls within the region of nonrejection. That is, the sample mean is only 1 standard deviation away from the hypothesized population mean. This is not too unlikely. **So we do not reject our null hypothesis. There is not sufficient evidence in our sample to conclude that the average procedure time is not 20 minutes.**

Again, note that we do not conclude that the mean time IS 20 minutes. We only fail to find evidence to disprove it.

2. Using the p-value

An alternative to choosing some cut off and comparing the z-statistic to this cut off, is to simply calculate the probability that one could obtain the given sample mean assuming the null hypothesis is true. For example.

In the above example with a mean of 15.2 we got a z statistic of -2.4 . If we use `=normsdist(2.4)`, we get .992, or there is a .008 probability that we could obtain a sample mean of 15.2 or lower if indeed the population mean is 20. This probability (.0082) is known as the p-value.

The p-value is the probability of obtaining a test statistic equal to or more extreme than the result obtained from the sample data, given that the null hypothesis is true.

The advantage of this approach is that it lets the reader decide what is acceptable. In this case we would likely conclude that this is such a low probability of occurrence that the null hypothesis is probably not true. Note that it is possible, but highly unlikely.

If instead our sample mean was 18, then the z-value of -1 translates into a p-value of .1587, or there is a 15.9 percent chance of getting a sample mean of 18 or lower if the population mean really was 20. This is not too unlikely and so we would fail to reject our null.

Note that both approaches are doing the same thing; we have to decide what is “too unlikely.” Often you will see the p-value reported, these are easy to interpret and do not need much other information to make inferences.

Note that there is a connection between hypothesis testing and confidence intervals. If we construct a 95% confidence interval for the average procedure time:

$$15.2 \pm 1.96(12/\sqrt{36})$$

$$15.2 \pm 3.924$$

or $11.27 < \mu < 19.124$ with 95 percent confidence

Note that since this interval does not contain 20, we conclude with 95% confidence (or with a 5% chance of a Type I error) that the true mean is not 20.

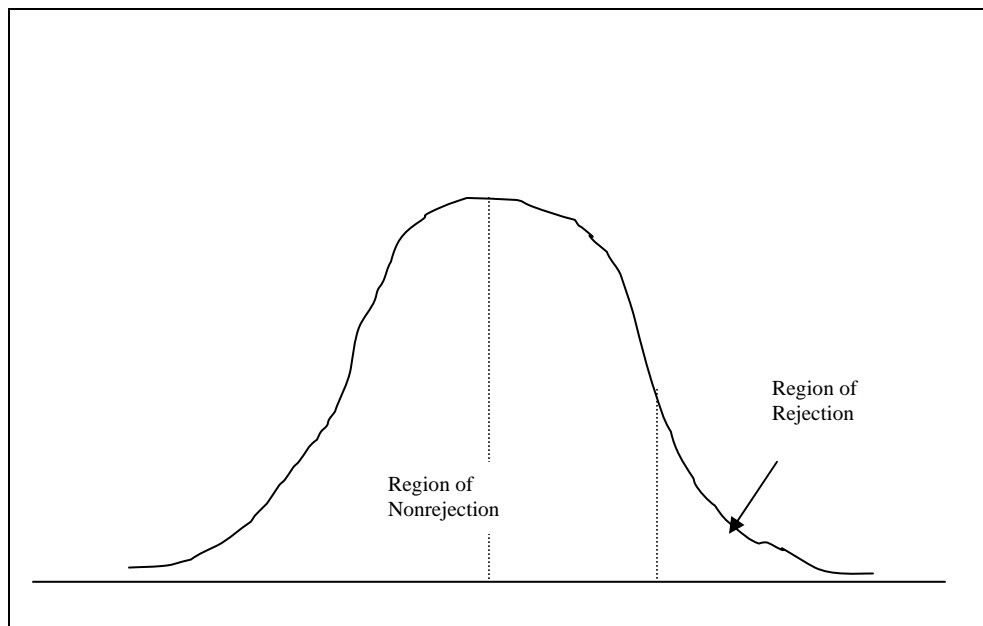
III. One-Tail Tests

Sometimes, the alternative hypothesis focuses in some specific direction. For example, suppose that a third party payer claims that a patient suffering from heart disease should experience a length of stay of no more than 6.5 days. Suppose we review a random sample of 40 cases of heart disease and get a mean of 8 days. If the population standard deviation is 5.2 days what can we conclude about these data at the 5% level of significance?

So here our null and alternative hypotheses are.

$$H_0: \mu \leq 6.5$$

$$H_1: \mu > 6.5$$



Now there is a single region of rejection. We are only concerned with “high” values of the variable. **This is known as a one-tailed test.**

Our z-statistic is the same:

$$Z = 8 - 6.5 / (5.2 / \sqrt{40}) = 1.5 / .822 = 1.82$$

But for our critical value we look up .05 since we are not dividing α into two regions.

So the critical value is 1.645. Thus we would reject our null hypothesis and conclude from this sample that there is evidence that the length of stay is **more than 6.5**.

The p-value associated with this Z is .0344 – there is about a 3.4% chance of getting a sample mean of 8 if the population mean is below 6.5.

IV. t-tests of Hypothesis for the Mean -- σ Unknown.

Generally, we do not know the population standard deviation. Just like we did in Chapter 6, when this happens we can use the sample standard deviation as an estimate for the population standard deviation. We then need to use the t distribution to account for this.

The same example above:

For example, suppose that a third party payer claims that a patient suffering from heart disease should experience a length of stay of no more than 6.5 days. Suppose we review a random sample of 40 cases of heart disease and get a mean of 8 days with a standard deviation of 5.2. What can we conclude about these data at the 5% level of significance?

So here our null and alternative hypotheses are.

$$H_0: \mu \leq 6.5$$

$$H_1: \mu > 6.5$$

$$t = 8 - 6.5 / (5.2 / \sqrt{40}) = 1.5 / .822 = 1.82$$

Now we look under the t-distribution under .05 and 39 d.f. Our critical value is 1.6849. Thus we (barely) reject our null hypothesis and conclude there is evidence that the true mean is greater than 6.5. Notice that the critical value here is slightly larger than the critical value for when σ was known. The logic is that since we are using an estimate for σ , we are a little more careful and need a bigger deviation from the null hypothesis to reach the same conclusion.

We can also calculate the p-value associated with this statistic. Note that the construction of the t-table at the end of the book does not work well in calculating the p-value. But we can do this easily in Excel with the command:

TDIST(x,degrees_freedom,tails)

Where X is your t-statistic, degrees_freedom is the degrees of freedom, and tails 2 for a two tailed test and 1 for a one tailed test. So if I do:
 =Tdist(1.82, 39, 1), excel returns .03822. This is the p-value. There is a 3.8 percent chance of getting the sample mean of 8 or more if the population mean is 6.5. Thus if we use the 5% level of significance we would reject the null and conclude there is evidence the mean is greater than 6.5.

V. Z Test of Hypothesis for the Proportion

This is not much different from tests for the mean. Our Z statistic is:

$$Z = \frac{P_s - P}{\sqrt{\frac{P(1-P)}{n}}}$$

Where $P_s = X/n$ or the proportion of successes in the sample, and P is the hypothesized proportion of successes in the population.

Ex: In 1990 it was determined that 20 percent of high school seniors smoked cigarettes. In a recent survey taken in 2001 of 60 high school seniors it was found that 23 percent of the sample smoked cigarettes. Is there sufficient evidence at the 5% level of significance to conclude that the proportion of high school seniors who smoke has changed?

$H_0: P = .20$

$H_1: P \neq .20$

$$Z = \frac{.23 - .20}{\sqrt{\frac{.2(1-.2)}{60}}} = .03/.052 = .5809$$

This will be within the nonrejection region, so we do not reject H_0 and conclude that there is not evidence that the proportion has changed.

Or the p-value associated with this z-statistic is .28 – that is there is a 28% chance we could get a sample proportion of 23 or more if the population proportion is 20.

VI. Hypothesis test for the correlation coefficient.

Recall that the sample correlation coefficient is:

$$r = \frac{\sum(X - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X - \bar{X})^2} \sqrt{\sum(Y_i - \bar{Y})^2}}$$

This indicates how the variables X and Y move together. Note that r is an estimator for the population correlation coefficient ρ , which is unknown. So if we calculate a correlation coefficient from a sample we would like to be able to make some inferences about the population correlation coefficient.

Suppose we calculate a correlation coefficient between two variables (X and Y) from a sample of 14 observations and obtain $r=.737$. Can we conclude that the population correlation coefficient is different from zero?

Our null and alternative hypothesis are:

$$H_0: \rho = 0$$

$$H_1: \rho \neq 0$$

The test statistic here is

$$t = \frac{r - \rho}{\sqrt{\frac{1 - r^2}{n - 2}}}$$

This follows a t-distribution with $n-2$ degrees of freedom

So in our example:

$$t = \frac{.737 - 0}{\sqrt{\frac{1 - .737^2}{14 - 2}}} = 3.777$$

Using the .05 level of significance the critical value (under 12 degrees of freedom) is 2.1788. So we would reject our null hypothesis and conclude that there is evidence of an association between the variables X and Y.

Note that this could be done as a one tailed test in the same way as before. It depends on how the question is framed.

VII. More on the Power of a Test

Recall, the **Power of a Test**, denoted as $1-\beta$. is the complement of a Type II error – this is the probability of rejecting the null hypothesis when in fact it is false.

Typically the calculation of power is used to plan a study before any data have been obtained.

Assume the standard deviation is known.

Example:

Suppose we want to test the hypothesis that mothers with low socioeconomic status (SES) deliver babies whose birth weights are lower than normal. To test this, a list is obtained of birth weights from 100 consecutive, full-term, live born deliveries from the maternity ward of a hospital in a low SES area. The mean birth weight is found to be 115

with a sample standard deviation of 24 oz. Suppose we know from nationwide surveys based on millions of deliveries that the mean birth weight in the US is 120oz. Assume the pop standard deviation is 24. What is the power of the test assuming the alternative is 115.

To determine this we need to do a power calculation. The power of the study is the probability that we will be able to declare a significant difference with a sample size of 100 if the true mean is 115. Usually we want a power of at least 80% to perform a study.

The Null and Alternative Hypothesis:

$$H_0: \mu = \mu_0$$

$$H_1: \mu = \mu_1 < \mu_0$$

We know from the Z-statistic, that H_0 is rejected if $Z < Z_\alpha$ and H_0 is not rejected if $Z > Z_\alpha$ (recall the Z's will be negative here). The probability of a Type I error is α (rejecting the null when it is true).

The power of the test $(1-\beta)$ can be written as:

$$\text{Power} = P(\text{Reject } H_0 | H_0 \text{ false}) = P(Z < Z_\alpha | \mu = \mu_1)$$

$$= P \left[\frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} < Z_\alpha \mid \mu = \mu_1 \right]$$

$$= P \left[\bar{X} < \mu_0 + Z_\alpha \frac{\sigma}{\sqrt{n}} \mid \mu = \mu_1 \right]$$

If H_1 is true then we know $\bar{X} \sim N(\mu_1, \frac{\sigma^2}{n})$

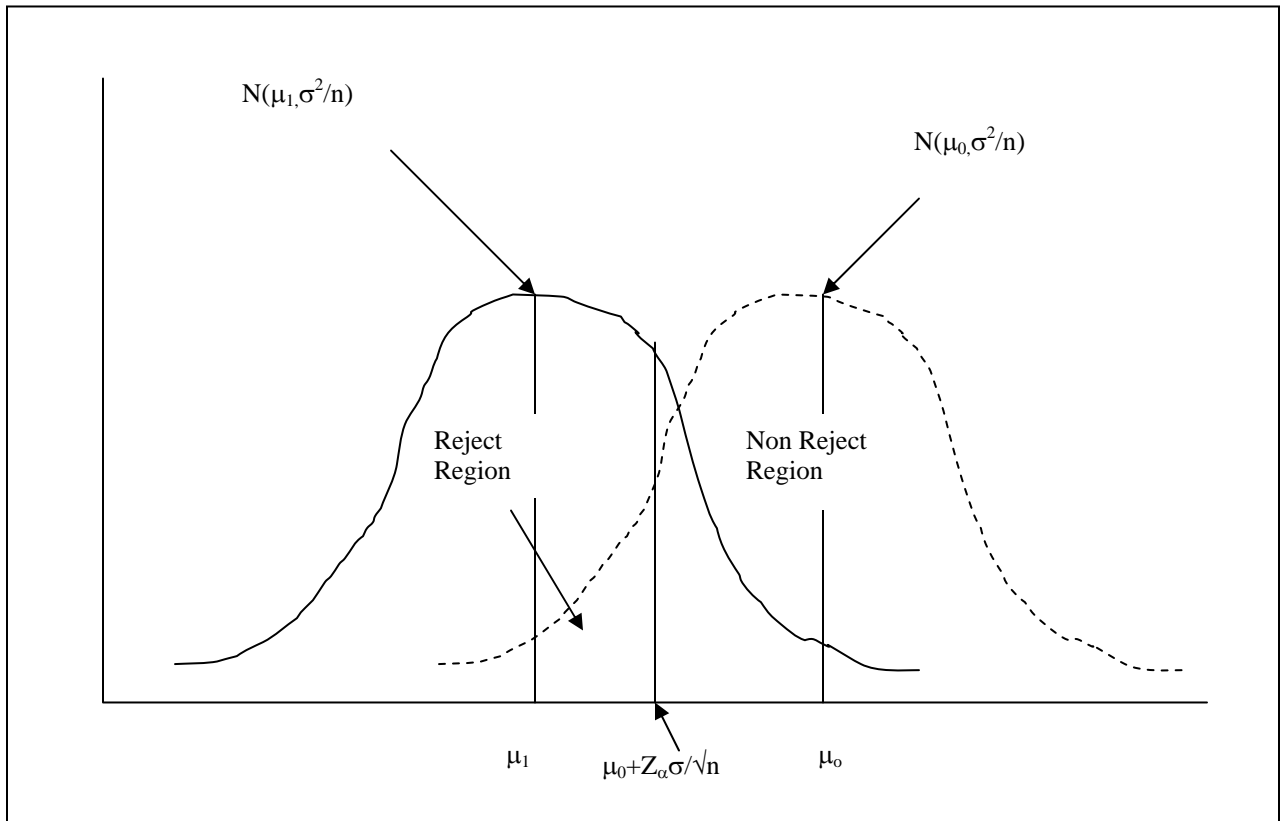
So if we standardize $\mu_0 + Z_\alpha \sigma/\sqrt{n}$ - subtract μ_1 and divide by the standard error. We get:

$$\text{Power} = \Phi \left[(\mu_0 + Z_\alpha \frac{\sigma}{\sqrt{n}} - \mu_1) / \frac{\sigma}{\sqrt{n}} \right]$$

The symbol Φ represents the normal distribution or the normal density function.

Which rearranged becomes:

$$= \Phi \left[Z_\alpha + \frac{(\mu_0 - \mu_1) \sqrt{n}}{\sigma} \right]$$



So if you can figure this out, you are doing great!

The dashed distribution is the distribution assuming the mean is μ_0 the solid distribution assumes μ_1 is the true mean. Under the null (μ_0 is true) we would pick our critical value based on α this is the term $\mu_0 + Z_\alpha \sigma / \sqrt{n}$. The area to the left of this but still under μ_0 distribution is α , the probability of a type I error. But suppose that μ_1 is true. Note now that the area to the left of $\mu_0 + Z_\alpha \sigma / \sqrt{n}$ under the μ_1 distribution is the probability rejecting H_0 when in fact H_0 is false. This is the power of the test. The area to the right of $\mu_0 + Z_\alpha \sigma / \sqrt{n}$ is β (prob of Type II). So to figure out the power of the test we calculate

$$P(X < \mu_0 + Z_\alpha \sigma / \sqrt{n}) = \Phi \left[Z_\alpha + \frac{(\mu_0 - \mu_1) \sqrt{n}}{\sigma} \right]$$

So back to our example,

$$\mu_0 = 120, \mu_1 = 115, \alpha = .05, \sigma = 24, n = 100$$

$$\text{So Power} = \Phi \left[-1.645 + \frac{(120 - 115)}{24} \sqrt{100} \right] = \Phi(-1.645 + (5/24)10) = \Phi(.438)$$

Using NORMSDIST in excel I get prob= .669

Or there is about a 67% chance of detecting a significant difference using a 5% significance level with this sample size.

A standard rule of thumb is to have the power of the test be at least .8, but note that it is difficult to do since we really do not know μ_1 . If the null is not true, what specifically is the alternative? We generally do not know this.

Factors affecting the Power:

1. If the significance level is made smaller (α decreases), Z_α decreases and hence the power decreases – and therefore β increases!
2. If the alternative mean is shifted further away from the null mean, then the power increases.
3. if the standard deviation of the distribution of individual observations increases, the power decreases
4. If the sample size increases then the power increases.

Generally what one does is selects α and then makes the sample size big enough to get the power to some acceptable level. Of course this is all assuming the alternative is known.

VIII. Some Issues in Hypothesis Testing

1. Random Samples – it must be the case that the data being used is obtained from a random sample. If the sample is not random then the results will be suspect. Respondents should not be permitted to self-select (selection bias) for a study, nor should they be purposely selected.
2. One Tail or Two? – if prior information is available that leads you to test the null against an alternative in one direction, then a one tailed test will be more powerful. Previous research or logic often will establish the difference in a particular direction, allowing you to conduct a one-tailed test. If you are only interested in differences from the null, but not the direction of the difference, then the two-tailed test is the appropriate procedure.
3. Choice of Significance α -- typically it is standard to use $\alpha=.05$, but this is somewhat arbitrary. The important thing is to establish α before you conduct your test. Suppose you conduct your test and get a p-value of .075. If you use $\alpha=.05$ you will not reject the null, whereas if you use $\alpha=.1$ then you will reject the null. The danger is to select the level of significance that allows you to make the conclusion you would like. This is where the p-value is useful in that you can simply report the p-value. Alternatively you will often see results presented where they specify some

results are significant at the .05 level while others are only significant at the .10 level, etc.

4. Cleansing Data – Here the danger is throwing out observations you do not like under the grounds that they are “outliers.” In order to throw out observations you must verify that they are not valid, mistakes, incomplete, etc. But just because an observation is “extreme” doesn’t mean you can delete it.