7 Hypothesis Testing

7.1 Introduction

Statistical hypothesis testing is a formal procedure for conducting scientific inquiry according to the scientific method. In the scientific method, a scientist poses a hypothesis based on observations from nature. She then conducts an experiment to test that hypothesis, and the results of that experiment either support or refute the hypothesis. From our mathematical standpoint, a *hypothesis* is a claim about one or more parameters of a population (or more than one population) of interest. For example, we might claim that a parameter equals a certain value or falls within a certain range. We then take a sample from our population and use an appropriate estimator to estimate the parameter of interest. We then use statistical techniques to either support or refute our hypotheses, and to determine the probability that we made the correct decision.

Here are two examples of hypothesis tests.

1. You are a pollster for a major news outlet, and are interested in the preferences of urban vs. rural voters in Pennsylvania. Your hypothesis is that rural voters are less likely to prefer Clinton. The parameter of interest is $p_1 - p_2$, the difference in the proportion of Clinton supporters in the two populations. You then sample registered voters from both populations and construct the estimator $\hat{p}_1 - \hat{p}_2$. Using statistics, you will be able quantify the probability that you correctly accept or reject your hypothesis based on the value of the estimator. In particular, you will be able to determine the probability that you made the *wrong* decision in accepting or rejecting your hypothesis.

2. You are the principal investigator for a trial of a new drug to treat hypertension (high blood pressure). You claim that your new drug will reduce systolic blood pressure by 10 mmHg compared to a placebo pill. The parameter of interest is $\mu_1 - \mu_2$, the difference in mean blood pressure reductions between the mythical populations of all people with hypertension who would receive the drug and all people with hypertension who would receive a placebo. Your hypothesis is that $\mu_1 - \mu_2 \geq 10$. To test this, you will do a double-blind study and randomly assign 100 patients to receive the drug and 100 patients to receive the placebo pill. You will use the estimator $\bar{Y}_1 - \bar{Y}_2$ to test your hypothesis. Using statistics, you will be able quantify the probability that you correctly accept or reject your hypothesis based on the value of the estimator.

Hypothesis testing uses the ideas we learned in the previous sections – sampling, estimators, and confidence levels – to

7.2 Elements of a Hypothesis Test

There are four elements of any hypothesis test. We will go through an example as we discuss them.

Example. Suppose we are interested in the voting preference in Pennsylvania for the 2016 presidential election.

1. We start with a hypothesis that we would like to support. This is called the *alternative hypothesis*. For example, our alternative hypothesis could be "More than 50% of the voters in Pennsylvania support Clinton". In mathematical terms, our population of interest is the registered voters in Pennsylvania, and our parameter of interest is p, the proportion of registered voters who support Clinton. The alternative hypothesis can be stated in terms of p as $p > 0.5$.

2. We will obtain support for our alternative hypothesis by showing (to a specified degree of confidence) that the *null hypothesis*, the converse of the null hypothesis, is false. You may have heard of this as "rejecting the null hypothesis". The null hypothesis in this case is "50% or fewer voters in Pennsylvania support Clinton". Mathematically, we could state this as $p \leq 0.5$ if we like. However, it turns out that this is not a useful way to state the null hypothesis. Recall that we wish to find evidence to *reject* the null hypothesis. The only way to reject the hypothesis $p \leq 0.5$ is for our estimator $\hat{p}$ to produce a value over 0.5. Consider the hypothesis $p = 0.5$. Suppose our estimator $\hat{p}$ is 0.6, and we decide that this is sufficient evidence to reject the hypothesis $p = 0.5$. If we will reject $p = 0.5$, we will certainly also reject $p \leq 0.5$, since if we are willing to reject $p = 0.5$, we are even more willing to reject anything lower. Thus it suffices to take $p = 0.5$ as our null hypothesis.

3. A *test statistic* is something we can actually measure which we will use to either reject or not reject the null hypothesis. In general, the test statistic will be one of our common estimators, such as $\bar{Y}$ or $\hat{p}$, or the equivalent estimators for the difference between two populations. In this case, since we are interested the population proportion p, we will use $\hat{p} = Y/n$ for our estimator, where Y is the number of people out of a sample of size n who prefer Clinton.

4. Armed with a test statistic, the only thing left is to decide when we will reject the null hypothesis. A *rejection region* (RR) specifies the values of the test statistic for which the null hypothesis will be rejected. The RR is given as a range a values. In this case, since we will reject the null hypothesis if $\hat{p}$ is *high*, the rejection region will look like $\hat{p} \geq k$, where k is a threshold we will choose. A lower value of k means that we are more likely to reject the null hypothesis; in particular, we are more likely to reject the null hypothesis when it is in fact true, so we are more likely to commit a false positive error. Conversely, a higher value of k means that we are less likely to reject the null hypothesis; in fact, we are more likely to fail to reject the null hypothesis when in fact the null hypothesis is false, thus we are more likely to commit a false positive error. Thus there is no idea value of k for us to choose. We will discuss later how to pick a k based on the amount of false positives and false negatives we are willing to accept.

We can summarize the elements of any hypothesis test in the following table. Although the null hypothesis is "more fundamental", I list the alternative hypothesis first, since that lets us find the form of the null hypothesis.

Elements of a Hypothesis Test

---

1. Alternative hypothesis, H_a
2. Null hypothesis, H_0
3. Test statistic
4. Rejection region (RR)

Before we go any further, we will take several real-world examples of hypothesis tests, and specify their four parameters. We will discuss how to choose rejection regions and how to quantify our confidence in rejecting the null hypothesis in the next sections. For completeness, the first one will be the example we did above. We will use these examples throughout this section.

1. You are a pollster who is interested in the voting preference in Pennsylvania for the 2016 presidential election. The population of interest is the number of registered voters in Pennsylvania, and the parameter of interest is the p, the proportion of voters who are Clinton supporters.
   (a) Alternative hypothesis, $H_a : p > 0.5$
   (b) Null hypothesis, $H_0 : p = 0.5$
   (c) Test statistic, $\hat{p} = Y/n$
   (d) Rejection region (RR), $\{\hat{p} > k\}$
2. You are the principal investigator for a trial of a new drug to treat hypertension. The populations of interest are all patients with hypertension who receive you drug and all patients with hypertension who receive a placebo. The parameter is interest is the difference in mean blood pressure reduction $\mu_1 - \mu_2$ between these two populations.
   (a) Alternative hypothesis, $H_a : \mu_1 - \mu_2 > 10$
   (b) Null, $H_0 : \mu_1 - \mu_2 = 10$
   (c) Test statistic, $\bar{Y_1} - \bar{Y_2}$
   (d) Rejection region (RR), $\{\bar{Y_1} - \bar{Y_2} > k\}$
3. You are a mechanical engineer and have designed a ball bearing machine which produces ball bearings which are 5 mm in diameter. Since your customers demand precision, the machine needs to be recalibrated if the average ball bearing diameter deviates from 5 mm by more than 1 percent. You suspect there might be something wrong with the machine. The population of interest is all ball bearings produces by your machine. The parameter of interest is μ, the average ball bearing diameter.

(a) Alternative hypothesis, $H_a : \mu \neq 5$

(b) Null hypothesis, $H_0 : \mu = 5$

(c) Test statistic, $\bar{Y}$

(d) Rejection region (RR), $|\bar{Y} - 5| > k$

Note that there is something unusual about the third example. For the first two examples, the alternative hypothesis was that the parameter of interest is *above* a certain value. These are known as upper-tail hypothesis tests, since we will reject the null hypothesis if the test statistic falls in the upper tail of the appropriate probability distribution. Similarly, we could have a lower-tail hypothesis test if the alternative hypothesis is that the parameter of interest is *below* a certain value. In this case, the null hypothesis would be rejected if the test statistic falls in the lower tail of the appropriate probability distribution.

The final example is known as a *two-tailed hypothesis test*. The alternative hypothesis is that the parameter of interest is not equal to some value, i.e. that it is either above or below that value. The null hypothesis, as always, is that the parameter of interest is equal to a specific value (in this case, we do not require an additional argument to show that this makes sense). The rejection region is stated as $|\bar{Y} - 5| > k$, indicating that we will reject the null hypothesis if the test statistic deviates from a specific value by more than a certain amount.

We will first discuss hypothesis tests where the sample we take is large, then we will consider the case where the sample is small. Before we do that, we will define the two types of error which can result from a hypothesis test.

Types of Error

---

1. A *type I error* is made if the null hypothesis is rejected when in fact the null hypothesis is true. The probability of making a type I error is denoted by α:

$$\begin{aligned}\alpha &= \mathbb{P}(\text{reject null hypothesis when null hypothesis is true}) \\ &= \mathbb{P}(\text{test statistic lies in rejection region when null hypothesis is true})\end{aligned}$$

2. A *type II error* is made if the null hypothesis is accepted when the alternative hypothesis is true. The probability of making a type II error is denoted by β:

$$\begin{aligned}\beta &= \mathbb{P}(\text{accept null hypothesis when alternative hypothesis is true}) \\ &= \mathbb{P}(\text{test statistic lies outside rejection region when alternative hypothesis is true})\end{aligned}$$

Some statisticians (especially biostatisticians) will use the term *power*, which is defined as $1 - \beta$.

7.3 Large Sample Hypothesis Tests

In this section, we will test hypotheses about the mean μ or proportion p of a population. This includes the cases where we are interested in the difference between two populations. We will take a large enough sample that we can assume (by the central limit theorem) that the test statistic is normally distributed. Thus we can use the Z distribution (standard normal) in our computations.

Let's look at one-tailed hypothesis tests first. We will do an upper-tailed test, but this is similar for lower-tailed tests. Suppose we have a parameter of interest θ (e.g. μ, p, or the equivalent for the difference of two populations). We wish to test the alternative hypothesis $\theta > \theta_0$ (upper-tailed test), so the null hypothesis is $\theta = \theta_0$. The test statistic is the estimator $\hat{\theta}$, where we choose the appropriate estimator based on the parameter of interest. Since this is an upper-tail test, the rejection region will be of the form $\hat{\theta} > k$, where k will be chosen later. Thus the parameters for the test are:

1. Alternative hypothesis, $H_a : \theta > \theta_0$
2. Null hypothesis, $H_0 : \theta = \theta_0$
3. Test statistic, $\hat{\theta}$
4. Rejection region (RR), $\hat{\theta} > k$

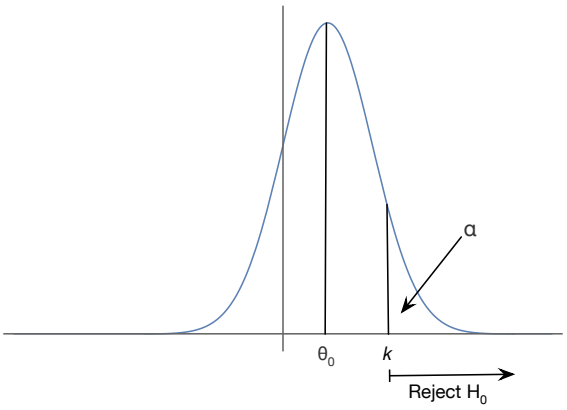
To determine the parameter k in the rejection region, we fix the level α of type I error which we are willing to accept. If the null hypothesis is true, then the estimator $\hat{\theta}$ has a normal distribution with mean θ_0 and standard deviation $\sigma_{\hat{theta}}$. Since α is the probability of incorrectly rejecting the null hypothesis when it is in fact true, we want to choose the rejection region such that the probability that $\hat{\theta} \geq k$ to be α. Converting to the standard normal random variable, we want:

$$\begin{aligned} \mathbb{P}(\hat{\theta} \geq k) &= \alpha \\ \mathbb{P}\left(\frac{\hat{\theta} - \theta_0}{\sigma_{\hat{\theta}}} \geq \frac{k - \theta_0}{\sigma_{\hat{\theta}}}\right) &= \alpha \\ \mathbb{P}\left(Z \geq \frac{k - \theta_0}{\sigma_{\hat{\theta}}}\right) &= \alpha \end{aligned}$$

Looking at the Z table, we choose z_α such that $\mathbb{P}(Z \geq z_\alpha) = \alpha$. (On our Z table, this is equivalent to $\mathbb{P}(Z \leq -z_\alpha) = \alpha$). Thus we have:

$$\begin{aligned} \frac{k - \theta_0}{\sigma_{\hat{\theta}}} &= z_\alpha \\ k &= \theta_0 + z_\alpha \sigma_{\hat{\theta}} \end{aligned}$$

This is shown in the diagram below:



Recall that $\sigma_{\hat{\theta}}$ is the standard deviation of the estimator, which is computed from the the population standard deviation and the sample size. If we do not know the population standard deviation, we can use the sample standard deviation S in place of the population standard deviation σ since the sample size is large.

Example. You are a pollster who is interested in the voting preference in Pennsylvania for the 2016 presidential election. The population of interest is the number of registered voters in Pennsylvania, and the parameter of interest is the p, the proportion of voters who are Clinton supporters. Suppose you sample 100 voters and 60 of them favor Clinton. Does the evidence support Clinton being favored at a level of 0.05?

The parameters of the test are given above. The test statistic is $\hat{p} = 0.6$. To find the appropriate rejection region at a level of $\alpha = 0.05$ we need to find the appropriate value of z_α from the Z table. Looking at the table, we find that $z_\alpha = 1.64$ (we could also have chosen 1.65). We do not know the population standard deviation, but we can estimate it by using $\hat{p}$ in place of the true value p:

$$\sigma_{\hat{\theta}} = \sqrt{\frac{p(1-p)}{n}} \approx \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{(0.6)(0.4)}{100}} = 0.049$$

From this we calculate:

$$k = 0.5 + z_\alpha \sigma_{\hat{\theta}} = 0.5 + 1.64(0.049) = 0.58$$

Thus the rejection region is

$$\{\hat{p} \geq 0.58\}$$

Since our test statistic is 0.6, it falls inside the rejection region, thus we can reject our null hypothesis with a level of 0.05, so we are 95% confident that Clinton is favored in the population at large.

Similarly, we can do this for a two-tailed hypothesis test. Here we "split" the α between the

upper and lower tails of the normal distribution.

$$\begin{aligned} \mathbb{P}(|\hat{\theta} - \theta_0| \geq k) &= \alpha \\ \mathbb{P}\left(\left|\frac{\hat{\theta} - \theta_0}{\sigma_{\hat{\theta}}}\right| \geq \frac{k}{\sigma_{\hat{\theta}}}\right) &= \alpha \\ \mathbb{P}\left(|Z| \geq \frac{k}{\sigma_{\hat{\theta}}}\right) &= \alpha \end{aligned}$$

Splitting the α evenly between the upper and lower tails, we want

$$\mathbb{P}\left(Z \leq -\frac{k}{\sigma_{\hat{\theta}}}\right) = \alpha/2$$

and

$$\mathbb{P}\left(Z \geq \frac{k}{\sigma_{\hat{\theta}}}\right) = \alpha/2$$

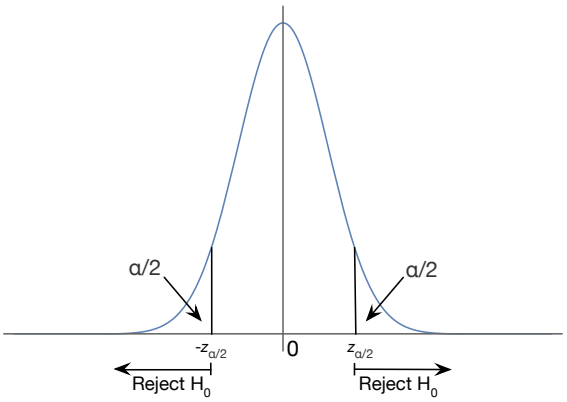
In the same way as before, we can use the Z table to find the appropriate value of $z_{\alpha/2}$. Thus we have:

$$\begin{aligned} \frac{k}{\sigma_{\hat{\theta}}} &= z_{\alpha/2} \\ k &= z_{\alpha/2}\sigma_{\hat{\theta}} \end{aligned}$$

Our rejection region is:

$$\{|\hat{\theta} - \theta_0| \geq z_{\alpha/2}\sigma_{\hat{\theta}}\}$$

The diagram below shows the rejection region for a two-tailed test in terms of the standard normal Z distribution.



Let's do our ball bearing example.

Example. You are a mechanical engineer and have designed a ball bearing machine which produces ball bearings which are 5 mm in diameter. You suspect there might be something wrong with the machine. You sample 64 ball bearings from the machine, and obtain a sample mean of 4.98 mm and a sample standard deviation of 0.1 mm. Can you conclude at a level of 0.95 that there is something wrong with the machine?
The parameters of the test are given above. The test statistic is $\bar{Y} = 4.98$. To find the appropriate rejection region at a level of $\alpha = 0.05$, since this is a two-sided hypothesis test, we need to find the appropriate value of $z_\alpha/2$ from the Z table. Looking at the table, we find that $z_\alpha = 1.96$. We do not know the population standard deviation, but we can estimate it by using S in place of σ. The standard deviation of the estimator is therefore approximately:

$$\sigma_{\bar{Y}} \approx \frac{S}{\sqrt{n}} = \frac{0.1}{\sqrt{64}} = \frac{0.1}{8} = 0.0125$$

Thus our value of k is:

$$k = z_{\alpha/2}\sigma_{\bar{Y}} = 1.96(0.0125) = 0.0245$$

The rejection region is therefore:

$$\{|\bar{Y} - 5| \geq 0.0245\} = \{\bar{Y} \leq 4.9755 \text{ or } \bar{Y} \geq 5.0245\}$$

Since out test statistic does not fall within our rejection region, we do not reject the null hypothesis, so for now you conclude that you do not have to do maintenance on the machine.

7.4 Large Sample Hypothesis Tests and Type II Error

We have discussed how to choose the rejection region based on the level of type I error we are willing to accept. Now we will discuss type II error. We will only quantify type II error for one-tailed hypothesis tests. For two-tailed tests, this process is arduous, thus will be omitted. The discussion below will concern upper-tail hypothesis tests. Lower-tail tests are similar, except everything is "flipped".

Recall that a type II error is made if we accept the null hypothesis when in fact the null hypothesis is false. In other words, our test statistic falls *outside* the rejection region, even though the null hypothesis is false and should be rejected. When evaluating type II error, we need to make an additional decision. We need to specify a *specific* value of the alternative hypothesis to be a *positive test threshold* (my own term) which we will use to compute β.

Let's use the same upper-tail setup as before. Suppose we have a parameter of interest θ. We wish to test the alternative hypothesis $\theta > \theta_0$ (upper-tailed test), so the null hypothesis is $\theta = \theta_0$. The test statistic is the estimator $\hat{\theta}$, and the rejection region is of the form $\{\hat{\theta} > k\}$. For now, assume we have chosen k according to our desired α using the methods of the previous section. To reiterate, our test has the following parameters:

1. Alternative hypothesis, $H_a : \theta > \theta_0$
2. Null hypothesis, $H_0 : \theta = \theta_0$

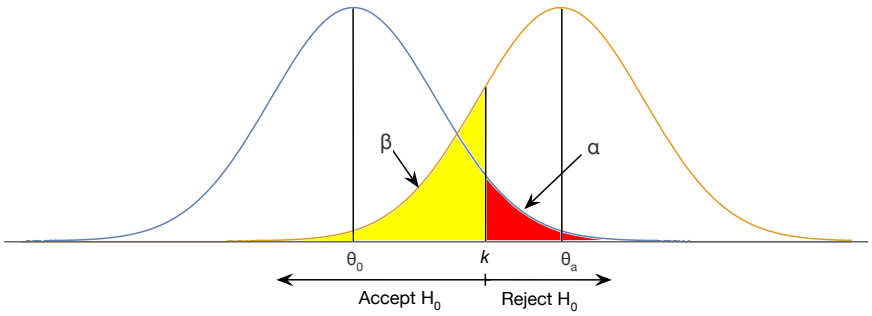
3. Test statistic, $\hat{\theta}$

4. Rejection region (RR), $\{\hat{\theta} > k\}$

We will choose a specific value θ_a for the alternative hypothesis (our positive test threshold). Since this is an upper tail test, we must have $\theta_a > \theta_0$. Then β, the probability of a type II error, is defined as the probability of accepting the null hypothesis when the true value of the parameter is in fact θ_a.

It seems odd that we must specify a value for θ_a in order to do this, but this has a nice interpretation. The probability of accepting the null hypothesis if the true parameter is θ_a is β. If the true value of θ is greater than θ_a, the probability of accepting the null hypothesis is less than β. Thus if the true value of θ is θ_a or greater, the probability that we will incorrectly accept the null hypothesis is at most β.

Now that we've gotten that out of the way, how we we actually find β? Take a look a the following picture:



β is the probability that the test statistic lies outside the rejection region when the true value of the parameter is θ_a. This time, the estimator has a normal distribution with mean θ_a and standard deviation $\sigma_{\hat{\theta}}$. Thus, for the upper tail test here, we want the probability that our test statistic is less than k, the threshold of the rejection region.

$$\begin{aligned}\beta &= \mathbb{P}(\hat{\theta} \leq k) \\ &= \mathbb{P}\left(\frac{\hat{\theta} - \theta_a}{\sigma_{\hat{\theta}}} \leq \frac{k - \theta_a}{\sigma_{\hat{\theta}}}\right) \\ &= \mathbb{P}\left(Z \leq \frac{k - \theta_a}{\sigma_{\hat{\theta}}}\right)\end{aligned}$$

Since we know θ_a and $\sigma_{\hat{\theta}}$, and since we have already computed the boundary of the rejection region k, we can use the Z table to solve for k
Let's go back and do our polling example from above.

Example. You are again a pollster who is interested in the voting preference in Pennsylvania for the 2016 presidential election. The population of interest is the number of registered voters in Pennsylvania, and the parameter of interest is the p, the proportion of voters who are Clinton supporters. You sample 100 voters and 60 of them favor Clinton. You create a hypothesis test with $\alpha = 0.05$, for which we have seen that the rejection region is $\{\hat{p} \geq 0.58\}$. For a value of the alternative hypothesis $p_a = 0.60$, calculate β, the probability of a type II error.

Given our discussion above, plugging in the values we determined in the example above ($k = 0.58$, $\sigma_{\hat{\theta}} = 0.049$), and using $p_a = 0.60$:

$$\begin{aligned}\beta &= \mathbb{P}\left(Z \leq \frac{k - p_a}{\sigma_{\hat{p}}}\right) \\ &= \mathbb{P}\left(Z \leq \frac{0.58 - 0.60}{0.049}\right) \\ &= \mathbb{P}(Z \leq -0.41) \\ &= 0.3409\end{aligned}$$

7.5 Sample Size Selection

If you are a scientist devising a hypothesis test, you don't want to just run the test and calculate β after the fact the way we did in the example above. If you did that, β might be too large, and your test could be meaningless! You would like a procedure where, if you specify the maximum values of α and β you are willing to tolerate, you can compute the size of the sample n you need to attain these values. Along the way, you will also compute k, the threshold for the rejection region. The following derivation is for an upper-tail hypothesis test. A similar derivation will work for a lower-tail test. The two-tailed test will not be discussed.

For simplicity, we will do the computation for a hypothesis test involving the sample mean. We can similarly do this for a hypothesis test involving the sample proportion. As an experimenter, you are conducting an upper tail hypothesis test with the following four parameters:

1. Alternative hypothesis, $H_a : \mu > \mu_0$
2. Null hypothesis, $H_0 : \mu = \mu_0$
3. Test statistic, $\bar{Y}$
4. Rejection region (RR), $\{\bar{Y} > k\}$

In addition, you need to choose three more parameters:

1. θ_a, a specific value of the alternative hypothesis. As before, we require $\theta_a > \theta_0$
2. α, the maximum type I error you are willing to accept

3. β, the maximum type II error you are willing to accept

The population standard deviation is denoted σ. If we do not know σ, we can estimate it by the sample standard deviation S. In the real world, to obtain S, we can do a pilot study for the express purposes of computing S. We just need to make sure that the pilot study is large enough that the S we obtain is a good estimator for σ.

As above, we will write the appropriate equations for α and β. Since we will have two equations and only two unknowns (k and n), we can solve them for the unknowns. Using the definition of α and referring to the picture above:

$$\begin{aligned}
\alpha &= \mathbb{P}(\text{test statistic is in rejection region when null hypothesis is true}) \\
&= \mathbb{P}(\bar{Y} \geq k \text{ when } \mu = \mu_0) \\
&= \mathbb{P}\left(\frac{\bar{Y} - \mu_0}{\sigma/\sqrt{n}} \geq \frac{k - \mu_0}{\sigma/\sqrt{n}}\right) \\
&= \mathbb{P}(Z \geq z_\alpha)
\end{aligned}$$

where z_α is chosen from our Z-table so that $\mathbb{P}(Z \geq z_\alpha) = \alpha$. Using the definition of β and again referring to the picture above:

$$\begin{aligned}
\beta &= \mathbb{P}(\text{test statistic is outside rejection region when alternative hypothesis is true}) \\
&= \mathbb{P}(\bar{Y} \leq k \text{ when } \mu = \mu_a) \\
&= \mathbb{P}\left(\frac{\bar{Y} - \mu_a}{\sigma/\sqrt{n}} \leq \frac{k - \mu_a}{\sigma/\sqrt{n}}\right) \\
&= \mathbb{P}(Z \leq -z_\beta)
\end{aligned}$$

where z_β is chosen from our Z-table so that $\mathbb{P}(Z \leq -z_\beta) = \beta$. We use the negative sign for convenience, since the value of z we are looking for will always fall to the left of 0 on a graph of the standard normal distribution. We then have two equations we can solve simultaneously:

$$\begin{aligned}
\frac{k - \mu_0}{\sigma/\sqrt{n}} &= z_\alpha \\
\frac{k - \mu_a}{\sigma/\sqrt{n}} &= -z_\beta
\end{aligned}$$

If we solve both equations for k, we get:

$$\begin{aligned}
k &= \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}} \\
k &= \mu_a - z_\beta \frac{\sigma}{\sqrt{n}}
\end{aligned}$$

We can use the first equation in the set above to find k, the boundary of the rejection region. This is the same equation we derived in the section on large sample hypothesis tests. Setting

these two equations equal to each other:

$$\begin{aligned}
\mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}} &= \mu_a - z_\beta \frac{\sigma}{\sqrt{n}} \\
(z_\alpha + z_\beta)\frac{\sigma}{\sqrt{n}} &= \mu_a - \mu_0 \\
\sqrt{n} &= \frac{(z_\alpha + z_\beta)\sigma}{\mu_a - \mu_0} \\
n &= \frac{(z_\alpha + z_\beta)^2 \sigma^2}{(\mu_a - \mu_0)^2}
\end{aligned}$$

7.6 p-values

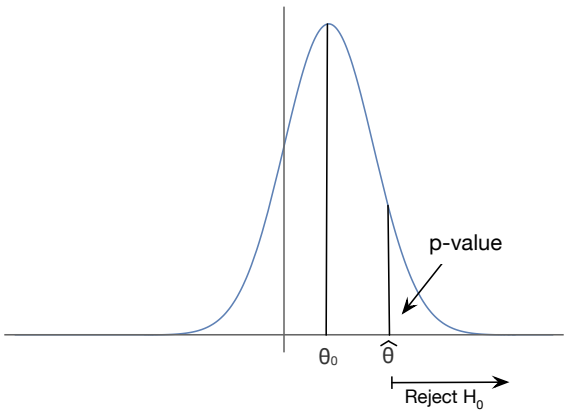
The parameter α is the probability of making a type I error, i.e. rejecting the null hypothesis when it is in fact true. Generally, experimenters choose small values of α to maximize their confidence that a positive result is not a false positive, i.e. due to chance alone. Typical values of α you will see in the scientific literature are 0.05 and 0.01, although this is a little arbitrary.

Another way of reporting the probability of a type I error is the p-value. This is what is most often given in the scientific literature.

p-value

The p-value, or attained significance level, of a test is the smallest level of significance α for which the test statistic indicates that the null hypothesis should be rejected.

In other words, for any α greater than or equal to the p-value, the null hypothesis can be rejected using the test statistic. For α less than the p-value, the null hypothesis cannot be rejected. This provides more information than just saying that the null hypothesis was rejected for a certain value of α. The following picture may be useful. It shows an upper-tail hypothesis test, but this can be computed for any hypothesis test. The null hypothesis is given by θ_0, and the test statistic is $\hat{\theta}$.



Let's look at our voter polling example from before.

Example. You are a pollster who is interested in the voting preference in Pennsylvania for the 2016 presidential election. The population of interest is the number of registered voters in Pennsylvania, and the parameter of interest is the p, the proportion of voters who are Clinton supporters. Suppose you sample 100 voters and 60 of them favor Clinton. What is the p-value for this test? (It is a little confusing that p is the sample proportion and we also have a p-value for the test, but this notation is standard in both cases; we will write the p-value as "p-value" rather than just p to avoid this confusion).

We want the smallest value of α for which we will reject the null hypothesis. The value we obtained for $\hat{p}$ is 0.6. We estimated the standard deviation of the estimator $\hat{p}$ as 0.049 above (this include the factor of $\sqrt{n}$ in the denominator). Thus we have:

$$\begin{aligned} p\text{-value} &= \mathbb{P}(\hat{p} \geq 0.6 \text{ given null hypothesis is true}) \\ &= \mathbb{P}(\hat{p} \geq 0.6 \text{ given } p = 0.5) \\ &= \mathbb{P}\left(\frac{\hat{p} - 0.5}{\sigma_{\hat{p}}} \geq \frac{0.6 - 0.5}{\sigma_{\hat{p}}}\right) \\ &= \mathbb{P}\left(Z \geq \frac{0.1}{0.049}\right) \\ &= \mathbb{P}(Z \geq 2.04) \\ &= 0.0207 \end{aligned}$$

Thus we have a p-value of 0.0207, which is the smallest value of α for which the test statistic indicates that we can reject the null hypothesis. Above, we showed that we can reject the null hypothesis with an α of 0.05. This is consistent with our p-value, since 0.05 is greater than the p-value.

7.7 Small sample hypothesis testing for the population mean

In the previous section, we discussed hypothesis testing procedures for large samples. The large sample assumption is important because that guarantees that the test statistic has a normal distribution (by the central limit theorem). The large sample size also allows us to estimate the population standard deviation using the sample standard deviation.

What happens when the sample size is not large? As in the section on confidence intervals, if we have a population that is normally distributed with unknown standard deviation, we can use the t-distribution in place of the Z distribution. Suppose we have a normally-distributed population, and we take a small sample ($n < 30$) from that population. The population standard deviation is unknown, so we estimate it with the sample standard deviation S. Suppose we are are doing an upper-tailed hypothesis test. Let $\mu = \mu_0$ be the null hypothesis, and $\mu > \mu_0$ be the alternative hypothesis. The test statistic is $\bar{Y}$, the standard unbiased estimator for μ. Then to find the appropriate rejection region based on our desired level α, we follow the same procedure above except that we use the t distribution with $n - 1$ df in

place of the Z distribution. This works because

$$T = \frac{\bar{Y} - \mu_0}{S/\sqrt{n}}$$

has a t distribution with $n - 1$ df. For example, if we are doing an upper-tail test with desired level α, then the rejection region is given by:

$$\begin{aligned} k &= \mu_0 + t_\alpha \sigma_{\bar{Y}} \\ &= \mu_0 + t_\alpha \frac{S}{\sqrt{n}} \end{aligned}$$

All we did to get this formula was replace the z with a t. We will return to our rocket science example from the section on confidence intervals.

Example. You are a rocket scientist, and you conduct an experiment which involves measuring the launch velocity of a model rocket. You claim that the launch velocity of the model rocket is more than 29 m/s. Suppose 8 measurements are taken. The sample mean is 29.59 m/s, and the sample standard deviation is 0.391 m/s. Is the claim supported at the 0.025 level of significance?

This is an upper-tailed hypothesis test with the following parameters:

1. Alternative hypothesis, $H_a : \mu > 29$
2. Null hypothesis, $H_0 : \mu = 29$
3. Test statistic, $\bar{Y}$
4. Rejection region (RR), $\{\bar{Y} > k\}$

We assume that the launch velocities are normally distributed (a normally distributed population is essential for use to use the t-distribution). For the rejection region, we have:

$$\begin{aligned} k &= \mu_0 + t_\alpha \sigma_{\bar{Y}} \\ &= 29 + t_\alpha \frac{S}{\sqrt{n}} \\ &= 29 + t_{0.025} \frac{0.391}{\sqrt{8}} \\ &= 29 + 2.365(0.138) \\ &= 29.326 \end{aligned}$$

The rejection region is $\{\bar{Y} \geq 29.326\}$. Since our measurement lies in the rejection region, we can reject the null hypothesis with a level of 0.025. Thus the claim is supported at the 0.025 level of significance.

7.8 Power of Hypothesis Tests

In the previous sections, we have discussed large-sample and small-sample hypothesis tests for various test statistics. We learned how to specify a rejection region for a desired α and how to compute β for a specific value of the alternative hypothesis. How did we decide on those test statistics, and how do we know we selected the best rejection region. In other words, how "good" are the tests?

So far, we have used the parameters α and β, the probabilities of type I and type II error, to measure the "goodness" of a hypothesis test. Recall that to compute β we had to choose a specific value for the alternative hypothesis θ_a. We would like a function which gives us the error of the test given the true value of the parameter. The function we use is called the power of a hypothesis test.

Power of a Hypothesis Test

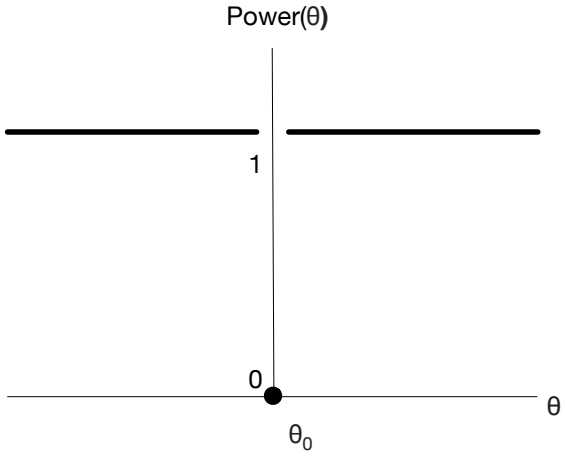
Suppose we have a hypothesis test, with test statistic $\hat{\theta}$ and rejection region RR. Then the *power* of the hypothesis test, denoted $Power(\theta)$ is the probability that the test statistic $\hat{\theta}$ will lie in the rejection region if the true parameter value is θ, i.e.

$$Power(\theta) = \mathbb{P}(\hat{\theta} \text{ lies in RR when true parameter value is } \theta)$$

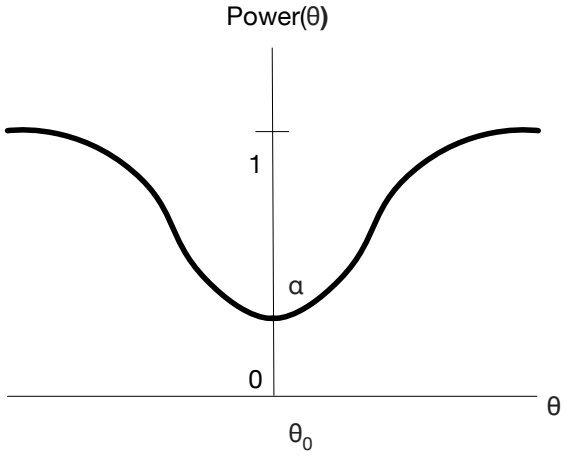
The power function relates α and β in the following way. Suppose our null hypothesis is θ_0, and we have chosen the specific value θ_a for our alternative hypothesis. Let α be the probability of a type I error, and let $\beta(\theta_a)$ be the probability of a type II error when the true value is θ_a. Then we observe the following:

1. $Power(\theta_0) = \alpha$ since $Power(\theta_0)$ is the probability of rejecting the null hypothesis when it is true.
2. $Power(\theta_a) = 1 - \beta(\theta_a)$, since $\beta(\theta_a)$ is the probability of accepting the null hypothesis when θ_a is true.

For an ideal test, the power function would be 0 at θ_0 and 1 for all possible values of the alternative hypothesis θ_a. Here is what this would look like graphically for a two-tailed hypothesis test.



No hypothesis test, however, is perfect. Realistically, the power curve for a two-tailed hypothesis test will look more like this:



What we would like is a test which maximizes the power function for a given α. This is called the *most powerful α-level test*. Before we state the condition under which we have such a test, we need one more set of definitions.

Simple and Composite Hypotheses

Suppose we have a random sample taken from a population with parameter θ, and consider a hypothesis specifying the value of θ. If the hypothesis uniquely specifies the distribution of the population, we call our hypothesis a *simple hypothesis*. Otherwise, we call it a *composite hypothesis*.

Let's look at examples of simple and composite hypotheses.

Example. 1. Suppose we have a population which is normally distributed with mean μ and variance 1. Since the distribution of the population is uniquely specified by μ, a hypothesis involving μ such as the null hypothesis $\mu_0 = 0$, is a simple hypothesis.

2. Suppose we have a population which is normally distributed with mean μ and unknown variance σ^2. Here we need both μ and σ_2 to specify the distribution of the population, so a hypothesis involving μ such as the null hypothesis $\mu_0 = 0$, is a composite hypothesis.

3. Suppose we have a population with is exponentially distributed with parameter λ. Then a hypothesis involving λ is a simple hypothesis since λ uniquely specifies the distribution of the population. Since the population mean $\mu = 1/\lambda$, a hypothesis involving the mean μ is also a simple hypothesis.

Consider a hypothesis test where we are testing a simple null hypothesis $\theta = \theta_0$ against a simple alternative hypothesis $\theta = \theta_a$. We would like to choose a rejection region such that:

1. $Power(\theta_0) = \alpha$
2. $Power(\theta_a)$ is as large as possible

In other words, we are looking for the most powerful α-level test. The following theorem tells us how to derive the most powerful α-level test in this case.

Neyman-Pearson Lemma

Suppose we have a population whose distribution parameterized by θ. Suppose we have a hypothesis test in which we wish to test the simple null hypothesis $H_0 : \theta = \theta_0$ against the simple alternative hypothesis $H_a : \theta = \theta_a$. We do this by taking a random sample $Y_1, \ldots, Y_n$ drawn from the population. Let $L(Y_1, \ldots, Y_n|\theta)$ be the likelihood function for the sample when the value of the parameter is θ. Then, for a given α, the test that maximizes the power for the alternative hypothesis θ_a has a rejection region given by:

$$\frac{L(Y_1, \ldots, Y_n|\theta_0)}{L(Y_1, \ldots, Y_n|\theta_a)} < k$$

where the value of k is chosen so that the test has the desired α. The ratio of likelihood functions is called a *likelihood ratio*. Such a test is the most powerful α-level test for H_0 versus H_a.

Let's look at an example application of this theorem.

Example. Suppose Y is a single observation from a population parameterized by θ with probability density function

$$f_\theta(y) = \begin{cases} \theta y^{\theta-1} & 0 < y < 1 \\ 0 & \text{otherwise} \end{cases}$$

Find the most powerful hypothesis test with $\alpha = 0.05$ to test the null hypothesis $H_0 : \theta = 2$ against the alternative hypothesis $H_a : \theta = 1$.

Since the distribution of the population is uniquely determined by the parameter θ, both hypotheses are simple, so we can use the Neyman-Pearson lemma to derive the most powerful test. Since there is only one sample, the likelihood ratio is given by:

$$\frac{L(Y|\theta_0)}{L(Y|\theta_a)} = \frac{f_{\theta_0(Y)}}{f_{\theta_a(Y)}} = \frac{2(Y)}{1(Y^0)} = 2Y$$

So the rejection region is of the form $2Y < k$, where we will determine k based on the desired level $\alpha = 0.05$. Dividing by 2, we get $Y < k/2$, and letting $m = k/2$, the rejection region is of the form $Y < m$. We determine m based on the definition of α.

$$\begin{aligned} 0.05 = \alpha &= \mathbb{P}(Y \text{ lies in RR when the null hypothesis is true}) \\ &= \mathbb{P}(Y \text{ lies in RR when } \theta = 2) \\ &= \mathbb{P}(Y < m \text{ when } \theta = 2) \\ &= \int_0^m 2y dy \\ &= m^2 \end{aligned}$$

Thus we have $m^2 = 0.05$, so $m = \sqrt{0.05} = 0.2236$. Thus the rejection region for the 0.05-most powerful test is:

$$\{Y < 0.2236\}$$

In other words, among all hypothesis tests for $H_0 : \theta = 2$ versus $H_a : \theta = 1$ based on a sample size of 1 and $\alpha = 0.05$, this test has the largest possible value for $Power(\theta_a) = Power(1)$. Equivalently, this test has the smallest type II error given a sample size of 1 and $\alpha = 0.05$ for this specific pair of alternative and null hypotheses. What is the actual value of $Power(1)$ in this case. Using the definition of the power of a hypothesis test,

$$\begin{aligned} Power(1) &= \mathbb{P}(Y \text{ lies in RR when } \theta = 1) \\ &= \mathbb{P}(Y < 0.2236 \text{ when } \theta = 1) \\ &= \int_0^{0.2236} 1 dy \\ &= 0.2236 \end{aligned}$$

The value 0.2236 is the maximum value of the power of the test among all tests with $\alpha = 0.05$. But for this test, $\beta = 1 - 0.2236 = 0.7764$, which is very large. So this test is not very good. However, no other test with these same parameters is any better.

What about our hypothesis tests from the previous section. It turns out that they are the uniformly best hypothesis tests for the given situations. We give one example (without proof) below.

Example. Suppose $Y_1, \ldots, Y_n$ are samples drawn from a normal distribution with unknown mean μ and known variance σ^2. We wish to test the null hypothesis $H_0 : \mu = \mu_0$ against the alternative hypothesis $H_a : \mu > \mu_0$. Then the α-most powerful test is given by $\bar{Y} > k$, where

$$k = \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}}$$

This is the test statistic and rejection region we used earlier for large-sample hypothesis testing. Thus our large sample hypothesis test using the Z-distribution is in fact the best hypothesis test we can construct. The proof of this result is an application of the Neyman-Pearson lemma. It is relatively straightforward, but the calculus and algebra are messy, so we will omit that here.

7.9 Likelihood Ratio Tests

In the previous section, we learned how to construct the most powerful α-level test for simple hypotheses. In this case, the distribution of the population is known except for the value of a single parameter θ, and the null and alternative hypothesis are specified in terms of θ. The Neyman-Pearson lemma shows us how to use a likelihood ratio test to construct the most powerful α-level test.

In many cases, we are interested in testing hypotheses involving one parameter, but the population has more than one unknown parameter. We have already encountered this in the section on small-sample hypothesis tests for the mean of normally distributed populations where the population variance is unknown. In this case, the parameter of interest is the population μ. We don't care about the unknown population standard deviation σ, so it is called a *nuisance parameter*. We are also interested in cases where we do not have to choose a specific value for the alternative hypothesis, like we did when we used the Neyman-Pearson lemma. In both these cases (multiple unknown parameters and more complicated alternative hypotheses), we can use a *likelihood ratio test*.

Here is the setup for a likelihood ratio test:

1. We have a population which is parameterized by a set of parameters $\Theta = (\theta_1, \ldots, \theta_n)$. For example, if the population is normally distributed, it is parameterized by $\Theta = (\mu, \sigma)$.

2. We take a sample of size n of independent samples $Y_1, \ldots, Y_n$ from the population.

3. Given specific values of the parameters $\Theta = (\theta_1, \ldots, \theta_n)$, the likelihood function for our sample is denoted $L(Y_1, \ldots, Y_n|\Theta)$. For a normally distributed population, our likelihood function will depend on $\Theta = (\mu, \sigma)$.

4. The null hypothesis states that Θ lies in a particular set of values Ω_0, where the alternative hypothesis states that Θ lies in another set of values Ω_a, where Ω_0 and Ω_a must be disjoint (it does not make sense otherwise). Note that the null and alternative hypotheses no longer need to be single points. They do not have to be simple hypotheses

since they can contain unknown parameters or multiple values of a parameter. For example, if we have a population which is exponentially distributed with parameter λ, then if we want to test the null hypothesis $H_0 : \lambda = \lambda_0$ versus the alternative hypothesis $\lambda \neq \lambda_0$, then we would have $\Omega_0 = \{\lambda_0\}$ and $\Omega_a = \{\lambda > 0 : \lambda \neq \lambda_0\}$.

5. The *parameter space* is defined to be $\Omega = \Omega_0 \cup \Omega_a$, which is all possible values of the parameters. In the exponential example $\Omega = \{\lambda > 0\}$, which is all possible values of an exponential parameter.

6. We define:

$$L(\hat{\Omega}_0) = \max_{\Theta \in \Omega_0} L(Y_1, \ldots, Y_n | \Theta)$$
$$L(\hat{\Omega}) = \max_{\Theta \in \Omega} L(Y_1, \ldots, Y_n | \Theta)$$

   We can think of $L(\hat{\Omega}_0)$ as the "best explanation" for the observed data given the null hypothesis is true, i.e. $\Theta \in \Omega_0$. $L(\hat{\Omega}_0)$ is the "best explanation" for the observed data given all possible values of Θ. If $L(\hat{\Omega}_0) = L(\hat{\Omega})$, then the "best explanation" of the observed data is the null hypothesis, so we should accept the null hypothesis. If $L(\hat{\Omega}_0) < L(\hat{\Omega})$, the "best explanation" of observed data is found inside Ω_a, and we should consider rejecting the null hypothesis in favor of the alternative hypothesis.

7. A *likelihood ratio test* is based on the likelihood ratio $L(\hat{\Omega}_0)/L(\hat{\Omega})$.

Likelihood Ratio Test

Given the setup above, define the likelihood ratio λ by

$$\lambda = \frac{L(\hat{\Omega}_0)}{L(\hat{\Omega})} = \frac{\max_{\Theta \in \Omega_0} L(Y_1, \ldots, Y_n | \Theta)}{\max_{\Theta \in \Omega} L(Y_1, \ldots, Y_n | \Theta)}$$

A likelihood ratio test of the null hypothesis $H_0 : \Theta \in \Omega_0$ versus the alternative hypothesis $H_a : \Theta \in \Omega_a$ has the likelihood ratio λ as a test statistic, and the rejection region is given by $\{\lambda \leq k\}$. The specific value of k is chosen so that α is a desired level. It can be shown that $0 \leq \lambda \leq 1$. A value of λ close to 0 indicates that the likelihood of the sample is much smaller under H_0 than H_a, which favors rejection of the null hypothesis.

The following example of a likelihood ratio test is presented without proof, and justifies our t-test for small samples drawn from a normally distributed population with unknown variance.

Example. Suppose $Y_1, \ldots, Y_n$ are samples drawn from a normal distribution with unknown mean μ and *unknown* variance σ^2. In this case, n is small. We wish to test the null hypothesis $H_0 : \mu = \mu_0$ against the alternative hypothesis $H_a : \mu > \mu_0$. Then if we use the likelihood

ratio test described above, we obtain a hypothesis test with test statistic $\bar{Y}$ and rejection region $\bar{Y} > k$. The value is k is given by

$$k = \mu_0 + t_\alpha \frac{S}{\sqrt{n}}$$

where S is the sample standard deviation computed from our unbiased estimator for the sample variance. Thus the likelihood ratio test for this scenario is exactly the t-test we discussed earlier. The proof of this is several pages of messy calculus and algebra, and will be omitted.