Introduction to Statistics - Young Researchers Fellowship Program

Lecture 6 - Hypothesis Testing - Two Sample tests

Daniel Sánchez Pazmiño

Table of contents

More on hypothesis testing

T-tests (when we don't know the population standard deviation)

As you might remember, when we cannot obtain enough information for a reliable estimate of the population standard deviation, we can use the sample standard deviation as an estimate of the population standard deviation. In this case, we don't know σ_x and we use s_x instead. This is the case when we have a small sample size, or when we don't know the population standard deviation, and it is the most common case in practice.

Much like when we want to compute confidence intervals without knowledge about the population standard deviation, we will need to use another probability distribution for this specific type of hypothesis testing. We will again use the t distribution and must use the t statistic instead of the z statistic. The t statistic is defined as:

$$t = \frac{\bar{x} - \mu_0}{s_x / \sqrt{n}}$$

here μ_0 is the hypothesized value of the population mean. The denominator of the t statistics is the standard error of the sampling distribution of $\bar{x}$, $\sigma_{\bar{x}}$.

The t statistic is used in the same way as the z statistic, but we use the t distribution instead of the normal distribution. The t distribution is similar to the normal distribution, but it has heavier tails. The t distribution is defined by its degrees of freedom, which is equal to $n-1$.

Once again, there are three different kinds of hypothesis tests that we can perform with the t statistic: left-tailed, right-tailed, and two-tailed. The t statistic is used in the same way as the z statistic, but we use the t distribution instead of the normal distribution.

Left-tailed test

Let us do an example. Imagine we pull out a sample of 25 observations from a population with an unknown mean and standard deviation. We want to test the hypothesis that the population mean is equal to 10. We want to test this hypothesis at the 5% significance level. The sample mean is 9.5 and the sample standard deviation is 2.5.

The null hypothesis is that the population mean is equal to 10. The alternative hypothesis is that the population mean is less than 10. Write this down in mathematical notation:

$$H_0 : \mu >= 10$$

$$H_1 : \mu < 10$$

The test statistic is the t statistic:

$$t = \frac{\bar{x} - \mu_0}{s_x / \sqrt{n}}$$

The degrees of freedom are $n-1 = 25-1 = 24$. As before, we have two ways of proceeding from here. We can either compute the t statistic and compare it to the critical value, or we can compute the p-value. Let us do both.

First, we compute the t statistic:

$$t = \frac{9.5 - 10}{2.5/\sqrt{25}} = -2$$

The critical value is the value of the t statistic that corresponds to the 5% significance level and 24 degrees of freedom. We can find this value in a table of the t distribution, by looking at the "left portion" or "greater portion". However, it is common that we don't find the value for such a small probability to the left side. In this case, we simply look for the number of standard errors that correspond to the 5% significance level to the right side, and then we multiply that number by -1. This is because the t distribution is symmetric. In this case, the 5% significance level corresponds to 1.711 standard errors. We multiply this number by -1 to get the critical value. The critical value is -1.711.

For a left-tailed test, we reject the null hypothesis if the t statistic is less than the critical value. Since the t statistic is less than the critical value, we reject the null hypothesis. Thus, there is enough statistical evidence to conclude that the population mean is less than 10.

The p-value is the probability of observing a t statistic less than -2. We can find this probability in a table of the t distribution. The p-value is 0.029. Since the p-value is less than 0.05, we reject the null hypothesis.

Right-tailed test

Let us do another example. Imagine we pull out a sample of 25 observations from a population with an unknown mean and standard deviation. We want to test the hypothesis that the population mean is equal to 10. We want to test this hypothesis at the 5% significance level. The sample mean is 9.5 and the sample standard deviation is 2.5.

The null hypothesis is that the population mean is equal to 10. The alternative hypothesis is that the population mean is greater than 10. Write this down in mathematical notation:

$$H_0 : \mu <= 10$$
$$H_1 : \mu > 10$$

The test statistic is the t statistic:

$$t = \frac{\bar{x} - \mu_0}{s_x/\sqrt{n}}$$

The degrees of freedom are $n - 1 = 25 - 1 = 24$. The test statistic will be

$$t = \frac{9.5 - 10}{2.5/\sqrt{25}} = -2$$

The critical value for a right-tailed test is the value of the t statistic that corresponds to the 5% significance level and 24 degrees of freedom. We can find this value in a table of the t distribution by looking at the "right area" or "smaller portion". The critical value is 1.711. The critical value is positive because we are doing a right-tailed test.

Because the test statistic is greater than the critical value, we fail to reject the null hypothesis. Thus, there is not enough statistical evidence to conclude that the population mean is greater than 10.

Two-tailed test

Let us do another example. Imagine we pull out a sample of 25 observations from a population with an unknown mean and standard deviation. We want to test the hypothesis that the population mean is equal to 10. We want to test this hypothesis at the 5% significance level. The sample mean is 12 and the sample standard deviation is 2.5.

The null hypothesis is that the population mean is equal to 10. The alternative hypothesis is that the population mean is different than 10. Write this down in mathematical notation:

$$H_0 : \mu = 10$$

$$H_1 : \mu \neq 10$$

The test statistic is the t statistic:

$$t = \frac{\bar{x} - \mu_0}{s_x/\sqrt{n}}$$

The degrees of freedom are $n - 1 = 25 - 1 = 24$. The test statistic will be

$$t = \frac{12 - 10}{2.5/\sqrt{25}} = 4$$

As before, critical values of the t statistic can be found in a table of the t distribution. We need to find the two critical values, which correspond to the number of standard errors necessary to have a 5%/2 = 2.5% probability on each side of the distribution. The critical values are -2.064 and 2.064. We need to compare the test statistics to the critical values. If the test statistic is less than -2.064 or greater than 2.064, we reject the null hypothesis. If the test statistic is between -2.064 and 2.064, we fail to reject the null hypothesis.

In this case, the test statistic is 4, which is greater than 2.064. Thus, we reject the null hypothesis. There is enough statistical evidence to conclude that the population mean is different than 10.

For a two-tailed test, the p-value is the probability of observing a t statistic less than -4 or greater than 4. Because the t distribution is symmetric, we can compute the probability of observing a t statistic less than -4 and multiply it by 2. We can find this probability in a table of the t distribution. The p-value is 0.0002. Since the p-value is less than 0.05, we reject the null hypothesis.

Proportions and their sampling distribution

When we have categorical data, we cannot compute a "mean", but we can calculate the percentage of people falling in certain category. When we remove the % of the percentage, it becomes a proportion. The statistical theory for proportions is very similar to the statistical theory for means, we just need to perform a few adjustments.

First of all, we will know that the proportion will also have a sampling distribution that is approximately normal with large enough samples. The mean of the sampling distribution will be the population proportion.

How to calculate the standard error of the proportion? We need to have in mind that a sample proportion is calculated as the number of successes divided by the sample size. This, in formula notation, is

$$\hat{p} = \frac{x}{n}$$

where x is the number of successes and n is the sample size. In here, successes means the number of observations that fall into the category that we want to build the proportion for. For instance, if we want to calculate the proportion of students with mental health issues, we'd divide the number of studentes who reported having mental health issues by the total number of students in the sample.

Because we know that the sample proportion has a sampling distribution with mean equal to the population proportion, we say that the expected value of the sample proportion is equal to the population proportion. In formula notation, this is

$$E(\hat{p}) = p$$

where p is the population proportion.

The standard error of the sample proportion is calculated as

$$SE(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}$$

where p is the population proportion and n is the sample size.

With this information, we can build confidence intervals with the same method we've seen before. The only difference is that we will use the standard error of the proportion instead of the standard error of the mean, and we will always use the z distribution instead of the t distribution, no matter what we know or what we don't know.

Hypothesis testing for proportions

We will also use the z distribution for hypothesis testing for proportions. The test statistic is calculated as

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

where $\hat{p}$ is the sample proportion, p_0 is the null hypothesis value for the population proportion, and n is the sample size. Note how the test statistic or "z-score" for the proportion follows the same formula as the test statistic for the mean, except that we use the standard error of the proportion instead of the standard error of the mean. We substract the hypothesized value of the population proportion from the sample proportion, and we divide by the standard error of the proportion.

We would also have left tailed, right tailed and double tailed tests. The critical values would be found in the z distribution table. The p-value would be calculated as the probability of observing a z statistic less than the test statistic for a left-tailed test, greater than the test statistic for a right-tailed test, and less than the absolute value of the test statistic for a two-tailed test.

Let's do an example to see how this works. Imagine we want to test the hypothesis that the proportion of students with mental health issues is equal to 0.2. We pull out a sample of 100 students and we find that 25 of them have mental health issues. The sample proportion is 0.25. The null hypothesis is that the population proportion is equal to 0.2. The alternative hypothesis is that the population proportion is different than 0.2. Write this down in mathematical notation:

$$H_0 : p = 0.2$$

$$H_1 : p \neq 0.2$$

The test statistic is the z statistic:

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

The test statistic will be

$$z = \frac{0.25 - 0.2}{\sqrt{\frac{0.2(1-0.2)}{100}}} = 1.58$$

We can also either use p-values or critical values to test the hypothesis. Let's use critical values. The critical values for a 0.05 significance level are -1.96 and 1.96. We need to compare the test statistic to the critical values. If the test statistic is less than -1.96 or greater than 1.96, we reject the null hypothesis. If the test statistic is between -1.96 and 1.96, we fail to reject the null hypothesis.

In this case, the test statistic is 1.58, which is between -1.96 and 1.96. Thus, we fail to reject the null hypothesis. There is not enough statistical evidence to conclude that the population proportion is different than 0.2.

The interval approach for hypothesis testing

We can also use the confidence interval approach for hypothesis testing for proportions. The interval approach is a way of testing hypotheses that is equivalent to the p-value approach. The interval approach is based on the idea that if the null hypothesis is true, then the hypothesized value of the population parameter should be a plausible value for the population parameter. If the null hypothesis is not true, then the hypothesized value of the population parameter should not be a plausible value for the population parameter.

This implies that if the null hypothesis is true, then the hypothesized value is inside the confidence interval. If the null hypothesis is not true, then the hypothesized value is not inside the confidence interval. Thus, we can test the null hypothesis by checking whether the hypothesized value is inside the confidence interval or not.

This is important to know, especially for data visualization. Often, a sample estimate is drawn with a confidence interval around it. If the hypothesized value is inside the confidence interval, then we fail to reject the null hypothesis. If the hypothesized value is not inside the confidence interval, then we reject the null hypothesis.

Political polls often use this approach to test hypotheses about the proportion of people who support a certain candidate. If two candidates are running for office, and the confidence interval for the proportion of people who support one candidate does not include the proportion of people who support the other candidate, then the pollster will conclude that the first candidate is leading in the polls. Look at the graph below:

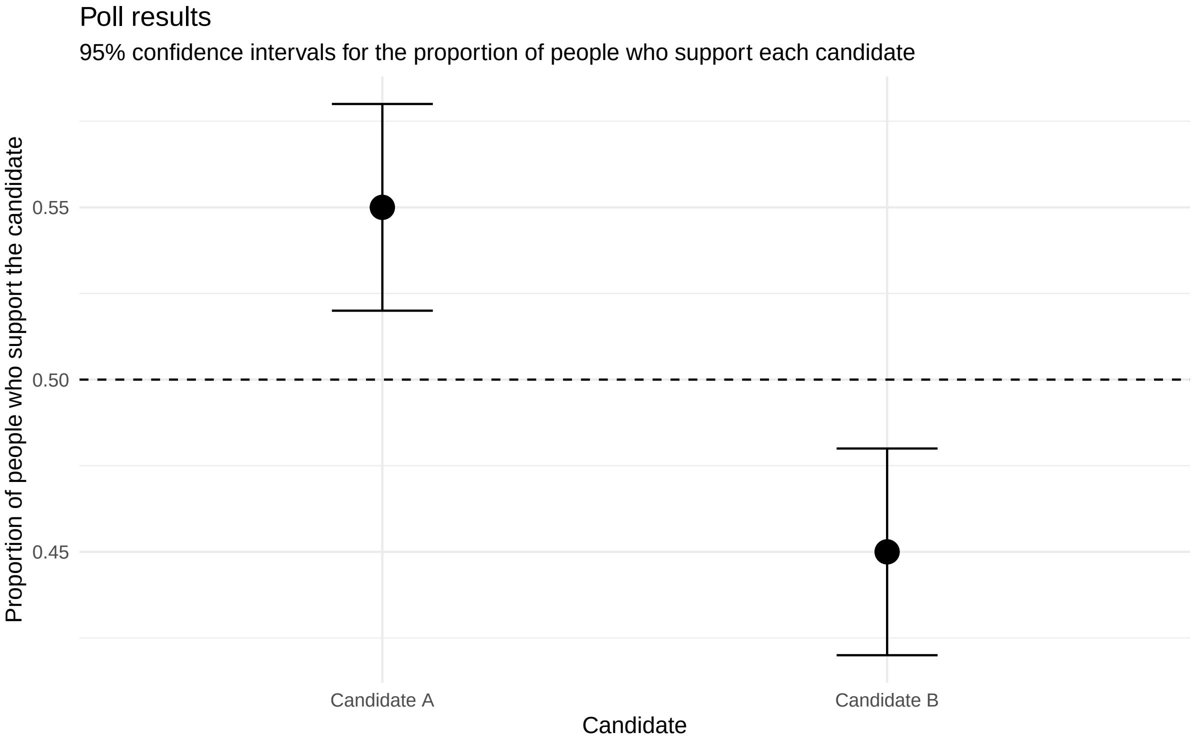

Figure 1: Confidence interval for the proportion of people who support each candidate

In this case, the confidence interval for the proportion of people who support Candidate A does not include the proportion of people who support Candidate B. Thus, the pollster would conclude that Candidate A is leading in the polls, because the proportion of people who support Candidate A is statistically different than the proportion of people who support Candidate B. We will talk some more about this in the following sections.

Hypothesis testing for means of two populations

So far, we've been performing hypothesis tests about statements that involve only one population. For instance, we've been testing whether the population mean is equal to a certain value, or whether the population proportion is equal to a certain value. These are often called the one sample t-test and the one sample z-test.

However, we can also test statements that involve two populations. We can test whether the population mean of one population is equal to the population mean of another population. There are several cases that we can consider, and we'll review below the most common ones.

Difference between two population means with known population standard deviations (independent samples)

In this case, we are in a situation where we want to know whether or not a population mean is different to another. For instance, we might want to know if the score on an exam for a group of students is different than the score on an exam for another group of students. We might want to know if the average salary of a group of employees is different than the average salary of another group of employees. We might want to know if the average number of hours of sleep of a group of students is different than the average number of hours of sleep of another group of students.

For this, we will focus on making a hypothesis test on a difference between two population means. If $\mu_1 - \mu_2$ is statistically different from zero, we know that the means are different. If $\mu_1 - \mu_2$ is not statistically different from zero, we know that the means are not different.

We will assume that the population standard deviations are known. We will also assume that the samples are independent. This means that the observations in one sample are not related to the observations in the other sample. For instance, the observations in one sample are not the same observations as the observations in the other sample.

Independent samples are cool because we can make cool tricks with them. We can define the difference between the two sample means as

$$\bar{x}_1 - \bar{x}_2$$

where $\bar{x}_1$ is the sample mean of the first sample and $\bar{x}_2$ is the sample mean of the second sample. We will assume that this difference is a random variable, which follows a probability distribution which is a sampling distribution. The mean of that sampling distribution will be the difference between the two population means. The standard error of that sampling distribution will be

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

where σ_1 is the population standard deviation of the first population, σ_2 is the population standard deviation of the second population, n_1 is the sample size of the first sample, and n_2 is the sample size of the second sample.

With this information, we can build confidence intervals with the same method we've seen before. The only difference is that we will use the standard error of the difference between the two sample means instead of the standard error of the mean, and we will always use the z distribution instead of the t distribution, no matter what we know or what we don't know.

Because we're using the z distribution, we will also use the z distribution for hypothesis testing. The test statistic is calculated as

$$z = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

where $\bar{x}_1$ is the sample mean of the first sample, $\bar{x}_2$ is the sample mean of the second sample, μ_1 is the population mean of the first population, μ_2 is the population mean of the second population, σ_1 is the population standard deviation of the first population, σ_2 is the population standard deviation of the second population, n_1 is the sample size of the first sample, and n_2 is the sample size of the second sample.

There are sometimes variations of this test statistic, especially when we don't know the population standard deviations and/or the sample sizes are small. We will cover those variations in the next sections. Further, notice that still this test statistic is no different than the test statistic we've seen before. We still take the result from our samples ($\bar{x}_1 - \bar{x}_2$), subtract the value we expect to see if the null hypothesis is true ($\mu_1 - \mu_2$), and divide by the standard error of the sampling distribution of the difference between the two sample means.

Often, we simply want to know if the population means are different to each other, or if one is bigger than the other. In this case, we are basically trying to know if

$$\mu_1 = \mu_2$$

which means that

$$\mu_1 - \mu_2 = 0$$

So, we will try to test if the difference between the two population means is equal to zero. If μ_1 is bigger than μ_2, then this means that the difference is positive (greater than zero) and if μ_1 is smaller than μ_2, then this means that the difference is negative (smaller than zero).

This is why we can use the same test statistic we've seen before, by setting the hypothesized difference to zero. The test statistic is then calculated as

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

This is the most common type of hypothesis test for two independent population means.

We can then define all three types of hypothesis tests for two independent population means as follows:

$$H_0 : \mu_1 - \mu_2 = 0$$
$$H_1 : \mu_1 - \mu_2 \neq 0$$

$$H_0 : \mu_1 - \mu_2 \leq 0$$
$$H_1 : \mu_1 - \mu_2 > 0$$

$$H_0 : \mu_1 - \mu_2 \geq 0$$
$$H_1 : \mu_1 - \mu_2 < 0$$

The first one is a two-sided test, the second one is a left-sided test, and the third one is a right-sided test.

If we want to test whether the population means are equal to each other, we will use the two-sided test. If we want to test whether one population mean is bigger than the other, we will use the right-sided test. If we want to test whether one population mean is smaller than the other, we will use the left-sided test.

How do we perform the test? As you can see from the test statistic, we need to know the population standard deviations. We also need to know the sample sizes. We will then calculate the test statistic, and we will use the z distribution to calculate the critical value for the given significance level. We will then compare the test statistic to the critical value, and we will make a decision. All of this is done in the same way as we've seen before, the only difference is that we use the z distribution instead of the t distribution and we need to use the scary formulas for the standard error and the test statistic. A computer is best suited to do this.

Difference between two population means with unknown population standard deviations (still with independent samples)

What if we don't know the population standard deviations? We can estimate them with the sample standard deviations. This means that we will use the t distribution instead of the z distribution. The standard error of the difference between the two sample means is then calculated as

$$SE(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

where s_1 is the sample standard deviation of the first sample, s_2 is the sample standard deviation of the second sample, n_1 is the sample size of the first sample, and n_2 is the sample size of the second sample.

The test statistic is calculated as

$$t = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

where $\bar{x}_1$ is the sample mean of the first sample, $\bar{x}_2$ is the sample mean of the second sample, μ_1 is the population mean of the first population, μ_2 is the population mean of the second population, s_1 is the sample standard deviation of the first sample, s_2 is the sample standard deviation of the second sample, n_1 is the sample size of the first sample, and n_2 is the sample size of the second sample.

We will also need to calculate the degrees of freedom. The degrees of freedom are calculated as

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1-1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2-1}}$$

where s_1 is the sample standard deviation of the first sample, s_2 is the sample standard deviation of the second sample, n_1 is the sample size of the first sample, and n_2 is the sample size of the second sample.

We can then perform left sided, right sided, and two sided tests. We would need to compute critical values from the t distribution with the degrees of freedom we calculated. We would then compare the test statistic to the critical value, and we would make a decision, with the same rules as the one sample t-tests.

This type of test is the most common type of hypothesis test for two independent population means. The complicated formulas make it a bit more difficult to perform, but a computer can do it easily.

Difference between two population means with dependent samples

What if we have dependent samples? This means that we have two samples, but the samples are not independent. This means that the samples are related to each other. For example, we could have a sample of students, and we could measure their test scores before and after a training. The samples are related to each other, because the same students are measured before and after the training. We often call them paired samples, matched samples or dependent samples.

If we do this, we need to define something known as the difference between the measured statistic between the two groups. Consider the table below:

Student	Test score before training	Test score after training	Difference
1	80	90	10
2	70	85	10
3	90	100	10
4	60	71	10
5	50	43	10

Matched samples like these allow us to perform a one sample t-test on the difference between the two groups (the third column). Take the average and standard deviation of the difference, and perform a one sample t-test on the difference. Is it significantly different from zero? If it is, then the two groups are significantly different from each other. Is it positive? Then the test score before training is significantly higher than the test score after training. Is it negative? Then the test score before training is significantly lower than the test score after training.

We can define left-tailed, right-tailed, and two-tailed tests in the same way as we've seen before. The only difference is that we use the difference between the two groups instead of the sample mean of one group. The test statistic for the difference between two groups is calculated as

$$t = \frac{\bar{d} - \mu_d}{\frac{s_d}{\sqrt{n}}}$$

where $\bar{d}$ is the sample mean of the difference between the two groups, μ_d is the hypothesized population mean of the difference between the two groups, s_d is the sample standard deviation of the difference between the two groups, and n is the sample size of the difference between the two groups.

We use the t distribution to calculate the critical value, and we compare the t-statistic to the critical value. We make a decision based on the comparison.

Summary

- When we don't know the population standard deviation, we use the t distribution instead of the z distribution for a one sample t-test.
- Proportions are a special case of means, so we can use the same methods for proportions as we use for means. We only need to calculate the standard error differently and then use the z distribution instead of the t distribution.
- When we have two independent samples, we can use the z distribution for the difference between the two sample means if we know the population standard deviations. If we don't know the population standard deviations, we can use the t distribution.

- The test statistic for the difference between two independent sample means is different, but calculating the critical value and making a decision is the same as for the one sample t-test.
- When we don't know the population standard deviations, we use the t distribution instead of the z distribution for a two sample t-test. We also need to calculate the degrees of freedom differently, but the methods of calculating critical values and making decisions are the same as for the one sample t-test.
- When we have two dependent samples, we can essentially do a one sample t-test on the difference between the two groups. We take the difference between the two groups, and we perform a one sample t-test on the average difference. We can then make a decision based on the result of the one sample t-test.