Lecture 13

Two-Sample Tests

Sometimes we want to compare two samples, collected in different ways or from two different sub-populations, to each other, rather than just one sample compared to previous data. In this case, we generally assume that the methods have no meaningful impact and their difference is zero. We then set as our alternative that the difference is not zero (or greater or less depending on what we expect). This is a common strategy for most two-sample tests. Although, it is possible to set some non-zero difference (suppose the difference must be large enough to justify the expense of one technique being compared), but these situations are somewhat uncommon. We will, nonetheless, present the formulas in the most general case in case you do encounter the situation.

We are going to start here with the proportion case.

There are two versions of the test statistic that we can use for proportions. One formula treats all the data with a common proportion for the standard error (this is okay when both samples are large and about the same size), and the other is a little more precise if the sample sizes are a bit smaller or two two samples are of different sizes. Both proportions we are comparing must satisfy the same condition introduced previously for proportions on their own.

The large sample (estimated) test statistic is

$$z = \frac{\hat{p}_1 - \hat{p}_2 - (p_{01} - p_{02})}{\sqrt{pq\left(\frac{1}{m} + \frac{1}{n}\right)}}$$

Let's talk about the notation here. The "hat" notation is the estimate from the sample. The subscripts 1 and 2 are the samples the proportions come from. The 0 subscripts indicate that these are part of the null hypothesis. Typical $(p_{01} - p_{02}) = 0$ in most tests. The p without the subscript is the proportion of successes in both samples (think of the data as pooled) and is equal to $p = \frac{p_1 m + p_2 n}{m+n}$, while $q = 1 - p$. The samples sizes for the two populations are m and n respectively.

You may also see some versions of this formula which uses the null hypothesis value (or the mean of it) in place of the weighted average. The formulation of p and q may depend on how much you know.

There is a little less computation involved here, but it is more of an estimate.

A slightly more precise version of the test statistic is

$$z = \frac{\hat{p}_1 - \hat{p}_2 - (p_{01} - p_{02})}{\sqrt{\frac{\hat{p}_1\hat{q}_1}{m} + \frac{\hat{p}_2\hat{q}_2}{n}}}$$

There is more variability if the proportions are different from each other, and if the sample sizes are smaller, so this will cover more cases and capture more variability. It is a little bit more computationally involved, but it is more accurate. I will use the more accurate value when doing our next example.

Example.

	PC	Mac	Row Totals
Male	66	40	106
Female	30	87	117
Column Totals	96	127	223

In the table above, we have some data on gender and computer preference. Let's think of the males and females as our two populations. We want to conduct a test to determine if the computer preference differs between men and women. We will pick on computer type to be our "success". I'm going to pick PC to be the "success" case. We can calculate our sample proportions then. For men, $p_1 = \frac{66}{106}$, and $p_2 = \frac{30}{117}$, and we have $m = 106$ and $n = 117$.

Our hypothesis test is

$$H_0: p_1 - p_2 = 0$$
$$H_a: p_1 - p_2 \neq 0$$

Alternatively, we could say

$$H_0: p_1 = p_2$$
$$H_a: p_1 \neq p_2$$

This is a two-sided test. If you are doing a one-sided test, be extra careful about the order of subtract combined with your inequality. It may be easier to make the statement with the two populations on different sides.

Our test statistic then becomes

$$z = \frac{\hat{p}_1 - \hat{p}_2 - (p_{01} - p_{02})}{\sqrt{\frac{\hat{p}_1\hat{q}_1}{m} + \frac{\hat{p}_2\hat{q}_2}{n}}} = \frac{\frac{66}{106} - \frac{30}{117} - 0}{\sqrt{\frac{\frac{66}{106}\left(1 - \frac{66}{106}\right)}{106} + \frac{\frac{30}{117}\left(1 - \frac{30}{117}\right)}{117}}} \approx \frac{0.36623}{\sqrt{0.003846}} \approx 5.905 \ldots$$

Let's convert this to a P-value using a standard normal distribution. (Recall that for a two-sided test, you need to calculate one side and then multiply by 2).

$$2P(Z \geq 5.905) = 3.5 \times 10^{-9}$$

This value is much less than 0.05, so we can reject the null. There is sufficient evidence to think that men and women have different computer preferences.

It's worth noting that we can use the denominator of our test statistic as the standard error we use in construction confidence intervals as well.

Our interval would then be

$$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{m} + \frac{\hat{p}_2\hat{q}_2}{n}}$$

In this case we can see that the margin of error for a 95% confidence interval would be $1.96 \times 0.062 = 0.122$

The difference in our proportions is 0.31, so our confidence interval would be $(0.188, 0.432)$. We can see that this is consistent with our hypothesis test result since the confidence interval does not contain 0.

We can calculate the power of the test in a similar manner that we did before, but one useful application is determining the sample size needed to obtain a given power level.

$$n = \frac{\left[z_\alpha\sqrt{(p_1 + p_2)(q_1 + q_2)/2} + z_\beta\sqrt{p_1q_1 + p_2q_2}\right]^2}{d^2}$$

Here z_α is the standard score associated with the significance level, and z_β is the standard score associated with the desired power. p_1 and p_2 are the assumed proportion values for each population (these could be estimated from a smaller sample or by other means), and where the corresponding $q = 1 - p$. The denominator d is short for $d = p_1 - p_2$, just the difference of the two proportions. This formula also assumes that $m = n$, which is to say the two sample sizes will be equal.

If samples are small (they don't meet our normality estimate requirements) then things can become complex and computational methods may be required.

Let's consider now tests for means.

When doing tests for means, we generally use the t-test. There are circumstances when the sample size is sufficiently large, but since we are comparing two samples, we generally work from the sample data and don't usually have information about the population (for instance the population standard deviations would be based on all comparisons of the same type).

The first kind of two-sample t-test is the paired t-test. A paired sample has two observations that are matched or paired. Sometimes, these measurements are taken of the same subject but after changing some circumstance. For instance, perhaps you are testing if some training technique improves performance on a test. A researcher might take measurements before the training and again on the same subjects after the training and compare the results. Samples might also be paired by matching subjects for various characteristics. In such a circumstance, the results of observations in each sample are considered dependent. When identifying such a problem, the sample sizes must be the same, and look for keywords like "paired" or "matched". For this test, we must also have raw data to work from.

The reason we need raw data is that we are going to calculate the differences between each pair of data, and then follow the one-sample test procedure. We will use the mean of the differences as our sample mean, and the standard deviation of the differences as the basis for our standard error.

Our test statistic will be

$$t = \frac{\bar{\delta} - \delta_0}{\frac{s}{\sqrt{n}}}$$

The value $\bar{\delta}$ is the mean of the differences. The value δ_0 is the null hypothesis value. n is the number of pairs.

Let's look at a small example.

Example.

Subject #	Score 1	Score 2
1	3	20
2	3	13
3	3	13
4	12	20
5	15	29
6	16	32
7	17	23
8	19	20
9	23	25
10	24	15
11	32	30

A university tested 11 non-native English speakers for their English skills upon arriving at the university. The students were then offered a training program to improve their English skills and then tested again. The table above shows their test scores.

Our hypotheses are

$$H_0: \delta = 0$$
$$H_a: \delta > 0$$

Since we are testing to see if the training improved test scores.

The first step is to find the differences in each pair.

Subject	1	2	3	4	5	6	7	8	9	10	11
Difference	17	10	10	8	14	16	6	1	2	-9	-2

We need both the mean and the standard deviation. $\bar{\delta} = 6.636, s = 8.041$. Now we can calculate our test statistic.

$$t = \frac{6.636 - 0}{\frac{8.041}{\sqrt{11}}} \approx 2.737 \ldots$$

We calculate the P-value from this using 10 degrees of freedom. We get 0.01047 which is less than 0.05, so we can reject the null hypothesis. There is good evidence to think the training improved test scores.

Sometimes we don't collect data in matched pairs and we are just comparing sample means. This kind of test is considered an independent test and there are two ways to go about conducting this test: using a pooled standard deviation and an unpooled standard deviation. This is somewhat similar to the proportion test discussed above. The pooled procedure is a little bit easier to compute, but the unpooled procedure is more accurate. Let's look at the test statistic formulas for both methods.

The pooled test assumes that the variances for the two samples are the same, and combines them into a single value. Sometimes this test is referred to as the equal variance t-test.

$$S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2} = (\frac{n_1-1}{n_1+n_2-2})S_1^2 + (\frac{n_2-1}{n_1+n_2-2})S_2^2$$

The test formula shown here assumes that the hypothesized difference is 0. If the assumption for the null is not zero, you need to subtract it off in the numerator as in other tests.

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{S_p^2(\frac{1}{n_1}+\frac{1}{n_2})}} = \frac{\bar{X}_1 - \bar{X}_2}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

The degrees of freedom used with this test is $n_1 + n_2 - 2$.

The unpooled case allows for unequal variances.

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_{01} - \mu_{02})}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

The test statistic itself isn't bad, but the calculation for the degrees of freedom is a little messy and can produce decimal degrees of freedom.

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

In statistical software, this figure is typically calculated for you, but it's one of the reasons the pooled test is sometimes used despite the fact that it makes an extra assumption about the variances. In general, of the two tests, the unequal variance (unpooled) case is better. We'll see later on that there is a non-parametric test that we can also use to compare means.

Let's look at an example.

Example.

Suppose we want to test whether men and women have the same or different body fat measurements. We have the following summary statistics. From a sample of 13 men, they had a mean of 14.95% and a standard deviation of 6.84. From a sample of 10 women, they had a mean 22.29% and a standard deviation of 5.32.

Our hypotheses are

$$H_0: \mu_1 = \mu_2$$
$$H_a: \mu_1 \neq \mu_2$$

Let's calculate the test statistic using the unpooled test.

$$t = \frac{(14.95 - 22.29) - 0}{\sqrt{\frac{(6.84)^2}{13} + \frac{(5.32)^2}{10}}} \approx -2.8948 \ldots$$

Before we can find the P-value (or the bounds on the rejection region), we need to calculate the degrees of freedom.

$$df = \frac{\left[\frac{(6.84)^2}{13} + \frac{(5.32)^2}{10}\right]^2}{\frac{\left(\frac{(6.84)^2}{13}\right)^2}{12} + \frac{\left(\frac{(5.32)^2}{10}\right)^2}{9}} \approx \frac{(3.59889 + 2.83024)^2}{\frac{3.59889^2}{12} + \frac{2.83024^2}{9}} \approx \frac{41.3337}{1.96936} \approx 20.988$$

Note that the outcome for this calculation is similar to $n_1 + n_2 - 2 = 13 + 10 - 2 = 21$.

It's a two-tailed test, so we find the P-value as $2P(t \leq -2.8948) = 0.00867 \ldots$

This is much less than 0.05, so we can reject the null hypothesis. We can conclude that men and women have different amounts of body fat.

As with the other cases, we can also construct a confidence interval using the standard error (from the denominator of our test statistic), and the degrees of freedom from our degrees of freedom formula.

If the samples are large enough, then we can convert to a z-test to estimate the p-value.

The last two-sample parametric case we are going to look at is a comparison of variances (or standard deviations). This test uses an F-statistic which requires two degrees of freedom.

Example.
Two coworkers commute from the same building. They are interested in whether or not there is any variation in the time it takes them to drive to work. They each record their times for 20 commutes. The first worker's times have a variance of 12.1. The second worker's times have a variance of 16.9. The first worker thinks that he is more consistent with his commute times. Test the claim at the 10% level. Assume that commute times are normally distributed.

Let's state our hypotheses.

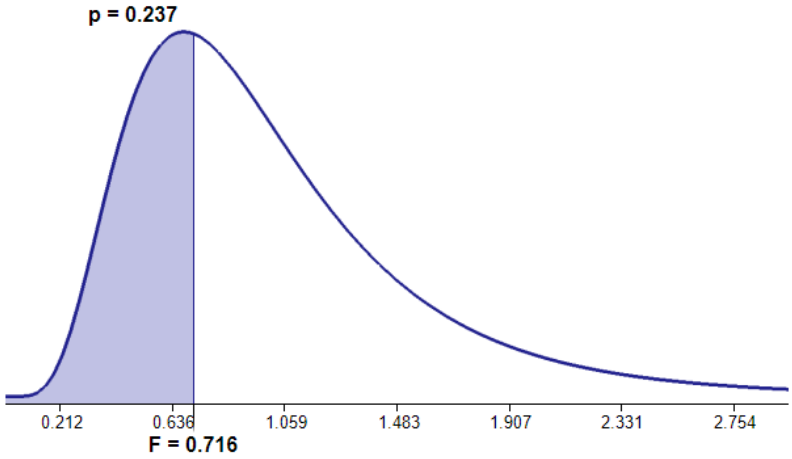
$$H_0: \sigma_1^2 = \sigma_2^2$$
$$H_a: \sigma_1^2 < \sigma_2^2$$

If the first worker's times are more consistent, then their variance is smaller, so we do a one-sided test. Our test statistic is

$$F = \frac{\frac{s_1^2}{\sigma_1^2}}{\frac{s_2^2}{\sigma_2^2}} = \frac{s_1^2}{s_2^2} = \frac{12.1}{16.9} \approx 0.716$$

Since we are assuming $\sigma_1^2 = \sigma_2^2$, they cancel in the test statistic formula.

The degrees of freedom for the numerator is $\nu_1 = m - 1$ and for the denominator is $\nu_2 = n - 1$. Since this is left-tailed, we want $P(F \leq 0.716) = 0.237$.



The P-value is higher than the significance level stated in the problem at 0.10 so we fail to reject the null hypothesis. There is not enough evidence to support the claim that the first worker has a less variable commute than the second worker.

References:

1. https://faculty.ksu.edu.sa/sites/default/files/probability_and_statistics_for_engineering_and_the_sciences.pdf
2. https://assets.openstax.org/oscms-prodcms/media/documents/IntroductoryStatistics-OP_i6tAI7e.pdf
3. https://statisticsbyjim.com/probability/contingency-tables-probabilities/
4. https://www.statisticshowto.com/probability-and-statistics/t-test/

5. https://stats.libretexts.org/Bookshelves/Applied_Statistics/Book%3A_Natural_Resources_Biometrics_(Kiernan)/04%3A_Inferences_about_the_Differences_of_Two_Populations/4.02%3A_Pooled_Two-sampled_t-test_(Assuming_Equal_Variances)
6. https://www.statsdirect.co.uk/help/parametric_methods/utt.htm
7. https://www.jmp.com/en_us/statistics-knowledge-portal/t-test/two-sample-t-test.html
8. https://opentextbc.ca/introbusinessstatopenstax/chapter/test-of-two-variances/