Two Independent Samples: Conditions and Two-Sample t-test

Learning Objective: In a given context, carry out the inferential method for comparing groups and draw the appropriate conclusions.

Step 2: Check Conditions, and Summarize the Data Using a Test Statistic

The two-sample t-test can be safely used as long as the following conditions are met:

1. The two samples are indeed independent.
2. We are in one of the following two scenarios:
    a. Both populations are normal, or more specifically, the distribution of the response Y in both populations is normal, and both samples are random (or at least can be considered as such). In practice, checking normality in the populations is done by looking at each of the samples using a histogram and checking whether there are any signs that the populations are not normal. Such signs could be extreme skewness and/or extreme outliers.

    b. The populations are known or discovered not to be normal, but the sample size of each of the random samples is large enough (we can use the rule of thumb that > 30 is considered large enough).

Assuming that we can safely use the two-sample t-test, we need to summarize the data, and in particular, calculate our data summary—the test statistic.

The two-sample t-test statistic is:

$$t = \frac{(\overline{y_1} - \overline{y_2}) - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Where:

$\overline{y_1}$, $\overline{y_2}$ are the sample means of the samples from population 1 and population 2 respectively.

s_1, s_2 are the sample standard deviations of the samples from population 1 and population 2 respectively.

n_1, n_2 are the sample sizes of the two samples.

Comment

Let's see why this test statistic makes sense, bearing in mind that our inference is about $\mu_1 - \mu_2$.

- $\overline{y_1}$ estimates μ_1 and $\overline{y_2}$ estimates μ_2, and therefore $\overline{y_1} - \overline{y_2}$ is what the data tell me about (or, how the data estimate)

  $\mu_1 - \mu_2$.
- 0 is the "null value" — what the null hypothesis, H_o, claims that $\mu_1 - \mu_2$ is.
- The denominator $\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ is the standard error of $\overline{y_1} - \overline{y_2}$. (We will not go into the details of why this is true.)

We therefore see that our test statistic, like the previous test statistics we encountered, has the structure:

$$\frac{sample\ \ estimate - null\ \ value}{standard\ \ error}$$

and therefore, like the previous test statistics, measures (in standard errors) the difference between what the data tell us about the parameter of interest $\mu_1 - \mu_2$ (sample estimate) and what the null hypothesis claims the value of the parameter is (null value).

Example

Let's first check whether the conditions that allow us to safely use the two-sample t-test are met.

1. Here, 239 students were chosen and were naturally divided into a sample of females and a sample of males. Since the students were chosen at random, the sample of females is independent of the sample of males.

2. Here we are in the second scenario — the sample sizes (150 and 85), are definitely large enough, and so we can proceed regardless of whether the populations are normal or not.

In order to avoid tedious calculations, we use statistics software to find the test statistic. The output (edited) is shown below:

Two Sample T – Test and CI: Score (Y), Gender (X)

Summary statistics for Score (Y):

Gender (X)	n	Mean	Std. Dev.	Std. Err.
Female	150	10.733334	4.254751	0.347399
Male	85	13.3294115	4.0189676	0.43591824

Hypothesis test results:
μ_1 : mean of Score (Y) where X = Female
μ_2 : mean of Score (Y) where X = Male
$\mu_1 - \mu_2$: mean difference
$H_0 : \mu_1 - \mu_2 = 0$
$H_A : \mu_1 - \mu_2 \neq 0$

Difference	Sample Mean	Std. Err.	DF	T-Stat	P-value
$\mu_1 - \mu_2$	-2.5960784	0.55741435	182.97267	-4.657358	<0.0001

95% confidence interval results:

Difference	Sample Mean	Std. Err.	DF	L. Limit	U. Limit
$\mu_1 - \mu_2$	-2.5960784	0.55741435	182.97267	-3.6958647	-1.4962921

As you can see we highlighted the "ingredients" needed to calculate the test statistic, as well as the test statistic itself. Just for this first example, let's make sure that we understand what these ingredients are and how to use them to find the test statistic.

Learn By Doing

1/1 point (graded)
Bearing in mind that sample 1 is the sample of females, and sample 2 is the sample of males, according to the output what is n_1?

150 ✔

Answer
Correct: 150 is the sample size of sample 1, the sample of females.

Submit

Learn By Doing

1/1 point (graded)

Bearing in mind that sample 1 is the sample of females, and sample 2 is the sample of males, according to the output what is n_2?

85

Answer

Correct: 85 is the sample size of sample 2, the sample of males.

Submit

Learn By Doing

1/1 point (graded)

Bearing in mind that sample 1 is the sample of females, and sample 2 is the sample of males, according to the output what is y_1-bar?

10.73

Answer

Correct: 10.73 is the sample mean importance score in sample 1, the sample of females.

Submit

Learn By Doing

1/1 point (graded)

Bearing in mind that sample 1 is the sample of females, and sample 2 is the sample of males, according to the output what is y_2-bar?

13.33

Answer

Correct: 13.33 is the sample mean importance score in sample 2, the sample of males.

Submit

Learn By Doing

1/1 point (graded)

Bearing in mind that sample 1 is the sample of females, and sample 2 is the sample of males, according to the output what is s_1?

4.25

Answer

Correct: 4.25 is the sample standard deviation of sample 1, the sample of females.

Submit

Learn By Doing

1/1 point (graded)

Bearing in mind that sample 1 is the sample of females, and sample 2 is the sample of males, according to the output what is s_2?

4.02

Answer

Correct: 4.02 is the sample standard deviation of sample 2, the sample of males.

Submit

And when we put it all together we get that indeed,

$$t = \frac{(\overline{y}_1 - \overline{y}_2) - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{10.73 - 13.33}{\sqrt{\frac{4.25^2}{150} + \frac{4.02^2}{85}}} = -4.66$$

The test statistic tells us what the data tell us about $\mu_1 - \mu_2$. In this case that difference (10.73 - 13.33) is 4.66 standard errors below what the null hypothesis claims this difference to be (0). 4.66 standard errors is quite a lot and probably indicates that the data provide evidence against H_o.

We have completed step 2 and are ready to proceed to step 3, finding the p-value of the test.