Hypothesis Testing

Task 5

Your last task is determining whether the average number of minutes watched in the US is similar to that in India.

Understanding the differences in usage patterns can help in product localization. The platform might need to tailor its content, features, or user interface to better fit the preferences or needs of users in different regions.

You'll focus only on free-plan students in 2022. Use the Excel sheet Task 5 to perform your calculations.

Your null hypotheses should (respectively) include the following:

- The engagement (minutes watched) in the US is higher than or equal to that in India ($\mu 1 \geq \mu 2$). We test only free-plan students.
- The engagement (minutes watched) in the US is lower than that in India ($\mu 1 < \mu 2$). We test only free-plan students.

Additionally, perform a two-sample t-test assuming unequal variances.

Optional: *Perform a two-sample f-test for variances to support the assumptions.*

What conclusion can you draw from this test? Is the engagement in the US higher than that in India?

Tip: Note that the degrees of freedom are calculated using the following formula for independent samples with unknown variances which are assumed to be unequal:

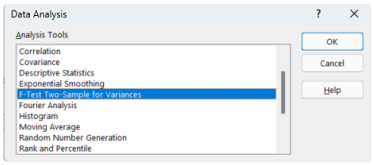
$$df = \frac{\left(\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}\right)}{\frac{\left(s_x^2/n_x\right)^2}{n_x - 1} + \frac{\left(s_y^2/n_y\right)^2}{n_y - 1}}$$

Note: Assume that the degrees of freedom are equal to 11,001.

The t-test is equal to the following:

$$T = \frac{(\bar{x} - \bar{y}) - \mu_d}{\sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}}$$

First, you must perform a two-sample f-test for variances to prove that assumption of unequal variances between the samples (minutes watched by free-plan subscribers in the US and India):



Excel will perform the f-test for variances and provide the test statistic (f-value) and the p-value. The p-value would indicate the probability of obtaining the observed f-value if the null hypothesis (equal variances) were true.

As mentioned, compare the p-value to your chosen significance level (alpha) to determine if the variances are significantly different. If the p-value is less than or equal to your alpha level, you can reject the null hypothesis of equal variances.

The next step is to perform a t-test and compare it to the critical value from the t-distribution.

Consider the following steps for hypothesis testing where the variances are assumed unequal:

1. Specify the significance level:

$$\alpha = 1 - Confidence\ Level = 1 - 0.95 = 0.05$$

2. Calculate the t-statistic using the following formula:

$$T = \frac{(\bar{x} - \bar{y}) - \mu_d}{\sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}}$$

3. Look up the critical t-value using a t-distribution table (https://365datascience.com/calculators/tables/std-table/) to correspond to your chosen significance level (commonly 0.05) and calculated degrees of freedom.

4. Note that the degrees of freedom are calculated using the following formula for independent samples with unknown variances which are assumed to be unequal:

$$df = \frac{\left(\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}\right)}{\frac{\left(s_x^2/n_x\right)^2}{n_x - 1} + \frac{\left(s_y^2/n_y\right)^2}{n_y - 1}}$$

Note: Assume that the degrees of freedom are equal to 11,001.

5. Compare t-statistic to critical t-value. Interpret the magnitude and the sign of your t-statistic. The decision rule—based on the critical value approach—is as follows:

$$If\ T \leq -t_{df,0.05}\ ,\ reject\ H_0$$

$$If\ T > -t_{df,0.05}\ ,\ fail\ to\ reject\ H_0$$

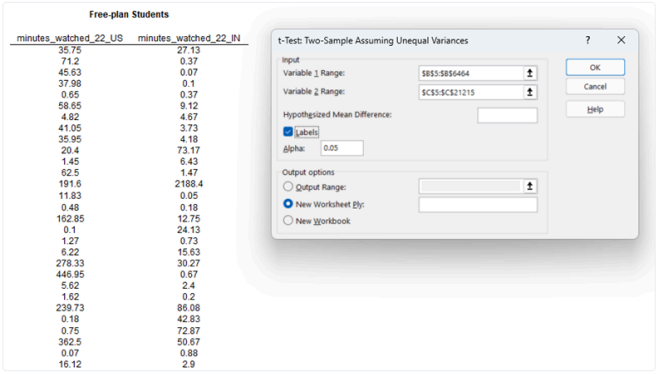
Assume that the critical t-value is equal to 1.65.

The decision rule—based on the p-value approach—is as follows:

$$p - value\ \leq\ \alpha\ ,\ Reject\ H_0$$

$$p - value\ >\ \alpha\ ,\ Fail\ \ to\ reject\ H_0$$

Alternatively, you can use the Data Analysis ToolPak in Excel to obtain the result directly:



Free-plan Students

minutes_watched_22_US	minutes_watched_22_IN
35.75	27.13
71.2	0.37
45.63	0.07
37.98	0.1
0.65	0.37
58.65	9.12
4.82	4.67
41.05	3.73
35.95	4.18
20.4	73.17
1.45	6.43
62.5	1.47
191.6	2188.4
11.83	0.05
0.48	0.18
162.85	12.75
0.1	24.13
1.27	0.73
6.22	15.63
278.33	30.27
446.95	0.67
5.62	2.4
1.62	0.2
239.73	86.08
0.18	42.83
0.75	72.87
362.5	50.67
0.07	0.88
16.12	2.9

Excel will perform the two-sample t-test assuming unequal variances and provide the results, including the t-statistic, degrees of freedom, and the p-value. The p-value would indicate the probability of obtaining the observed t-statistic if the null hypothesis (no difference between means) were true.

Compare the p-value to your chosen significance level (alpha) to determine if the difference between the means of the two samples is statistically significant. If the p-value is less than or equal to your alpha level, you can reject the null hypothesis of similar means.