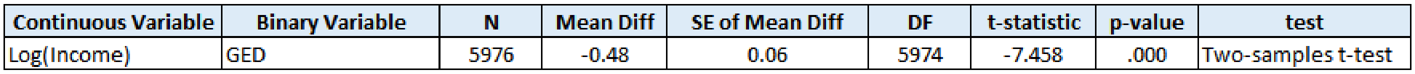
GR 5072 Activity

1. Your task is to create a function which will conduct a two-samples independent t-test. The independent t-test, or the two-sample t-test, is a procedure that tests whether there is a statistically significant difference between the means of two groups. In short, the t-test is a bivariate statistical test which compares a continuous variable to a binary variable to assess whether the two are statistically associated. If this test is significant, then the two variables are deemed to be associated. Otherwise, they are independent. If you need a refresher about what the t-test is and how it is conducted, read here.

You will write a function called **t_test** which will return a 9 x 1 pd.DataFrame with the following information:

- The continuous variable's name
- The binary variable's name
- The total sample size
- The mean difference between the two groups (rounded to 2 decimal places)
- The SE of the mean difference (rounded to 2 decimal places)
- The DF
- The t-statistic (rounded to 3 decimal places)
- P-value (rounded to 3 decimal places, with no leading 0)

Visually, this table will look like this if you are using Log(income) and GED as the two variables you are comparing:



Continuous Variable	Binary Variable	N	Mean Diff	SE of Mean Diff	DF	t-statistic	p-value	test
Log(Income)	GED	5976	-0.48	0.06	5974	-7.458	.000	Two-samples t-test

Name your function **t_test**, and allow it to take two arguments as inputs:

- num_var: This is the continuously distributed variable
- bin_var: This is the categorical binary variable

To create this function, you'll need the following formulas:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \qquad s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

Where *t* is the *t-statistic*, $\bar{x}_1$ and $\bar{x}_2$ are the sample means of the two groups, s_1^2 and s_2^2 are the sample variances of the two groups, n_1 and n_2 are the sample sizes of the two groups. These quantities are all functions of the input data, and you should compute them using pandas functions. s_p refers to the pooled standard deviation and is a function of the sample sizes and sample variances in the two groups. The mean difference is equal to the numerator of the t-statistic, $\bar{x}_1$ - $\bar{x}_2$. The standard error of the mean difference is equal to the denominator of the t-statistic, or $s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$. Finally, the DF is equal to $n_1 + n_2 - 2$.

2. Once you’ve created your t-test function, we’re going to test it. First, import the dataset “ged_data.csv”. Second, call your function with the following arguments: **t_test(num_var=data[“income_log”], bin_var==data[“ged”].** To compare the accuracy of your function, import the scipy.stats module to conduct a t-test using the ttest_ind function using the code below. Compare the results to the function you created to make sure you get the same result. If you didn’t, then fix your t_test function!

```
import scipy.stats as stats

output =stats.ttest_ind(data['income'][data['ged'] == 1],

          data[‘income_log’][data['ged'] == 0])

output
```

3. Once you’ve verified your t-test function is working, write a 3 sentence description of the results as if you were reporting them in a thesis or publication. Again, use income as the numeric variable (the DV in this case), and the GED as the binary IV. This write-up should conform to APA standards.

4. Call your **t_test** function and conduct these three additional tests. Just run them using your function and save the output, no need to write up the results.

- Income and HSG
- Income and female
- Post_sec and female

5. Fit a simple linear regression of Log(income) on GED status. Once you’ve estimated this regression, what are the t-statistic and p-value? Are they similar to what you found in number 2 and 3? Why or why not?