

## Below is information to share with your candidate

### Irrigation Impact

One thing we noted from the two farmers who gave statements was that irrigation seemed to be an important factor to stabilize crop yields, given the regression output below that looks at the last 3 years of data, what is the total yield (dependent variable) for a **100 acre plantation with and without irrigation in a drought year**? And how might you improve this basic model?

Coefficients:					
	Estimate	Std. Error	t value	Pr(> t )	Significant
(Intercept)	8.247	1.972	4.182	4.76E-05	***
Drought	-1.999	1.285	-1.556	0.122	
Irrigation	6.007	1.096	5.482	1.61E-07	***
Drought and Irrigation	0.427	1.512	0.283	0.778	
Acres	0.103	0.007	14.029	2.00E-16	***
Residual standard error: 40.14 on 160 degrees of freedom					
F-statistic: 249.7 on 4 and 160 DF, p-value: < 2.20E-16					
Multiple R-squared: 0.8619,					
Adjusted R-squared: 0.8585					
Signif. codes:	*** 0.001				
	** 0.01				
	* 0.05				
	. 0.1				

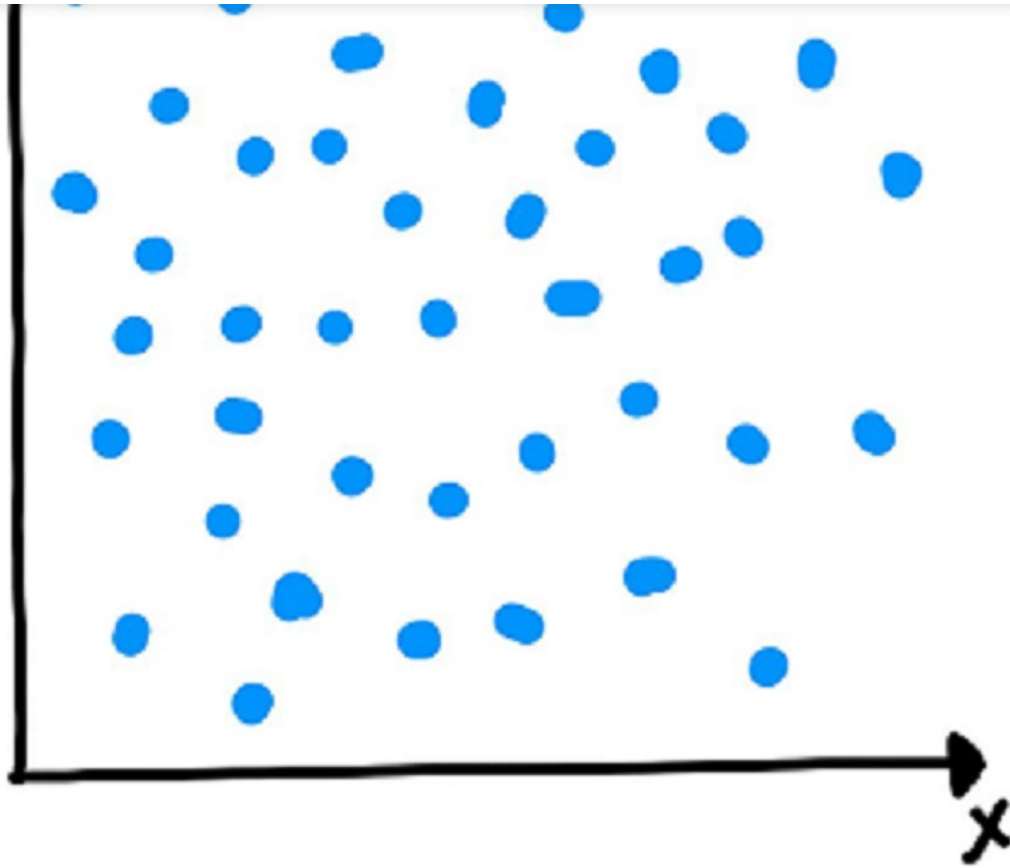
Table 1 – Regression output – Impact of Irrigation for Crop Yield

# Hypothesis Testing in Linear Regression

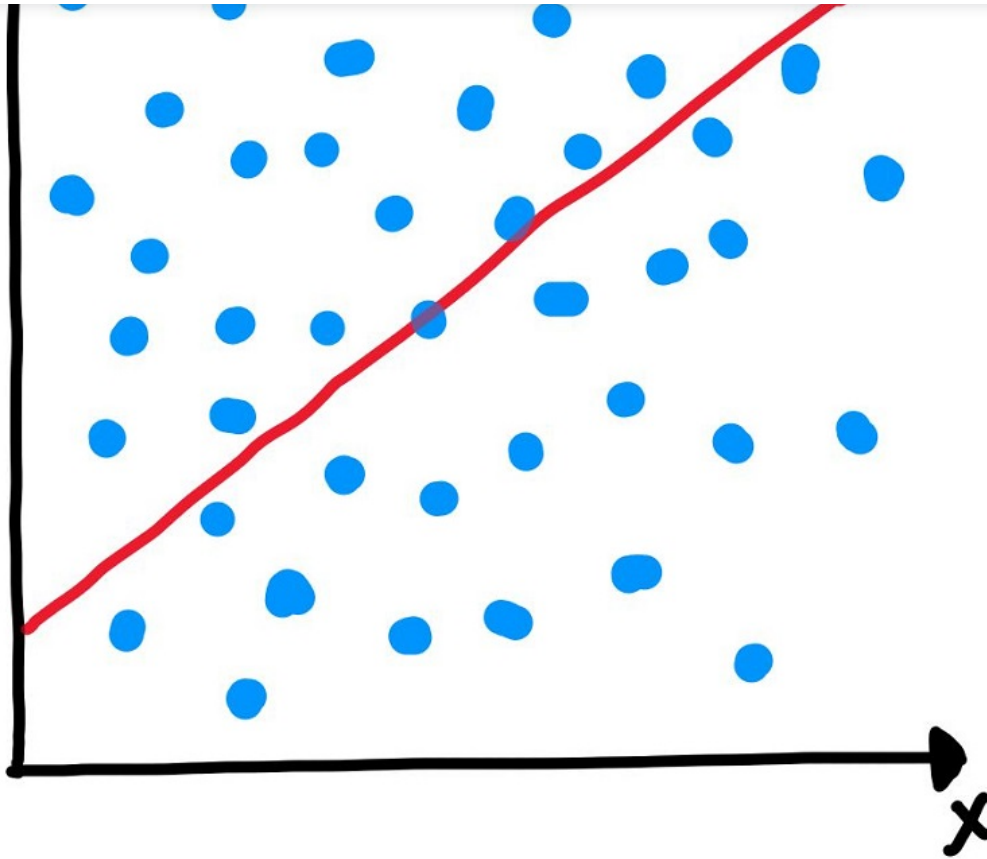
Before you move on to the model building part, there is still one theoretical aspect left to be addressed - the significance of the derived beta coefficient. When you fit a straight line through the data, you'll obviously get the two parameters of the straight line, i.e. the intercept ( $\beta_0$ ) and the slope ( $\beta_1$ ). Now, while  $\beta_0$  is not of much importance right now, but there are a few aspects surrounding  $\beta_1$  which need to be checked and verified.

The first question we ask is, "Is the beta coefficient significant?" What does this mean?

Suppose you have a dataset for which the scatter plot looks like the following:



Now, if you run a linear regression on this dataset in Python, Python will fit a line on the data which, say, looks like the following:



Now, you can clearly see that the data is randomly scattered and doesn't seem to follow a linear trend or any trend, in general. But Python will anyway fit a line through the data using the least squared method. But you can see that the fitted line is of no use in this case.

Hence, every time you perform a linear regression, you need to test whether the fitted line is a significant one or not or to simply put it, you need to test whether  $\beta_1$  is significant or not. And in comes the idea of Hypothesis Testing on  $\beta_1$ . **Please note** that the following text will assume the knowledge of hypothesis testing, which was covered in one of the earlier modules. Please revisit the [module on hypothesis testing](#) in case you need to brush up.

$\beta_1$  is 0. And the alternative hypothesis thus becomes  $\beta_1$  is not zero.

- **Null Hypothesis** ( $H_0$ ):  $\beta_1 = 0$
- **Alternate Hypothesis** ( $H_A$ ):  $\beta_1 \neq 0$

Let's first discuss the implications of this hypothesis test. If you fail to reject the null hypothesis that would mean that  $\beta_1$  is zero which would simply mean that  $\beta_1$  is insignificant and of no use in the model. Similarly, if you reject the null hypothesis, it would mean that  $\beta_1$  is not zero and the line fitted is a significant one.

Now, how do you perform the hypothesis test? Recall from your hypothesis testing module that you first used to compute the **t-score** (which is very similar to the **Z-score**) which is given by  $\frac{X - \mu}{s / \sqrt{n}}$  where  $\mu$  is the population mean and  $s$  is the sample standard deviation which when divided by  $\sqrt{n}$  is also known as standard error.

Using this, the t-score for  $\hat{\beta}_1$  comes out to be (since the null hypothesis is that  $\beta_1$  is equal to zero):

$$\frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)}$$

Now, in order to perform the hypothesis test, you need to derive the p-value for the given beta. If you're hazy on what **p-value** is and how it is calculated, it is recommended that you revisit the [segment on p-value](#). Please note that the formula of  $SE(\beta_1)$  provided in the t-score above is out of scope of this course.

hypothesis that we have stated) on the distribution

- Calculate the **p-value** from the cumulative probability for the given t-score using the t-table
- Make the decision on the basis of the p-value with respect to the given value of  $\beta$  (significance level)

Now, if the p-value turns out to be less than **0.05**, you can reject the null hypothesis and state that  $\beta_1$  is indeed significant.

Please note that all of the above steps will be performed by Python automatically, which you'll learn in the very next segment.

**Question 1/5**

Mandatory

**Hypothesis Test**

What does it mean if you fail to reject the Null hypothesis in the case of simple linear regression?

☐  $\beta_1$  and thus, the independent variable it is associated with is significant in the prediction of the dependent variable.

☒  $\beta_1$  and thus, the independent variable it is associated

✓ Correct

$$\beta_1 = 0$$

Thus, if we fail to reject the Null hypothesis, it means that  $\beta_1$  is indeed zero, and thus insignificant for the prediction of the dependent variable.

- ☐  $\beta_0$  and thus, the independent variable it is associated with is significant in the prediction of the dependent variable.
- ☐  $\beta_0$  and thus, the independent variable it is associated with is insignificant in the prediction of the dependent variable.



Your answer is Correct.

Attempt 1 of 2

Continue

## Coming up

Now that you know how to determine whether your beta is significant or not, you'll start building the model in the next segment

### Additional Reading

Why does the test statistic for  $\beta_1$  follow a t-distribution instead of a normal distribution? ([here](#))



[Report an error](#)

PG Diploma in  
Data Science

 **Learn**

 **Live**

 **Careers**

 **Discussion**

Aug 2020

 **Navigate**



**Q&A**





# Building a Linear Model

Since 'TV' is very strongly correlated to 'Sales', let's first build a simple linear regression model with 'TV' as the predictor variable.



The first important step before building a model is to perform the test-train split. To split the model, you use the **train\_test\_split** function.

From now on, you will always use the SKLearn library to perform a test-train split before fitting a model on any data.



After you import the **statsmodel.api**, you can create a simple linear regression model in just few steps.

```
import statsmodels.api as sm
X_train_sm = sm.add_constant(X_train)
lr = sm.OLS(y_train, X_train_sm)
lr_model=lr.fit()
```



Now, let's take a look again at the summary statistics that was outputted by the model.

## Summary Statistics

Now, let's take a look at the summary statistics that was outputted by the model again.

### OLS Regression Results

Dep. Variable:	Sales	R-squared:	0.816
Model:	OLS	Adj. R-squared:	0.814
Method:	Least Squares	F-statistic:	611.2
Date:	Tue, 18 Jun 2019	Prob (F-statistic):	1.52e-52

	coef	std err	t	P> t	[0.025	0.975]
const	6.9487	0.385	18.068	0.000	6.188	7.709
TV	0.0545	0.002	24.722	0.000	0.050	0.059
Omnibus: 0.027 Durbin-Watson: 2.196						
Prob(Omnibus): 0.987 Jarque-Bera (JB): 0.150						
Skew: -0.006 Prob(JB): 0.928						
Kurtosis: 2.840 Cond. No. 328.						

## Summary Statistic

## F-statistic

You were introduced to a new term named **F-statistic** and **Prob(F-statistic)**. Now, recall that in the last segment, you did a hypothesis test for beta to determine whether or not the coefficient  $\beta_1$  outputted by the model was significant or not. Now, F-statistic is similar in the sense that now instead of testing the significance of each of the betas, it tells you whether the overall model fit is significant or not. This parameter is examined because many a time it happens that even though all of your betas are significant, but your overall model fit might happen just by chance.

The heuristic is similar to what you learnt in the normal p-value calculation as well. If the '**Prob (F-statistic)**' is less than **0.05**, you can conclude that the overall model fit is significant. If it is greater than 0.05, you might need to review your model as the fit might be by chance, i.e. the line may have just luckily fit the data. In the image above, you can see that the p-value of the F-statistic is **1.52e-52** which is practically a zero value. This means that the model for which this was calculated is definitely significant since it is less than 0.05.

## R-squared

Like you studied earlier as well, R-squared value tells you exactly how much variance in the data has been explained by the model. In our case, the R-squared is about 0.816 which means that the model is able to explain 81.6% of the variance which is pretty good.

## Coefficients and p-values:

The p-values of the coefficients (in this case just one coefficient for TV) tell you whether the coefficient is significant or not. In this case, the coefficient of TV came out to be 0.0545 with a standard error of about 0.002. Thus, you got a t-value of 24.722 which lead to a practically **zero p-value**. Hence, you can say that your coefficient is indeed significant.



Apart from this, the summary statistics outputs a few more metrics which are not of any use as of now. But you'll learn about some more of them in multiple linear regression.











Let's see how the model actually looks by plotting it.



You visualised the predicted regression line on the scatter plot of the training data which is one of the things you should do as a part of model evaluation.



  **Question 4/4** Mandatory

	Question 1	Incorrect	
	Question 2	 Correct	
	Question 3	 Correct	
	Question 4	Incorrect	

## Coming up

Now that you have fit the straight line, you will analyse the residuals and make predictions on the test set, in the next segment.

## Additional Reading

PG Diploma in  
Data Science

 **Learn**

 **Live**

 **Careers**

 **Discussion**

Aug 2020

 **Navigate**

 **Q&A**

 [Report an error](#)



PREVIOUS

Hypothesis Testing in  
Linear Regression

NEXT

Residual Analysis and  
Predictions



# The t-Test

Now that you have learnt all the basics of hypothesis testing, you are now well equipped to frame a hypothesis, test it, and make a decision to reject or not reject the null hypothesis. (This is done considering the fact that the population standard deviation for the data is known and the sample size is greater than 30.)

But how will you test the hypothesis if these conditions are not fulfilled? Let's find out. The t-distribution, as you studied earlier, is kind of a normal distribution; it is also symmetric and single peaked but less concentrated around its peak. In layman's terms, a t-distribution is shorter and flatter around the centre than a normal distribution. It is used to study the mean of a population that has a distribution fairly close to a normal distribution (but not an exact normal distribution).

Two simple conditions to determine when to use the t-statistic are as follows:

1. **The population standard deviation is unknown.**
2. **The sample size is less than 30.**

Even if one of them is applicable in a situation, you can comfortably go for a t-test. The formula to determine the t-statistic is:

$$t = \frac{x - \mu}{s / \sqrt{n}}$$



The NATIONAL HIGHWAYS AUTHORITY OF INDIA (NHAI) stated that the **average number of accidents per month** on national highways is **12,000**. A researcher wanted to test this claim. To that end, he collected **25 samples** for 25 months and found out that the **sample mean** was **13,105** and the **sample standard deviation** was **1638.4**.

Let's now try to solve this problem according to the steps we discussed earlier.

The hypothesis for this case will be:

$$H_o: \mu = 12000$$

$$H_a: \mu \neq 12000$$

In this case, the population standard deviation is not given. So, you will calculate the t-statistic.

$$\begin{aligned} t &= (x - \mu) / (s/\sqrt{n}) \\ &= (13105 - 12000)/(1638.4/\sqrt{25}) \\ &= 1105/327.68 \\ &= 3.37 \end{aligned}$$

Now, as in the case of a normal test, you need to compare the value you calculated with the tabular value.

For a 90% confidence interval and a sample size of 25, the **critical t value is 1.71**.

(Here is a link to the tutorial of critical t-value calculation:

<http://www.dummies.com/education/math/statistics/how-to-find-t-values-for-confidence-intervals/>.)

Thus, our acceptance region lies between +1.71 and -1.71.

the number of accidents is equal to 12,000 per month on the highways.

With this example, you have a complete understanding of the one-sample t-test. Let's now focus on the **two sample t-test**. As the name suggests, this test is conducted on two sets of sample data in order to **compare the means of two samples**.

Note that a two-sample test can be performed for multiple statistical parameters, but you are going to focus only on the two-sample test for means, where the standard deviations of both the samples are unknown.

The formula for the two-sample t-test is:

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

df = smaller of  
 $n_1 - 1$  or  $n_2 - 1$

Suppose that you want to come up with a hypothesis test regarding the mean age difference between men and women. You can use the two-sample t-test in such a case.

 [Report an error](#)



PREVIOUS  
The Z-Test

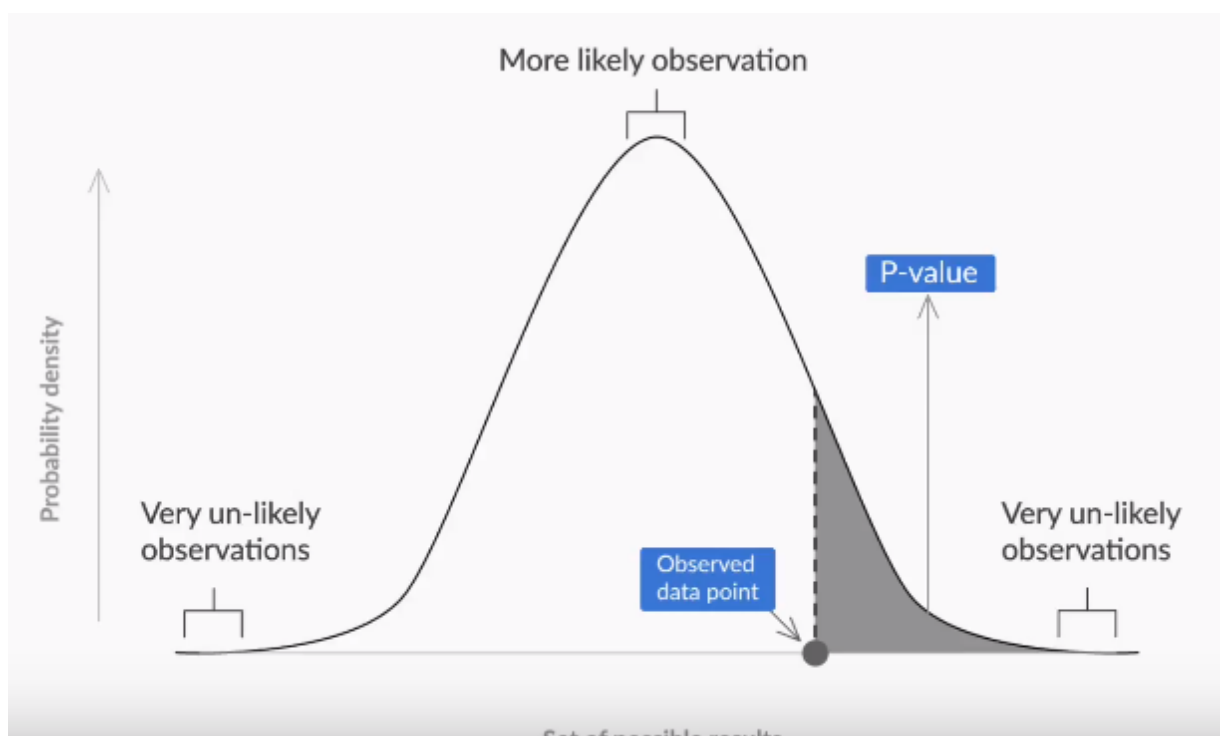
NEXT  
Chi-Squared Test



# The p-Value Approach

The concept of p-value is very important in the field of statistics because of one solid advantage it has over the critical value method; you don't have to state the significance level before conducting the hypothesis test in the case of the p-value method. It is easier to understand intuitively whether or not you are going to reject the null hypothesis. In this segment, we will be looking at a very typical problem of testing whether a coin is fair or not using the concept of p-value.

Recall the definition of p-value: It states the probability of observing a similar or more extreme observation, given that the null hypothesis is true.



Let's try to understand the definition a little better here because you may not have noticed, but this definition allows us to conduct hypothesis testing on distributions that are **not normal** in nature. ( In fact, hypothesis testing can be done on non-normal distributions. However, given the concepts that you learnt in the previous sessions, only the p-value method is within the scope of what we can discuss here.)

This method is best explained using an example. This is a very common type of question asked in interviews.

### Demonstration

Suppose you toss a coin, the nature of which (whether it is biased or unbiased) you are not aware of. After tossing for 10 times, you observed 8 heads and 2 tails. Now you are asked to test the hypothesis of whether the coin is biased or unbiased. You are also asked to measure the p-value at a 0.05 significance level and make a decision.

Now, the solution methodology for this case may not seem straightforward at first glance, but as a matter of fact, it is quite neat and intuitive.

First, as we always do while conducting a hypothesis test, let's define the null and the alternative hypotheses.

So, what would the null hypothesis be in this case?

Well, according to the question, the null hypothesis of this test is that the coin is unbiased, i.e.,  $P(H) = P(T) = 0.5$ .

$$H_0 : P(H) = 0.5$$

$$H_1 : P(H) \neq 0.5$$

(You can also use  $P(T)$  to denote the null and alternative hypotheses in the case above.)

Now, as stated in the problem, we have observed 8 heads.

Recall what the p-value definition states: It is the probability of observing a similar or more extreme observation, given that the null hypothesis is true.

Let's use this definition in our solution methodology to get the answer.

The solution methodology using the definition of p-value would look somewhat like this:

### Solution methodology

1. Assume the null hypothesis to be true, i.e.,  $P(H)=0.5$ .
2. Here, a similar or more extreme observation would be denoted by  $(\text{Heads} \geq 8)$  and its probability would be given by  $P(\text{Heads} \geq 8)$ .
3. Calculate the probability of  $P(\text{Heads} \geq 8)$ , given that  $P(H) = 0.5$ .
4. Observe that the hypothesis-test is two-tailed. Hence, multiply the previous probability by 2. This would be the p-value of this test.

### Explanation

First, we assumed that the null hypothesis is true. Then we checked the current observation and tried to deduce what the extreme version of this observation might be from the given null hypothesis.

(But you can also say that observing 1 or 2 heads can also be an extreme observation. How do we take that into consideration? You will see how in a short while.)

Step 3 is the most crucial step. Here, we leverage the definition to calculate the p-value. Given that the null hypothesis is true, i.e.,  $P(H) = 0.5$ , we are about to calculate the probability of getting similar or extreme observations, which is the probability as given by  $P(\text{Heads} \geq 8)$ .

If you observe carefully, you will see that it is equivalent to calculating the probability of observing 8 or more heads in a coin toss experiment where the unbiased coin is flipped 10 times.

Or, the aforementioned problem can be reduced to that of calculating the cumulative probability of a binomial distribution, with  $p = 0.5$ ,  $n = 10$  and  $r = 8$ .

Thus,  $P(\text{Heads} \geq 8) = P(X \geq r) = P(X \geq 8) = P(X = 8) + P(X = 9) + P(X = 10) = {}^{10}C_8 0.5^8 0.5^2 + {}^{10}C_9 0.5^9 0.5^1 + {}^{10}C_{10} 0.5^{10} = 0.055$ .

Thus, the probability of  $P(\text{Heads} \geq 8)$  is now calculated. Now, note that this would be analogous to a two-tailed test because from the null hypothesis, we can infer that the extreme observations can occur at both ends, i.e., it can be biased towards the tails or heads. (Take a look at the image above to understand the position of the extreme observations.)

So, we can have observations of 2 or 3 heads as another extreme. What do we do now?

And voila! We have the p-value as  $2 * P(\text{Heads} \geq 8) = 2 * 0.055 = 0.11$ .

Given the significance level of 0.05 and the calculated p-value, we can safely say that we fail to reject the null hypothesis.

Now, try to answer the following question to understand an alternative approach to solving this problem. You can learn more about this method [here](#).

< > **Question 1/1**

Mandatory



## The p-Value Approach

Let's say that you want to calculate the p-value using  $P(T) = 0.5$  as the null hypothesis. How would your approach change here? Does the solution, i.e., the p-value, change? Please write the answer below. Use the step-wise methodology mentioned above.

Word Count 0

Word Limit 30 - 150

PG Diploma in  
Data Science

 **Learn**

 **Live**

 **Careers**

 **Discussion**

Aug 2020

 **Navigate**

 **Q&A**

**Attempt 0 of 1**

**Submit**

 [Report an error](#)



**PREVIOUS**  
**Assessments - I**

**NEXT**  
**F-Test**

