**[Associative Statistics – One sample T-Test with TSQL and R](http://www.sqlservercentral.com/blogs/malas-data-blog/2016/09/21/associative-statistics-one-sample-t-test-with-tsql-and-r/)**

Posted on 21 September 2016

FROM: <http://www.sqlservercentral.com/blogs/malas-data-blog/2016/09/21/associative-statistics-one-sample-t-test-with-tsql-and-r/>

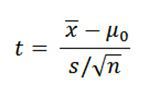
In this post I am going to attempt to explore a statistical procedure called ‘One Sample T Test’.

A T-Test is used to test the mean value of a data sample against a known mean of the entire population from where the sample came from. An example would be, if the average voting preference of the USA is Democrat, I want to test the average voting preference in KY to see if that corresponds to the national average. Another simpler example is if all coke bottles produced are 150 ml on an average – I want to see if this quantity is exactly true (on an average) of the 10 coke bottles I have in the fridge.

For this post I decided to go with a simple example of how many steps I walked with my per day for the month of August. My goal is 10,000 steps per day – that has been my average over the year but is this true of the data I gathered in August? I have a simple table with two columns – day and steps. Each record has how many steps I took in August per day, for 30 days. So – *SELECT AVG(steps) FROM [dbo].[mala-steps]* gives me 8262 as my average number of steps per day in August. I want to know if am consistently under performing my goal, or if this is a result of my being less active in August alone. Let me state my problem first – or state what is called ‘null hypothesis’:

***I walk 10,000 steps on an average per year.***

How is T value calculated? The formula for T value is a bit complex –



The numerator x bar is the mean of the sample. The mew-zero as it is spelled is the hypothesised mean – or what I say I expect of the sample value – in my case , 10,000 steps.. The denominator – s, is the standard deviation of the sample or the square root of sum of difference between mean and each value, and n refers to sample size. Without pulling hair out on what this stands for etc – what it really means is the ratio of differences between sample values to the mean compared to ‘inner noise’, or the difference within the sample set itself. If you get a high value it means the sample set is probably not fitting my hypothesis, or there are too many differences between values and the hypothesised mean. A low value means the opposite, that the differences are internal to the sample.

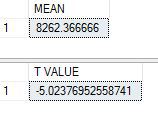
My goal is to prove or disprove this with the sample selected from August. I am using a significance level of 0.01 or what is called 99% confidence level.

Using TSQL:

SELECT AVG(steps) AS 'MEAN' FROM [dbo].[mala-steps]

SELECT (AVG(STEPS)-10000)/(SQRT(VAR(STEPS))/SQRT(COUNT(\*))) AS 'T VALUE'

FROM [dbo].[mala-steps]



We get a mean of 8262.36 and a T value of -5.023.

Calculating the p value or probability for this T value can be done via calculator [here](https://graphpad.com/quickcalcs/PValue1.cfm) – unfortunately this is not possible with TSQL. If we stick in the values of 5.023 and 29 degrees of freedom (30 values – 1) we get a really low p value

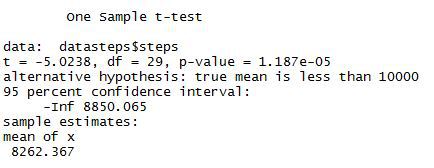
R does the entire math for us in one simple step as below:

install.packages("RODBC")

library(RODBC)

datasteps <- sqlQuery(cn, 'SELECT steps FROM [dbo].[mala-steps]')

t.test(datasteps$steps, mu=10000, alternative="less", conf.level=0.95)



The same R code can be called from within TSQL too – giving results as below:

-- calculate one way T value

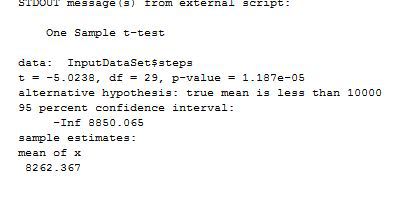
EXEC sp\_execute\_external\_script

@language = N'R'

,@script = N' tvalue <-t.test(InputDataSet$steps, mu=10000, alternative="less", conf.level=0.95);

print(tvalue)'

,@input\_data\_1 = N'SELECT steps FROM [dbo].[mala-steps];'



From the output, we can see that the mean number of steps I have taken for the month of August is 8262. The t value is 5.02 (the sign does not matter). Which means that difference between individual values and the mean is higher than the ‘inner noise’ or the difference between values in the sample sample set.

The p-value ,or the probability of getting this t value from this sample is really really low –  1.187e-05.So, it is more likely this dataset/sample is probably not a good one – or in other words, i can’t accept the null hypothesis that I did walk an average of 10,000 steps a day based on this sample. Maybe August was not the best month to judge my commitment to exercise…or maybe i’d have to try more samples than August alone! More on those in next post!