T-test with two-sample assuming equal variances using Excel

Submitted Research to Obtain the Bachelor's Degree in Applied Mathematics

Preparation by

سارة علي عبد الرحمن الهاجري

Supervised by

Dr. Salem A. Alyami

Department of Mathematics and Statistics – College of Science

Imam Mohammad Ibn Saud Islamic University

1444 H - Semester 2

بسم الله الرحمن الرحيم

Table of Contents

1. Definitions

1.1. Empirical Data

Empirical data is an information collected by scientists and practitioners through observational learning or doing experimentation. Practically, gathering empirical data play a crucial role in real life applications to get a better understanding of phenomenal beliefs and concepts.

Typically, there are two types of empirical data: qualitative and quantitative. The later type is a data that can be measurable and countable in numerical values. Quantitative data represented by two main types: discrete data such as number of patients, number of surgeries in a hospital, and number of car accidents; and continuous data such as weight, height, and time. Whereas qualitative data gathered based on qualities which involves descriptions of observed data and not expressed numerically nor measurable.

1.2. Random variable

A random variable is a measurable function from a sample space described by a set of possible events or outcomes to a measurable space. We use uppercase letters such as X or Y to mathematically refer random variables. That is, if X and Y are random variables, then $f(X)$ and $g(Y)$ are used to indicate their real functions for all possible values of X and Y, respectively.

1.3. Population and sample

A population is the entire group of items that a researcher targets to draw data from, for practical study and conclusions. A population can be a set of people, elements, organisations, species, events, or any other objects, in which they belong to a specific area and time. Populations can be either limited e.g. patients who diagnosed with heart disease in a hospital, or unlimited e.g. Hamour fish in the Red Sea, which is intractable; and a sample must be generated.

A sample is a subset of population representing by specific group; collected data from a particular population. The sample is therefore smaller and more manageable compared to the population drawn from it. Further, the sample can be generated using different techniques that often unbiased. One common technique is to generate a sample randomly in which all elements in the population have the same probability to be drawn.

1.4. Mean and standard deviation

The population mean μ is a mathematical measurement to calculate the average of a given population data. We use the symbol $\bar{x}$ to refere to the sample mean of a given sample $x_1, x_2, \dots, x_n$, as follows

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}(x_1 + x_2 + \cdots + x_n)$$

Where n is the sample size or number of data, x_i is an observed value i in the sample data.

The standard deviation of a population σ is a mathematical measurement to compute dispersion. It is calculated mathematically by taking the square root of the variance σ^2. Below is the formulation to compute the standard deviation for a population:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(X_i - \mu)^2}$$

So that σ measures the amount of variation of a set of data from the mean. When σ is small, the data tend to be close to the mean of the data. When σ is high, then the data spread away from its mean. The standard deviation of a sample is referred by s and is formulated by:

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

1.5. Central Limit Theorem

If X follow a distribution with two measurements μ and σ, with sufficiently large n, then the variable

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

will have, approximately, standard normal distribution.

As we increase the value of n, the approximation is going to be quite good. It is statistically well-recognized to choose $n \geq 30$, to satisfy the condition of normality of data.

1.6. Hypotheses testing

The null hypothesis (H_0) is defined as the hypothesis that we have by default or the hypothesis that we believe in fact. It is achieved if the true difference between the target two-group means is zero.

The alternate hypothesis (H_1) is the hypothesis that we want to test. It is achieved if the true difference between the two-group means is different from zero.

There are two possible outcomes of hypothesis testing: Reject the null hypothesis; or fail to reject it.

1.7. T-test

A t-test is used as an inferential statistic to know if there is a significant difference between the means of two groups.

To apply the Two-sample T-test with equal variance, there are some requirements e.g.:

1- The samples are normally distributed. One possible test to check normality is called Kolmogorov-Smirnov Test. It requires some famous measurements include skewness, kurtosis, data count, standard deviation, and median.
2- The standard deviation of both populations are unknown and assumed to be equal,
3- The size of sample is sufficiently large and over 30.

The t-test is used to answer hypothesis testing in statistical studies.

Calculating t-test, typically, requires three fundamental data values:

1- The difference between the two mean values from each data group.
2- The standard deviation of each data group.
3- The number of data values of each group.

One example to implement t-test, suppose we have a sample of students from class A and another sample of students from class B, assuming both samples would not have the same mean but their standard deviation are equals. Then a t-test is applied to compare the average values of the two data sets and to determine whether or not they came from the same population.

1.8. P-Value

P-value is commonly reported in statistical tests. It is a measurement used to validate a hypothesis against observed data. It measures the probability of obtaining the observed results, assuming that the null hypothesis is true.

In order to address greater statistical significance of an observed difference, the p-value has to be as low as possible.

2. Methodology

This section provides a mathematical framework to apply two samples t-test. Section 2.1 briefly defines some mathematical formulas. Section 2.2 outlines the general formulation of data. Section 2.3 addresses the general hypothesis assumption. Section 2.4 explains how to reach "Data Analysis" tool in Excel to perform the analysis. Section 2.5 lists the general steps on how to implement two samples t-test using "Data Analysis" tool. Section 2.6 helps reader to read the outputs obtained after implementing the test.

2.1. Mathematical formulas

We propose that $\bar{x}$ and $\bar{y}$ are the sample means of two data sets with sizes n_x and n_y, respectively. If x and y are normally distributed, or their n_x and n_y are sufficiently large following the principle of central limit theorem, plus x and y have the same variance, then the t-score is theoretically given by:

$$t = \frac{(\bar{x} - \bar{y}) - (\mu_x - \mu_y)}{s\sqrt{\frac{1}{n_x} + \frac{1}{n_y}}}$$

with a distribution $T(n_x + n_y - 2)$ where

$$s = \sqrt{\frac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{(n_x - 1) + (n_y - 1)}}$$

Note that the t-score is typically used later to be compared with the level of alpha value.

2.2. Data Formulation

The table below shows a simple formulation for the dataset of two variables as it is appeared in most real life studies. The dataset in the below table includes two columns. Each column contains the values belong to one variable. Please note that the number of values in each column is not necessary to be equal.

Individual No.	Variable X	Variable Y
1	x_1	y_1
2	x_2	y_2
3	x_3	$\vdots$
4	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$
n	x_n	y_n

2.3. Hypothesis assumptions

In this project, we can formulate our hypothesis testing as follows:

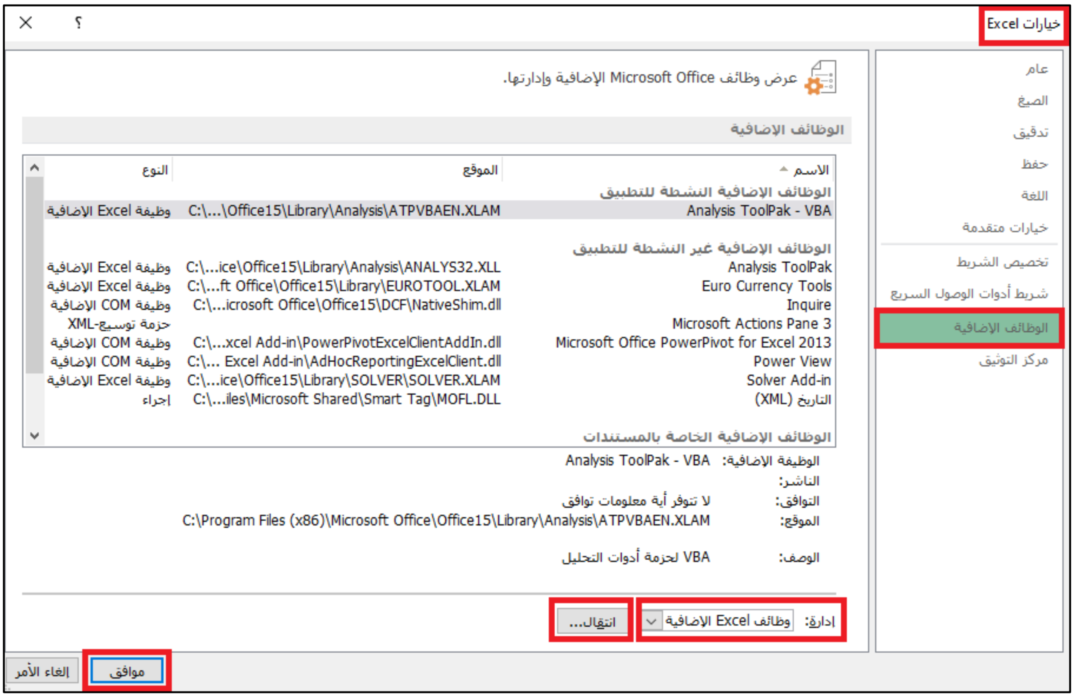
H_0 — There is no a significant difference in the means of those two populations belong to variable X and those belong variable Y, $\mu(X) = \mu(Y)$.

H_1 — There is a significant difference in the means of those two populations belong to variable X and those belong variable Y, $\mu(X) \neq \mu(Y)$ or $\mu(X) > \mu(Y)$ or $\mu(X) < \mu(Y)$.

2.4. Data Analysis ToolPak in Excel

"Data Analysis ToolPak" is a powerful tool available in Excel. It provides several mathematical tests that can be used in applications to analyse data. One mathematical test listed in the TollPak is called "Two-sample t-test". It comes with two assumptions: assuming two samples with equal variance; and two samples with unequal variance. In this project, we assume equality of variance for the two samples.

The figure below illustrates how to add "Data Analysis ToolPak" if it's not already loaded in your Excel. Reader can easily follow the red rectangles shown on the figure.



عرض وظائف Microsoft Office الإضافية وإدارتها.

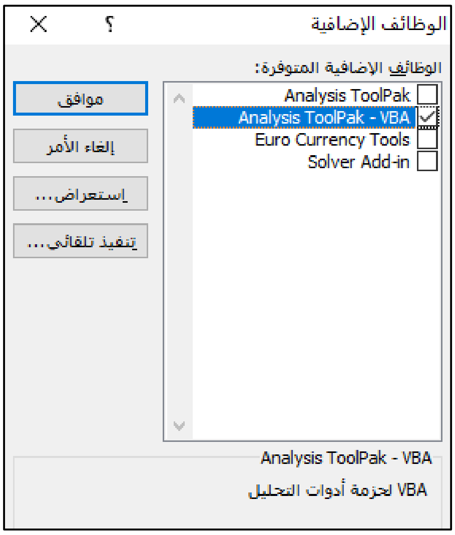
الوظائف الإضافية

الاسم	الموقع	النوع
الوظائف الإضافية النشطة للتطبيق		
Analysis ToolPak - VBA	C:\...\Office15\Library\Analysis\ATPVBAEN.XLAM	وظيفة Excel الإضافية
الوظائف الإضافية غير النشطة للتطبيق		
Analysis ToolPak	C:\...ice\Office15\Library\Analysis\ANALYS32.XLL	وظيفة Excel الإضافية
Euro Currency Tools	C:\...ft Office\Office15\Library\EUROTOOL.XLAM	وظيفة Excel الإضافية
Inquire	C:\...icrosoft Office\Office15\DCF\NativeShim.dll	وظيفة COM الإضافية
Microsoft Actions Pane 3		حزمة توسيع-XML
Microsoft Office PowerPivot for Excel 2013	C:\...xcel Add-in\PowerPivotExcelClientAddIn.dll	وظيفة COM الإضافية
Power View	C:\... Excel Add-in\AdHocReportingExcelClient.dll	وظيفة COM الإضافية
Solver Add-in	C:\...ice\Office15\Library\SOLVER\SOLVER.XLAM	وظيفة Excel الإضافية
التاريخ (XML)	C:\...iles\Microsoft Shared\Smart Tag\MOFL.DLL	إجراء
الوظائف الإضافية الخاصة بالمستندات		

الوظيفة الإضافية: Analysis ToolPak - VBA
الناشر:
التوافق: لا تتوفر أية معلومات توافق
الموقع: C:\Program Files (x86)\Microsoft Office\Office15\Library\Analysis\ATPVBAEN.XLAM

الوصف: VBA لحزمة أدوات التحليل

After clicking OK, the below box will pop up, and simply reader can choose "Analysis ToolPak" by checking on its small square as shown in the figure below, and then press OK, so that the tool will appear on your Excel under "Data" tap.



الوظائف الإضافية المتوفرة:

- □ Analysis ToolPak
- ☑ Analysis ToolPak - VBA
- □ Euro Currency Tools
- □ Solver Add-in

Analysis ToolPak - VBA

VBA لحزمة أدوات التحليل

2.5. Two sample t-test in Excel

We briefly describe the implementation of two samples t-test in Excel in ten steps as follows:

1- Input your two datasets into Excel, by typing each dataset in one separated column.

2- Write down your null and alternate hypothesis.

3- Go to the “Data” tab and then click “Data analysis.” If you do not see the "Data Analysis" option, go to Section 2.4 and follow the description to load it.

4- Choose “t test two sample for means assuming equal variance” from the options window, then press “OK”.

5- Select the first variable list by clicking the “Variable 1 Range” box.

6- Select the second variable list by clicking the “Variable 2 Range” box.

7- Enter zero number into the Hypothesized Mean Difference box. This is because our null hypothesis stated no difference between the means.

8- Importantly, check the “Labels” box in case you have labels in the first row.

9- In the "alpha level box", we choose an alpha level of 0.05 or0.01, which is a standard in hypothesis testing.

10- Select a cell where you like to save the results, by clicking the Output Range box, and lastly click “OK.”

2.6. Reading the results

After running the two-sample t-test, analyst only looks to two values: t-score and alpha level. Then, one of these two comparisons is carried out:

1- Comparing the alpha level chosen by the analyst (commonly set to 0.05) to the p-value resultant in the output. So that, if the p-value is smaller than the alpha level, the null hypothesis is then rejected.

2- Comparing the t-critical value resultant in the output - which can be one-tail or two-tail - with the calculated t-value. If the calculated t-value is larger than the t-critical value, the null hypothesis is then rejected.

3. Application

This Section provides practical application on how to analyse and implement two-sample t-test using Excel to real life data. Therefore, we use real life data published by LibereTexts STATISTICS.

3.1. Empirical Data

My empirical data describes the cholesterol levels of two sets. The first set is a group of patients who had diagnosed with heart attacks, in which their cholesterol levels were measured two days after the heart attack disease. The second set is a group of healthy people who show no signs of heart disease. My goal is to see if patients in the first group have higher cholesterol levels over healthy people in the second group.

3.1.1. Hypothesis assumption

H_0 — There is no difference in the means of cholesterol levels between the heart attack patients μ_1 and healthy adults μ_2; $(\mu_1 = \mu_2)$.

H_1 — There is a difference in the means of cholesterol levels between the heart attack patients μ_1 and healthy adults μ_2; $(\mu_1 > \mu_2)$.

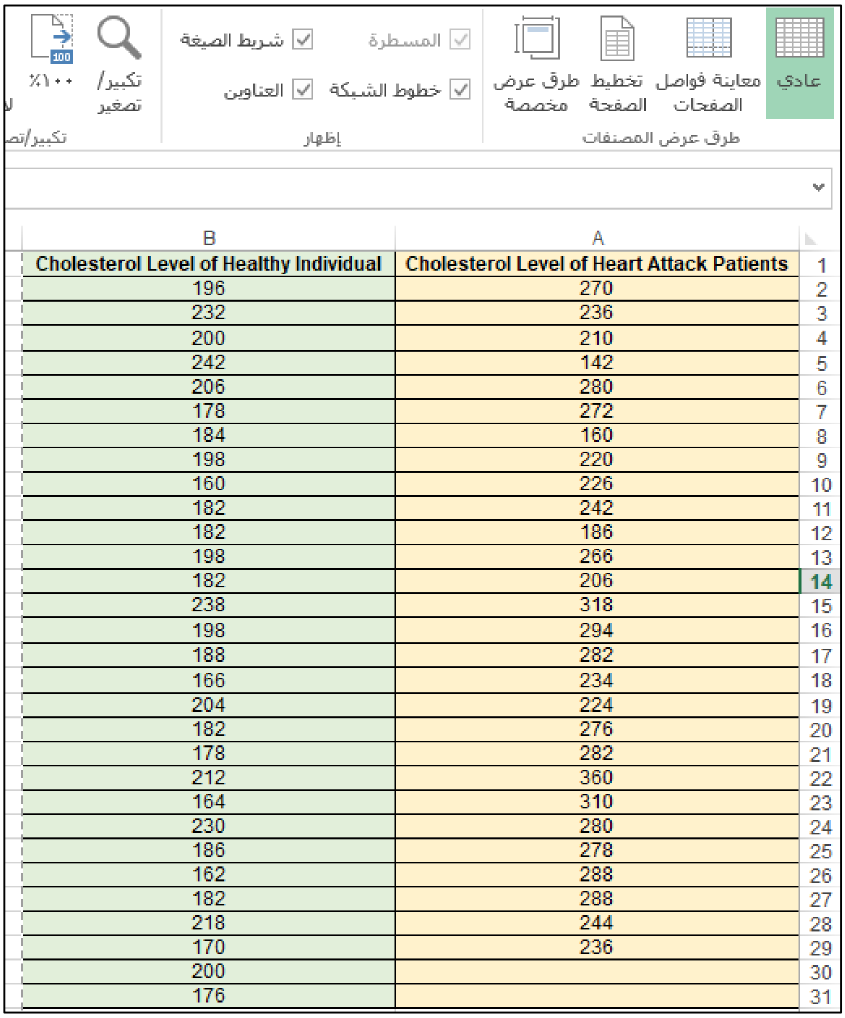
3.1.2.Input empirical data into Excel

I typed all data in Excel sheet as shown in the table below.

1- The data typed in column A represents the cholesterol levels of patients who had heart attack.

2- The data typed in column B represents the cholesterol levels of healthy adults.

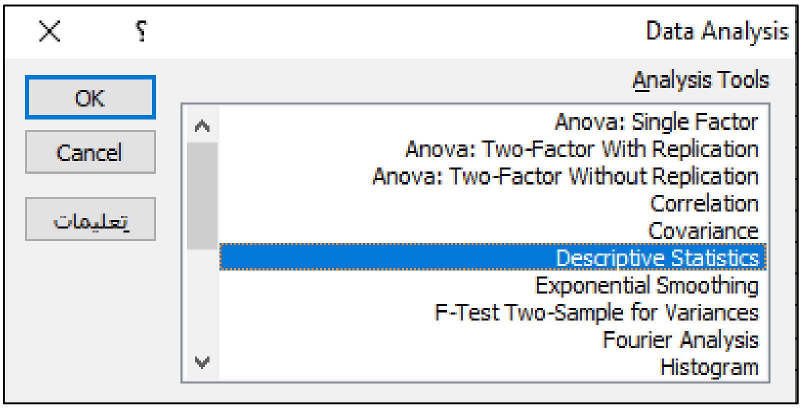
3- The sample size of patients who had heart attack typed in column A is 28, whereas the number of healthy adults typed in column B is 30. As we mentioned earlier in this project, the sample sizes of two-sample are not necessary equal.



B	A	
Cholesterol Level of Healthy Individual	Cholesterol Level of Heart Attack Patients	1
196	270	2
232	236	3
200	210	4
242	142	5
206	280	6
178	272	7
184	160	8
198	220	9
160	226	10
182	242	11
182	186	12
198	266	13
182	206	14
238	318	15
198	294	16
188	282	17
166	234	18
204	224	19
182	276	20
178	282	21
212	360	22
164	310	23
230	280	24
186	278	25
162	288	26
182	288	27
218	244	28
170	236	29
200		30
176		31

3.1.3. Descriptive statistics

In order to get better view of our data, I calculated some statistical measurements by running "Descriptive Statistics" tool available in "Data Analysis" in Excel, as it appeared in the below box:



The means and standard deviations of both groups are highlighted in green color.

Cholesterol Level of Healthy Individual		*Cholesterol Level of Heart Attack Patients*	
193.1333	Mean	253.9285714	Mean
4.071412	Standard Error	9.016435401	Standard Error
187	Median	268	Median
182	Mode	236	Mode
22.30004	Standard Deviation	47.71049156	Standard Deviation
497.292	Sample Variance	2276.291005	Sample Variance
-0.17908	Kurtosis	0.472748497	Kurtosis
0.661633	Skewness	-0.351336484	Skewness
82	Range	218	Range
160	Minimum	142	Minimum
242	Maximum	360	Maximum
5794	Sum	7110	Sum
30	Count	28	Count

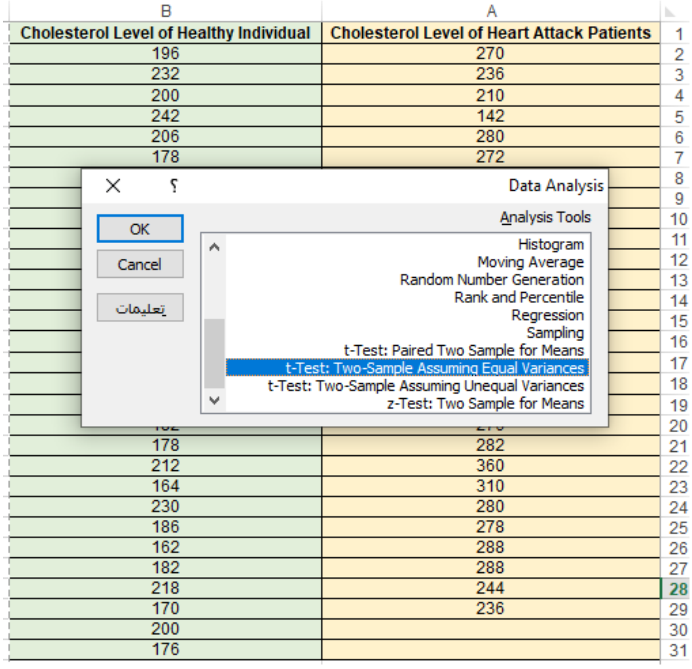
3.2. Performing two-sample t-test in Excel

I applied the two-sample t-test available in "Data analysis" tool in Excel to the data presented in Section 3.1.2. The steps illustrated in Section 2.5 are then followed to perform the analysis.

However, before applying the t-test with equal variance, it is necessary to meet the t-test requirements mentioned in Section 1.7. Specifically, we check the normality of our data. For that, I used the Kolmogorov-Smirnov Test mentioned in Section 1.7. This is done using the online tool available via:

https://www.socscistatistics.com/tests/kolmogorov/default.aspx

The results show that the values of the K-S test statistic of first group and second group are 0.13738 and 0.12817, respectively, which indicate that our data of both groups are normally distributed.



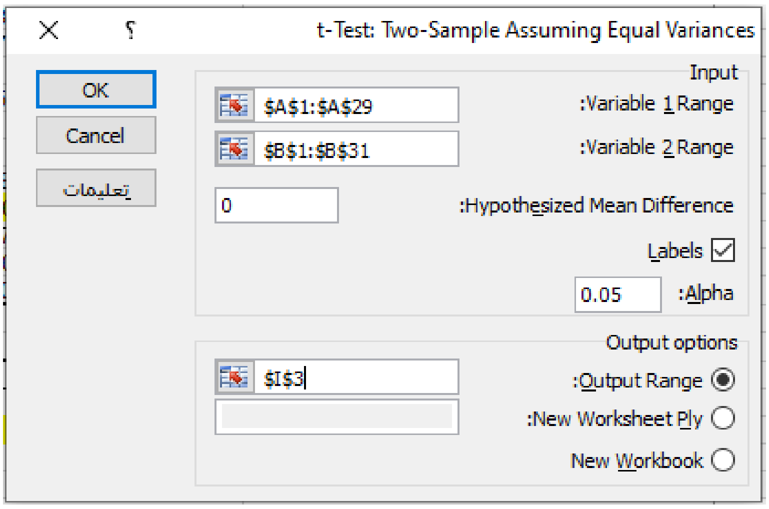
B	A	
Cholesterol Level of Healthy Individual	Cholesterol Level of Heart Attack Patients	1
196	270	2
232	236	3
200	210	4
242	142	5
206	280	6
178	272	7
[illegible]	[illegible]	20
178	282	21
212	360	22
164	310	23
230	280	24
186	278	25
162	288	26
182	288	27
218	244	28
170	236	29
200		30
176		31

Data Analysis

Analysis Tools: t-Test: Two-Sample Assuming Equal Variances

؟
OK
Cancel
تعليمات
Histogram
Moving Average
Random Number Generation
Rank and Percentile
Regression
Sampling
t-Test: Paired Two Sample for Means
t-Test: Two-Sample Assuming Equal Variances
t-Test: Two-Sample Assuming Unequal Variances
z-Test: Two Sample for Means

Above figure shows the access to the t-Test: Two-Sample Assuming Equal Variances. The window box, below, clarifies how the blanks were filled. It is noticeable that "Variable 1 Range" matches all data belong to the heart attack patients, "Variable 2 Range" matches all data belong to the healthy adults. The "labels" tap was checked, and "Alpha" level was set to 0.05 by default.



t-Test: Two-Sample Assuming Equal Variances

Input

:Variable 1 Range $A$1:$A$29

:Variable 2 Range $B$1:$B$31

:Hypothesized Mean Difference 0

Labels ☑

:Alpha 0.05

Output options

:Output Range ● $I$3

:New Worksheet Ply ○

New Workbook ○

The analysis outputs as obtained from Excel are introduced below:

Cholesterol Level of Healthy Individual	*Cholesterol Level of Heart Attack Patients*	
193.1333333	253.9285714	Mean
497.291954	2276.291005	Variance
30	28	Observations
	1355.023639	Pooled Variance
	0	Hypothesized Mean Difference
	56	df
	6.285238672	t Stat
	2.60104E-08	P(T<=t) one-tail
	1.672522303	t Critical one-tail
	5.20208E-08	P(T<=t) two-tail
	2.003240719	t Critical two-tail

3.3. Outputs Interpretation

Given the analysis outputs in Section 3.2, we can address the following:

1- The p-value (0.05) in comparison with alpha level one tailed (2.60104E-08), we can see that: 2.60104E-08 < 0.05.

2- Given that (alpha level < p-value), then we reject the null hypothesis H_0 and accept the alternative hypothesis H_1.

Taking into account the values of means in Section 3.1.3, the analytical comparison proves that there is a difference in the means and therefore the heart attack patients in the first group have higher cholesterol levels over healthy people in the second group.

4. References

Overholser, Brian R; Sowinski, Kevin M (2017). "Biostatistics Primer: Part I". Nutrition in Clinical Practice. 22 (6): 629–35.

Guo, Beibei; Yuan, Ying (2017). "A comparative review of methods for comparing means using partially paired data". Statistical Methods in Medical Research. 26 (3): 1323–1340.

Casella, George, (2008). "Statistical Inference". Springer Texts in Statistics.