



University of Wales
Trinity Saint David

Title of report

Data Analysis and Visualisation

ASSIGNMENT 2

Module Code: ACCB4002

RANGEL GEORGIEV STOITSOV

Student ID: 2029065

Lecturer: Koushik Modak

Date submission:

06/05/2022

Table of Contents

INTRODUCTION:	3
Description of the data:	4
Mean:	5
Median:	6
What is a Hypothesis?.....	7
Null hypothesis and alternative hypothesis.	7
Null hypothesis (H_0)	7
Alternative hypothesis (H_1)	7
Testing of hypotheses to various significance level.....	8
Type I errors and Statistical significance	10
Null hypothesis testing	10
Conclusion	13
References:.....	14

INTRODUCTION:

In the given assessment task, we are required to analyze the given set of data and to analyze the information provided to us about the tests of the classes in the primary school. We were given the tests of 40 students who took two tests, the first result being a test and the students were prepared in the standard way of learning. The second is how to prepare with the help of the new online application. The school received a free math learning app from a local online math learning company developed by it. The company wants to analyze the use of the application as a tool for teaching mathematics in school and hopes that the findings of the analysis and the results will help them decide whether to continue to develop the online application. I will present a description of the data, including a set of descriptive statistics.

Description of the data:

In our representative information we have 40 students who took a test first in the standard way of learning and then with the help of the online application and in both cases were evaluated by the same criteria. The purpose of the data collection is to determine with the help of the online application whether students show better results in tests.

Student	Test score before using app	Test score after using app
1	12	14
2	9	9
3	8	10
4	9	11
5	12	15
6	14	15
7	15	15
8	17	10
9	18	20
10	15	20
11	14	18
12	14	16
13	14	17
14	12	14
15	10	20
16	11	15
17	8	4
18	7	10
19	6	11
20	6	19
21	5	16
22	17	17
23	19	19
24	20	18
25	17	20
26	16	19
27	17	20
28	12	14
29	13	14
30	11	13
31	12	12
32	6	9
33	8	20
34	10	17
35	15	17
36	17	17
37	17	19
38	18	20
39	12	13
40	17	15

Tab 1. Student and results of test

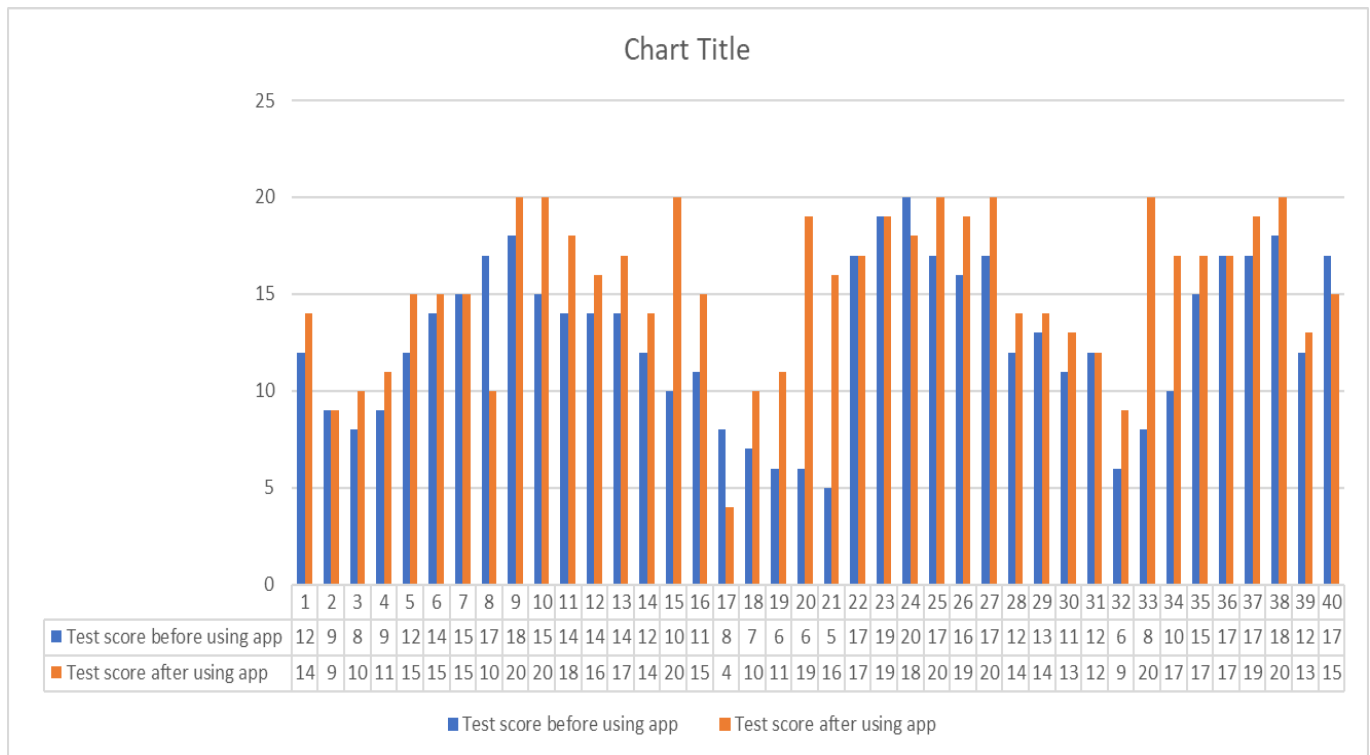


Figure 01: Data presentation by graph

Mean:

The average value is probably the most used method for expressing a central trend. To calculate the average, we simply sum all the values and divide the total by the total. [2] Take for example the following values of the test results:

Mean data before using the application:

$$\begin{aligned} \text{Mean} &= \text{total data} / \text{total number of data} \\ &= 508/40 \\ &= 12.7 \end{aligned}$$

Mean data after using the application:

$$\begin{aligned} \text{Mean} &= \text{total data} / \text{total number of data} \\ &= 614/40 \\ &= 15.35 \end{aligned}$$

Median:

The median is the average result is the result that is right in the center of the set of values. The simplest way to calculate the average is to list all the data in ascending order and then find the exact average, which is the average.[1] There are a total of 40 data in the table. This data must be organized in ascending order to obtain the median, you can use an Excel worksheet to organize them in an expanding range.

Before using the application the median is:

Here 20 and 21 are 12 and 13.

So, the median is $(12 + 13) / 2 = 12.5$

After using the application, the median is:

And the 20th and 21st value is $(16 + 16) / 2 = 16$

Shape of data

The figure shows the data format according to the condition of the set results before using the application and after it is used. The blue bar represents the data before using the application, and the orange bar shows the data after using the application. The graph clearly shows the ups and downs of the data and the variation of the data.[3]

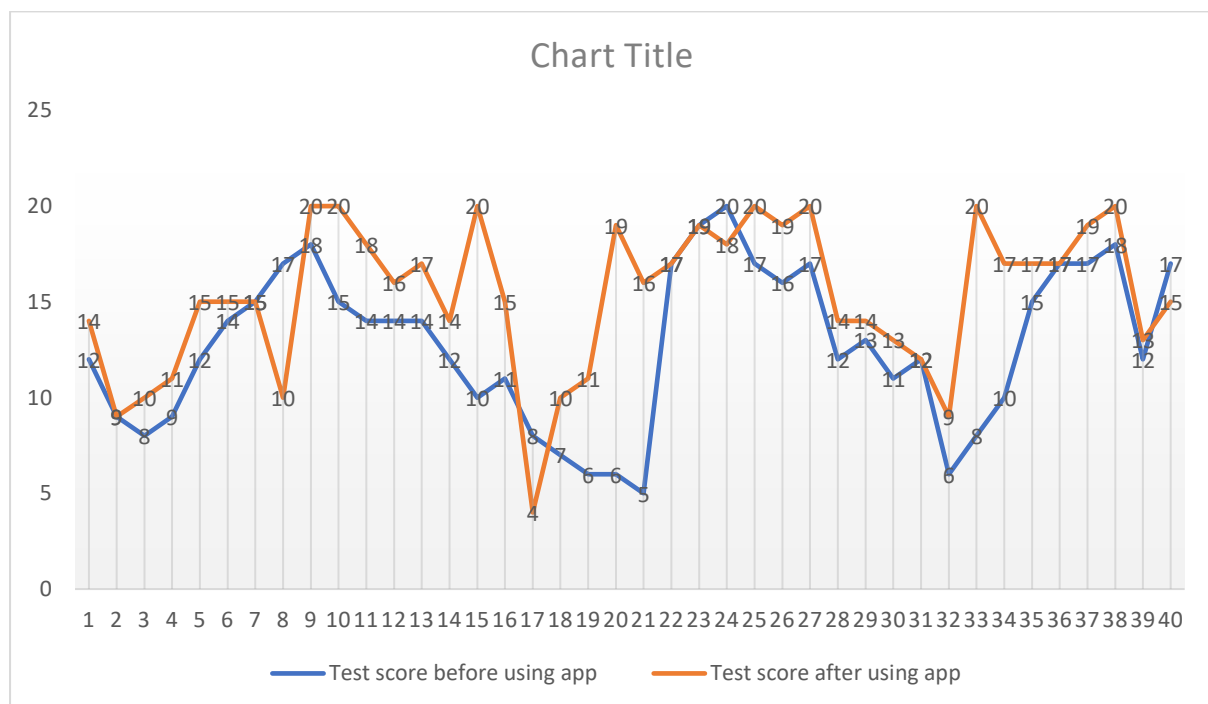


Figure 02: Shape of Data by graph

What is a Hypothesis?

Hypothesis testing is an act in statistics whereby an analyst tests an assumption regarding a population parameter. The methodology employed by the analyst depends on the nature of the data used and the reason for the analysis. Hypothesis testing is used to assess the plausibility of a hypothesis by using sample data.

Null hypothesis and alternative hypothesis.

Testable hypotheses for the experiment.

Different hypotheses can be fully used to analyze the survey data. The most common hypotheses are the null hypothesis and the alternative hypothesis.

Null hypothesis (H_0)

The null hypothesis is a statistical hypothesis in which there is no significant difference between the many variables. This is the original or standard statement, without effect, often represented by H_0 (H-zero). The hypothesis is always tested. Means the defined value of the population parameter as μ , s , p . The null hypothesis can be rejected, but it cannot be accepted based on one test alone. When it comes to hypothesis testing, the null hypothesis is often an initial statement that is supported by research or specialized knowledge.[4]

Alternative hypothesis (H_1)

A statistical hypothesis used in hypothesis testing that states that there is a significant difference between the many variables. It is often called a hypothesis other than the null hypothesis, often denoted by H_1 (H-one). This is what the researcher seeks to prove indirectly by using the test. It refers to a certain value of the sample statistics, for example $\bar{x}$, s , p . [5]

The acceptance of an alternative hypothesis depends on the rejection of the null hypothesis, ie. until and if the null hypothesis is rejected, an alternative hypothesis cannot be accepted.

Testing of hypotheses to various significance level

a) $P = 0.10$

The hypotheses are calculated for various significance level, as the significance level is 0.10.

- 1) If the t Stat is smaller than the critical two-tail, and here it is true as the -2.8465 is smaller than 1.6646, the value of null hypotheses cannot be rejected, and that data must support the null hypothesis.

t-Test: Two-Sample Assuming Unequal Variances

	<i>Test score before using app</i>	<i>Test score after using app</i>
Mean	12.75	15.3
Variance	16.80769231	15.29230769
Observations	40	40
Hypothesized Mean Difference	0	
df	78	
t Stat	-2.846542418	
P(T<=t) one-tail	0.002822714	
t Critical one-tail	1.292499597	
P(T<=t) two-tail	0.005645427	
t Critical two-tail	1.664624645	

Fig. 1 Test for $p = 0.10$

b) $P = 0.05$

- 1) The null hypotheses cannot be rejected if the value of t stat is less than the critical value with two tails. The value of t stat in this case is -2.8465, which is less than the value with two queues, which is 1.9908, so the null hypotheses cannot be rejected and are fulfilled.

t-Test: Two-Sample Assuming Unequal Variances

	<i>Test score before using app</i>	<i>Test score after using app</i>
Mean	12.75	15.3
Variance	16.80769231	15.29230769
Observations	40	40
Hypothesized Mean Difference	0	
df	78	

t Stat	-2.846542418
P(T<=t) one-tail	0.002822714
t Critical one-tail	1.664624645
P(T<=t) two-tail	0.005645427
t Critical two-tail	1.990847069

Fig. 2 Test for $p = 0.05$

2) The zero value cannot be rejected if the value of p with two tails is bigger than the value of significance. In this case, p with two tails is 0.00564, which is greater than α .

c) $P = 0.01$

If is found after the hypotheses that:

1) If the value of t stat is less than the value of t critical two-tail value, then the null hypotheses cannot be rejected. So, from the result, it can be observed that - **2.8465** is less than **2.6403** so the null hypothesis is not avoided.

2) Also, it is also stated that, if the value of the probability of two-tail is greater than the significance value, from the observations it is seen that the **0.0056** is greater than 0.01 then null hypotheses will not be rejected.

t-Test: Two-Sample Assuming Unequal Variances

	<i>Test score before using app</i>	<i>Test score after using app</i>
Mean	12.75	15.3
Variance	16.80769231	15.29230769
Observations	40	40
Hypothesized Mean Difference	0	
df	78	
t Stat	-2.846542418	
P(T<=t) one-tail	0.002822714	
t Critical one-tail	2.375110958	
P(T<=t) two-tail	0.005645427	
t Critical two-tail	2.640340015	

Fig. 3 Test for $p = 0.01$

Type I errors and Statistical significance

Statistical estimates are usually taken to reduce the risk of type I errors

-this can lead to many type II errors.

Statistical findings are often referred to as "statistically significant at 0.05"

-this means that there is a 5% probability of making a type I error - an error in data collection or analysis [3].

This error will occur 5 times out of every 100 cases. To increase the likelihood of making a type II error, you need to increase the level to one percent.

Null hypothesis testing

The null hypothesis for a paired t-test is $H_0: \mu_1 = \mu_2$.

The three alternate hypotheses become:

$$\mu_1 \geq \mu_2$$

$$\mu_1 \leq \mu_2$$

$$\mu_1 \neq \mu_2$$

We assume that the standard deviations of the two populations are the same. To confirm this assumption, you can use the F-test hypothesis test.

We calculate test statistics using:
$$t = \frac{\bar{x}_1 - \bar{x}_2}{sp \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where sp is the pooled standard deviation and calculated from the samples by

$$sp = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}$$

The degrees of freedom used to determine the critical value is $df = n_1 + n_2 - 2$.

The critical value (or rejection region) for the three tests given a $(1-\alpha)100\%$ confidence level becomes:

Reject H_0 if $t > t_{\alpha, df}$

Reject H_0 if $t < t_{\alpha, df}$

Reject H_0 if $|t| > t_{\alpha/2, df}$

We have data from two different views of 40 students in this section. One set of data contains the results of a math test taken before using the program, while the other set of data contains the results of a math exam taken after using the application.

The means are as follows: $\bar{X}_1 = 12.7$ and $\bar{X}_2 = 15.35$; $s_1 = 4.0585$ and $s_2 = 3.9193$; $s_1 = 4.0585$ and $s_2 = 3.9193$. $Df = 40 + 40 - 2 = 78$ degrees of freedom are obtained from the forty samples obtained from each procedure.

The critical value is

- 2) $t_1(.05, 78) = 1.990$, given an $\alpha = 0.05$ or a 95% confidence level
- 3) $t_2(.01, 78) = 2.639$, given an $\alpha = 0.01$ or a 99% confidence level
- 4) $t_3(.1, 78) = 1.664$, given an $\alpha = 0.1$ or a 90% confidence level

Calculation for t_1 :

$$Sp = \sqrt{\frac{(40-1)4.0585^2 + (40-1)3.9193^2}{40+40-2}}$$
$$= 3.989$$

$$t = \frac{15.35 - 12.7}{3.989 \sqrt{\frac{1}{40} + \frac{1}{40}}}$$
$$= 2.970$$

In this situation, because the value of 2,970 exceeds each part of the test, the null hypothesis is rejected, as stated below: This implies that there is compelling evidence that the two techniques do not have the same influence on the product. This demonstrates that the application's commitment to the educational process has the potential to improve the school system and produce better results.

Deviations in the data

As these data are exam scores, there is no possibility of extraordinary values of these data. There may be deviations in the study itself during its construction. Because the outcome is dependent on the evaluation of the tests, they must be specified or judge properly to avoid data discrepancies. [5] If there are any discrepancies in the data, they may be quickly removed in the following ways:

The shortest method: Because the list contains values from many areas, sorting can assist you effortlessly remove all superfluous values from the list. Truncation can assist you set data from the lowest to the highest values. This can aid in the verification of the data.

Conclusion

The company is developing software for a math training application. They test their application in different ways at school. They conducted a survey among 40 5-year-old students to determine the exact result. For the survey, they applied the same test to 40 students without the application and the same test after providing the application for use. The table shows how the application has a positive effect on knowledge and test results. However, careful study and hypotheses can lead to an accurate assumption and risk-free probability. The data provided to us were reviewed in the report and the results were presented in varying degrees of significance. I believe that the application has contributed and will contribute if it is introduced into the training system for permanent use.

References:

- [1] Democracy and education (Dewey, 2001) Dewey, J. (2001), The Pennsylvania State University
- [2] Toward collaborative open data science in metabolomics using Jupyter Notebooks and cloud computing. Mendez, K.M., Pritchard, L., Reinke, S.N. and Broadhurst, D.I., 2019.
- [3] (Elmasri & Navathe, 2011) Elmasri, R., Navathe, S. B. (2011) Fundamentals of Database Systems - Sixth Edition, Addison Wesley, ISBN-13: 978-0-136-08620-
- [4] Determinants of cloud computing adoption by Romanian SMEs in the digital economy. Dincă, V.M., Dima, A.M. and Rozsa, Z., 2019. *Journal of Business Economics and Management*
- [5] The science behind why our brains craves infographics (In an Infographic). Agency Post. Retrieved from <http://blog.hubspot.com/agency/science-brains-crave-infographics>. Oetting, J. (2015, February).