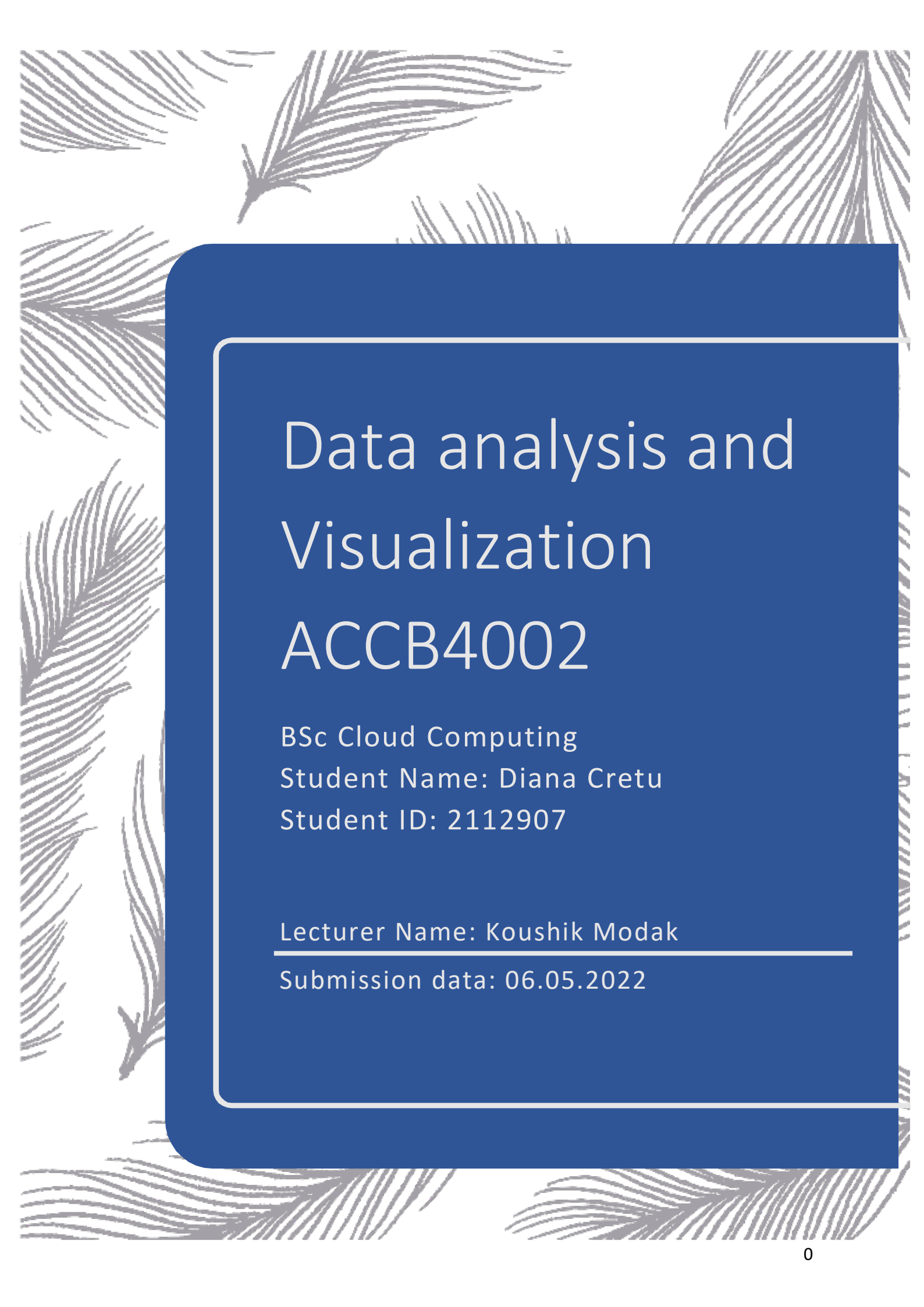




Data analysis and Visualization ACCB4002

BSc Cloud Computing

Student Name: Diana Cretu

Student ID: 2112907

Lecturer Name: Koushik Modak

Submission data: 06.05.2022

Table of Contents

INTRODUCTION.....	2
DESCRIPTION OF THE DATASET INCLUDING DESCRIPTIVE STATISTICS	2
VISUALIZATIONS OF DATASET	4
PROPOSE AN APPROPRIATE HYPOTHESIS FOR STATISTICAL ANALYSIS.....	5
HYPOTHESIS STATEMENT:.....	5
THE NULL HYPOTHESIS:	5
THE ALTERNATIVE HYPOTHESIS:	6
FACTORS AFFECTING HYPOTHESES TESTING.....	11
RESULTS.....	11
CONCLUSION	11
BIBLIOGRAPHY.....	12
APPENDIX	13
ABOUT EQUAL VARIANCE	13
ABOUT REMOVING THE OUTLIERS.....	13
ABOUT PROBABILITY:	15

Table of figures

Figure 1 Stages for the hypothesis	2
Figure 2 Data set given (including Sparkline)	3
Figure 3 Descriptive statistic.....	4
Figure 4 The Dataset visualisation bar chart	4
Figure 5 t-Test when $\alpha = 0.1$	7
Figure 6 t-Test when $\alpha = 0.05$	8
Figure 7 t-Test when $\alpha = 0.01$	9
Figure 8 Graphical interpretation	10
Figure 9 Two-paired T-test assuming equal variance	13
Figure 10 Impact chart.....	14
Figure 11 T-distribution sample.....	15

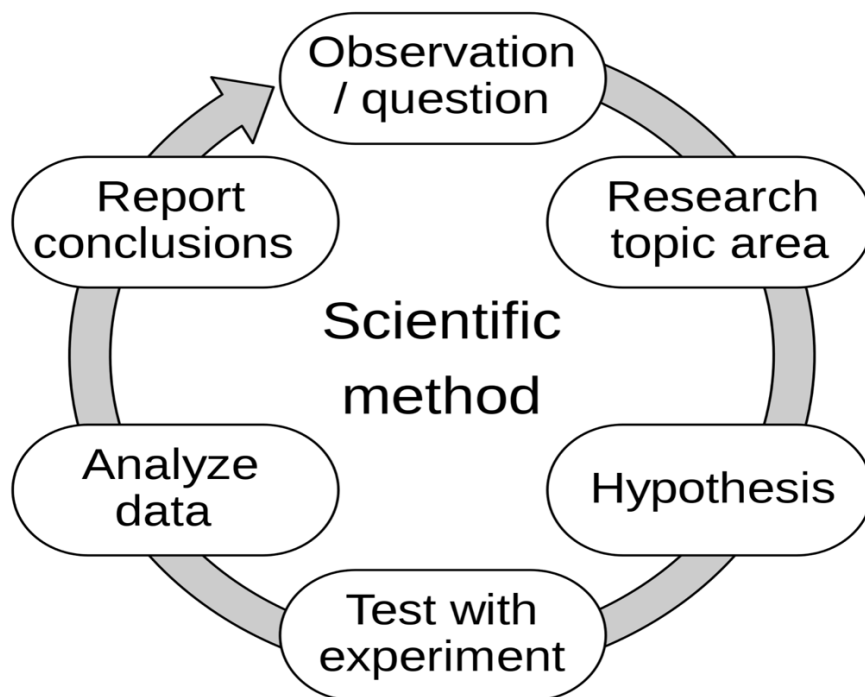
Introduction

In this report, I will discuss hypotheses and how we can make decisions based on descriptive statistics.

Based on the scenario, I need to discuss and analyze the given data set, develop the required testable hypothesis, and select and justify my hypothesis including data set visualizations. Based on the definition that “The tentative answer to a scientific question is called a hypothesis, I will suggest some hypotheses. The hypothesis which can be demonstrated or invalidated, because of testing, and data collection means that is a testable hypothesis.” [1]

Also, I will provide some recommendations for a company that wants to know if their developed application can be used instead of traditional study materials.

Figure 1 Stages for the hypothesis



Description of the dataset including descriptive statistics

The collected data comes from a local primary school and displays the test results of children who are studying mathematical concepts. The school has been given a learning tool from an application development company. The data set found is about class test results before and after using the app. The data set is describing the difference between the test scores achieved. From this table, we can state that the columns (Before and After) are not similar. That gives us an opportunity to develop a hypothesis, make various tests on the hypothesis and, in conclusion, make a decision based on this.

I will try to establish a cause-effect relationship by testing some hypotheses for a qualitative understanding, also, the sparkline is helping to visualize much better the impact.

Figure 2 Data set given (including Sparkline)

	Student	Before	After	
	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	
n	40	40	40	
mean		12.75	15.90	
standard deviation		4.10	3.91	

To summarise a given data set, I did a descriptive statistic. “The Descriptive statistic allows us to proceed through min., max., sum., and measures of variability (or central tendency) such as the median, mean and mode, which are commonly used at all levels of statistic and math.” [2]

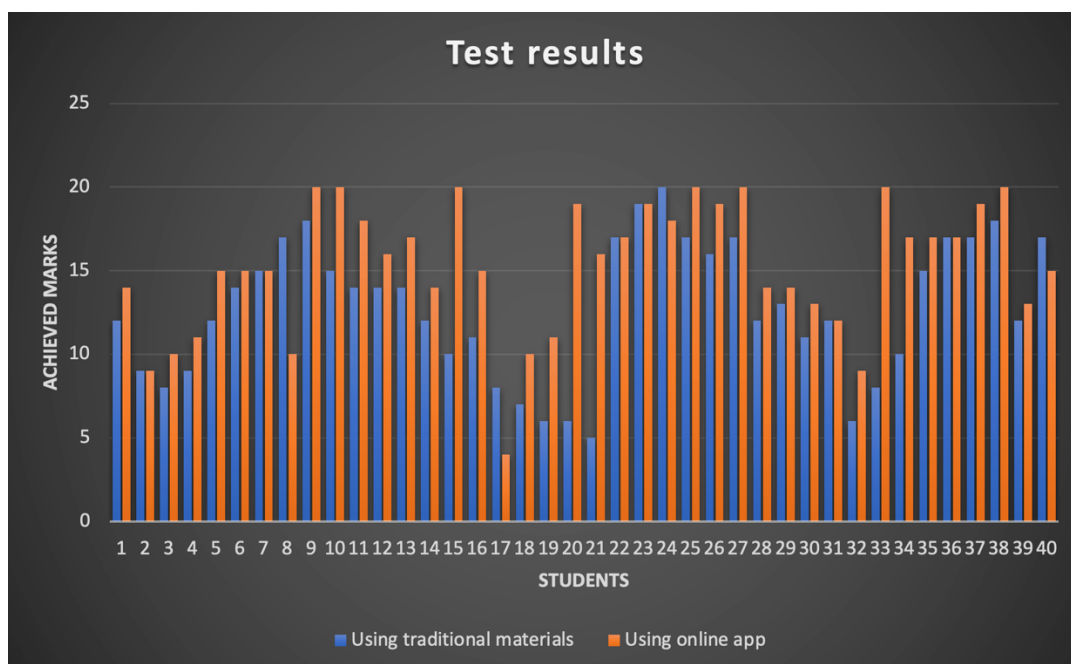
Figure 3 Descriptive statistic

<i>Before</i>		<i>After</i>	
Mean	12.75	Mean	15.3
Standard Error	0.648222421	Standard Error	0.61831035
Median	12.5	Median	15.5
Mode	17	Mode	20
Standard Deviation	4.099718564	Standard Deviat	3.91053803
Sample Variance	16.80769231	Sample Variance	15.2923077
Kurtosis	-0.992832012	Kurtosis	0.20853923
Skewness	-0.208284596	Skewness	-0.73290128
Range	15	Range	16
Minimum	5	Minimum	4
Maximum	20	Maximum	20
Sum	510	Sum	612
Count	40	Count	40

Visualizations of dataset

In this part, from the collected data I did a bar chart to visualize the dataset to make some discoveries from it.

Figure 4 The Dataset visualisation bar chart



Analysis

There are many different statistical tests to choose for example, the test for a proportion, the difference between two proportions, the Chi-sq test for independence, regression analysis and the difference between two means (independent samples) etc. To choose the most appropriate test I conclude that we have independent samples.

To select the suitable test to be used, I was analyzing the given data for what level of measurement it is or what type of data is, how many samples I have and what is the purpose of our investigation.

The given data set can be interpreted quantitatively, based on DePaul University Center for teaching and learning. "Possible analysis may include measures of relationship, distributions spread, mean comparisons and central tendency." [3] the most common summary value for interval/ratio data is a mean. In the section below, I will perform the Tests that involve interval/ratio data to present the given data set.

Next, we ask how many samples are involved. So, we have two samples which are being compared with each other: we are comparing test scores before using the app and test scores after using the app. In this way, the test score before using the app is independent data, and the test score after using the application is dependent data.

The purpose of the analysis is to make a decision based on the result of testing, to continue with the development and implementation of using the application for study.

Propose an appropriate hypothesis for statistical analysis

For the first, I will formulate a hypothesis. "A good hypothesis should look like statements expressing correlation of two or more quantifiable variables". [4] My hypothesis will be based on the existing knowledge from the data set. After examining the data table, I will come up with the following statement, trying to perform a specific and researchable question.

Hypothesis statement:

- Does using the app leads to **better** exam results?
- **If** all children will often use the online portal to learn, **then** their exam scores will **increase**?

To do the Hypothesis testing, I will state:

The Null Hypothesis:

The Null Hypothesis is the hypothesis that we believe or expect to be true. Based on current knowledge, I can say that the null hypothesis is that using the app is the same as traditional materials, which means the test score of children is equal when using the app or traditional approach. "The concept of the null hypothesis has a crucial function in a testing of hypothesis" [5]

- H_0 : Using a learning tool does not affect their final exam scores. ($H_0 = \mu_1 - \mu_2 = 0$, no impact)

The Alternative Hypothesis:

"The alternative hypothesis presents a new idea which could overturn or reject the null hypothesis in favour of accepting the alternative hypothesis." [6] In other words, the alternative hypothesis consists of the correlation between test scores before using the app and after using the app.

- H_a : Using a learning tool has **a positive effect** on their final exam scores. ($H_a > H_0$)

T-test: 2-sample assuming unequal variances:

"The T-test is the statistical analysis that shows us the significant difference between group means or two group means. Also, it allows understanding if those differences in means could have happened by chance and decide if I should support or reject the null hypothesis." [7] Below is the formula presented for the t-test (Student's test) and T stat/obt:

$$t = \frac{\sum(x_1 - x_2)}{\frac{s}{\sqrt{n}}}$$

Where: $X_1 - X_2$: difference means of the pairs, S: standard deviation, n: sample size.

$$t_{obt} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{SD_1^2 + SD_2^2}{n}}}$$

Where: $X_1 - X_2$: difference means of the pairs, SD: the estimates of the variances of the two populations, n: sample size.

Assuming that, the data set has a normal distribution, the same number of elements ($n=40$), and unsimilar variance. I will propose to execute the following tests:

- Paired samples t-test (dependent), to investigate the change in the mean before and after the experiment for the same children (using a new learning tool).

I did a calculation of the variance of the given 2 samples, and they are not the same ($x=16.81$ and $y=15.29$). This result supports my choice to do the Two-sample T-test assuming Unequal variances.

Paired sample at alpha Level 0.1

Figure 5 t-Test when $\alpha = 0.1$

t-Test: Two-Sample Assuming Unequal Variances		
Note: alpha level=0.1		
	<i>Before app</i>	<i>After app</i>
Mean	12.75	15.3
Variance	16.80769231	15.2923077
Observations	40	40
Hypothesized Mea	0	
df	78	
t Stat (obt)	-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	

p-value

0.00564542

Step 1:

I will compare the alpha level to the P-value:

Result: $0.1 > 0.00564542$ ($\alpha > P - value$)

Decision:

I can reject H_0 at the significance level of 0.1 because my p-value does not surpass 0.1.

Step 2:

I will compare the T critical value (one tail and two-tail) to the T-stat value:

Result one-tailed: $1.29 < |-2.85|$ ($T_{crit} < T_{obt}$) we take the absolute value of the T value in this case because I want to know if there is a difference, not even the direction.

Result two-tailed: $1.66 < |-2.85|$ ($T_{crit} < T_{obt}$)

Decision: I can reject H_0 at the significance level of 0.1 because T_{crit} is less than T_{stat} for both cases with one-tail and two-tail distributions.

With 90% significance, I believe that using the application for learning mathematical concepts improves the test score of children.

Paired sample at Level significance 0.05

Figure 6 t-Test when $\alpha = 0.05$

t-Test: Two-Sample Assuming Unequal Variances		
Note: alpha level=0.05		
	<i>Before app</i>	<i>After app</i>
Mean	12.75	15.3
Variance	16.80769231	15.29230769
Observations	40	40
Hypothesized Mean	0	
df	78	
t Stat (Obt)	-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	

p-value

0.00564542

Step 1:

I will compare the alpha level to the P-value:

Result: $0.05 > 0.00564542$ ($\alpha > P - value$)

Decision:

I can reject the null hypothesis (H_0) at the significance level of 0.05 because my p-value does not surpass 0.05.

Step 2:

I will compare the T critical value (one tail and two-tail) to the T-stat value:

Result one-tailed: $1.66 < |-2.85|$ ($T_{crit} < T_{obt}$) we take the absolute value of the T value in this case because I want to know if there is a difference, not even the direction.

Result two-tailed: $1.99 < |-2.85|$ ($T_{crit} < T_{obt}$)

Decision: I can reject H_0 at the alpha level of 0.1 because T_{crit} is less than T_{stat} for both cases with one-tail and two-tail distributions.

With 95% probability, I believe that using the application for learning mathematical concepts improves the test score of children.

Paired sample at Level significance 0.01

Figure 7 t-Test when $\alpha = 0.01$

t-Test: Two-Sample Assuming Unequal Variances		
Note: alpha level=0.01		
	<i>Before app</i>	<i>After app</i>
Mean	12.75	15.3
Variance	16.80769231	15.29230769
Observations	40	40
Hypothesized Mean	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	

p-value

0.00564542

Step 1:

I will compare the alpha level to the P-value:

Result: $0.01 > 0.00564542 (\alpha > P - value)$

Decision:

I can reject H_0 at the significance level of 0.05 as my p-value does not surpass 0.05.

Step 2:

I will compare the T critical value (one tail and two-tail) to the T-stat value:

Result one-tailed: $2.38 < |-2.85| (T_{crit} < T_{obt})$ we take the absolute value of the T value in this case because I want to know if there is a difference, not even the direction.

Result two-tailed: $2.64 < |-2.85| (T_{crit} < T_{obt})$

Decision: I can reject H_0 at the significance level of 0.1 because T_{crit} is less than T_{stat} for both cases with one-tail and two-tail distributions.

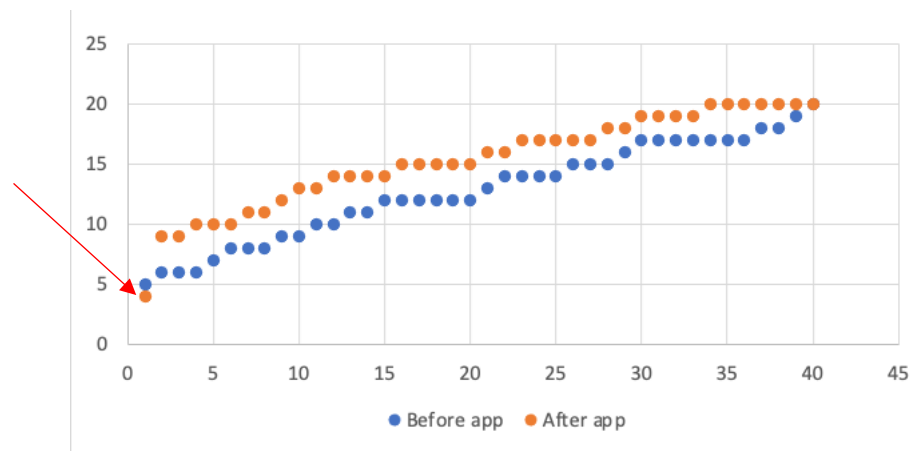
With 99% significance, I believe that using the application for learning mathematical concepts improves the test score of children.

Outliers in dataset

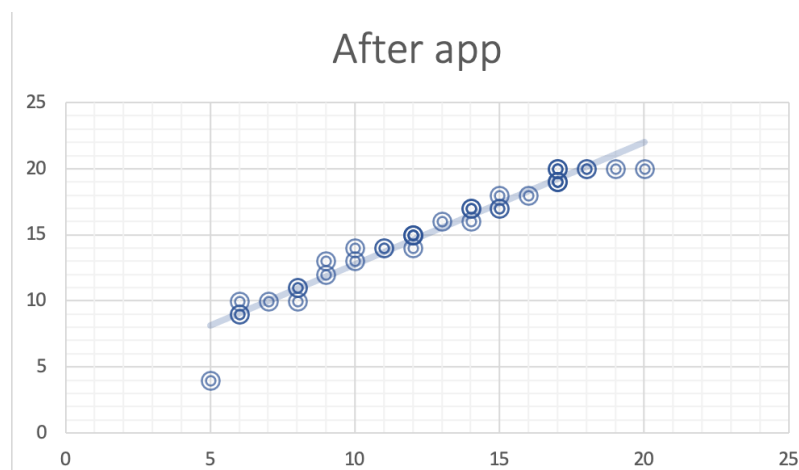
“Outliers are exceptions from the data set that differs from the most other data points” [8]. Outliers have a big impact on the result of any hypothesis test, especially when they don't represent a true value or represent a mistake. This extreme can come from incorrect data entry or some measurement error. Therefore, it can lead to an incorrect result of the tested hypothesis.

To identify the outliers of the dataset, need to perform some visualizations and sorting methods. By applying an ascending filter to the dataset value, I have got one outlier, seen in Figure 4.

Figure 8 Graphical interpretation



Analyzing: The outlier comes from the “After app” column, to ensure if I will remove this or not, I did a regression plot before deciding. And continue in the Appendix.



Factors affecting Hypotheses testing

Usually, factors affecting hypothesis testing are the factors which we cannot control. To test the initial hypothesis and to get more realistic results from the experiment it's not enough to test our hypothesis at the different significant levels and using different hypothesis tests, but also, we can repeat the experiment and then compare the result of the children's scores. Keep repeating the experiment may reject the initial hypothesis.

Theoretically, based on statistics teacher "there are four primary factors affecting power, significance level, sample size, variability or variance, and degree of the effect of the variable" [9]

Results

An independent-sample two-paired-test was used to check the effectiveness of using applications for educational purposes. Also at a different significance level, the experiment has significant differences found.

The test score before using the application ($M=12.75$, $SD=4.10$) compared to the same number of students after using the app ($M=15.30$, $SD=3.91$) demonstrated better test results, $t(78)=2.85$, $p=.005$.

The p-value (0.0056) and $M(78) = 2.85$ in investigation has considered statistically significant. The p-value which is less than 0.05 and the Mean value of the degree of freedom 78 ($df = 2.85$) are considered reliable. As a result, the app improves the score on the test.

Conclusion

Based on Jim Frost "It's critical that we understand the scientific method and knowledge to design a scientific investigation which includes statistical analysis. Every study has an outcome which is represented by the dependent variables." [10]

In conclusion, this report shows up how changes in one or more independent variables affect the dependent variable.

The output of the tested hypothesis is proving whether the application have positive impact, and the company must continue to implement its application in process of studying for children.

Bibliography

- [1] P. By Anne Marie Helmenstine, "ThoughtCo.," Dotdash Meredith, January 2019. [Online]. Available: <https://www.thoughtco.com/testable-hypothesis-explanation-and-examples-609100>. [Accessed 1.05.2022].
- [2] A. HAYES, "Investopedia," 04 03 2022. [Online]. Available: https://www.investopedia.com/terms/d/descriptive_statistics.asp. [Accessed 01.05.2022].
- [3] D. U. C. f. T. &. Learning, "Analysis & Interpretation," in *Assessing Learning*, Chicago, APA Style Handbook, 2001, pp. 20-21.
- [4] Sendil Mourougan¹, Dr. K. Sethuraman², Hypothesis Development and Testing, India: IOSR Journal of Business and Management (IOSR-JBM), 2017.
- [5] D. C. Howell, Statistical Methods for Psychology, Belmont: Cengage Wadsworth, 2010.
- [6] G. R. Loftus, The Null Hypothesis, Seattle: Department of Psychology, 2009.
- [7] C. H. Goulden, Methods of Statistical Analysis, 2nd ed., New York: Wiley, 1959, pp. 50-55.
- [8] P. Bhandari, "How to Find Outliers," Scribbr , 30 11 2021. [Online]. Available: <https://www.scribbr.com/statistics/outliers/>. [Accessed 04.05.2022].
- [9] Angela L.E. Walmsley and Michael C. Brown, "Statistic teacher," ASA Community Teacher Group, 15 09 2017. [Online]. Available: <https://www.statisticteacher.org/2017/09/15/what-is-power/>. [Accessed 04.05.2022].
- [10] J. Frost, "5 Steps for Conducting Scientific Studies with Statistical Analyses," 2022. [Online]. Available: <https://statisticsbyjim.com/basics/scientific-studies/>. [Accessed 01.05.2022].
- [11] G. Francis, Hypothesis Testing Reconsidered, Cambridge: Cambridge University Press, 2019.

Appendix

About equal variance

Also, I propose to execute a t-test for an equal variance to see what will happen if our variance was equal or similar:

Figure 9 Two-paired T-test assuming equal variance

t-Test: Two-Sample Assuming Equal Variances		
	<i>Before app</i>	<i>After app</i>
Mean	12.75	15.3
Variance	16.80769231	15.29230769
Observations	40	40
Pooled Variance	16.05	
Hypothesized Mean	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	

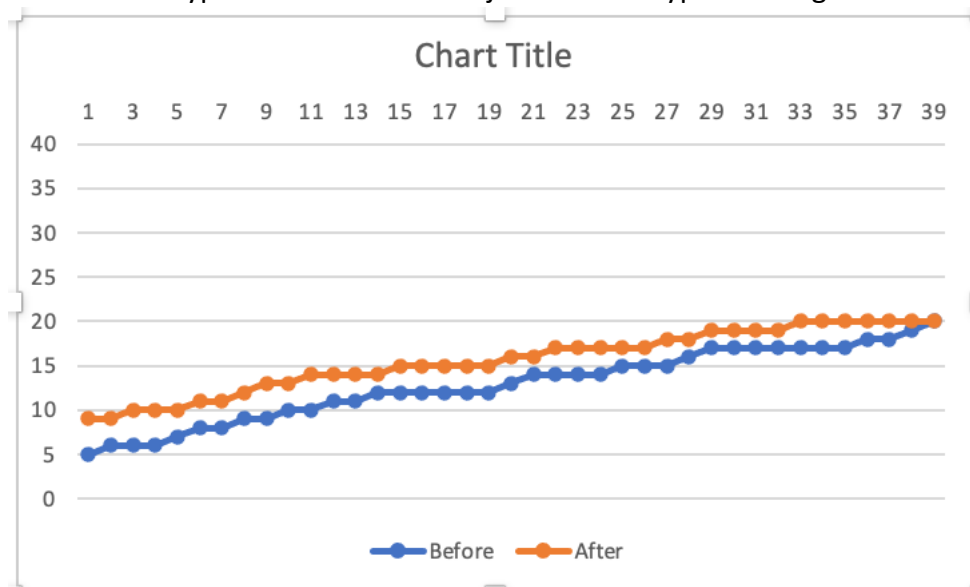
As I see, based on this output the result of the tested hypothesis is the same.

About removing the outliers

Let's see if we will remove the corresponding outliers from both samples. To remove the outliers: score 4 from after using the app column, I will delete the corresponding to this outlier score 8 from before using the app column. Then in the figure below I will perform a new T-test without outliers:

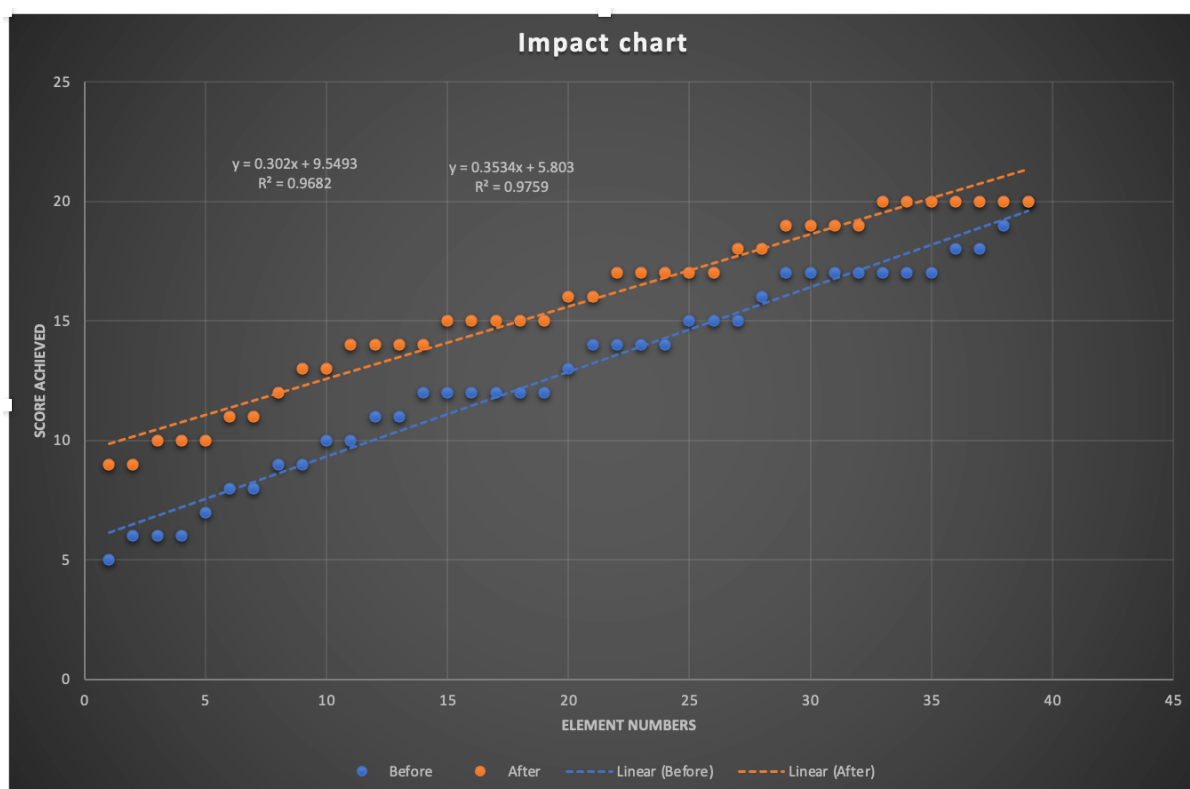
t-Test: Two-Sample Assuming Unequal Variances		
Note: without outliers		
	<i>Before</i>	<i>After</i>
Mean	12.87179487	15.58974359
Variance	16.64102564	12.24831309
Observations	39	39
Hypothesized Mean	0	
df	74	
t Stat	-3.157946449	
P(T<=t) one-tail	0.001150046	
t Critical one-tail	1.665706893	
P(T<=t) two-tail	0.002300092	
t Critical two-tail	1.992543495	

In the result, at level alpha 0.05 and removing the extreme of the data set I just confirmed the hypothesis above and reject the null hypothesis again.



In my opinion, using the have a positive impact on test scores of children, as is seen in the figure below:

Figure 10 Impact chart



About probability:

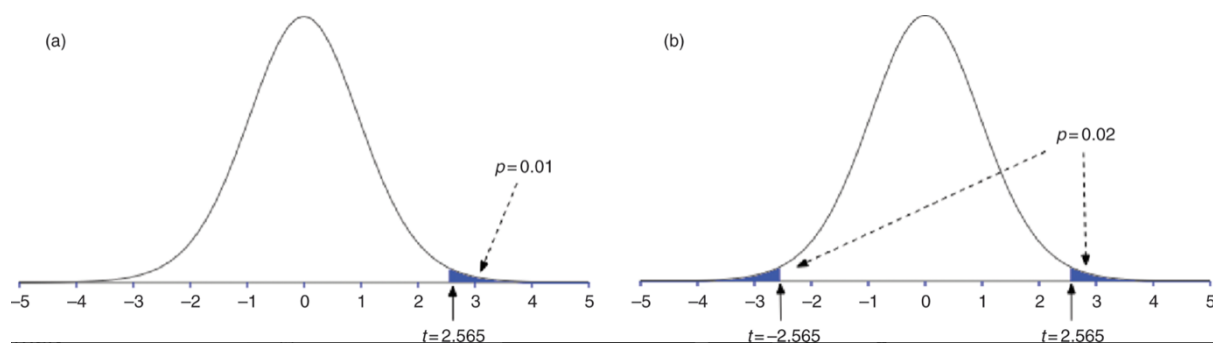
The p-value represents the probability or chance to prove the hypothesis. For example, if the p-value = 0.1, then there is a 10% chance of happening. In most cases we consider reliable or statistically significant is a p-value of 0.05 and below.

Let's assume that the variance between these two groups is approximately equal, then I will perform a new t-test to confirm the result of the tested hypothesis.

- Independent (unpaired) samples is used to compare the means of two groups. This case describes the situation when the test score before and after using the app it's performed by the different students.

Here is the sampling distribution for the one-tailed (unpaired) test and two-tailed test including the P-value (shaded area)

Figure 11 T-distribution sample



[11]