

Assignment 1: Data Analysis

Mircea-Radu-Şerban Dobrău

M7X01702

Dr. Mabrouka Abuhmida

Contents

1	Introduction	4
2	Hypothesis	4
3	Student's t test	4
4	Other measures	7
5	Data outliers and its effects	8
6	Other factors	11
7	Conclusion	12
	References	13

List of Figures

1	One-tailed t test for $p = 0.10$, $t_{crit} = 1.330$	6
2	One-tailed t test for $p = 0.01$, $t_{crit} = 2.552$	6
3	One-tailed t test for $p = 0.02$, $t_{crit} = 2.214$	6
4	Bar chart (young students)	8
5	Bar chart (mature students)	8

Abstract

The use of statistical studies and concepts can be often found in a large number of fields, such as science [1] and medicine [2], as well as data mining [3] and big data analysis [4]. As a result, an adequate knowledge of statistics and the proper application and interpretation of statistical concepts is essential towards working with and understanding data. This report presents an attempt at demonstrating an adequate knowledge of these concepts. It begins with a brief presentation of the data-sets provided by the study. Next, the processes suggested by the study in dealing with the provided data are undertaken and its results analysed and discussed. Finally, the report concludes with a summary of the report's findings.

1 Introduction

The data-sets provided by the study are presented below. The values ranging from 0 (low) to 60 (high) represent the level of university experience satisfaction:

Mature students	Young students
45	34
38	22
52	15
48	27
25	37
39	41
51	24
46	19
55	26
46	36

The following chapters detail the processes necessary towards discovering differences between the two groups, followed by the analysis and discussion of other features and factors of the data-set.

2 Hypothesis

The chosen hypothesis to be proven as being true is that mature students are significantly more satisfied with the university experience in comparison to young students. More formally, the hypothesis states: $\mu_{Mature} > \mu_{Young}$.

A method of proving this hypothesis as true is by stating the opposite case, also known as the null hypothesis H_0 , and subsequently proving it to be false. The null hypothesis states: $\mu_{Mature} \leq \mu_{Young}$.

To prove the null hypothesis as false and, as a result, the original hypothesis as true, the Student's t test is used.

3 Student's t test

The t test process [5] proceeds as follows:

1. Calculate t_{obt} , the obtained t value. As the sample size does not differ between the data-sets, the "equal sample size" formula is used:

$$t_{obt} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{SD_1^2 + SD_2^2}{N}}} = \frac{44.5 - 28.1}{\sqrt{\frac{75.38 + 72.98}{10}}} = \frac{16.4}{\sqrt{\frac{148.37}{10}}} = \frac{16.4}{\sqrt{14.83}} = \frac{16.4}{3.85} = \mathbf{4.25}$$

Where: $\bar{X}_1$ – is the mean of the mature student data-set

$\bar{X}_2$ – is the mean of the young student data-set

SD_1^2 – is the variance of the mature student data-set

SD_2^2 – is the variance of the young student data-set

N – is the number of samples of both data-sets, in this case **10**

2. Calculate df , the number of degrees of freedom:

$$df = N_1 + N_2 - 2 = 10 + 10 - 2 = \mathbf{18}$$

Where: N_1 – is the number of samples of the mature student data-set

N_2 – is the number of samples of the young student data-set

3. Find t_{crit} , the critical t value:

The critical t value t_{crit} is a value which represents the area into which a significant value must fall.

For this, we refer to a t_{crit} table, which binds together pairs of specific df (degrees of freedom) and p (probability of type I error) values to a t_{crit} value. In this particular case, we look for the t_{crit} values which correspond to $df = 18$ and $p = 0.10$, $p = 0.01$ and $p = 0.02$:

t_{crit}	p
1.330	0.10
2.552	0.01
2.214	0.02

4. Decide whether to use a directional or non-directional test:

As the mean value of the mature student sample is higher than that of the young student sample, what is necessary to prove, as the hypothesis states, is whether or not the difference is significant. Thus, the directional (one-tailed) t test is chosen as the test method.

The t_{obt} value falls within the t_{crit} range in all three scenarios, meaning that the H_0 hypothesis can be rejected at p levels 0.10, 0.01 and 0.02. Figures of the one-directional t tests are presented on the following page with the scope of illustrating this fact.

As a result, it can be confidently asserted that the mature student data-set's mean value is significantly higher than the young student data-set's mean value (or that $\mu_{Mature} > \mu_{Young}$) at p values 0.10, 0.01 and 0.02, meaning that mature students are significantly more satisfied with the university experience than young students, with the lowest probability of type I error (wrongly rejecting the contrary, H_0) being 1%.

Figure 1: One-tailed t test for $p = 0.10$, $t_{crit} = 1.330$

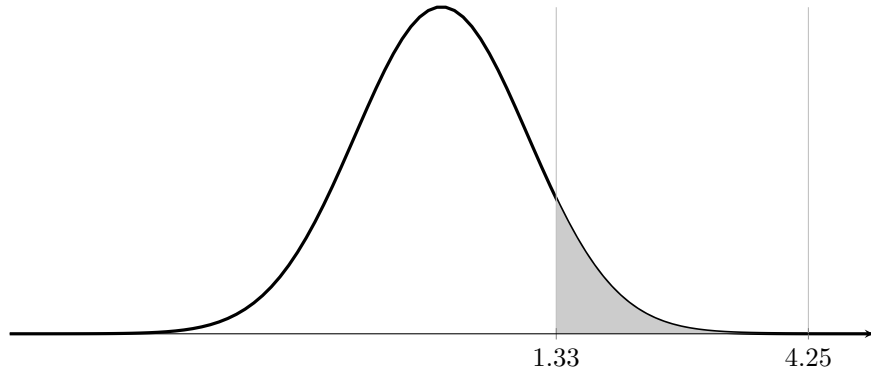


Figure 2: One-tailed t test for $p = 0.01$, $t_{crit} = 2.552$

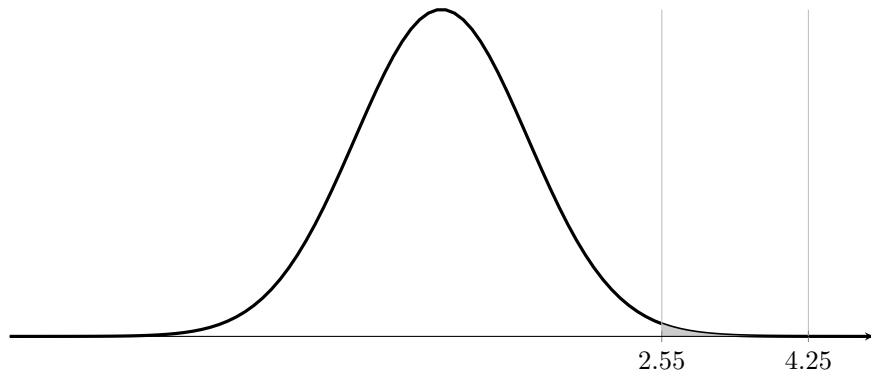
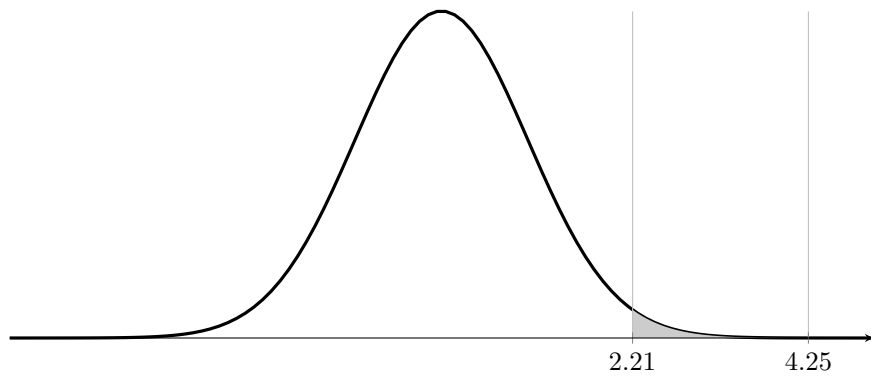


Figure 3: One-tailed t test for $p = 0.02$, $t_{crit} = 2.214$



4 Other measures

For analysis and comparison purposes, a table comprising of various other measures is presented below:

Measures	Young students	Mature students
Variance	72.98	75.38
Standard Deviation	8.54	8.68
Range	26	30
Median	26.5	46
Mode	none	46 (twice)
Z scores below mean	6	3
Z scores above mean	4	7

The following observations have been made:

Variance – The mature student data-set’s variance value (**75.38**) is only slightly larger than the young student data-set’s variance value (**72.98**). Even though mature students’ satisfaction levels vary between each other only slightly more in comparison to the young students, both young and mature students’ satisfaction levels present a roughly similar degree of variance.

Standard Deviation – The same conclusion can be drawn in regards to both data-sets’ standard deviation value, meaning that both young (**8.54**) and mature (**8.68**) students’ university experience satisfaction levels deviate to the roughly same degree from that particular group’s average university experience satisfaction level.

Range – The mature student data-set presents a slightly higher range value (**30**) than the young student data-set (**26**). From this it can be inferred that whilst mature students’ range of university experience satisfaction levels is slightly higher, both groups’ range of satisfaction levels is similar.

Median – The mature student data-set’s median value (**46**) is significantly larger in comparison to the young student data-set (**26.5**), meaning that, on average, mature students’ satisfaction level is higher than the young students’ satisfaction level.

Mode – The most frequently occurring value in the mature student data-set is **46** at a number of two occurrences. On the other hand, the young student data-set does not have a mode value, as each value is unique.

Z scores – The number of Z scores occurring below the particular data-set’s mean value is **3** for young students and **6** for mature students, whilst values occurring above the mean value is **7** and **4**, respectively. From this it can be deduced that, in contrast to young students (**3**), a higher number (**6**) of mature students’ satisfaction levels fall below the group’s average satisfaction level.

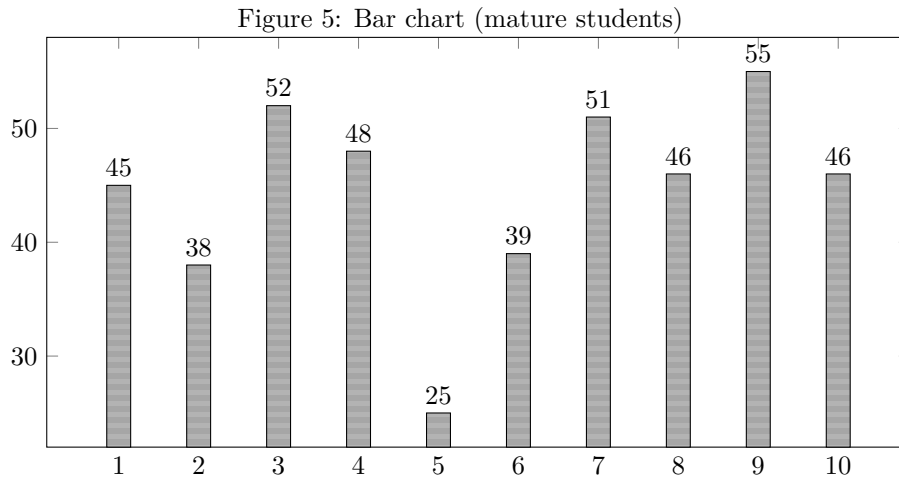
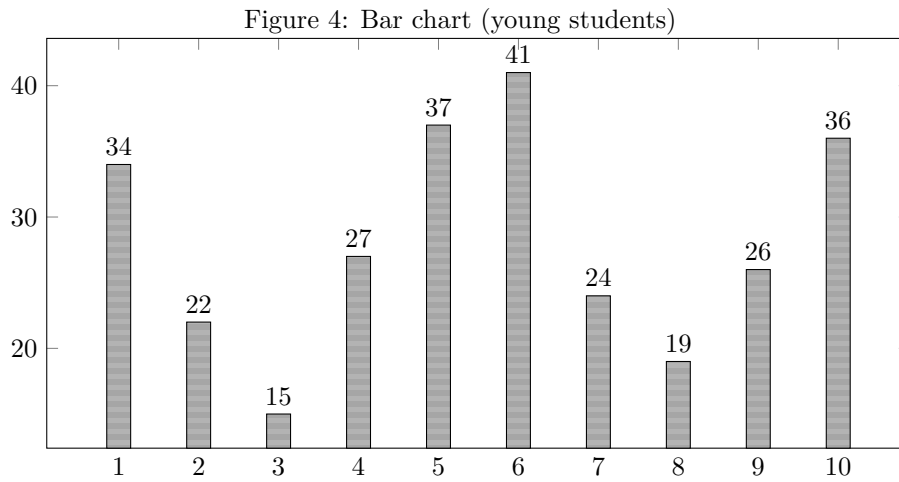
As such, there is not much difference in other measures between the two groups except for the number of Z scores below and above the mean, as well as the median value.

5 Data outliers and its effects

An outlier [5, p. 96-97] is identified as the value of a data-set which stands out to a significant degree from the other values of the data-set.

When it comes to analyzing data, proper identification and analysis of outliers is an important task, as outliers can lead to inaccurate results if they are not taken into consideration before doing further analysis of the data-set.

One way of identifying outliers is by deducing them with the help of a data graph, such as a bar chart:



Referring to the bar charts presented above, one could suppose the **15** (young students) and **25** (mature students) values to be outliers.

There is, however, no universally agreed upon process of properly identifying outliers. A common, but not standard, method of defining outliers is chosen for this particular scenario, which states that:

- Mild outliers are:
 - smaller than: the 25th percentile $- 1.5 \cdot IQR$
 - larger than: the 75th percentile $+ 1.5 \cdot IQR$
- Extreme outliers are:
 - smaller than: the 25th percentile $- 3.5 \cdot IQR$
 - larger than: the 75th percentile $+ 3.5 \cdot IQR$

Where IQR is the *interquartile range*, which is defined as “the range of numbers that contains the central 50% of the values of a variable”. [5, p. 547] It is calculated [5, p. 91] as the difference between the 75th and 25th percentile values.

The 25th and 75th percentile along with the interquartile range for each group is presented below, as well as per-group criteria for outliers according to the above definition:

Measures	Young students	Mature students
25 th percentile	22.5	40.5
75 th percentile	35.5	50.25
Interquartile range	5	5
Mild outliers	< 15 or > 43	< 33 or > 58

According to the criteria above, there is only one outlier: the value of **25** in the mature student data-set, as it is smaller than **33**. To find out whether the omission of the outlier value from the data-set has any substantial effect on the outcome of the t test, a comparison table of values before and after the omission is presented below:

Measures	Before	After
df	18	17
t_{obt}	4.25	12.19
t_{crit} at $p = 0.10$	1.330	1.333
t_{crit} at $p = 0.01$	2.552	2.567
t_{crit} at $p = 0.02$	2.214	2.224
Reject H_0 at $p = 0.10$?	Yes	Yes
Reject H_0 at $p = 0.01$?	Yes	Yes
Reject H_0 at $p = 0.02$?	Yes	Yes

A brief analysis of the data presented above leads to the following observations:

- The obtained t value t_{obt} of **12.19**, upon omission of the outlier, is nearly three times the original value of **4.25**.
- Due to the omission of the outlier value, the number of degrees of freedom df has been reduced from **18** to **17**. For this reason, different t_{crit} values have been extracted for the purpose of the t test. It can be seen however, that ultimately, this difference produces only a marginal increase of the t_{crit} values.

- As deduced from the previous points, the end-result of the t test is ultimately the same, with H_0 being rejected at all three p levels due to t_{obt} being larger than t_{crit} .

In conclusion, the omission of the outlier value from the mature student data-set does not ultimately alter the outcome of the t test.

Next, a comparison table of other measures is presented, including the young student data-set's values for a comparison with the new values:

Measures	Before	After	Young students
Mean	44.5	46.6	28.1
Variance	75.38	31.99	72.98
Standard Deviation	8.68	5.65	8.54
Range	30	17	26
Median	46	46	26.5
Mode	46 (twice)	46 (twice)	none
Z scores below mean	3	5	6
Z scores above mean	7	4	4

The following deductions can be made:

Mean – The data-set's mean value has slightly increased from **44.5** to **46.6**. This is due to the omission of the unusually low value of **25** from the data-set. As a result, the mature student data-set now sports a slightly larger value in comparison to the young student data-set's mean value (**28.1**).

Variance – The data-set's variance has been roughly halved from **75.38** to **31.99**. As a result, the young student data-set presents a doubly larger degree of variance (**72.98** v. **31.99**) in comparison to what was previously perceived as a roughly equal degree of variance (**72.98** v. **75.38**).

Standard Deviation – The data-set's standard deviation value of **5.65** is significantly lower in comparison to the previous value of **8.68**. In this case as well, the young student data-set's values now present a higher degree of deviation (**8.54**) from the data-set's mean value in comparison to the mature student data-set (**5.65**).

Range – The data-set's range has been nearly halved to **17** from the original value of **30**. Thus, the young student data-set now presents a substantially higher range of values (**26**) in comparison to the mature student data-set's range of values (**17**).

Median – The data-set's median value has not been affected by the omission of the outlier. After sorting the data-set in increasing order, the placement of the **46** value occurs in the middle (fifth) value of the data-set comprised of nine values, thus making it the median value.

Mode – The data-set's mode still remains to be the value of **46**, which occurs twice.

Z scores – The data-set's number of values occurring below and above the mean has changed from **3** and **7** to **5** and **4**, respectively. As a result, it can be concluded that there's now a more even balance of mature students'

university experience satisfaction occurring below and above the mean satisfaction value of **46.6** (**5** and **4**, respectively) in comparison to previous results (**3** and **7** with a mean value of **44.5**).

In conclusion, the omission of the outlier value does not result in any significant difference with regards to the outcome of the t test. However, the omission of the outlier value does produce significant differences in regards to other measures found in the mature student data-set, namely: the variance, standard deviation, range, and Z score values.

6 Other factors

Some factors which were not taken into consideration and are outside the control of the writer of this assignment are presented below:

Bias – It has not been taken into consideration whether any kind of bias (e.g. in measurement, sample selection, or the interview process [5, p. 16-17]) is or is not present.

For example, it can be argued that a young student could be likely to give a more negative response if he/she is interviewed by a person or persons of his age, or if the interviewer/s unknowingly encourage the student to answer in one specific way or another.

A proposed method of controlling these factors is by having more detailed information about the interview process. In the case of having found factors which might influence the study's results, the study should be redesigned so as to minimize the risks of these factors. Alternatively, the analysis of the provided data would take these biases into account.

Sample Size – It can be argued that the selected sample size may be too small for an accurate analysis of the survey and its results.

It could be possible, for example, for the survey to be lacking in responses from a significant number of young students who express a high satisfaction with the university. As a result, the sample might be inaccurately skewed on one side, thus preventing an accurate set of results.

To make sure that responses are either truly varied (some satisfied, some not satisfied) or do actually represent a rather narrow general consensus (most are satisfied or most are not satisfied), a re-taking of the survey would have to be done with a significantly larger sample size; for example, 80 students per group.

Personal Details – The study makes no mention of each individual's particular traits, e.g. nationality and/or whether he/she is studying full-time or part-time.

For example, if it is observed that part-time students have a lower general satisfaction of the university than full-time students, there would be a correlation between the type of course and student satisfaction levels. Thus, the university would know which area/s of the university experience need/s improvement.

A re-design of the survey with additional questions could help the university in identifying correlations which would help in identifying which areas of the university experience need improvement.

As such, it is suggested that taking into greater consideration a number of factors in regards to the engagement of the study, such as the ones mentioned above, would likely contribute towards broader and more accurate results.

7 Conclusion

The t test has succeeded in emphasizing the difference between the two groups. Ultimately, mature students have a significantly higher satisfaction with the university student experience in comparison to young students.

Furthermore, another significant difference between the two groups is the number of scores below and above the mean, where mature students express a high number of satisfaction levels above the mean value and young students a high number of satisfaction levels below the mean value.

Next, it has been proven that although removal of the outlier value from the mature student data-set does not ultimately affect the t test, it does significantly alter other measures such as the group's variance and standard deviation value.

Finally, it is suggested that bias, sample size and addition of other variable data to the study be taken into consideration as possible steps towards broader and more accurate results.

References

- [1] C. Aitken and D. Stoney, *The Use Of Statistics In Forensic Science*, ser. Ellis Horwood series in forensic science. CRC Press, 1991. [Online]. Available: <https://books.google.co.uk/books?id=wK9c4KttXj0C>
- [2] J. Bailar and D. Hoaglin, *Medical Uses of Statistics*. Wiley, 2012. [Online]. Available: <https://books.google.co.uk/books?id=K2MBZZhTLBkC>
- [3] S. Tufféry, *Data Mining and Statistics for Decision Making*, ser. Wiley Series in Computational Statistics. Wiley, 2011. [Online]. Available: <https://books.google.co.uk/books?id=5MTBlxZUKiIC>
- [4] K. Winson-Geideman, A. Krause, C. Lipscomb, and N. Evangelopoulos, *Real Estate Analysis in the Information Age: Techniques for Big Data and Statistical Modeling*. CRC Press, 2017. [Online]. Available: <https://books.google.co.uk/books?id=0b89DwAAQBAJ>
- [5] S. Boslaugh, *Statistics in a Nutshell*, ser. In a nutshell. O'Reilly Media, Incorporated, 2012. [Online]. Available: <https://books.google.co.uk/books?id=HZpoDjtKT0IC>