



DATA ANALYSIS AND STATISTICS

Lee Rodger – 1809707

Abstract

This essay outlines the importance of data analysis and applies these techniques to datasets representing student satisfaction within university. The analysis will tell us whether or not the age of a student has an effect on their level of satisfaction overall.

Contents

1 - Introduction	2
2 - Scenario and Null Hypothesis	2
3 - Outliers.....	3
4 - Averages.....	3
4.1 - Mean	3
4.2 - Median	4
5 - Standard Deviation	4
6 - T-test	4
7 - Results.....	5
8 - Conclusion.....	6
References	7
Appendix	8

1 - Introduction

Statistics and data analysis are techniques dated back thousands of years, used in order to understand what a set of data means and what understanding we can gain from it. Statistics have played an important role within society on vast data sets such as censuses [1].

The use of data analysis is important as we need to be able to interpret the data we collect, and then determine how reliable this data is by conducting a number of tests to assess its validity. For example, from a given data set of grades in a class, we can determine how well the students did overall, assessing both the results and the fairness of the sample given. Out of a class of 30 pupils, taking 5 results at random to find the average result score wouldn't be completely reliable as the 5 chosen could all have the highest/lowest results, given an average which is not in line with the rest of the students.

2 - Scenario and Null Hypothesis

The scenario in question looks at results from a student satisfaction survey, conducted on 10 young and 10 mature students. The goal being to see if there is a difference in levels of satisfaction between the two groups, through the use of statistical and data analysis.

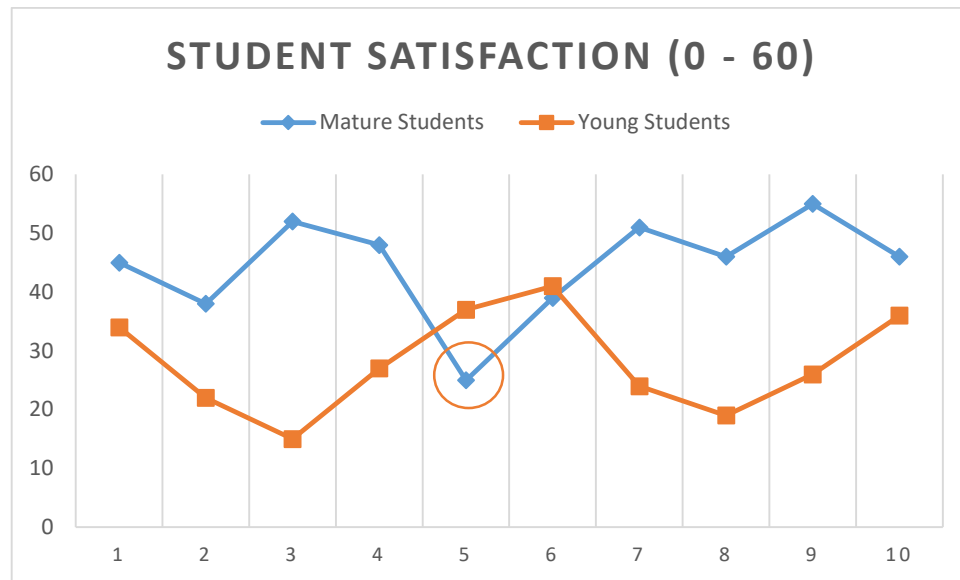
Mature Students	Young Students
45	34
38	22
52	15
48	27
25	37
39	41
51	24
46	19
55	26
46	36

[Figure 2.1] – Table of Student sets and satisfaction results

To begin with, I hypothesise that age has not effect on student satisfaction. This hypothesis will guide in using the correct data analysis techniques as this is the goal which I want to achieve. In this case, the alternative hypothesis would be that age does have an effect on student satisfaction within university.

3 - Outliers

Outliers are points of data will uncharacteristically fall out of the general range within a data set. These can cause discrepancies in results and therefore falsely prove/disprove the hypothesis.



[Figure 3.1] – Student results plotted on a line chart

From the generated line-chart of the student results, we can see that there is a potential outlier within the group of mature students. At first glance, the mature student results are almost all a considerable amount larger than that of the younger students. However, result number 5 for mature student dips significantly below the line of the young set. There are an endless number of possibilities to why this has occurred, so it would need to be determine whether or not this result should be omitted, to prevent an undesirable result occurring. But the validity of the result should still be considered beforehand.

For example, it is unlikely that this is the error of the person collecting the data as the result is perfectly valid and falls within the range of both sets of students combined. Therefore, it is likely that the result is genuine and should be used when analysing the data.

4 - Averages

In order to conduct more advanced statistical analysis, we must first begin by identifying the mean, median and mode of the data sets, which will give us a basic indication of the general results.

4.1 - Mean

The mean is calculated by adding all of the results together and dividing them by the amount of results in the set. In this study, to find the average result for mature students, we would add up all of the satisfaction scores and divide them by 10 (the number of results). We could then do the same for the young students and compare the results.

4.2 - Median

The median is a value which equally separates the upper and lower half of a set of data. We can use this to essentially determine a mid-point based on all of the data collected.

5 - Standard Deviation

The standard deviation is a measure of dispersion within a set of values. Calculating the standard deviation shows us the average difference between each value given and the mean value given in a set of data [2]. We can calculate the standard deviation using the following steps:

1. Calculate the mean for the set of values.
2. Subtract this result from each value in the set.
3. Eliminate potential problems by squaring any negative number and then finding the square root. This will essentially remove the minus sign.
4. Square each number and add them together.
5. Divide this by the size of the set – 1.
6. Calculate the square root.

The standard deviation of a set allows us to determine easily how spread out values are from one another. A large standard deviation value tells us the values are more than likely quite sparse, similarly a small value would tell us that all of the numbers in a set are close together. Having the standard deviation value also allows us to later conduct more advanced data analysis techniques which will give us more insight.

6 - T-test

The T-test was devised by William Gosset, an employee at Guinness in order to determine if there was a difference when comparing two sets of numbers, even if the number sets are quite small [2].

$$t_{obt} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{SD_1^2 + SD_2^2}{n}}} \quad (1)$$

[Figure 6.1] – T-Test formula for sets of equal size

The above equation (1) [3] shows the t-test formula for comparing two equal sets of values. We can use this in our study to determine whether or not the two data sets are equal or not, and therefore prove/disprove the null hypothesis. X 1 and 2 refer to the two sets of numbers. SD refers to the standard deviation of each and n is the number of values in the sets.

When conducting the T-test, we achieve a t Stat value and a t Critical two-tail value. If the t stat value is greater than the t critical two tail, the t test is significant in determining the hypothesis.

7 - Results

All of the above techniques were carried out on the dataset in question in order to disprove the null hypothesis. This gives insight into the data so we can apply a meaning behind it.

Firstly, we conduct the mean and median tests which will give us an immediately idea if we have met the conditions of the null or the alternative hypothesis.

The mean result for mature students is 44.5 and 28.1 for the young students. Therefore, by using the mean we can say the overall, the mature students have a higher level of satisfaction than the young students, supporting the alternative hypothesis to be true.

Also, in the case of this study, the mature students have a median of 46 and the young students have a median of 26.5. These are similar to the mean values but display a larger gap on the values, indicating a larger difference in student satisfaction between the groups, further supporting the alternative hypothesis.

The standard deviation results show us how far each set of numbers on average deviates from the mean of the set. The mature student set gave us a standard deviation of 8.682677518 and the younger students set gave us a standard deviation of 8.543353492. This indicates that both sets had an almost equal deviation from the mean, showing that both groups were well balanced. This supports the view that the sets were fairly made, likely randomly to not bias the results in any way.

It is important to take random samples when testing a hypothesis such as this otherwise the data can become misleading. For example, if the group of mature students were all achieving much higher grades than the younger group. This is likely to also then show a difference between the groups with regards to student satisfaction.

From looking at the data, this helps us to support the alternative hypothesis by showing that the data sets are fair and equal variance.

We can also conduct a t-test in order to determine whether or not our data is significant enough to be valid. We can test these against different Alpha levels. Tested at 0.10, 0.02 and 0.01, all of these values were deemed significant, as shown in the appendix (1, 2 and 3). Since the t stat is higher than the t critical two tail result, we can say confidently disprove the null hypothesis as the tests are significant. The chance of an outlying result which would drastically effect the outcome is very unlikely in this case.

8 - Conclusion

After analysing the data carefully, we can conclude by confidently rejecting the null hypothesis. All of the tests carried out support the alternative hypothesis by clearly indicating that the group of mature students are overall, as demonstrated in the results section above. However, the collected data sets may not be truly representative of the university demographic as a whole.

The first issue with the data sets is the sample size. It is likely that each demographics has hundreds, if not thousands of other students within it. A sample size of 10 would not be sufficient for understanding the opinion of the entire student base. A sample of 50 or 100 would give a clearer indication and allow for a more accurate result overall.

It is also of utmost importance that the sample is of a random selection of students in each demographic. This reduces the chances of the results being misleading and not entirely representative of the true group of students. For example, in this case, the mature student set could only consist of student who do well at university with their results, and the young student set may consist of students who are not doing well at university. This could have a clear impact on student satisfaction overall. Although the standard deviation results show that the sets are almost identical in variance from the mean result, both groups could have completely different capabilities. A random sample would likely give a set of students from all different levels of ability and give a fairer result.

A more detailed test could have been carried out by also separating each demographic into smaller sets based on students' results (averaging 1st, 2:1 2:2 etc.). By using this technique, trends can be spotted between both age demographics, whilst also eliminating the risk of misleading data based on student results.

References

- [1] - A Brief History of Data Analysis by Masashi Miyazaki – 03/11/2015 - <https://www.flydata.com/blog/a-brief-history-of-data-analysis/>
- [2] – Research Methods and Statistics – Ian Walker
- [3] – D. M. Abuhmida, RMDA Lecture 7, 2018.

Appendix

(1)

t-Test: Two-Sample Assuming Unequal
Variances

Alpha = 0.10

	<i>Mature Students</i>	<i>Young Students</i>
Mean	44.5	28.1
Variance	75.38888889	72.98888889
Observations	10	10
Hypothesized Mean Difference	0	
df	18	
t Stat	4.257546666	
P(T<=t) one-tail	0.000236832	
t Critical one-tail	1.330390944	
P(T<=t) two-tail	0.000473663	
t Critical two-tail	1.734063607	

(2)

t-Test: Two-Sample Assuming Unequal
Variances

Alpha = 0.02

	<i>Mature Students</i>	<i>Young Students</i>
Mean	44.5	28.1
Variance	75.38888889	72.98888889
Observations	10	10
Hypothesized Mean Difference	0	
df	18	
t Stat	4.257546666	
P(T<=t) one-tail	0.000236832	
t Critical one-tail	2.213703252	
P(T<=t) two-tail	0.000473663	
t Critical two-tail	2.55237963	

(3)

t-Test: Two-Sample Assuming Unequal Variances

Alpha = 0.01

	<i>Mature Students</i>	<i>Young Students</i>
Mean	44.5	28.1
Variance	75.38888889	72.98888889
Observations	10	10
Hypothesized Mean Difference	0	
df	18	
t Stat	4.257546666	
P(T<=t) one-tail	0.000236832	
t Critical one-tail	2.55237963	
P(T<=t) two-tail	0.000473663	
t Critical two-tail	2.878440473	