

Research Methods & Data Analysis

Assignment 1: Data Analysis

Edward Prasad (1809704)

TABLE OF CONTENTS

Abstract.....	2
Introduction and Background Research	2
Methodology.....	5
Results and Discussion	7
Conclusion.....	10
References	11

ABSTRACT

This report will analyse the data provided of satisfaction levels of mature and young students at a university. This report will propose a null and an alternative hypothesis and will perform statistical analysis on the data to safely reject one of the hypotheses. This report will outline the methods used in the analysis, including background information, and will discuss the results.

INTRODUCTION AND BACKGROUND RESEARCH

This document will analyse the data of a study conducted to examine the difference between mature and young student's satisfaction at a university, the scores of the students are displayed in table 1. This report will make use of analysis techniques including variance, standard deviation, correlation, the null hypothesis and t tests.

Table 1. Mature and young student satisfaction results.

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

Data analysis is used to judge the importance of data by means of statistics and evaluating those statistics to inform decision making to confirm or reject assumptions and hypotheses [1]. Outliers are data that are dissimilar to other data in a set. Outliers in a dataset may occur because of a number of different possibilities including human input error and corrupt data [2]. Outliers can be detected via a number of different techniques, including using standard deviation if the dataset represents a Normal curve. Measures of central tendency, measures of variability and the Normal curve will be discussed in this section as these are the stages that lead to the ability to detect outliers.

Measures of central tendency are the mean, median and mode of a data set and they are utilised when comparing an individual score to a larger group. The mean is the mathematic average of the data and can be calculated by the following equations:

Sample mean: $\bar{X} = \frac{\sum X}{N}$

Population mean: $\mu = \frac{\sum X}{N}$

Where X is the set of scores and N is the number of scores in the set. $\bar{X}$ represents the mean of a sample and μ the mean of a population [3]. The mean of a data set can be affected by extreme results and outliers. A population is a full set of data, for example all scores gathered from a study involving 100 people, while a sample is a portion of the population, such as taking 30 scores from the 100 of the previous example.

The median is the 50th percentile, or the very middle piece of data, in a sorted set and is not affected by outliers. The mode is the most commonly occurring value in a set of scores and is not affected by outliers.

Obtaining a single score from the measures of tendency can be misleading as different sets with a drastically larger range of scores could have the same mean, as shown by the following example.

Set A	Set B
90	62
80	61
55	60
45	59
30	58
$\bar{A} = \frac{\sum A}{N} = \frac{300}{5} = 60$	$\bar{B} = \frac{\sum B}{N} = \frac{300}{5} = 60$

Range, variance and standard deviation can be used to measure variability. Range represents the difference between the lowest and highest scores and does not take into account any skewness of the set of scores [4]. Variance is used to measure the distance between scores in a given data set and is calculated by the following equation. The difference between the score and the mean is squared to remove any negative values.

$$SD^2 = \frac{\sum(X - \bar{X})^2}{N - 1}$$

Where X is the current score, $\bar{X}$ is the mean of the data and N is the number of scores in a set. SD^2 represents variance here as it is the square of the standard deviation. The above equation must be performed on each score in a set to achieve the variance, so repeating the calculation for each score can become very tedious. Thankfully, there is an alternative formula to calculate the variance that also removes rounding errors.

$$SD^2 = \frac{\sum X^2 - \frac{(\sum X)^2}{N}}{N - 1}$$

$N - 1$ is known as the Bessel Correction and is used when the dataset in question is a sample of the population and to provide a fair estimate of the population. If the data set includes the entire population then N is used as the divisor [5].

Standard deviation, introduced by Karl Pearson in 1893, is used to measure how far a score differs from the mean and is denoted by sigma, σ [5]. 68% of scores will reside within one standard deviation from the mean, 95% within two and 99.7% within three [6].

The calculation for standard deviation is to simply square root the result of the variance equation, note the left hand side of the variance equation: SD^2 .

Data can be represented in a number of ways, particularly via bar charts to visually compare data. The Normal curve, also called Gaussian curve, is signified when presented data exhibits a bell curve shape. The Normal curve generally represents symmetrical data as the distribution either side of the mean is even [7], as shown in figure 1.

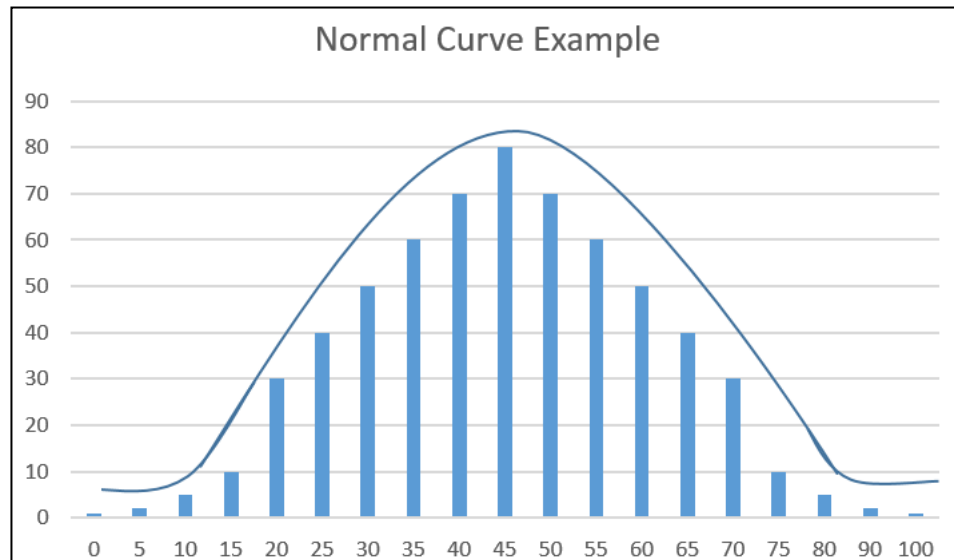


Figure 1. Normal Curve Example.

The Normal curve is a unimodal curve meaning it has a single defined peak at which the measures of central tendency all appear. Sample size can have a significant influence on data distribution. As a sample of a population of data is usually used to estimate the wider population, a small sample size will tend to not represent a Normal curve as there is not enough data to estimate an entire population's distribution [8].

METHODOLOGY

This section will outline and detail the methods used to conduct the analysis in this report including null hypothesis, t tests and how to detect outliers.

The null hypothesis was used in this report to suitably analyse the data provided (see Table 1). The null hypothesis, represented as H_0 , is a statement made about a set of data that is the opposite of what needs to be proven and an alternative (what needs to be proven) is posed against it. The null hypothesis will often be described by stating there is no difference between two data sets [4].

This report aims to analyse the data provided in Table 1 to determine if there is a difference between young and mature students' satisfaction levels. The null hypothesis is:

$$H_0 = \text{Age does not make a difference on satisfaction levels.}$$

The alternative hypothesis is:

$$H_A = \text{Age does make a difference on satisfaction levels.}$$

When rejecting a hypothesis there are two errors, known as Type I and Type II, that can be encountered if the rejection is not considered carefully. Type I error occurs when a true null hypothesis is rejected and Type II error occurs when a false null hypothesis is not rejected [9]. When studying statistical results, a level of significance threshold, α , needs to be applied to aid the rejection decision making process. For example, if $\alpha = 0.1$ then there is a probability of 10% of rejecting the null hypothesis. The probability can become more precise with lower alpha thresholds, such as $\alpha = 0.01$ meaning there is a 1% probability [10].

To be able to understand if a null hypothesis can be rejected, a t-test (also known as a student t-test) is used. There are various forms of the t-test that can be used to test the means of a single sample, two related samples and two independent samples. T-tests assume the populations are Normally distributed, if they are not, other tests can be performed to assist the decision to reject the null hypothesis [11].

A t-test to test the means of two independent samples was used, as the two samples in Table 1 are independent. The t-test will return a value t_{obt} which represents a statistical significance between two groups, if t_{obt} is large then there is a large statistical difference. The t_{obt} value can be calculated with the following equation, if the samples sizes are equal as they are in Table1:

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

Where: n is the number of scores, SD_1^2 and SD_2^2 are the variances of the samples and $\bar{X}_1$ and $\bar{X}_2$ are the means of the samples.

There are two types of results a t-test can produce, directional (one-tailed) and non-directional (two-tailed). One-tailed tests are used when wanting to prove if one sample is greater or smaller than another. As this report is trying to determine if there is a difference between the two samples, the two-tailed test was used.

Once a t_{obt} value is obtained, it needs to be compared to a value to determine if H_0 can be rejected; this value is called t_{crit} . To determine t_{crit} , the degree of freedom must be known which is calculated by $n - 2$, where n is the number of scores. For the scores in Table 1, the degree of freedom is calculated by $n - 2 = 20 - 2 = 18$. The degree of freedom is then compared against the level of significance, α , which must be chosen before the test is computed, and returns a value for t_{crit} . If t_{obt} is larger than t_{crit} , the null hypothesis can be rejected. The most common value for α is 0.05 but this report uses 0.10, 0.02 and 0.01 to determine the rejection of the null hypothesis.

As mentioned previously, outliers can be detected using standard deviation. As one standard deviation from the mean covers 68% of the data, two covers 95% and three covers 99.7% any data that lies beyond a certain number of standard deviations can be said to be an outlier if the data is distributed in a Normal fashion. If the sample is very large then four standard deviations could potentially be used, and if the sample is very small then two standard deviations could be used [2].

RESULTS AND DISCUSSION

This section will detail the results found from using the methods mentioned above and will discuss the impact of the findings.

Initially, grouped frequency diagrams were created to represent the data and used to see if it represented a Normal distribution. As this is a small sample two standard deviations from the mean was used to detect any outliers. Figure 2 shows the grouped frequency diagram for the mature students' results with an interval of four.

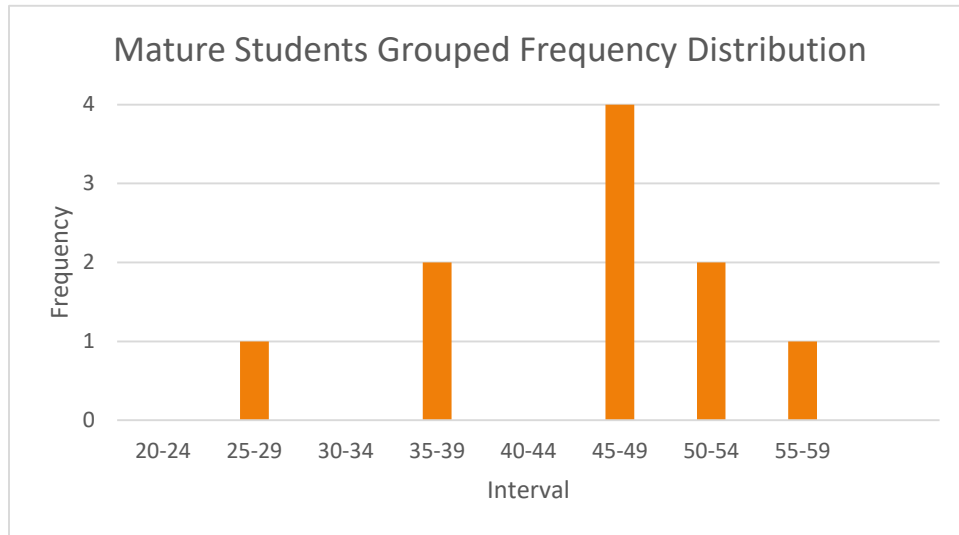


Figure 2. Grouped Frequency Diagram for Mature Students.

To detect any outliers in the mature students' data the standard deviation must be calculated and then each score is checked against two standard deviations from the mean. The variance of the mature students' data is 75.39 therefore the standard deviation is 8.68 and the mean is 44.5. Two standard deviations above the mean is 61.87 and two standard deviations below the mean is 27.13. As the value 25 is below two standard deviations from the mean, it is an outlier. This could occur because of a number of reasons, including personality traits, as the scores are based on personal satisfaction. When conducting experiments based on personal preferences outliers may be more prevalent as a candidate could have experienced other factors that led them to give a low score.

Figure 3 shows the initial grouped frequency diagram created from the young students' data with an interval of four.

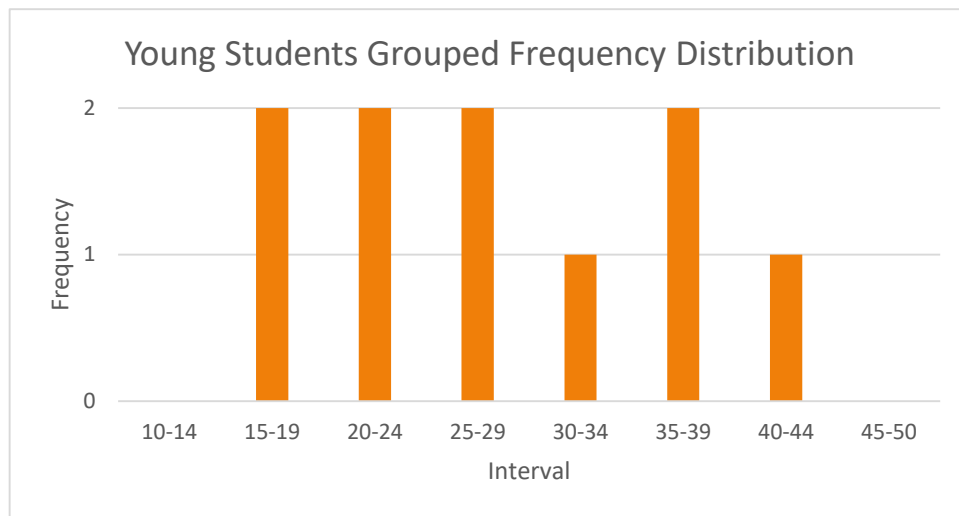


Figure 3. Grouped Frequency Distribution of Young Students with an interval of four.

Figure 3 shows a non-normal distribution of data which was a problem as a t-test cannot be performed on non-normal data as mentioned in the methodology section above. To resolve this, the interval size was changed from four to two, shown in figure 4.

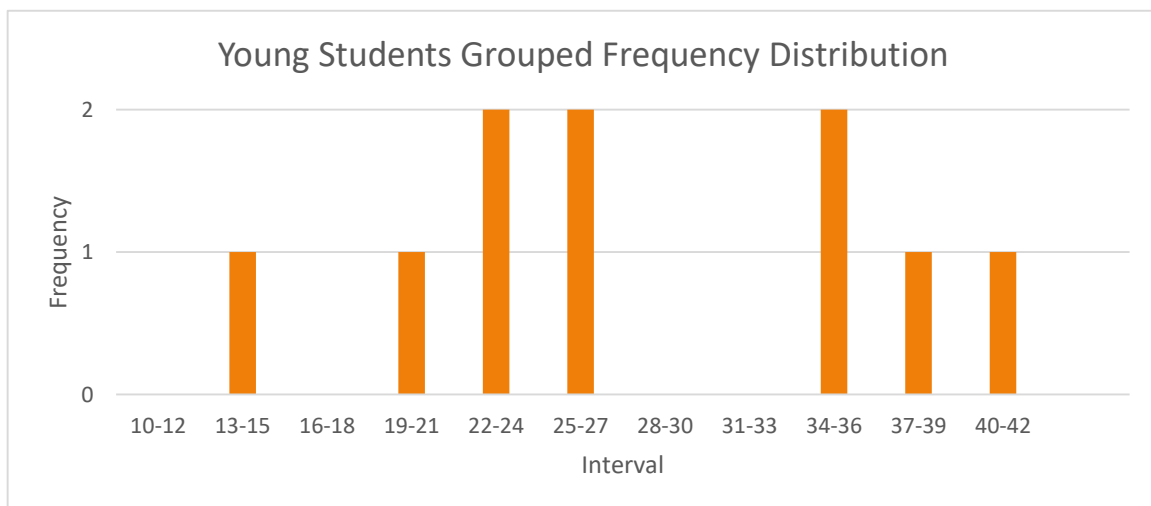


Figure 4. Grouped Frequency Distribution of Young Students with an interval of two.

Figure 4 now shows the data in a distribution that tends more towards the Normal curve. To detect outliers for the younger students the standard deviation is 8.54 so two standard deviations above the mean, of 28.1, is 45.19 and two below is 11.01. With these figures there are no scores that lie beyond two standard deviations from the mean, therefore there are no outliers.

With these distributions, t-tests can now be performed on the data. The t-tests shown below were all calculated using the data analysis tool in Microsoft Excel. Table 2 shows the t-test results from the data with an alpha level of 0.10.

Table 2. t-test Results with alpha of 0.10.

	<i>Variable 1</i>	<i>Variable 2</i>
Mean	44.5	28.1
Variance	75.38889	72.98889
Observations	10	10
Hypothesized Mean Difference	0	
df	18	
t Obt	4.257547	
P(T<=t) one-tail	0.000237	
t Critical one-tail	1.330391	
P(T<=t) two-tail	0.000474	
t Critical two-tail	1.734064	

As the t_{obt} value from Table 2 is larger than the t_{crit} value then the null hypothesis can be rejected without causing a Type I error. However, this is only proving 10% probability so to ensure a stronger rejection an alpha value of 0.02 was used as shown in Table 3.

Table 3. t-test Results with alpha of 0.02.

	<i>Variable 1</i>	<i>Variable 2</i>
Mean	44.5	28.1
Variance	75.38889	72.98889
Observations	10	10
Hypothesized Mean Difference	0	
df	18	
t Obt	4.257547	
P(T<=t) one-tail	0.000237	
t Critical one-tail	2.213703	
P(T<=t) two-tail	0.000474	
t Critical two-tail	2.55238	

Again the t_{obt} value is larger than the t_{crit} value so the null hypothesis can be rejected with an alpha of 0.02 without encountering a Type I error. This can be made even more robust by lowering the alpha level to 0.01 as shown in Table 4.

Table 4. *t*-test Results with alpha of 0.01.

	Variable 1	Variable 2
Mean	44.5	28.1
Variance	75.38889	72.98889
Observations	10	10
Hypothesized Mean Difference	0	
df	18	
t Obt	4.257547	
P(T<=t) one-tail	0.000237	
t Critical one-tail	2.55238	
P(T<=t) two-tail	0.000474	
t Critical two-tail	2.87844	

As with alpha levels of 0.10 and 0.02, the t_{obt} value is larger than the t_{crit} therefore the null hypothesis can once again be rejected.

These results show that as the null hypothesis can safely be rejected at each alpha level, particularly 0.01, age does in fact make a difference to satisfaction levels of students attending a university.

CONCLUSION

With the results gained from this analysis the null hypothesis can be safely rejected without creating a Type I error and the alternative hypothesis can be accepted. Performing a t-test with an alpha level of 0.01 confirmed it was safe to reject the null hypothesis at any alpha level greater than 0.01. Using two standard deviations from the mean an outlier was detected in the mature students' data but this was not omitted as it is still a possible score that a person in this study could give.

If this study were to be performed again in the future, a larger sample size would be needed to achieve a more robust rejection decision as a larger population will have a wider range of results. Personal circumstances can affect a score in this study so performing the study on multiple days and in different locations would provide a stronger dataset to analyse.

REFERENCES

- [1] M. Jeevan, "6 steps taken in data analysis," Edvancer Eduventures, 18 September 2017. [Online]. Available: <https://www.edvancer.in/6-steps-taken-data-analysis/>. [Accessed 02 January 2019].
- [2] J. Brownlee, "How to Use Statistics to Identify Outliers in Data," Machine Learning Mastery, 25 April 2018. [Online]. Available: <https://machinelearningmastery.com/how-to-use-statistics-to-identify-outliers-in-data/>. [Accessed 02 January 2019].
- [3] T. Payne, "X-Bar in Statistics: Theory & Formula," Study.com, [Online]. Available: <https://study.com/academy/lesson/x-bar-in-statistics-theory-formula.html>. [Accessed 02 January 2019].
- [4] R. Nerurkar, "Basics of statistics for postgraduates," *Indian Journal of Dermatology, Venereology and Leprology*, vol. 74, no. 6, pp. 691-695, 2008.
- [5] G. Upton and I. Cook, *A Dictionary of Statistics*, Oxford: Oxford University Press, 2008.
- [6] Statistics How To, "Bell Curve (Normal Curve): Definition," Statistics How To, 21 December 2014. [Online]. Available: <https://www.statisticshowto.datasciencecentral.com/bell-curve/>. [Accessed 02 January 2019].
- [7] The Editors of Encyclopaedia Britannica, "Normal distribution," Encyclopaedia Britannica, [Online]. Available: <https://www.britannica.com/topic/normal-distribution>. [Accessed 03 January 2019].
- [8] J. Krithikadatta, "Normal Distribution," *Journal of Conservative Dentistry*, vol. 17, no. 1, pp. 96-97, 2014.
- [9] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, New York: Lawrence Erlbaum Associates, 1988.
- [10] P. Wittkopf, "What are Type I and Type II Errors?," 21 April 2017. [Online]. Available: <https://www.students4bestevidence.net/type-i-and-type-ii-errors/>. [Accessed 03 January 2019].
- [11] R. M. Pandey, "Commonly used t-tests in medical research," *Journal of the Practice of Cardiovascular Sciences*, vol. 1, no. 2, pp. 185-188, 2015.