Data Analysis – Significant Difference Between Two Means

Introduction

Quantitative data produces numbers, which influence the way Businesses, Governments and Science work. Statistical analysis are also found in agriculture, astronomy, politics and education to name a few. [1] This report analyses two sets of data collected from a sample size of 10 relating to student satisfaction scores in a University pertaining to two groups; matured and young students. Tests were performed to determine if the samples were normally distributed. Two groups of data can be compared to determine if they are statistically different and a t-test was performed to determine the same.

Data Source

The survey was carried out in the form of a test with a sample size of ten from two groups of students in the age range 45 - 62 and 17 - 20. The scoring ranged between 0 – 60, where a higher score indicated higher satisfaction.

Population: University students with an unknown size n, mean μ and standard deviation σ.

Samples: Two randomly selected sets of size 10 each.

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

Assumptions:

Given the set of data:

- No background information available
- Test questions asked are unknown
- Nothing is known about the characteristics of the two groups except their age ranges
- No information on the two groups such as courses undertaken, year of study or services used

The following assumptions have been made regarding the two groups of data:

- The variances are homogenous.
- The two groups are independent.
- The distribution is normal
- There are no outliers.

Independent samples:

The samples satisfy all the criteria of independent variables as mentioned below:

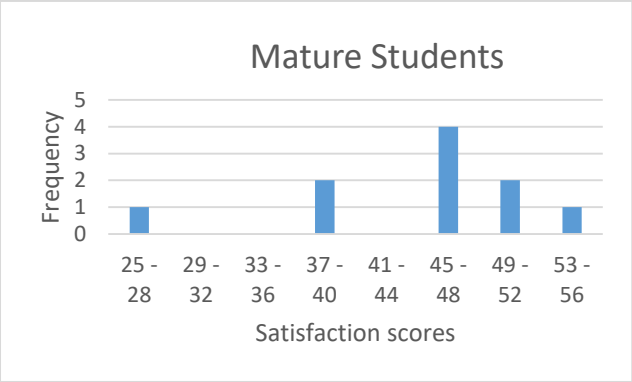
- The same set of students have not been part of the both groups. This is evident from their age ranges.
- The sampled data is a random sample.
- The population is normally distributed.

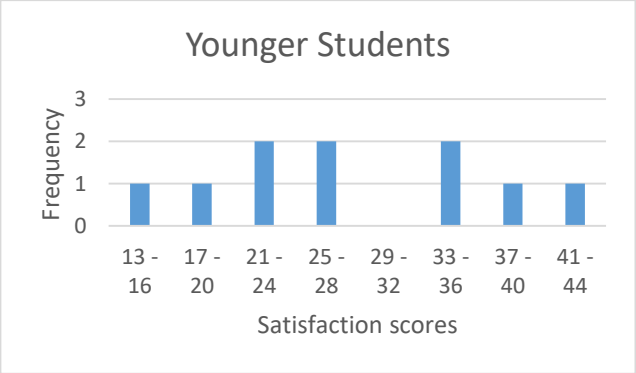
Normalcy of distribution

In order to evaluate the difference in the populations using inferential statistics, the data should follow a normal distribution. It is believed that all distributions are usually normal with a large population size. The normal distribution is represented in a bell shape curve.

The characteristics of normal distribution are [1]:

- It is symmetrical.
- The mean and the median are same.
- The interquartile range is 1.33 times of the standard deviation.

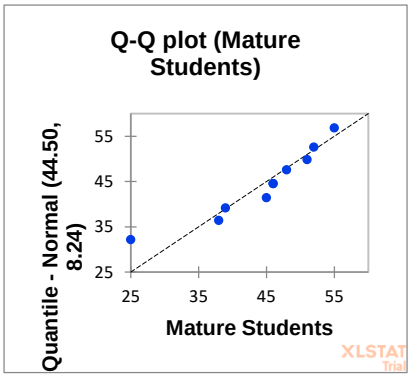




The Central Limit Theorem states that [2]:

> *"The mean of the sample means is the same as the population mean. The standard deviation of the sample means is always equal to the standard error. The distribution of sample means will become increasingly more normal as the sample size increases."*

It can be said that the data is normally distributed. However, the quantile-quantile plot (q-q plot) would ascertain this. The graph uses the scores in terms of quantiles and is plotted against a normal curve. The x axis being the observed position and y axis the expected position. As most of the points are on or very close to the line at 45 degrees angle, it can be concluded that the distribution is normal [3].



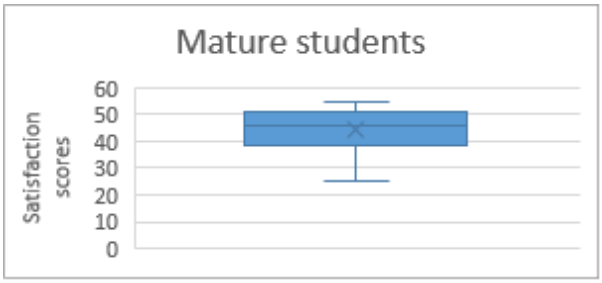
H_0: The sample distribution follows a normal distribution.

H_a : The sample distribution does not follow a normal distribution

The null hypotheses is accepted.

Outliers:

Any data that lies beyond ± 2 standard deviation from the mean is termed an outlier which affects the arithmetic mean adversely. Since the outlier contributes to the value of the mean, identifying the outlier using the mean + or – standard deviation is not a desired method. The deviation from the mean is considered better for identifying outliers [4]. A box and whisker plot used in descriptive analysis is used to give the distribution of the median and the quartiles [5]. The data set for Mature students is odd numbers and consideration is given as to whether median is to be included in the calculation of quartiles. The method chosen is critical wherever the boundaries of the data are to be fixed in order to decide [4]. In the data set, 25 appears to be a mild outlier, however when the median is excluded for the percentile calculations, it does fall within the fence of the box. Since it is not known if the 25 was recorded due to any errors, excluding 25 might result in giving a false confidence that no student is less satisfied. Simply removing the most extreme score could distort the sample of genuine data [3]. Both box and whiskers chart and the Grubbs test validate that there is no outlier.



Grubbs test for outliers / Two-tailed test:	
G (Observed value)	2.246
G (Critical value)	2.290
p-value (Two-tailed)	0.067
alpha	0.05

H_0: There data has no outlier
H_a: The extreme value is an outlier
p-value is greater than the significance level alpha=0.05, null hypothesis H0 is not rejected.

Hypotheses for t-test

The study aims to examine the difference between the two groups and is the research hypotheses referring to the sample. Hypotheses are guesses or estimates and cannot be accepted with confidence until it is proved. However, a benchmark is required to start the analysis and hence a null hypothesis which states otherwise is declared [6].

The null hypothesis and the alternate hypothesis are as follows:

H_0 : There is no significant difference between the two groups.
H_a : There is significant difference between the two groups.

To elaborate, the null hypotheses infers there is no significant difference between the two population means. The alternate hypothesis states that there is a significant difference between the two sample means.

H_0 : $\mu_1 = \mu_2$
H_a : $\mu_1 \neq \mu_2$

The significance level:

Although it is aimed to prove with confidence that either of the hypothesis is true, there are several factors that might influence the scores such as:

- The individual needs of the participants and the type of service provided
- Their preferences and use of facilities provided by the university
- Their individual skills and capabilities
- The amount of effort that attribute towards their success in the course.

It is a general assumption that success rates often lead to greater satisfaction. Consequently, there is leeway for some error in the inference of the data and it is not possible to say with absolute confidence of 100% that either of the hypothesis is accepted. There is always a probability of a Type I or Type II error.

A Type I error is rejecting the Ho, when it is true and Type II error is accepting a false Ho. The focus is to minimise the Type I error, whereby reducing the chances of a false positive [7].

The three different significance levels are set at 0.1, 0.01 and 0.02.

At probability $p< 0.1$, any difference between the groups would mean there is a 10% chance that it is due to the other factors that might have influenced the test score. At $p< 0.01$ and $p< 0.02$, the chance is even less at 1% and 2% respectively. These are the level of risk associated with an error being made in rejecting the null hypothesis when it should have been accepted. By setting them up as at lower levels we are reducing the risk of Type I error.

The test Statistic:

The measures of central tendency and variances for the set of data is shown in the table.

	Mature Students	Younger Students
Mean	44.5	28.1
Median	46.0	26.5
Mode	46.0	
Variance	75.38888889	72.98888889
Standard Deviation	8.682677518	8.543353492

The sample mean of the mature students differs from the mean of younger students by 16.4.

The median is considered as the measure of central tendency here as there is one value in mature student that could distort the arithmetic mean [1]

In order to establish whether there is a difference between the groups to reject the null hypotheses would depend on the variability around the mean. If there is a lot of noise around the data, it may prove difficult to evaluate the actual difference. A large difference might be a result of large noise, thereby reducing the level of confidence for the hypotheses to be rejected [8]. Hence the importance to prove that this difference did not occur by chance.

As the independence of the two samples is established, if the sample size were very large, the z distribution could have been appropriate. The independent sample t- test and the Wilcoxon-Mann-Whitney Test (WMW) are the most popular statistical tests [9]. To evaluate the hypotheses, a two tailed t-test for equal variances has been performed as the direction of the difference is not the objective.

The t value is obtained from the formula [7].

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

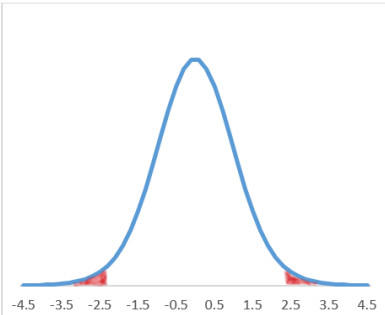
Even if their means are the same, there could still be a variability of data around this mean. The population variances are considered the same as the sample variances are assumed to be the similar. Since the variability between the two groups is the focus here, the estimate of the variances of the population is given by the denominator. This is the standard error of the difference between the means.

The t_{obt} values for the three chosen levels of significance and their critical values are shown in the table.

	α = 0.01	α = 0.02	α = 0.10
t_{obt}	4.257546666	4.257547	4.257547
t_{crit}	1.330390944	2.213703	2.55238

Discussion

The critical values are the values at the significance level α at degrees of freedom (df) 18. The two tailed t-test results are significant at all three levels where t_{obt} is more than t_{crit} and the null hypotheses can be rejected. To keep the risk level at 1%, divide it between the two sides of the curve so the result falls within a smaller area [3] It does not give an indication of the direction of the difference.



How powerful is this test? Three factors influence the power. The level of α, the effect and the size of the sample. By keeping the α at a high level of 0.10, the power of the test is increased [3]. However, this increases the chances of a Type I error. Rejecting the H0 at α of .01 is the least of the risk in committing a Type I error. This means that both the statistical power and the probability of a Type I Error are lower. This is 99% chance that there is no difference in the means when there actually isn't. The effect size is the magnitude of the difference in the means. The formula to calculate the Cohen effect size is [6].

$$\frac{\overline{X1} - \overline{X1}}{\sqrt{\frac{(SD1^2 + SD^2}{2}}}$$

where, the average of the standard deviations is the pooled SD. The effect size for this data set is calculated as 1.90. This deviates from zero largely and hence the effect is of a large magnitude. The two sets do not overlap. They are different.

Result

Since the obtained t scores are greater than the critical stores at the respective levels, the null hypotheses is rejected.

$$\mu_1 \neq \mu_2$$

The alternate hypotheses, that there is a significant difference between the two groups is accepted. The two sample means are significantly different and hence the population from which these were derived also have different means. Mature students and younger students have a difference in their satisfaction levels.

Conclusion

Satisfaction test and scoring of a random sample of 10 mature students and 10 younger students in a university were carried out to determine if there is a difference between their scores. A two tailed t-test for independent samples with equal variances was performed. The results show that there is a significant difference between the two groups. This test does not conclude which group is more or less satisfied as it was not the objective of the test. The University can conclude that there is clearly a difference in the satisfaction of the mature students as opposed to the younger students and should be looking to maintain a balance for overall satisfaction. It might be worthwhile to repeat the test with a larger sample size of at least 30.

References

[1] N. J. Salkind, in *Statistics for peoplewho hate Statistics*, USA, Sage Publications, 2014, p. 178.

[2] A. Rathi, “Towards Data Science,” 7 August 2018. [Online]. Available: https://towardsdatascience.com/inferential-statistics-for-data-science-b0075670fc8a. [Accessed 26 11 2018].

[3] P. R.Hinton, in *Statistics Explained*, Taylor & Francis, 2014, pp. p 92 - 105.

[4] E. Langford, “Quartiles in Elementary Statistics,” *Journal of Statistics Education,* vol. 14, no. 3, p. 2, 2006.

[5] &. V. H. Rietveld, “The t test and beyond: Recommendations for testing the central tendencies of two independent samples in research on speech, language and hearing pathology.,” *Journal of Communication Disorders,* pp. 58,158-168, 2015. [Online] Available: https://doi.org/10.1016/j.jcomdis.2015.08.002 [Accessed: 0/12/2018]

[6] J. Cohen, Statistical power analysis for the behavioral sciences, Hillsdale: Erlbaum, 1988.

[7] J. H. kranzler, in *Statistics for the Terrified*, USA, Rowman & Littlefield, 2018, pp. 122,123,130-136.

[8] W. M. Trochim, "Web centre for Social Research Methods -PostTest Analysis only," 20 10 2006. [Online]. Available: http://socialresearchmethods.net/kb/statsimp.php. [Accessed 5 12 2018].

[9] M. McGee, "Case for omitting tied observations in the two-sample t-test and the Wilcoxon-Mann-Whitney Test," *PLos One,* vol. 13, no. 7, 2018.[Online] Available: https://search.proquest.com/docview/2075522458?accountid=130472. DOI: http://dx.doi.org/10.1371/journal.pone.0200837 [Accessed 30/11/2018]

[10] N. J. salkind, "Statistics for people who hate statistics," USA, Sage publications, 2014, pp. 3-6.