

Research Methods & Data Analysis

Assignment 1: Data Analysis

Nathan Hughes - 1809708

January 7, 2019

Abstract

This report examines the samples collected from a university experience satisfaction survey. Descriptive statistics and figures are produced in order to formulate inferential tests.

A discussion of sample size and presence of outliers follows.

1 Introduction

A university experience satisfaction survey with a high reliability was used to gather the scores of 10 young and 10 mature students. In this report, the samples derived from two groups of students taking the survey are examined through descriptive and inferential methods. Measure of central tendency, dispersion and shape are first described, which then inform the later hypothesis tests into mean and variance equality. Tests are conducted on each of the points to determine if any qualify as an outlier to the main groups of data. The methodology of the study is criticised with regards to the information recorded and the size of both samples. Further tests are then recommended.

2 Descriptive Statistics

2.1 Measures of Central Tendency, Dispersion & Shape

As this was a trial study, it is correct to assume that the data provided is a sample of a larger population of students. Consequently, where appropriate, sample statistics are produced.

Using formula (1) for the *arithmetic mean* the sample means are $\bar{x}_M = 44.5$ and $\bar{x}_Y = 28.1$.

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n X_i \quad (1)$$

The *median* of each sample can be calculated as $\text{med}(x_M) = 46$ and $\text{med}(x_Y) = 41$.

The *minimum* and *maximum* of each of the samples are $\min(x_M) = 25$, $\min(x_Y) = 15$, $\max(x_M) = 55$, $\max(x_Y) = 41$.

The *lower* and *upper quartiles* of the samples are $Q_1(x_M) = 40.5$, $Q_1(x_Y) = 22.5$, $Q_3(x_M) = 50.25$ and $Q_3(x_Y) = 35.5$.

The *mode* of the mature student sample is 46, but no mode exists for the young student sample.

The *sample variances* are unbiased estimators of the population variance, which is a measure of the tendency of the data to disperse from the sample mean. The sample variances are calculated to be

$s_M^2 = 75.4$ and $s_Y^2 = 73.0$ using formula (2). The *sample standard deviations* are then the square root of these values, $s_Y = 8.68$ and $s_M = 8.54$.

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \quad (2)$$

The *standard error* of a sample is a measure of how different the sample mean is likely to be from the population mean. The standard error of the samples are $\sigma_{\bar{x}_M} = 2.75$ and $\sigma_{\bar{x}_Y} = 2.70$, calculated using formula (3).

$$\sigma_{\bar{x}} = \frac{s}{\sqrt{n}} \quad (3)$$

The *skewness* of a sample is a measure of asymmetry, where negative (cf. positive) skew indicating that the data tends to be more concentrated below (cf. above) the mean. A skew of 0 indicates the same symmetry as the standard $N(0, 1)$ distribution. The skewness of the samples are calculated as $G_{1,M} = -1.26$ and $G_{1,Y} = 0.0539$ using formula (4).

$$G_1 = \frac{n}{(n-1)(n-2)} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s} \right)^3 \quad (4)$$

The *kurtosis* of a sample is a measure of the peakedness of the data set. A positive value indicates concentration in the central body of the distribution, and a negative value indicates more concentration in the tails. The excess kurtosis is the kurtosis of the normal distribution, which is equal to 3, subtracted from the sample kurtosis. The excess kurtosis of the samples are calculated as $G_{2,M} = 1.98$ and $G_{2,Y} = -1.18$ using formula (5).

$$G_2 = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s} \right)^4 - \frac{3(n-1)^2}{(n-2)(n-3)} \quad (5)$$

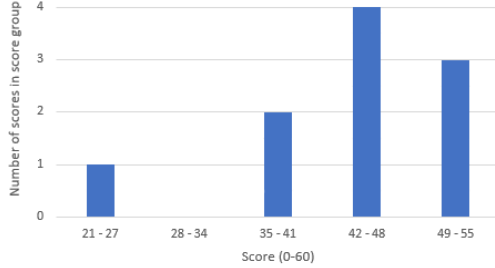
Table 1 below summarises the above calculation for ease of comparison.

Table 1: Descriptive statistics for mature and young student samples.

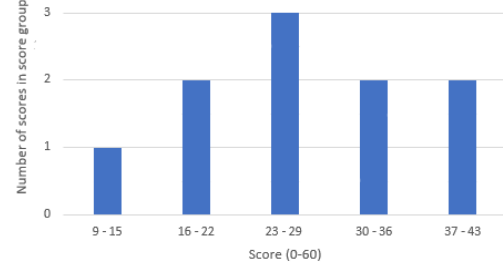
Property	Mature	Young
Sample Size	10	10
Arithmetic mean	44.5	28.1
Lowest Value	25	15
Lower quartile	40.5	22.5
Median	46	26.5
Upper quartile	50.25	35.5
Highest Value	55	41
Range	30	26
Mode	46	N/A
Variance	75.4	73.0
Standard Deviation	8.68	8.54
Standard Error	2.75	2.70
Skewness	-1.26	0.0539
Excess Kurtosis	1.98	-1.18

2.2 Data Visualisation

To visualise the data, two histograms have been produced which display the rough distribution of both samples. Additionally, a box and whisker plot has been created in order to visualise the comparative sizes of the maxima and minima, the interquartile ranges and the means.



(a) Histogram of mature student data



(b) Histogram of young student data

Figure 1: Histograms of both sets of data, displayed independently

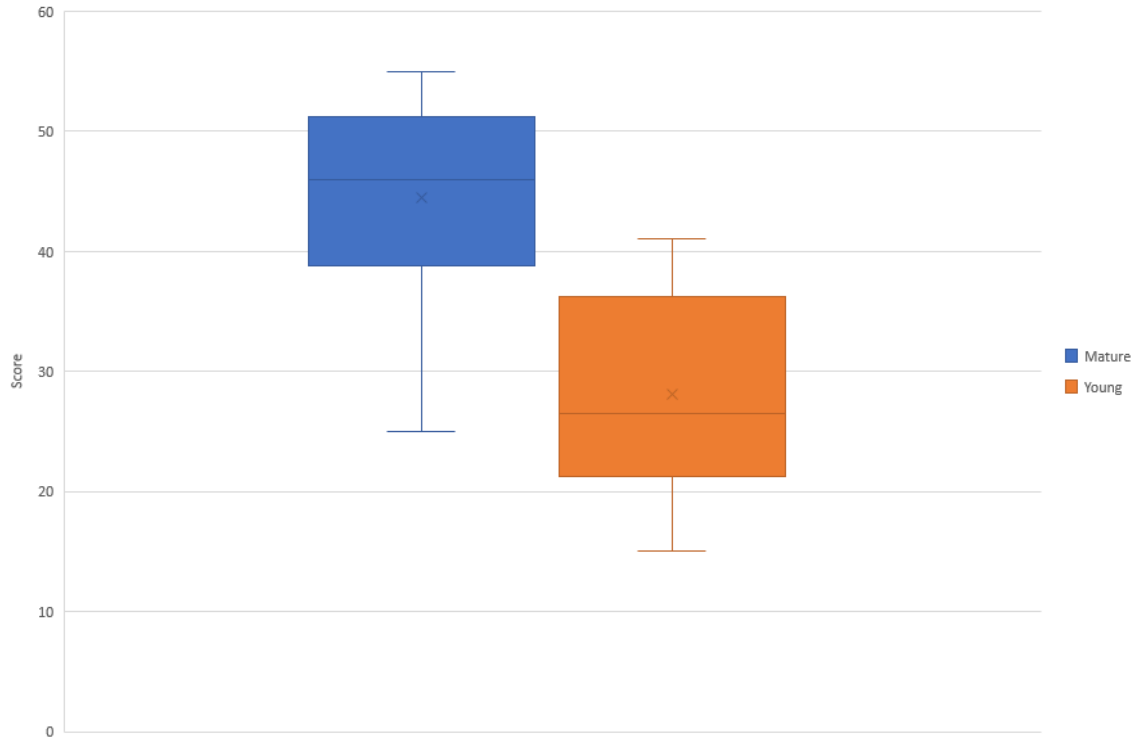


Figure 2: Box & Whisker plot of the young and mature student data

2.3 Size of the Sample

Prior to sampling from a population for the purpose of statistical testing, one might ask what effect the size of the sample. To explore this question, it is useful to consider the following theorem.

Theorem (Weak Law of Large Numbers). *Let $X_1, X_2, \dots$ be an infinite sequence of independent and identically distributed random variables, with $\mathbb{E}[X_i] = \mu$ for every $i \in \mathbb{N}$. Define $\bar{X}_n = \frac{1}{n}(X_1 + \dots + X_n)$. Then for every $\epsilon > 0$,*

$$\lim_{n \rightarrow \infty} \mathbb{P}(|\bar{X}_n - \mu| > \epsilon) = 0$$

A proof of this theorem can be found at [1]. The Weak Law of Large Numbers states that, as the size of the sample increases, the sample mean will tend to the population mean with probability 1. This proves that the sample mean becomes an increasingly better estimator of the population mean as the sample size increases.

Conversely, one might ask how large an error one might expect from a small sample. Again, the Weak Law of Large Numbers states that the error in the estimation of the population mean tends to be larger if the sample size is small; precisely, the standard error tends to be smaller for larger sample sizes, and therefore a smaller sample size will tend to have a larger standard error compared to large samples. Consequently, the size ($n = 10$) of each of the samples in question is likely to cause

a large margin for error in the calculations. This error is represented by the pre-factor $n^{-\frac{1}{2}}$ in the calculation of the standard error (See formula 3).

3 Inferential Statistics

One of the first tests to conduct is for equality of means. To do so, most tests require that the underlying distribution be identified. It is therefore necessary to justify any assumption relating to the underlying distribution of the data set. In this case, a normal distribution is the most likely candidate.

Additionally, the variances of the samples appear to be close, and so a test to validate their equality should be conducted.

Finally, as can be seen from the box and whisker plot in figure 2, the lower tail of the mature sample is comparatively large. Considering that the variances of both samples are of somewhat equal size, a test should be conducted to detect the presence of outliers in the samples.

3.1 Test for Normality

A *Shapiro-Wilk* test is carried out to test the normality of the two data sets. This test was chosen since it has higher *power* than similar distribution tests [2]. That is, the Shapiro-Wilk test is more likely than the other tests examined to correctly reject the null hypothesis.

The null hypothesis states that the data is sampled from a normal distribution, where the mean and variance are not specified, and the alternative hypothesis states that the data is not normally distributed. Specifically, for a population distribution X , the hypotheses under testing are

$$\begin{aligned} H_0: X \text{ is normally distributed,} \\ H_1: X \text{ is not normally distributed.} \end{aligned}$$

This test was conducted using R; the code can be found in appendix A. The values obtained are $W_M = 0.904$ and $W_Y = 0.960$. The critical values for the p -values 0.01, 0.02 and 0.10 are $W_{0.01} = 0.781$, $W_{0.02} = 0.804$ and $W_{0.10} = 0.869$ respectively [3]. Since W_M and W_Y are larger than these critical values, the null hypothesis is accepted.

The Shapiro-Wilk test, whilst being the most powerful test out of those considered, still has a relatively low power [2]. It is recommended that a sample size of at least 50 for each group is obtained to be confident that the hypothesis is true.

3.2 Testing for Equality of Means

The size of each sample ($n = 10$) is small enough to warrant the use of t -distributions in the hypothesis testing process. Since a test for equality of variance hasn't been conducted, Welch's generalisation of the t -test for unequal variances is employed here [4]. The goal of the study is to examine if there is any substantial difference between the groups, and as such the hypotheses used are

$$H_0: \mu_M = \mu_Y,$$

$$H_1: \mu_M \neq \mu_Y.$$

where μ_M, μ_Y are the population means of the mature and young students respectively.

$$t = \frac{\bar{x}_M - \bar{x}_Y}{\sqrt{\frac{s_M^2}{n_M} + \frac{s_Y^2}{n_Y}}} = \frac{44.5 - 28.1}{\sqrt{\frac{75}{10} + \frac{73}{10}}} = 4.26$$

$$df = n_M + n_Y - 2 = 10 + 10 - 2 = 18$$

The test is carried out with significance levels of 0.1, 0.02, 0.01. The corresponding critical values for the t -test are $t_{0.1} = 1.734$, $t_{0.02} = 2.552$, $t_{0.01} = 2.878$, which can be calculated using the code in appendix B. Since the t -statistic is greater than each of these critical values, the result is statistically significant, and the null hypothesis H_0 is rejected.

3.3 Confidence Intervals for Means

Since the standard deviation is unknown, to construct the confidence intervals for the mean t -statistics are used to model the uncertainty in the mean.

$$t_{0.1} = 1.8331$$

$$t_{0.02} = 2.8214$$

$$t_{0.01} = 3.2498$$

The confidence interval for a given critical t -value t^* is

$$(\bar{x} - \frac{t^* s}{\sqrt{n}}, \bar{x} + \frac{t^* s}{\sqrt{n}})$$

Table 2: Confidence intervals for significance levels.

Significance Level	Mature	Young
0.1	(39.5,49.5)	(23.1,33.1)
0.02	(36.8,52.2)	(20.5,35.7)
0.01	(35.6,53.4)	(19.3,36.9)

The confidence intervals only have non-empty intersection for $\alpha = 0.01$. This is consistent with the previous t -test asserting inequality of means at the confidence levels $\alpha \in \{0.1, 0.02\}$, but is also consistent with inequality for $\alpha = 0.01$; confidence intervals overlapping are not indicative of a lack of statistical significance [5].

Some caution is to be taken when using the confidence intervals in further inferential research [6]. Rather than being a measure of how accurate the sample mean describes the true population mean, the confidence intervals describe an interval in which the sample mean of $(100 - \alpha)\%$ further samples will lie, where α is the significance level.

3.4 Testing for Equality of Variances

Using the assumption of normality of the populations, an F -test can be employed to test for variance equality. Explicitly, the hypotheses are

$$\begin{aligned} H_0: \sigma_M &= \sigma_Y, \\ H_1: \sigma_M &\neq \sigma_Y. \end{aligned}$$

The F -statistic is given by

$$F = \frac{s_M^2}{s_Y^2}$$

For the samples in question, $F = 1.03$. This statistic is compared to the critical values of the F -distribution $F_{\frac{\alpha}{2}, 9, 9}$ and $F_{1-\frac{\alpha}{2}, 9, 9}$, for significance levels of $\alpha \in \{0.1, 0.02, 0.01\}$, given in table 3.

Table 3: Critical F -statistics for comparison in F -test.

Significance Level	$F_{\frac{\alpha}{2}, 9, 9}$	$F_{1-\frac{\alpha}{2}, 9, 9}$
0.1	2.44	0.410
0.02	4.33	0.231
0.01	5.35	0.187

Since F is between each pair of critical values, the result is statistically insignificant and the null hypothesis holds for each significance level.

4 Outliers

4.1 Detection of Outliers

When testing for the presence of outliers in a data set it not correct to use the mean as the measure of central tendency [7]. This is due to the sensitivity of the mean to the presence of said outliers. Rather, an expression dependent of the *median* is advisable.

Given a sample x , the *median absolute deviation*, written as $\text{MAD}(x)$, is defined as the median of the absolute deviation of the data points from the median of the data set. That is,

$$\text{MAD}(x) = \text{med}(|x_i - \text{med}(x)|)$$

This expression is used as a substitute for the standard deviation and the median is used as a substitute for the mean. A short calculation shows that $\text{MAD}(x_M) = 5.5$ and $\text{MAD}(x_Y) = 7.5$. The modified Z -score used in the test is

$$\tilde{z} = \frac{x_i - \text{med}(x)}{1.4826 \text{MAD}(x)}$$

where $1.4826 \approx \Phi^{-1}(0.75)^{-1}$ is the reciprocal of the 75th-percentile of the standard normal distribution. The moderately conservative critical value of 2.5 is used as a comparison for these values.

The calculated Z -scores can be found in table 4.

Table 4: Z -scores using median absolute deviation.

Mature		Young	
Data Point	$\tilde{z}$	Data Point	$\tilde{z}$
45	0.122	34	0.674
38	0.981	22	0.405
52	0.736	15	1.03
48	0.245	27	0.0450
25	2.58	37	0.944
39	0.858	41	1.30
51	0.613	24	0.225
46	0.00	19	0.674
55	1.10	26	0.0450
46	0.00	36	0.854

Comparing to the critical value of 2.5, the value 25 in the mature sample is flagged as an outlier. The effect of this outlier is most prevalent in the calculation of the mean. The mean of the mature student sample is 44.5, but with the outlier removed the mean is more than 2 points higher at 46.7. The standard deviation is also effected by the removal of this outlier, with a new standard deviation of 5.66 as compare to the original value of 8.68.

4.2 Ethics of Outlier Removal

The nature of the data means it would be unethical to remove the outliers from the data set. It is not reasonable to believe that any score is an outlier due to error since the method of obtaining said data is known to have high reliability. Since the data point is the opinion of a student, rather than a malfunction in testing, the hypothesis testing is not repeated with the data point omitted.

5 Discussion & Conclusion

The t -test carried out in section 3.2 concludes that the means of the underlying populations are not equal. The variances of the populations, on the other hand, were shown to be equal using an F -test. One hypothesis to be tested might be that the average score increases monotonically with age. This would require the age of each of the participants to be recorded along with the scores. Problems might arise with a lack of sample to be taken in certain age groups.

The assumption of normality was supported by the Shapiro-Wilk test, but since the sample size was comparatively low, it is still possible that the underlying population distribution is not normal [2]. A repeated test with a sample size of at least 50 will determine the validity of the normality assumption with a much higher power.

Appendices

A Shapiro-Wilk Test Code in R

```
matureSample <- c(45, 38, 52, 48, 25, 39, 51, 46, 55, 46)
youngSample <- c(34, 22, 15, 27, 37, 41, 24, 19, 26, 36)
shapiro.test(matureSample)
shapiro.test(youngSample)
```

This gives an output of:

```
Shapiro-Wilk normality test
```

```
data:  matureSample
W = 0.90337, p-value = 0.2385
```

```
Shapiro-Wilk normality test
```

```
data:  youngSample
W = 0.95959, p-value = 0.7812
```

B *t*-test Critical Values in R

```
qt(0.995,18)
qt(0.99,18)
qt(0.95,18)
```

This gives an output of:

```
[1] 2.87844
[1] 2.55238
[1] 1.734064
```

References

- [1] J. Laurie Snell and C. M. Grinstead, “Introduction to probability,” p. 307, <https://math.dartmouth.edu/prob/prob/prob.pdf>, last accessed on 2018-12-21.
- [2] N. Mohd Razali and B. Yap, “Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests,” *J. Stat. Model. Analytics*, vol. 2, no. 1, 2011.
- [3] S. S. Shapiro and M. B. Wilk, “An analysis of variance test for normality (complete samples),” *Biometrika*, vol. 52, no. 3/4, pp. 591–611, 1965. [Online]. Available: <http://www.jstor.org/stable/2333709>
- [4] B. Welch, “The generalization of ”Student’s” problem when several different population variances are involved,” *Biometrika*, vol. 34, pp. 28–35, 1 1947.
- [5] K. M. Bower, “Some misconceptions about confidence intervals,” *Six Sigma Forum-American Society for Quality*, 2003.
- [6] R. Morey, R. Hoekstra, J. Rouder, M. Lee, and E. Wagenmakers, “The fallacy of placing confidence in confidence intervals,” *Psychon Bull Rev.*, vol. 23, pp. 103–123, 2016.
- [7] C. Leys, C. Ley, O. Klein, P. Bernard, and L. Licata, “Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median,” *Journal of Experimental Social Psychology*, vol. 49, no. 4, pp. 764 – 766, 2013. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0022103113000668>