[Title]

SOR1232 – Hypothesis Testing

Supervised by Dr. David Suda

Chris Frendo (439600L) (Computer Science 1st Year)

Manwel Bugeja (454000L) (Computer Science 1st Year)

Domenico Agius (6500H) (Computer Science 1st Year

Table of Contents

[1 Introduction 2](#_Toc10562518)

[2 Aims and Objectives 3](#_Toc10562519)

[3 Descriptive Statistics & Illustrations 4](#_Toc10562520)

[3.1 Measurements of Location 4](#_Toc10562521)

[3.1.1 Rating 4](#_Toc10562522)

[3.1.2 Reviews 5](#_Toc10562523)

[3.1.3 Size 6](#_Toc10562524)

[3.2 Scatterplots 7](#_Toc10562525)

[3.2.1 Rating vs Reviews 7](#_Toc10562526)

[3.2.2 Rating vs Size 8](#_Toc10562527)

[4 Hypothesis Testing 9](#_Toc10562528)

[4.1 Rating vs. Installs 9](#_Toc10562529)

[4.2 Rating vs Content Rating 12](#_Toc10562530)

[5 Modelling 16](#_Toc10562531)

[6 Appendix 17](#_Toc10562532)

[6.1 References 17](#_Toc10562533)

# Introduction

The chosen dataset has to do with Google Play Store applications. It can be found at the following link: <https://www.kaggle.com/lava18/google-play-store-apps>. This dataset was chosen because it contains data that can provide actionable insight on what makes an application successful on this platform. This dataset contains data on around 10,000 Play Store applications which were scraped from the Google Play Store itself. The original dataset contains 13 attributes that describe each application however for the purpose of this assignment only 6 of these were kept. The variables that were used are listed below:

* Rating (Covariate and Dependent variable)
* Reviews (Covariate)
* Size (Covariate)
* Installs (Factor)
* Type (Factor)
* Content\_Rating (Factor)

The variable that is of most interest is *Rating* as it gives the best indication on how successful an app is. The *Reviews* attribute indicates how many reviews (positive or negative ones) an app has. The *Size* variable holds the size in kilobytes for each app. The *Installs* factor is used to indicate how many installs (based on a range) the app has. The *Type* factor indicates if the app is *Free* or *Paid* and the *Content\_Rating* factor indicates for which age group the app is targeted.

# Aims and Objectives

The objective of this assignment was to figure out if there were any correlations between the *Rating* and any of the other variables. This would be useful to identify what makes an application successful on the Google Play Store. Hypothetically it makes sense to assume that an application which is paid should have a higher rating. Moreover, if an application has a large number of installs it also makes sense to expect a higher rating. Also, through the tests the ideal demographical target of an app should be found by finding which factor in the *Content\_Rating* variable has the highest rating. Regarding *size* there are two possibilities, either an application with a large size gets a higher rating due to its better quality or else small sized apps get a higher rating because they do not take up as much of the space on their device (which can often be limited).

# Descriptive Statistics & Illustrations

## Measurements of Location

### Rating

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Rating | Mean | | 3.622 | .0145 |
| 95% Confidence Interval for Mean | Lower Bound | 3.594 |  |
| Upper Bound | 3.651 |  |
| 5% Trimmed Mean | | 3.751 |  |
| Median | | 4.200 |  |
| Variance | | 2.293 |  |
| Std. Deviation | | 1.5142 |  |
| Minimum | | .0 |  |
| Maximum | | 5.0 |  |
| Range | | 5.0 |  |
| Interquartile Range | | .8 |  |
| Skewness | | -1.765 | .024 |
| Kurtosis | | 1.561 | .047 |

The above table contains the measurements of location for the *Rating* covariate. The range, minimum and maximum clearly indicate that this rating ranges from 0 to 5. The average rating is 3.622 which shows that more applications in the dataset have a higher rating. In fact, this can be verified by the median which is 4.200 and by the skewness which is -1.765.

This negative skewness shows that the distribution of ratings is skewed to the right: towards the higher values. The kurtosis value (1.561) shows that people prefer to give either a very high or a very low rating instead of a medium rating. The 5% trimmed mean is 3.751 which shows that there is a higher number of lower rated extreme cases since this trimmed mean is greater than the actual mean. The standard deviation is relatively high considering the small range which shows that the ratings are also quite spread.

### Reviews

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Reviews | Mean | | 444193.87 | 28122.928 |
| 95% Confidence Interval for Mean | Lower Bound | 389067.79 |  |
| Upper Bound | 499319.95 |  |
| 5% Trimmed Mean | | 80483.37 |  |
| Median | | 2094.00 |  |
| Variance | | 8572554850856.179 |  |
| Std. Deviation | | 2927892.561 |  |
| Minimum | | 0 |  |
| Maximum | | 8E+007 |  |
| Range | | 78158306 |  |
| Interquartile Range | | 54760 |  |
| Skewness | | 16.449 | .024 |
| Kurtosis | | 341.029 | .047 |

Immediately it is noticeable that there is a large number of extreme cases within the *Reviews* covariate from the difference between the mean (444193.87) and the 5% trimmed mean (80483.37). The median continues to show the extreme cases because based on the median the average application has 2094 reviews whilst with the 5% trimmed mean the average application has 80483 reviews.

The range, as expected, is very large because there are applications that get no reviews and very popular applications that get millions of reviews from people all around the world. However, the skewness indicates that there are more applications that get few reviews than ones that get many reviews since the skewness value (16.449) is quite high: the distribution is shifted to the left. The Kurtosis (341.029) further amplifies the presence of outliers because it is very high which indicates that most of the values are found on the tails of the distribution curve.

### Size

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Size | Mean | | 18147.60 | 213.003 |
| 95% Confidence Interval for Mean | Lower Bound | 17730.08 |  |
| Upper Bound | 18565.13 |  |
| 5% Trimmed Mean | | 15452.56 |  |
| Median | | 9200.00 |  |
| Variance | | 491768450.859 |  |
| Std. Deviation | | 22175.853 |  |
| Minimum | | 0 |  |
| Maximum | | 100000 |  |
| Range | | 100000 |  |
| Interquartile Range | | 23400 |  |
| Skewness | | 1.704 | .024 |
| Kurtosis | | 2.508 | .047 |

mean size for the applications in this dataset is around 18Mb[[1]](#footnote-2). When the extreme cases are trimmed the average size drops to around 15Mb (5% trimmed mean) which shows that there are more outliers with larger sizes. The median is approximately 9.2Mb which is a better representation of the expected size of an application due to the great number of outliers in this dataset which comes from its relatively large size.

Application sizes vary from less than 1Mb to around 100Mb based on the range. The skewness shows that the distribution of the sizes is shifted to the left since it is positive (1.704) which implies that there are more applications with a small size than large applications. The kurtosis lies at 2.508 which indicates that there is some distribution of sizes along the tails, but it is not too great.

## Scatterplots

The following section will describe how scatterplots were used to visually inspect the data, to see if any relationships between the dependent variable being observed (i.e. *Rating*) and the other covariates (i.e. *Reviews* and *Size*) exist.

### Rating vs Reviews

In this case, the *Rating* variable (on the y-axis) is plotted against the *Reviews* variable (on the x-axis), along with a line of best fit and the output is given as follows:



Figure 1: Scatterplot of Ratings against number of reviews.

The output seen in the figure above suggests that a linear regression model might not be a good fit for the data, since many data points seem to deviate from the line of best fit. In fact, the scatterplot suggests that a quadratic model would be more adequate for the data in question. However, this has yet to be determined when performing regression modelling on the data (see Section 2). It is also of note that data points which have a larger number of reviews seem to be quite sparse when compared to those having much less reviews, which may suggest that they are outliers. Moreover, a lot of variability can be observed in the data when the app has no (or little) reviews. This is because when an app has very few reviews, each one has a lot more weight on the final rating of the app. Hence, a single bad or good review can cause the rating of the app to spike or plummet immediately. Nevertheless, as the number of reviews increases, the range of ratings that the app can have can be seen to decrease, usually lying somewhere in the range between 4 and 5.

### Rating vs Size

In this case, the *Rating* response variable (on the y-axis) is plotted against the *Size* variable (on the x-axis) along with a line of best fit, to check for any relationships between the two variables. The output is given as follows:



Figure 2 Scatterplot of Rating against Size (in megabytes).

As in the previous case, the above scatterplot also suggests that a linear regression model would not fit the data well, given that more points than before seem to deviate from the line of best fit. Moreover, just as before, this scatterplot also seems to show a quadratic relationship between the variables. However, as one might expect, the correlation between the two variables seems to be far less strong, which is made obvious by the fact that the data points are much more scattered when compared to the data points in the previous scatterplot. Yet, it can still be observed that as the file size of the application increases, the ratings seem to reduce down to a smaller range around the larger ratings, similarly to the previous scatterplot. In addition, it can also be seen that there is a large variability in size for applications with a low rating. Though there does not seem a very clear reason why this would be the case, one possible cause would be lack of correlation between the variables due to reasons such as inflated file sizes, or limited storage capacity on devices making it impossible for users to download the app etc.

# Hypothesis Testing

In this section, parametric/non-parametric tests are used to see if any of the fixed factor variables (i.e. *Installs*, *Type*, and *Content\_Rating*) have any significant impact on the mean (or median) of the variable of interest (i.e. the *Rating* variable).

## Rating vs. Installs

For the first test, the effect of the *Installs* variable on the mean (or median) of the *Rating* variable is tested. Since the *Installs* variable has five categories, all of which are independent from each other, only the *One-Way ANOVA* or *Kruskal Wallis* test could be used for this purpose; to determine which one to use, the data is tested to see if it respects the assumptions of the *One-Way ANOVA* test, which include the following:

1. For each group, the response variable must be normally distributed.
2. The variances of all groups must be equal.

The first assumption is tested using the *Kolmogorov-Smirnov* and *Shapiro-Wilk* tests, both of which test the following hypotheses:

**H0:** *Rating* follows a normal distribution for the given group of the *Installs variable*

**H1:** *Rating* does not follow a normal distribution for the given group of the *Installs variable*

The outputs of the tests were computed in SPSS and can be seen below:



Figure 3 Outputs of Normality tests for each different ‘Installs’ category.

For every category, the p-value of both tests is zero, which is far less than the level of significance (which is 0.05). Hence, the null-hypothesis is rejected for all cases, and the response variable does not follow a normal distribution for any category. This can also be confirmed by looking at the *Q-Q plot charts*, which shows that the data points for each category deviate a lot from the expected normal distribution:

Given that one of the assumptions of the *One-Way ANOVA* test is not upheld, the non-parametric version of the test (i.e. Kruskal Wallis) must be used to check the influence of the *Installs* variable on the *Rating*. In this case, the following hypotheses are tested:

Figure 4 Q-Q plots of all the 'Install' categories



**H0:** The median values of *Rating* are the same for all categories of *Installs*

**H1:** The median values of *Rating* are different for the categories of *Installs*

The test is performed in SPSS and it gives the following outputs:



Figure 5 Outputs of the Kruskal Wallis test

As be seen above, the p-value of the test, which is zero, is much less than the level of significance (0.05). Hence, the null-hypothesis is rejected, and the median *Rating* of the different groups are not the same. Now, since the medians of the groups are different from each other, a Post-Hoc analysis is conducted to verify where these differences lie. Each pairwise comparison consists of a *Mann-Whitney* test which tests the following hypotheses:

**H0:** The median ratings of the *Rating* of the two groups are the same.

**H1:** The median ratings of the *Rating* of the two groups are different.

The test is conducted in SPSS and the following output is obtained:



Figure 6 Post-hoc pairwise comparisons of each 'Install' category

Each of the tests conducted above can be seen to have a p-value far below the level of significance (0.05). Hence, for each pairwise comparison, the null-hypothesis is rejected, and each interval of the *Installs* variable has a distinct median value of the *Rating* variable.

Next, to confirm the result found above, the sample mean and median *Rating* of each group were computed:



Figure 7 Sample means and medians of each category in 'Installs'

The output above confirms the results found using the Post-hoc analysis, suggesting that as the number of installs increases, the rating of the application also increases.

## Rating vs Content Rating

In the second test, the effect of the *Content Rating* variable on the mean (or median) of the *Rating* variable is tested. Similarly to the previous test, since the *Content Rating* variable has four independent categories, either the *One-Way ANOVA* or *Kruskal Wallis* test should be used; the choice between the two is once again made by checking if the data fits the assumptions necessary to use the *One-Way ANOVA* test.

To check if the *Rating* variable is normal for each group in *Content Rating*, the *Kolmogorov-Smirnov* and *Shapiro-Wilk* tests are used to verify the following hypothesis:

**H0:** *Rating* follows a normal distribution for the given group of the *Installs variable*

**H1:** *Rating* does not follow a normal distribution for the given group of the *Installs variable*

The output of the tests given by SPSS can be seen below:



Figure 8 Output of normality tests for each group of the 'Content Rating' variable

The output above shows that none of the categories have a p-value above the level of significance (0.05). Therefore, the null-hypothesis is rejected, and the *Rating* variable does not follow a normal distribution for any of the *Content Rating* variable’s groups. To confirm this result, the *Q-Q plot* charts for the different categories were also created:

In the images above, the data points for all groups can be seen to deviate from the expected normal distribution, hence the result given by the *Kolmogorov-Smirnov* and *Shapiro-Wilk* tests is confirmed.

Figure 9 Q-Q plot charts for the different categories of 'Content Rating'



Hence, since one of the assumptions necessary for the *One-Way ANOVA* test is not upheld, the *Kruskal Wallis test* must be used to test if there is any significant difference between the medians of the different dependent variables. The *Kruskal Wallis test* verifies the following hypotheses:

**H0:** The median values of *Rating* are the same for all categories of *Content Rating*

**H1:** The median values of *Rating* are different for the categories of *Content Rating*

The output for the test can be seen in the image below:



Figure 10 Outputs of Kruskal Wallis test

Since the test gives a p-value of zero, at a level of significance of 0.05 the null-hypothesis is rejected. Thus, the median *Rating* is different for the categories of the *Content Rating* variable. To check where the discrepancies in *Rating* lie specifically, a Post-hoc analysis is performed, and the medians of each group were compared in a pairwise manner using the *Mann-Whitney* test. For each pair of groups, the following hypothesis were tested:

**H0:** The median *Rating* of the two groups are the same.

**H1:** The median *Rating* of the two groups are different.

When the tests are performed in SPSS, the following output is given:

The output above shows that at a 0.05 level of significance, only the tests considering the difference between the ‘*10+’* category and some other category reject the null-hypothesis. Hence, only the median of the ‘*10+’* category seems to have any significant difference from the other variables. On the other hand, the categories ‘*Everyone*’, ‘*Teen’* and ‘*17+*’ all seem to have the same median.

Figure 11 Output of the Post-Hoc analysis for each group in 'Content Rating'



To confirm these findings, the sample means and medians for all the independent groups have been calculated in SPSS, and can be seen in the image below:



Figure 12 Means and medians of all the categories in 'Content Rating'

The output shown in the above table seems to confirm the results of the Post-Hoc analysis; for the means and medians of the ‘*Everyone*’, ‘*Teen’* and ‘*17+*’ categories all appear to be similar to each other. Moreover, the mean and median of the ‘10+’ category is in fact higher than the other categories, further confirming the results found previously. In addition, this output also suggests that mobile applications targeting to this demographic may achieve a higher rating than if it were to target any other demographic of phone users.

## Rating vs. Type

In the third and final test, the data was to be verified to see if there is any difference in the mean (or median) *Rating* of paid and free apps, where ‘paid’ and ‘free’ are categories in the fixed factor variable called *Type*. Given that *Type* has only two independent categories, this could be done using either the *Independent Samples T-test* or the *Mann-Whitney* test; to determine which one was needed, the data was verified to see if it upheld the assumptions of the *Independent Samples T-test*, which include the following:

1. Both samples must come from normal populations.
2. Both populations must have equal variances.

The first assumption was verified by performing the *Kolmogorov-Smirnov* and *Shapiro-Wilk* tests on the *Rating* variable for both ‘paid’ and ‘free’ apps. The following hypotheses were tested:

**H0:** *Rating* follows a normal distribution for the given type of app (free or paid)

**H1:** *Rating* does not follow a normal distribution for the given type of app (free or paid)

The results are computed in SPSS and can be seen in the following table:

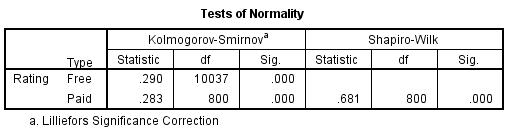
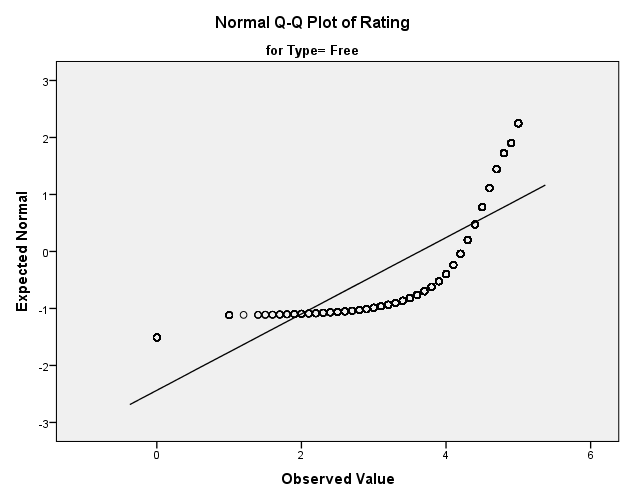
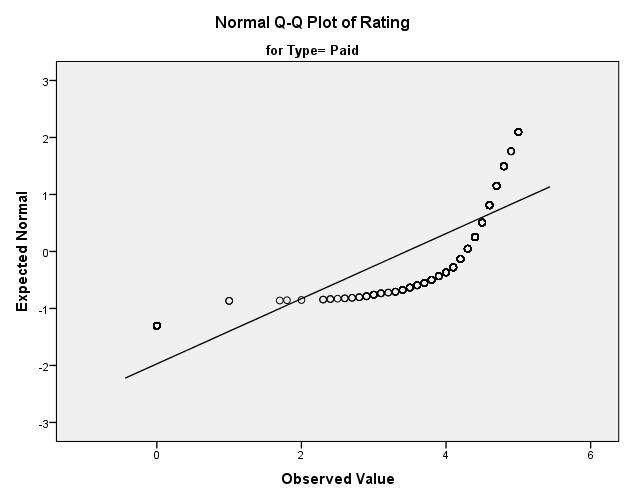


Figure 13 Outputs of Normality tests for 'paid' and 'free' apps

In this case we can observe that both sets of p-values are significantly less than the level of significance (0.05). Hence, the null-hypothesis is rejected, and the *Ratings* variable does not follow a normal distribution for ‘free’ and ‘paid’ apps. To confirm this result, the *Q-Q plot charts* of the two categories were also created:

Figure 14 Q-Q plot charts for free and paid apps.



As can be seen in the figure above, the data points deviate significantly from the expected normal distribution, confirming the results found by the *Kolmogorov-Smirnov* and *Shapiro-Wilk* tests.

Therefore, since one of the assumptions of the *Independent Samples T-test* is not upheld, to test if there is any difference between the medians of the two groups, the *Mann-Whitney* test must be used. The *Mann-Whitney* test checks for the following hypotheses:

**H0:** The median *Rating* of free and paid apps are the same.

**H1:** The median *Rating* of free and paid apps are different.

The results of the test are computed in SPSS, and are given as follows:

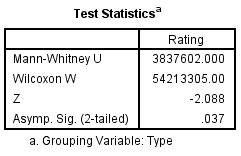


Figure 15 Outputs for the Mann-Whitney test

Since the p-value (0.037) of the test is less than the level of significance (0.05), the null-hypothesis is rejected, and the median *Rating* of free and paid applications is not the same.

To verify the results of the hypothesis test, the sample means and medians of free and paid apps were calculated in SPSS:

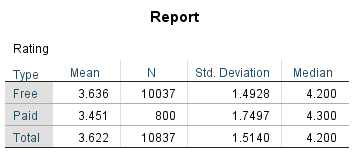


Figure 16 Table containing sample mean and median of paid and free apps.

Comparing the two medians, it is apparent that they are not the same, and that the median of paid apps is larger, confirming the result found by the *Mann-Whitney* test. However, when comparing the sample means, the mean of free apps is the highest, which contradicts the results of the test as well as the sample medians. As such, the output does not make it clear if a free or paid application is preferable to obtain a higher rating.

# Modelling

# Appendix

## References

M. B. Inguanez, F. Sammut, D. Suda , *Statistical analysis using SPSS and R software*, pages 108-111

1. Results are in Kb [↑](#footnote-ref-2)