Video Games Sales and Ratings Analysis

Rohith Narra

12/5/2019

Contents

1 Introduction

The gaming industry and community have grown quite a bit since a decade. In fact, according to a report by WePC, the video games market is expected to be worth over 90 billion U.S dollars by 2020 from nearly 78.61 billion in 2017. It has evolved as the biggest form of entertainment. It is bigger than it ever was, bigger than every other form of entertainment. This industry is witnessing a growth of 10.7% over the last decade, as opposed to other entertainment services like Television whose revenue fell by 8%. According to the market research firm Super Data, as of 2017, By region, Asia accounted for 71.4 billion dollars, North America for 25.4 billion dollars, Europe for 23 billion dollars and South America for 6.5 billion dollars. There are different publisher, platforms and genres for video games. This project relates to the sales of these video games based on different genres, publishers and platforms.

1.1 Purpose of study

Video games have evolved over the years to offer players almost any type of experience. According to the European Mobile market, there are more than 2.5 billion video gamers in the world. The adoration these people have for video games is infinite. One of the most addictive things the gamers do is to read about video games, especially the interesting facts behind some of their favourite games, or gaming publishers, or game developers. This study focuses on sales for different publishers, platforms and genres which would give the basic idea about the most popular genres, publishers and platforms amongst all. Also analyzing the effect of genres on sales.

1.2 Statement of the Problem

The purpose of the study is to explore the following questions:
1. Whether the global sales of the games released on the "PS2" platform and "Wii" platforms are equal.
2. According to the published articles and blogs, it is said that the average critic score of the video games is 70. Is that true?
3. Is it true that half of the games released by Electronic Arts publishers are sports genre?
4. Did Activision and Nintendo publishers released the same number of racing games between 1980 and 2011?
5. Is it true that the number of games released in 2008 with Platform, Fighting, Strategy genres are equal?

1.3 Dataset

This data is an extended version of Gregory Smith's web scrape of VGChartz Video Games Sales which contains the list of video games with sales. The sales data of over 1700 videogames releases are recorded in the data. Because the scraping algorithm focused on the sales of game units, the data is not the representation of the modern gaming industry. This data also favours consoles more since most of the big PC games are platform independent

1.3.1 Variables

1.3.1.1 Name

A categorical variable which contains the name of the Video game release.

1.3.1.2 Platform

A categorical variable which contains the name of the platform on which the game is released. The breakdown consisted of 31 unique values with Platform named "DS" as more frequent value with "2251" occurrences and "GG" as a less frequent value which has occurred only once in the data set

1.3.1.3 Year_of_Release

A categorical variable which contains the year of release of the video game. The breakdown consisted of 43 unique years ranging from 1976 to 2017 with "2009" as a most frequently occurred year with 1550 occurrences and 1976 as a less frequent year which has occurred only once in the dataset.

1.3.1.4 Genre

A categorical variable which contains the name of the genre of the game released. The breakdown consisted of 12 unique values with "Action" as the most frequently occurred genre with 3503 occurrences and puzzle as the less frequently occurred genre with 615 occurrences.

1.3.1.5 Publisher

A categorical variable which contains the name of the publisher of the game released. The breakdown consisted of 627 unique values with "Ubisoft" as the more frequently occurred publisher with 941 occurrences.

1.3.1.6 NA_Sales

A continuous variable which contains the sales of NorthAmerican region in millions. It is observed that the variable ranges from 0 to 41.36 million with a mean of 0.2545 million and a median of 0.700 million.

1.3.1.7 EU_Sales

A continuous variable which contains the sales of Europe region in millions. It is observed that the variable ranges from 0 to 28.96 million with a mean of 0.1407 million and a median of 0.0200 million.

1.3.1.8 JP_Sales

A continuous variable which contains the sales of Japan region in millions. It is observed that the variable ranges from 0 to 10.22 million with a mean of 0.07502 million and a median of 0 million.

1.3.1.9 Other_Sales

A continuous variable which contains the sales of regions other than North America, Europe and Japan in millions. It is observed that the variable ranges from 0 to 10.57 million with a mean of 0.4591 million and a median of 0.01 million.

1.3.1.10 Global_Sales

A continuous variable which contains the global sales in millions. It is observed that the variable ranges from 0 to 82.54 million with a mean of 0.5165 million and a median of 0.1600 million.

1.3.1.11 Critic_Score

A continuous variable which contains the critic scores of the video games released. It is observed that the variable ranges from 13 to 98 with a mean of 68.91 and a median of 71.

1.3.1.12 Critic_Count

A continuous variable which contains the critic count of the video games released. It is observed that the variable ranges from 3 to 113 with a mean of 26.19 and a median of 21.

1.3.1.13 User_Score

A continuous variable which contains the user scores of the video games released. It is observed that the variable ranges from 0 to 9.7 with a mean of 7.117 and a median of 7.500.

1.3.1.14 User_Count

A continuous variable which contains the user count of the video games released. It is observed that the variable ranges from 10766 to 4 with a mean of 162.7 and a median of 25.

1.3.1.15 Rating

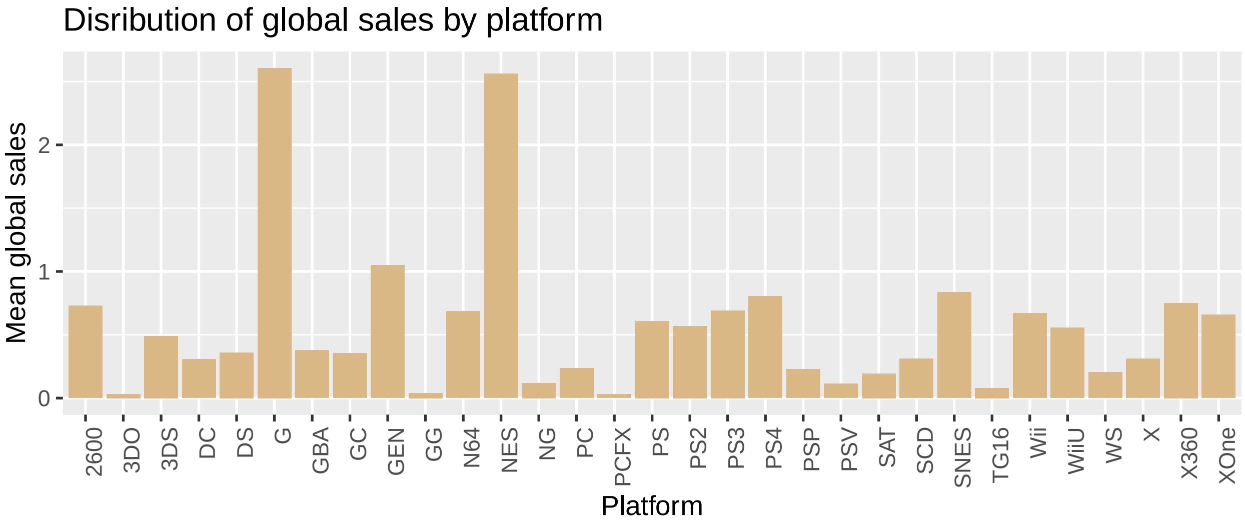
A categorical variable which contains the rating of which the game is released. The breakdown consisted of 8 unique values with "E" as more frequent value with "4120" occurrences and "AO" as a less frequent value which has occurred only once in the data set

2 Exploratory Data Analysis

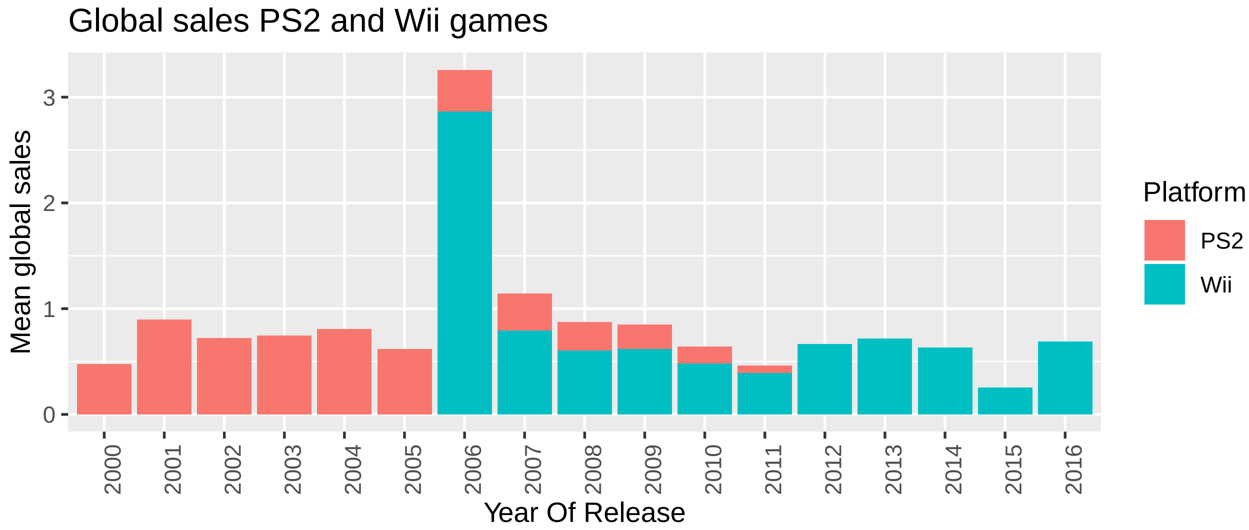
The findings of this study will be presented in four sections according to the genre, platform, publisher and critic scores.

2.1 Global sales by platform.

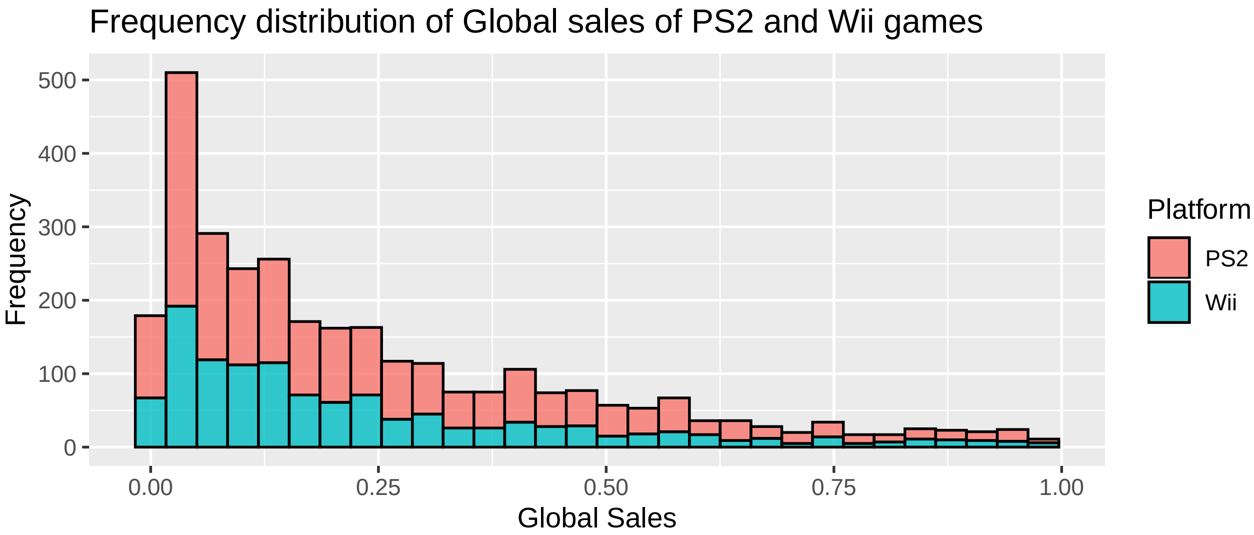
The global sales of the video games are analysed individually based on the platform they are released and the results are represented graphically in the following figure.



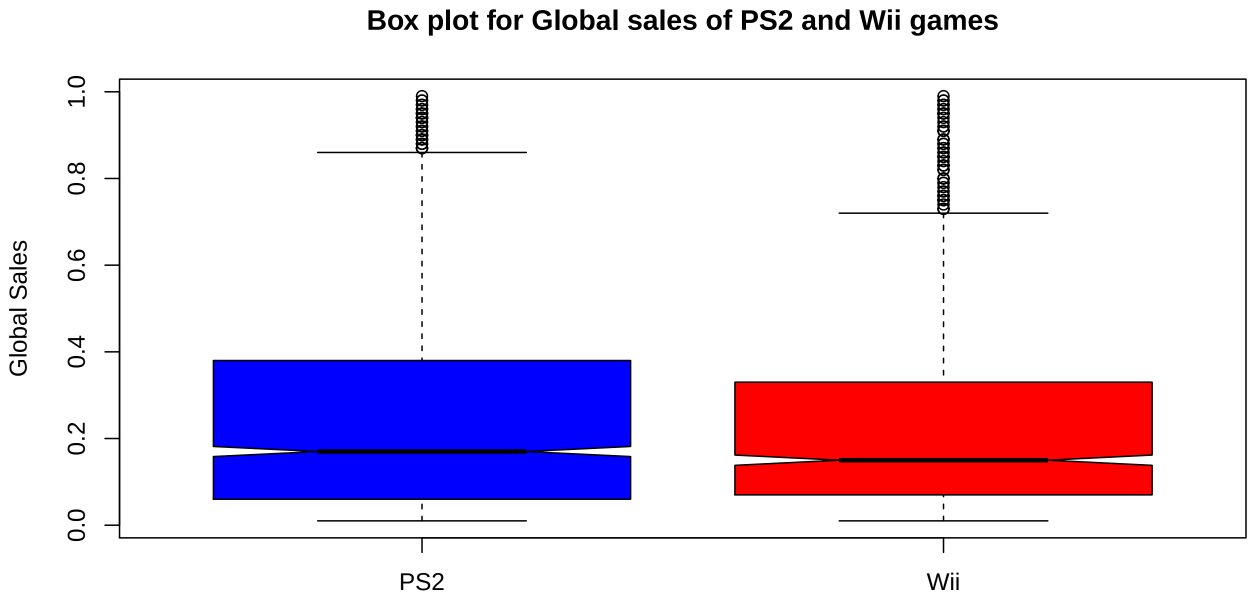
From the above graph, it can be noticed that the sales of the video games released on the G and the NES platforms are higher than games released on other platforms. As this study focuses on the sales of PS2 and Wii games, having closer attention on the sales of the games released on PS2 and Wii platforms it is observed that sales of the games of two these categories are almost equal. This can be visualized clearly by plotting the mean global sales of only the PS2 and Wii games



The above figure presents the mean global sales of PS2 and Wii games individually according to the year of release. It is observed that the popularity of the PS2 video games has decreased over the years. On the other hand, the popularity of Wii games has increased over the years.



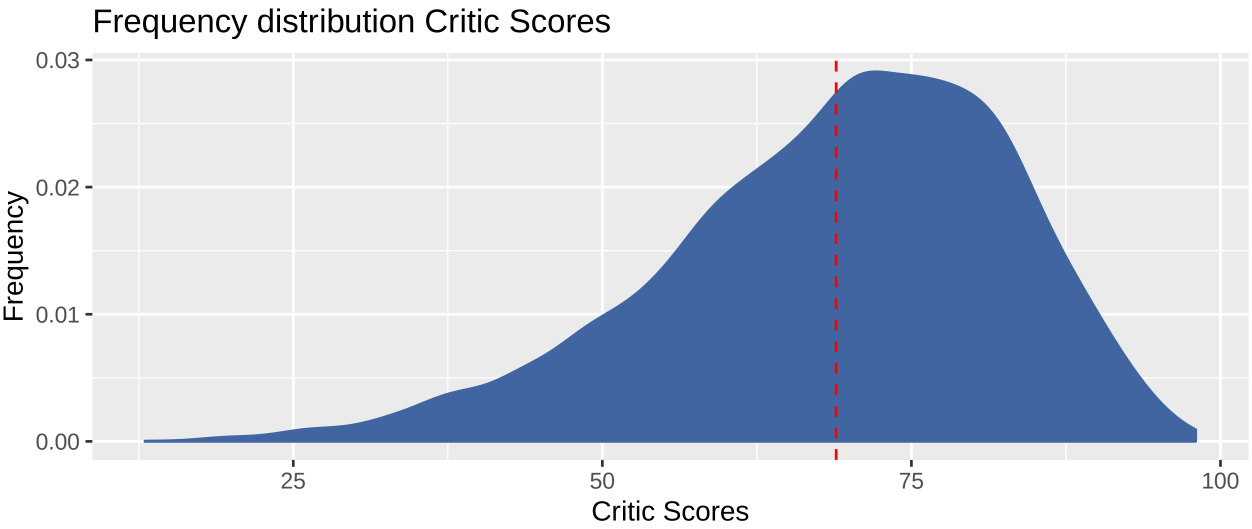
From the frequency distribution of global sales of PS2 and Wii games It can be noticed that the data is more skewed towards left. This might be the indication that there are outlies in the data. This can be understood clearly by plotting whisker graph.



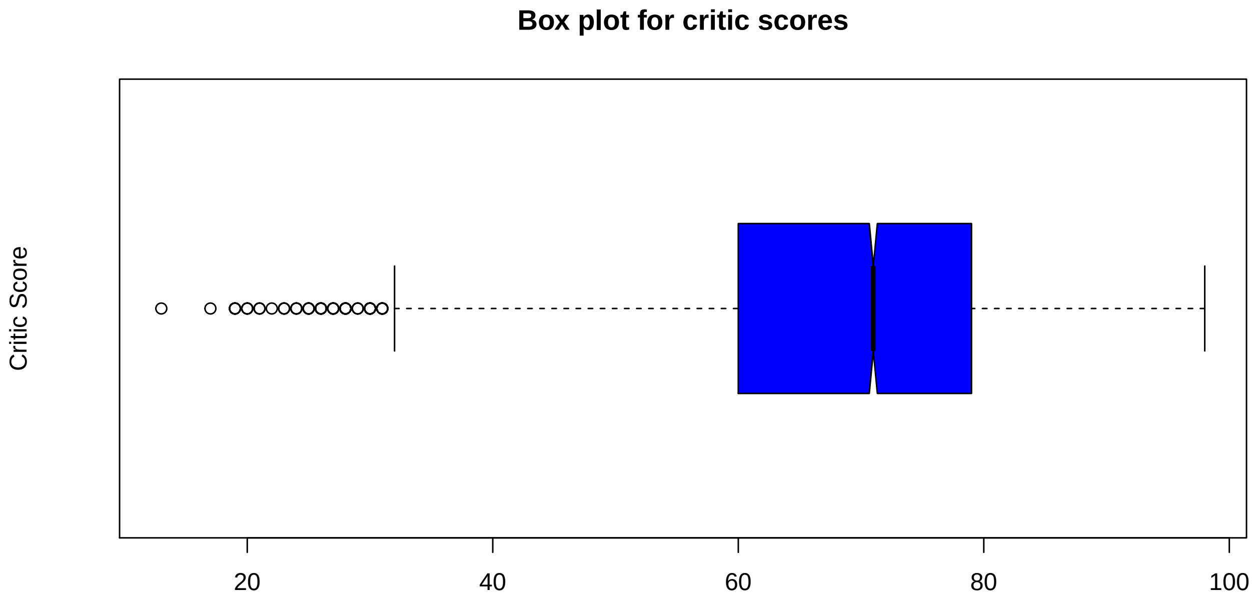
As mentioned above it is clear from the whisker plot that both the categories contain certain outliers, which are having a significant effect on the mean global sales of the video games of both the categories. This is an indication that there is a great amount of variation in the global sales of both the products.

2.2 Critic Score Analysis

The critic scores of the video games released from 1976 to 2017 are analyzed and the results are presented graphically.



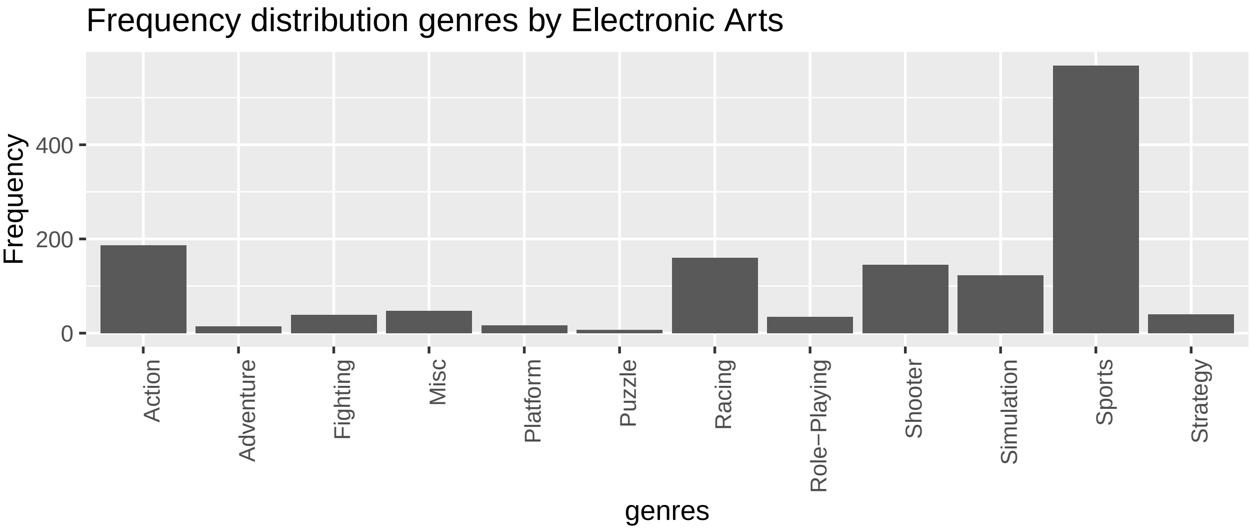
From the frequency distribution graph it can be noticed that the data is more skewed to right which indicates that there are outlies in the data. The line in the graph represents the mean critic score of the sample data. As, the data is skewed to the right we can't relay on the mean for our estimates. The spread of the data can be clearly understood by plotting whisker graph.



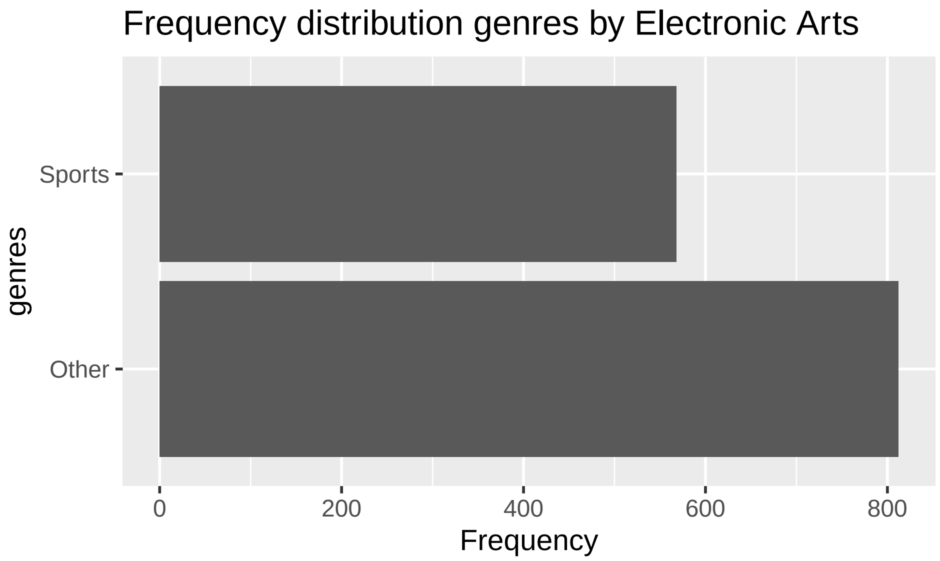
As mentioned, it is clear from the whisker plot that critic scores contain certain outliers, which are having a significant effect on the mean which indicates that there is a great amount of variation in the critic scores of the games. Having a closer look wIt can be observed that the average critic score is around 70. Further inferences to this claim can be made with statistical evidence by conducting hypothesis testing.

2.3 Games published by Electronic Arts

The games published by the Electronic Arts are analysed individually and the results are represented graphically in the following figure.



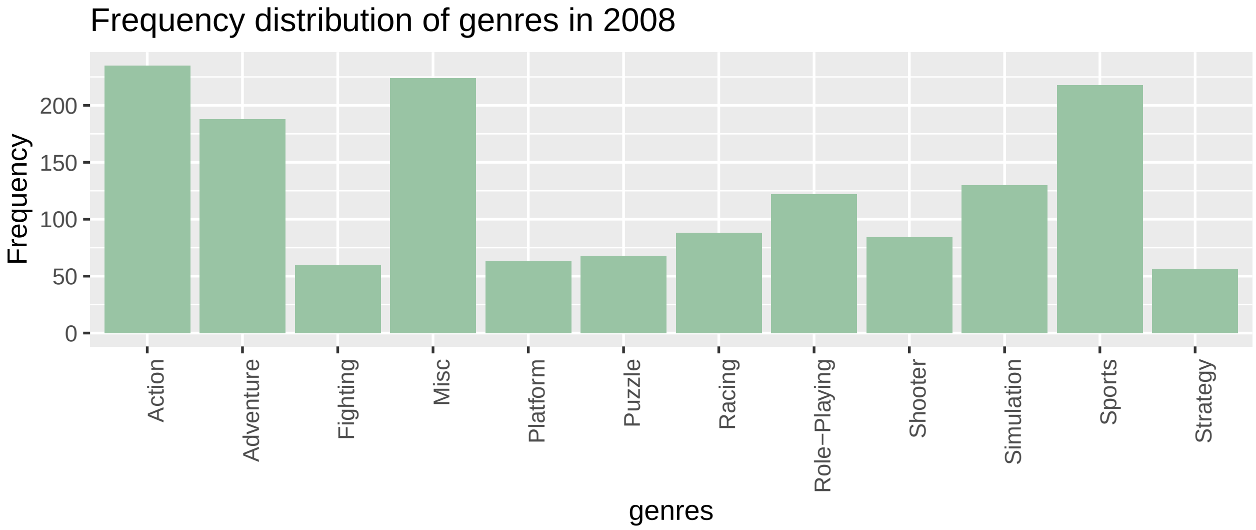
From the above graph, it can be noticed that the Electronic arts focus majorly on the games with sports theme. It is barely focusing on the games with Adventure, Platform and Puzzle genre. The true proportion of the games with a sports theme can be visualized clearly by combining all the categories excluding sports as a single entity.



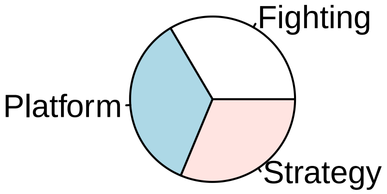
This graph elucidates the proportion of the sports genre games released by Electronic Arts. As we can observe that almost 40% of the releases from the Electronic arts is made up by the sports genre. Further inferences to this claim can be made with statistical evidence by conducting hypothesis testing.

2.4 Analysis of the games released in 2008

This part of analysis contemplates the genres released int the year of 2008.



The above graph depicts information about the genres released in 2008. It is noticed from the graph that Action, Misc and Sports were the major contributors, whereas, the contribution of games with fighting, Platform and Strategy genres is less. It is also observed that the proportion of games released with these genres are equal. This can be visualized clearly by plotting a pie graph with only fighting, Platform and Strategy genres.



The above pie chart strengthens the claim that the proportion of video games released with fighting, strategy and platform theme are equal. Further inferences can be made to this claim with statistical evidence by conducting hypothesis testing.

3 Statistical Analysis

The analysis till this point describes the sample dataset we had, but it doesn't give any information about the population. This section makes the inference on the population. Given the objectives of the study, analysis in this section is broken down into five parts.

3.1 Global Sales of PS2 and Wii Platforms

From the exploratory data analysis, it was observed that the global sales of the video games released on the "PS2" platform and "Wii" platform are almost equal. To make further inferences to this claim with statistical evidence we have to conduct a hypothesis test. The question of interest for this hypothesis test would be whether the global sales of video games released on PS2 and Wii platforms are same?

3.1.1 Selection of statistical test

From the question of interest, the goal is to compare the global sales of the video games released on "PS2" and "Wii" platforms. As the comparison is made between the samples from two different populations and we don't have any information about the population parameter, the test that needs to be selected is two-sample t-test.

3.1.2 Population parameter

As the comparison is made between the global sales of two different categories and it is clear from the sample data that global sales variable is a continuous variable, so the population parameter should be taken as the difference of means.

$$\mu_p - \mu_w$$

Where μ_p and μ_w are the population mean global sales of PS2 and Wii games respectively.

3.1.3 Null Hypothesis

Assumed null hypothesis is the mean global sales of the games released on the PS2 platforms is equal to the mean global sales of the games released on Wii platforms, which implies, the difference between the mean global sales of games released on PS2 and Wii platforms should be zero.

$$H_0 : \mu_p - \mu_w = 0$$

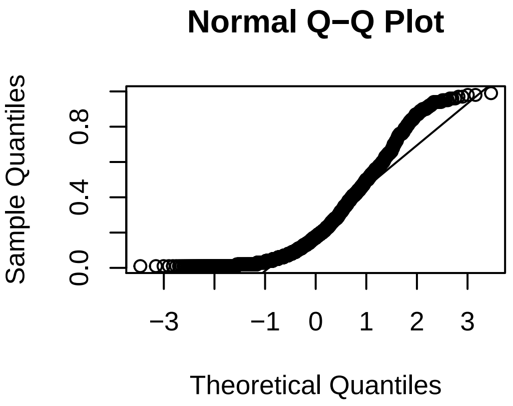
3.1.4 Alternate Hypothesis

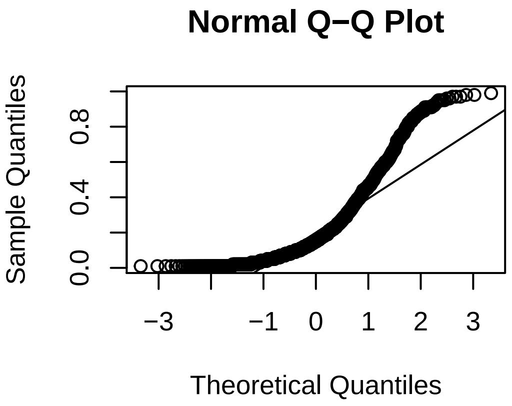
In contrast to the null hypothesis, the alternate hypothesis is assumed as the mean global sales of the games released on the PS2 platforms and Wii platforms are not equal. Which implies, the difference between then the mean global sales of PS2 games and Wii games should not be zero

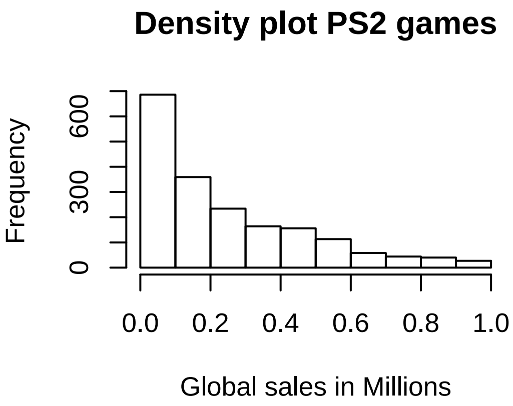
$$H_a : \mu_p - \mu_w \neq 0$$

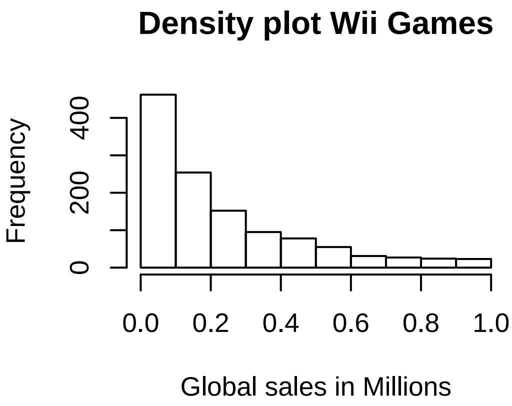
3.1.5 Conditions for applying the test

For appling the the two sided two sample t-test the sample data should satisfy the following conditions.
i. Question of interest has to do with the difference of means between two populations
As per null hypothesis we are intrested in calculating the difference in the mean global sales of PS2 and Wii games.
ii. Two independent samples from two populations
Yes, this analysis considers the PS2 games and Wii games to be different populations.
iii. The population data must be normally distributed
This condition can be verified by plotting QQplot of the sample data.







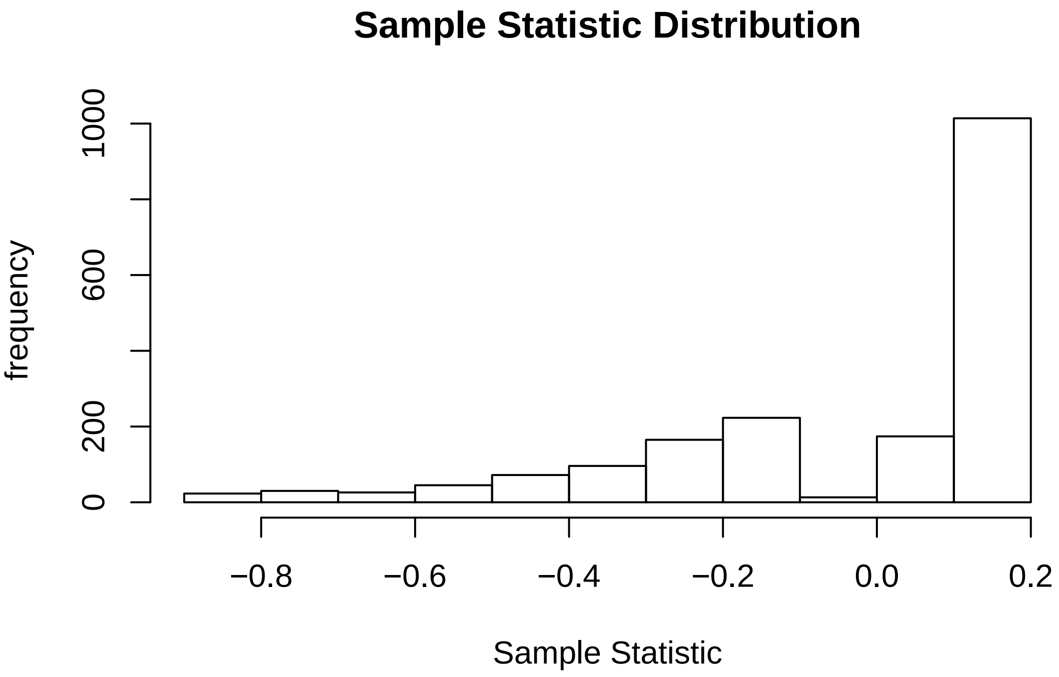


It can be observed from the above graphs that the data is exactly not aligned to the $x = y$ line. It might be happening because the data is skewed to the left. As the normality of the data is not assured we can say that the test will not be optimally powerful and the results are not completely reliable. In this case, the t-test will not perform as it should - i.e. that if the null hypothesis is true, it will falsely reject the null 5% (considered $\alpha = 0.05$) of the time.

3.1.6 Sample Statistic

With respect to the population parameter the sample statistic would be,

$$\bar{x}_p - \bar{x}_w$$



From the above graph, It can be observed that the distribution of sample static is skewed to the right, which might be an indication that the platform of the game will have an effect on global sales.

3.1.7 Test statistic

Adding to the testing statergy selected, the alternate hypothesis states that the difference in the means is not equals zero, so the two sided testing statergy is to be applied. Which mean, the test statistic should be

$$t_{min(n_p-1,n_w-1)} = \frac{(\bar{x}_p - \bar{x}_w) - (\mu_p - \mu_w)}{\sqrt{\frac{s_p^2}{n_p} + \frac{s_w^2}{n_w}}} \sim t_{min(n_p-1,n_w-1)}$$

3.1.8 Traditional Approach

3.1.8.1 Test Results

Before conducting the test the test significance values α needs to be chosen. In this case, it is considered the test significance $\alpha = 0.05$, so the confidence intervals for the test would be 95%.

From the sample means $\bar{x}_p$, $\bar{x}_w$ and the sample standard deviations s_p, s_w the test statastic values can be computed easily. The t-stat is calculated as 1.581814

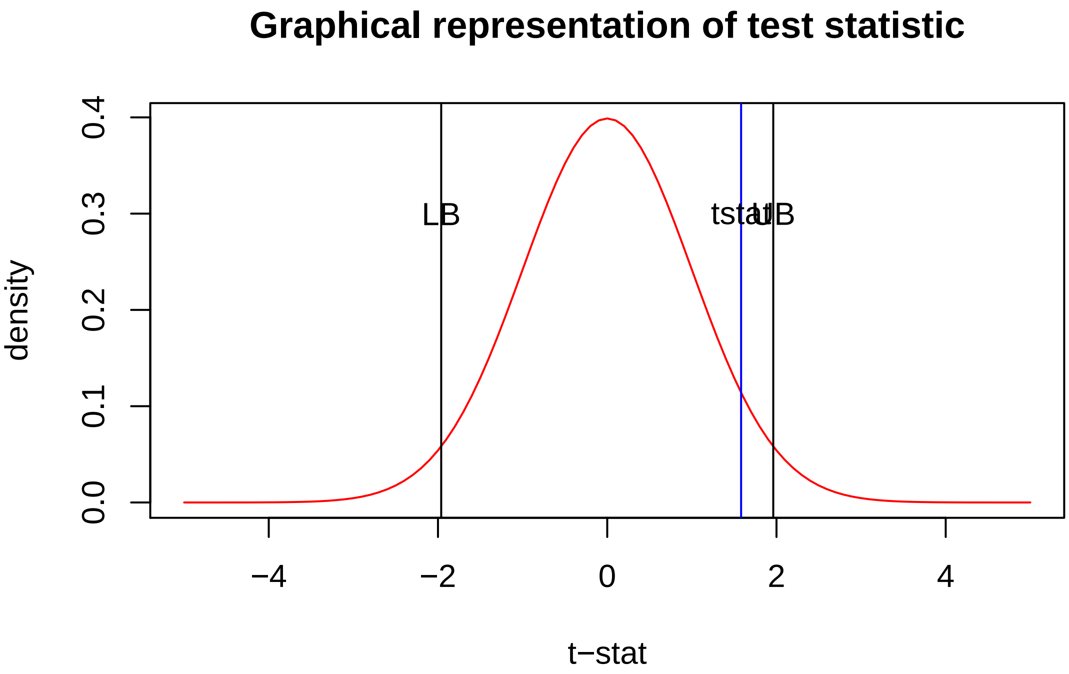
Under the null hypothesis, the test statistic has a t distribution with $min(n_p - 1, n_w - 1)$ degrees of freedom, which is 1200. To complete the test, p-value of the test needs to be calculated. Since this is a two-sided test with not equals alternative, The area under the right and left tails of the t-distribution with 1200 degrees of freedom needs to be considered.

$$P(t_{1200} > 1.581814) \times 2$$

The p-value is calculated as 0.1139556. As the p-value of the test calculated, which is greater than our significance level (0.05), so we fail to reject the null hypothesis.

3.1.8.2 Building confidence intervals

The one piece of information we are missing is the critical value $t_{1200}(\alpha/2) = t_{1200}(0.025)$, which is required for building the confidence intervals. This can be calculated by finding the quantile wherein the probability is 0.025 in the t distribution, which is calculated as -1.961943



The graph aboove describes the spread of confidence intervals, black lines represens the t-critical values and the blue line represents our actual t-stat value. The 95% Confidence Intervals for the difference in mean global sales of PS2 and Wii games is calculated by the following formula,

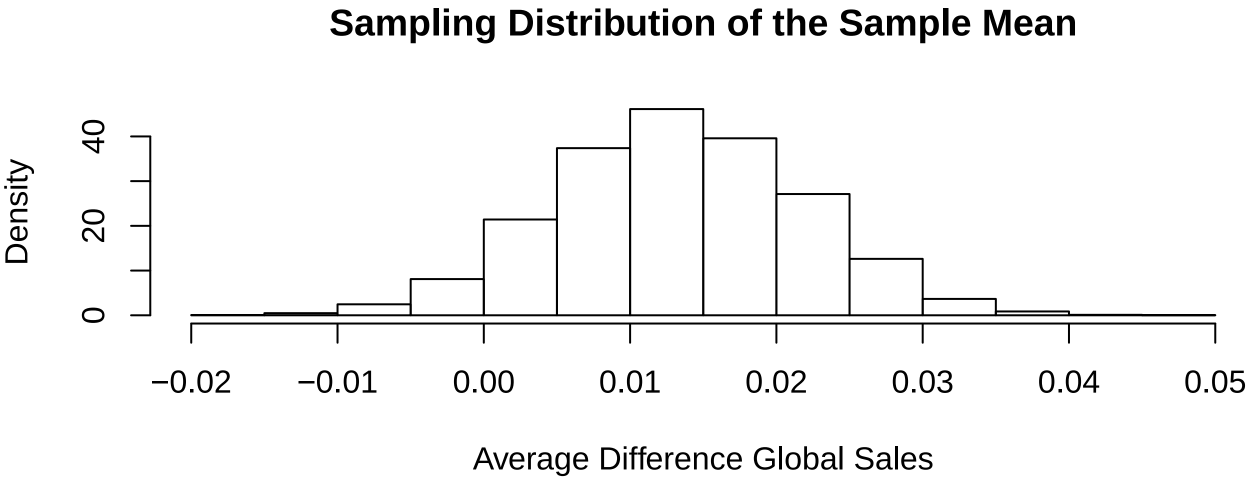
$$(\bar{x}_p - \bar{x}_w) \pm t_{1200}(0.025)\sqrt{\frac{s_p^2}{n_p} + \frac{s_w^2}{n_w}}$$

The calculated confidence intervals are -0.003223622 and 0.03005225. As the 95% Confidence interval extends from approximately -0.003223622 to 0.03005225 it can be observed that the null hypothesis value 0 is present in the calculated confidence intervals which strengthen our claim that we fail to reject the null hypothesis.

3.1.9 Bootstrap Approach

3.1.9.1 Bootstrap confidence intervals

For performing the bootstrap two sample t-test test a random samples from the two groups with the same size is to be taken from the original sample with replacement and the means are computed on these samples. The distribution of difference in means is to be plotted by performing this finate numberof times.



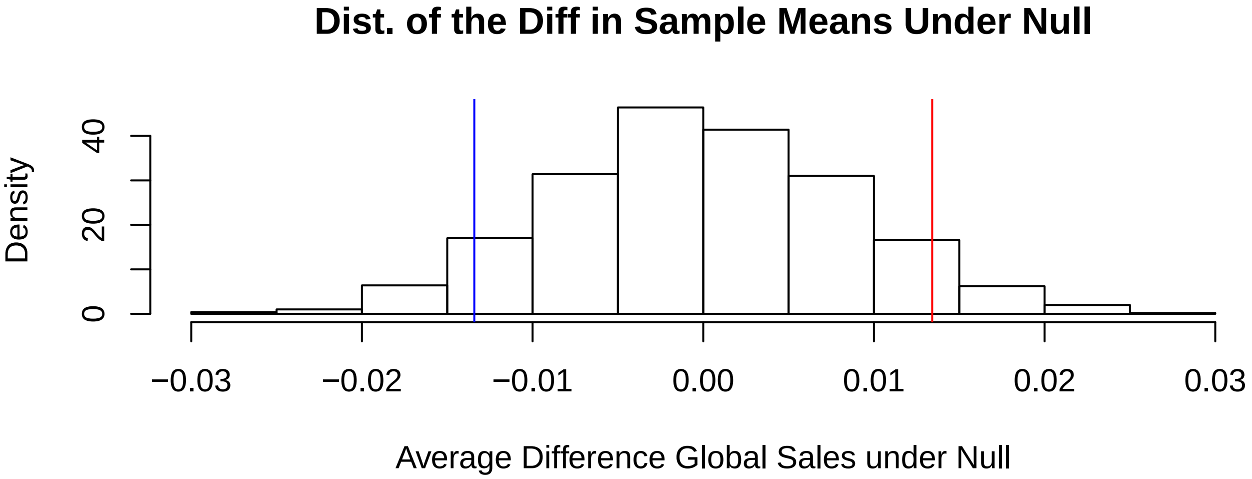
As the calculation of sample means is simulated for 1000 times it can be observed that the distribution has become more symmetric. The confidence intervals are calculated by finding the 0.025 and 0.975 quantiles on the sampling distribution of the sample mean which are calculated as -0.0030268 and 0.0297171.

3.1.9.2 Bootstrap p-value

From the definition of the p-value: The p-value is the probability that we observe a test statistic as or more unusual than the one we observed, given that the null hypothesis is true.

If the null hypothesis is true, then the true mean difference between the two groups is 0. We need to shift our distribution so that this is true or we can create a randomization distribution

In the case of randamization distribution, we have to simulate many samples assuming the null hypothesis is true. As the Null hypothesis is true, there is no relationship between two variables.
1. Create many samples were the treatment group is shuffled (randomized).
2. Compute the difference in groups for each of the samples.
3. Create a histogram of the randomized statistics. This is an approximation of the null distribution.



The lines in the graphs represent the difference in the mean global sales calculated from the sample. The p-value is calculated as the counts of values more extreme than the test statistic in our original sample, given the null hypothesis is true, which is calculated as 0.111.

As the bootstrap p-value of the test is known, which is greater than the significance level (0.05), so we fail to reject the null hypothesis.

3.1.10 Interpretation

There is no evidence (p-value = 0.1139556) to suggest that the true population mean global sales of video games hosted on the PS2 platforms is different from the video games hosted on the Wii platforms. We fail to reject the null hypothesis that the true mean global sales of the PS2 video games is equal to the mean global sales of the Wii video games. With 95% confidence, the true difference in the mean global sales of PS2 and Wii video games is between -0.003223622 and 0.03005225. The null hypothesized difference between mean global sales is zero and zero is in the 95% confidence interval which agrees with our failure to reject the null hypothesis.

3.2 Critic Score Analysis

Every time the video game is released, Critics release articles providing their opinions on the video game. After playing the video game and criticizing it, everyone has a general agreement on its performance. So, these scores will have a significant impact on the sales of the game. According to the published articles and searched queries it is said that the average critic score of the video games is 70. Is that true? this part of the analysis focuses on proving the theoretical average critic score with statistical evidence.

3.2.1 Selection of statistical test

From the selected question, the goal is to compare the statistical average critic score with the theoretical value. As the comparison involves only one sample and we don't have any information about the population parameter, the test that needs to be selected is one-sample t-test.

3.2.2 Population parameter

As the comparison is made between statistical critic score with theoretical value and it is clear from the sample data that critic score is a continuous variable, so the population parameter should be taken as the mean critic score.

$$\mu_c$$

3.2.3 Null Hypothesis

The selected question says that "Weather average critic score of video games is 70?". So the null hypothesis can be assumed as the mean critic score of the video games is equal to 70.

$$H_0 : \mu_c = 70$$

Where μ_c is the population mean critic score of video games.

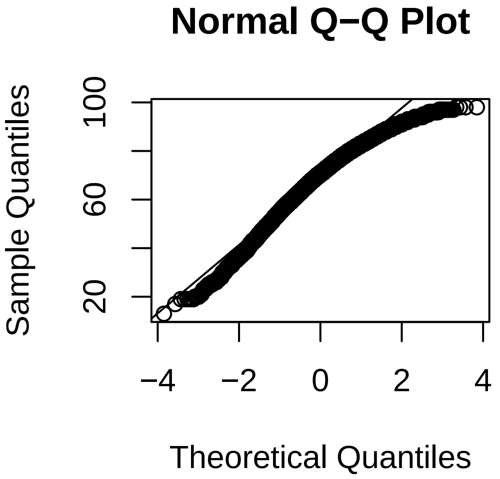
3.2.4 Alternate Hypothesis

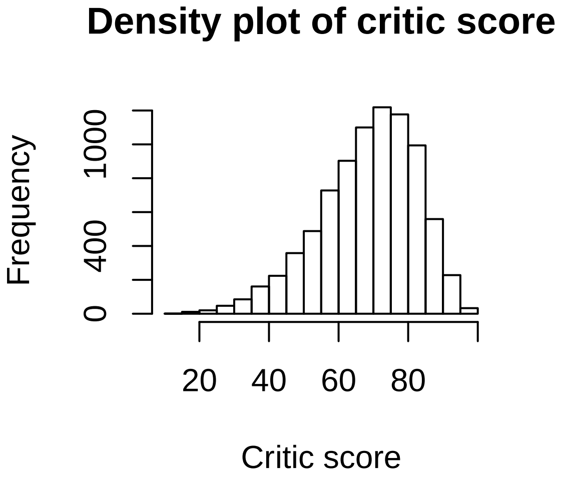
In contrast to the null hypothesis, the alternate hypothesis is assumed as the mean critic score of video games is not equals to 70.

$$H_a : \mu_c \neq 70$$

3.2.5 Conditions for applying the test

For applying the two sided one-sample t-test the sample data should satisfy the following conditions.
i. One quantitative variable of interest
Yes, the Critic Score variable is quantiative
ii. The sample should come from a single population
Yes, the sample comes from a single population.
iii. The population data must be normally distributed
This condition can be verified by plotting QQplot of the sample data.



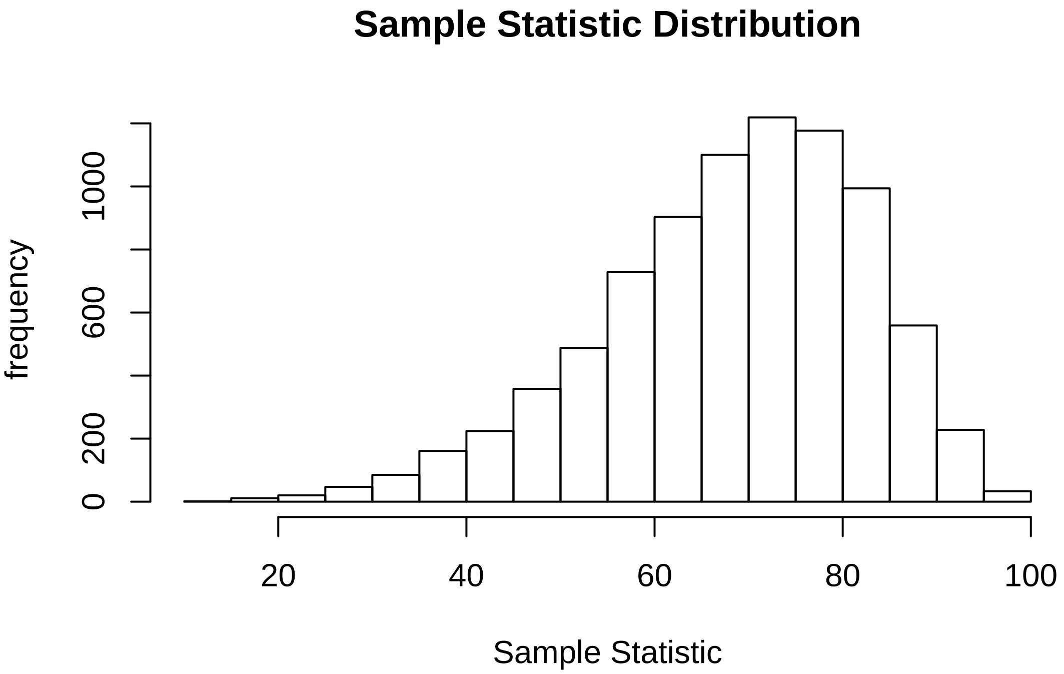


It can be observed from the above graphs that the data is exactly not aligned to the $x = y$ line. It might be happening because the data is skewed to the right, As the normality of the data is not assured we can say that the test will not be optimally powerful and the results are not completely reliable. In this case, the t-test will not perform as it should - i.e. that if the null hypothesis is true, it will falsely reject the null 5% (considered $\alpha = 0.05$) of the time.

3.2.6 Sample Statistic

With respect to the population parameter the sample statistic would be,

$$\bar{x}_c$$



From the above graph, it can be observed that the distribution of sample static is peaked at 70, which might be an indication that the mean critic score lies some where around 70.

3.2.7 Test statistic

Adding to the testing statergy selected, the alternate hypothesis states that the mean critic score is not equals 70, so the two sided testing statergy is to be applied. Which mean, the test statistic should be

$$t_{n-1} = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

3.2.8 Test Results

Before conducting the test the test significance values α needs to be chosen. In this case, we considered the test significance $\alpha = 0.05$, so the confidence intervals for the test would be 95%.

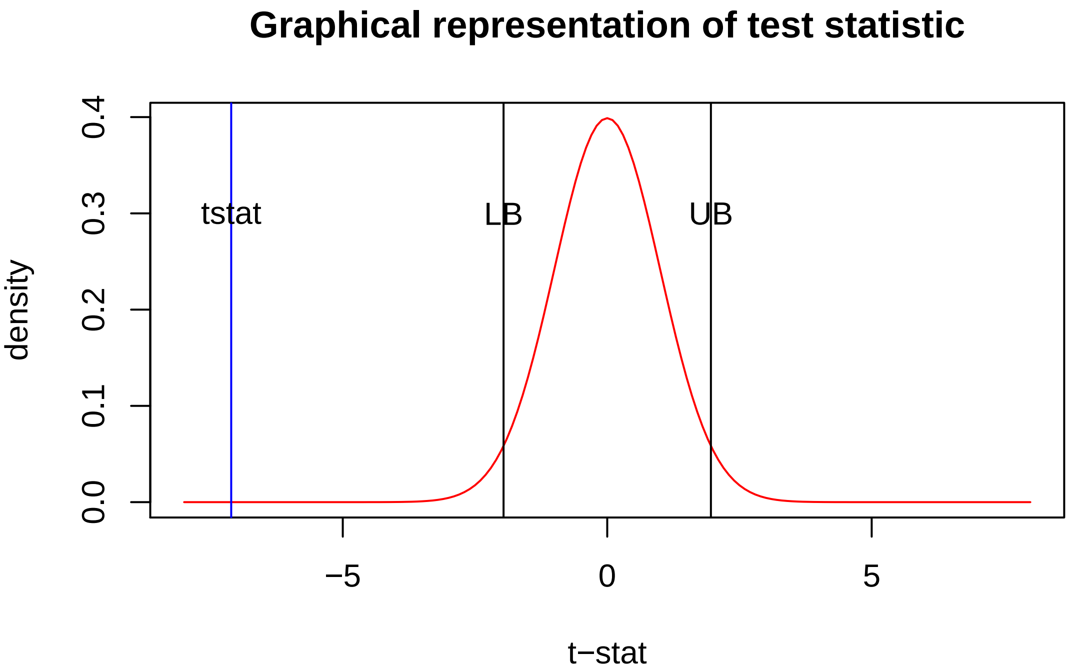
From the sample mean $\bar{x}$ and the sample standard deviations s we can easily compute test statastic. The t-stat is calculated as -7.1096787. Under the null hypothesis, the test statistic has a t distribution with $n-1$ degrees of freedom, which is 8335. To complete the test, the p-value of the test needs to be obtained. Since this is a two-sided test with not equals alternative, the area under the right and left tails of the t-distribution with 8335 degrees of freedom needs to be considered.

$$P(t_{8335} < -7.1096787) \times 2$$

The calculated p-value is 1.259158×10^{-12}. As the p-value of the test is known, which is less than our significance level (0.05), so we reject the null hypothesis.

3.2.9 Building confidence intervals

The one piece of information we are missing is the critical value $t_{8335}(\alpha/2) = t_{8335}(0.025)$, which is required for building the confidence intervals. This can be calculated by finding the quantile wherein the probability is 0.025 in the t distribution, which is calculated as -1.9602486



The graph aboove describes the spread of confidence intervals, black lines represens the t-critical values and the blue line represents our actual t-stat value. The 95% Confidence Intervals for the mean critic score is calculated by the following formula,

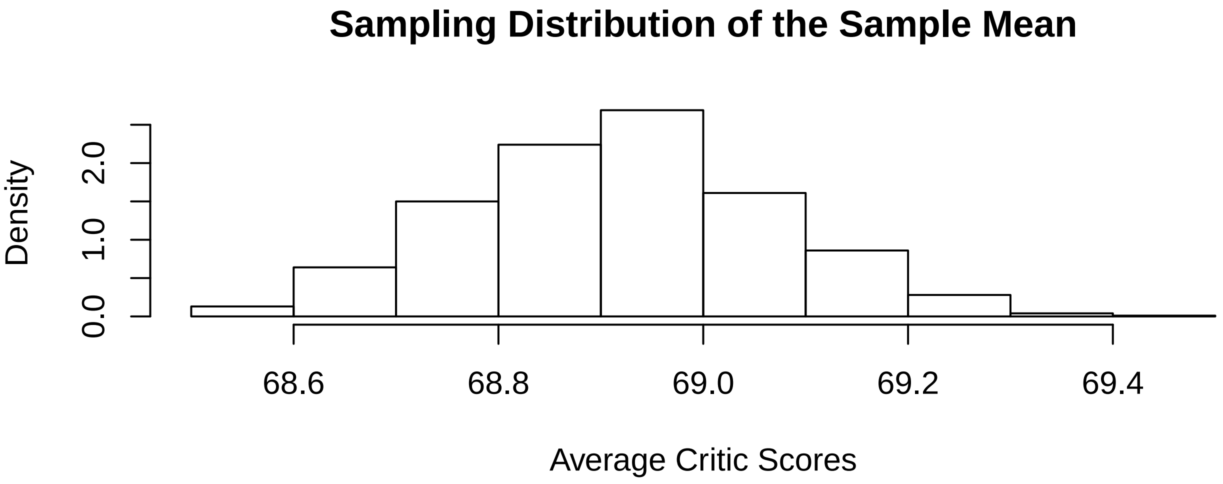
$$(\bar{x}) \pm t_{8335}(0.025)\frac{s}{\sqrt{n}}$$

The calculated confidence intervals are -0.4606588 and 1.084468. As the 95% Confidence intervals extends from approximately 68.6127206 to 68.6127206. We can see our null hypothesis value 70 is not present in our calculated confidence intervals which strengthen our claim that we reject the null hypothesis.

3.2.10 Bootstrap Approach

3.2.10.1 Bootstrap confidence intervals

For performing the bootstrap one sample t-test a random sample with the same size is to be taken from the original sample with replacement and the mean is computed on these samples. The distribution of sample mean is to be plotted by performing this finate number of times.

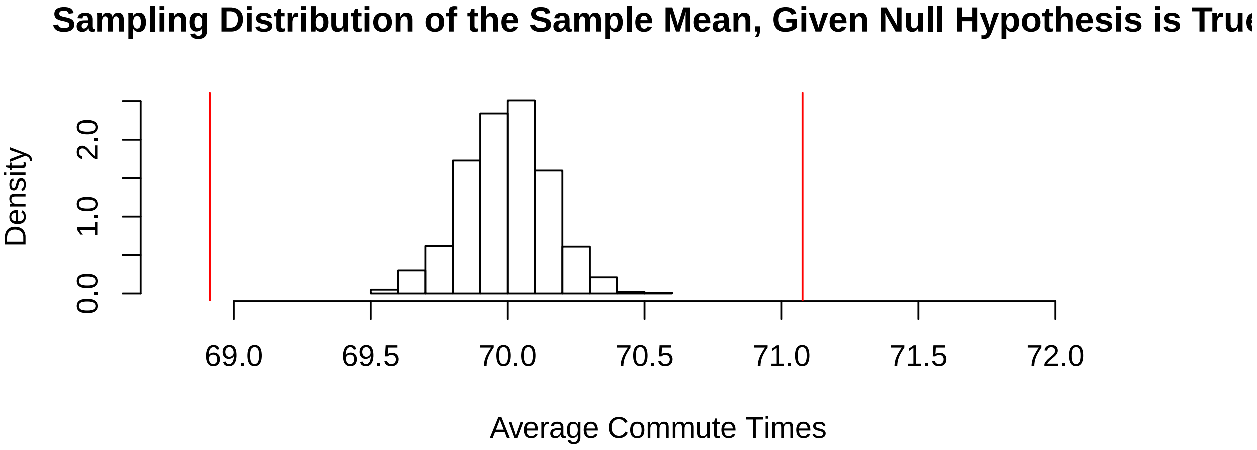


As the calculation of sample means is simulated for 1000 times it can be observed that the distribution has become more symmetric. The confidence intervals are calculated by finding the 0.025 and 0.975 quantiles on the sampling distribution of the sample mean which are calculated as 68.6107497 and 69.2143463.

3.2.10.2 Bootstrap p-value

From the definition of the p-value: The p-value is the probability that we observe a test statistic as or more unusual than the one we observed, given that the null hypothesis is true.

If the null hypothesis is true, then the true mean critic score is 70. We need to shift our distribution so that this is true.



The lines in the graphs represent the mean critic scores calculated from the sample. The p-value is calculated as the counts of values more extreme than the test statistic in our original sample, given the null hypothesis is true, which is calculated as 0. As the bootstrap p-value of the test is known, which is less than our significance level (0.05), so we reject the null hypothesis.

3.2.11 Interpretation

There is strong evidence (p-value=1.259158×10^{-12}) to suggest that the true mean critic score for video games is different than 70. We reject the null hypothesis that the true critic score is 70 at the level. With 95% confidence, the true mean critic score is between 68.6127206 and 69.2123754 which suggests that the true mean critic score is less than 70.

3.3 Racing games released by Activision and Nintendo

From the exploratory data analysis, it is found that between 1980 and 2016 the number of racing games released by the Activision and Nintendo are almost equal. To make further inferences to this claim with statistical evidence we have to conduct a hypothesis test.

3.3.1 Selection of statistical test

Did Activision and Nintendo publishers release the same number of racing games between 1980 and 2011? From the question of interest, the goal is to compare the number of racing games released by the Activision and Nintendo. As the comparison is made between the samples from two different populations and deals with the proportion of racing games released, the test that needs to be selected is the two-sample test of proportions.

3.3.2 Population parameter

As the comparison is made between the number of racing games released by two different publishers and it is clear from the sample data that genre is a categorical variable, so the population parameter should be taken as the difference of proportions

$$p_a - p_n$$

Where p_a and p_n are the population proportion of racing games released by the Activision and Nintendo respectively.

3.3.3 Null Hypothesis

Assumed null hypothesis is the proportion of racing games released by the Activision and Nintendo are equal, which implies, the difference between proportions should be zero.

$$H_0 : p_a - p_n = 0$$

3.3.4 Alternate Hypothesis

In contrast to the null hypothesis, the alternate hypothesis is assumed as the proportion of racing games released by the Activision and Nintendo are not equal, which implies, the difference between proportions should not be zero.

$$H_a : p_a - p_n \neq 0$$

3.3.5 Conditions for applying the test

For applying the two-sided two-sample test of proportions the sample data should satisfy the following conditions.
i. Question of interest has to do with the difference of proportions between two populations
As per null hypothesis we are interested in calculating the difference in the proportion of racing games released by Activision and Nintendo publishers
ii. Two independent samples from two populations
Yes, we are considering the Activision games and Nintendo games to be different populations.
iii. Sample needs to be representative of the population
Yes, the sample is the representative of the population.
iv. Categorial response variable with 2 categories

Yes, the genre variable consists of two categories racing games and other games.
v. $np \geq 10$ and $n(1-p) \geq 10$ for both the populations
Yes, the sample had a sufficient number of records to satisfy this condition.

3.3.6 Sample Statistic

With respect to the population parameter the sample statistic would be,

$$\hat{p}_a - \hat{p}_n$$

3.3.7 Test statistic

Adding to the testing statergy selected, the alternate hypothesis states that the difference in the proportions is not equals zero, so the two sided testing statergy is to be applied. Which mean, the test statistic should be

$$z = \frac{(\hat{p}_a - \hat{p}_n) - (p_a - p_n)}{\sqrt{\frac{\hat{p}_a(1-\hat{p}_a)}{n_a} + \frac{\hat{p}_n(1-\hat{p}_n)}{n_n}}} \sim N(0,1)$$

3.3.8 Test Results

The test significance values α needs to be chosen before conducting the test. In this case, the test significance value is considered as $\alpha = 0.05$, so the confidence intervals for the test would be 95%.

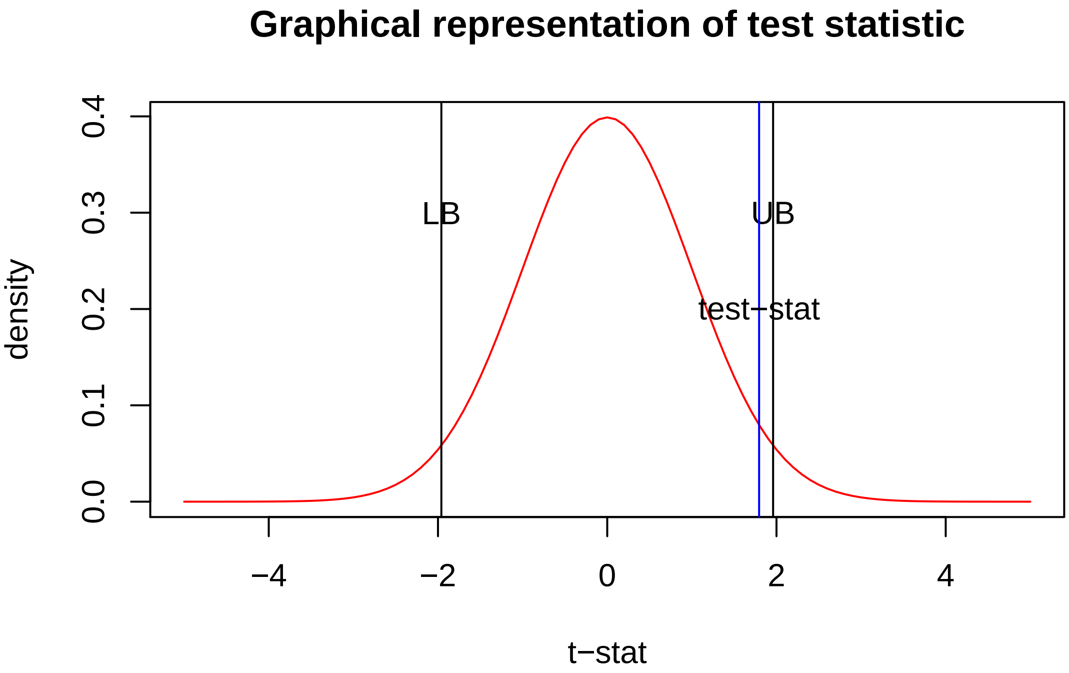
From the sample proportions $\hat{p}_a$, $\hat{p}_n$ and the sample sizes n_a, n_n we can easily compute test statistic. The z is calculated as 1.7946249. Under the null hypothesis, the test statistic has a normal distribution. To complete the test, the p-value of the test needs to be obtained. Since this is a two-sided test with not equals alternative, we need the area under the right and left tails of the normal curve.

$$P(z > 1.7946249) \times 2$$

The calculated p-value is 0.0727135. As the p-value of the test is known, which is greater than our significance level (0.05), so we fail to reject the null hypothesis.

3.3.9 Building confidence intervals

The one piece of information we are missing is the critical value $z(\alpha/2) = z(0.025)$, which is required for building the confidence intervals. This can be calculated by finding the quantile wherein the probability is 0.025 in the normal distribution, which is calculated as -1.959964



The graph aboove describes the spread of confidence intervals, black lines represens the z-critical values and the blue line represents our actual test statastic value. The 95% Confidence Intervals for the difference in proportion of racing games released by Activision and Nintendo is calculated by the following formula,

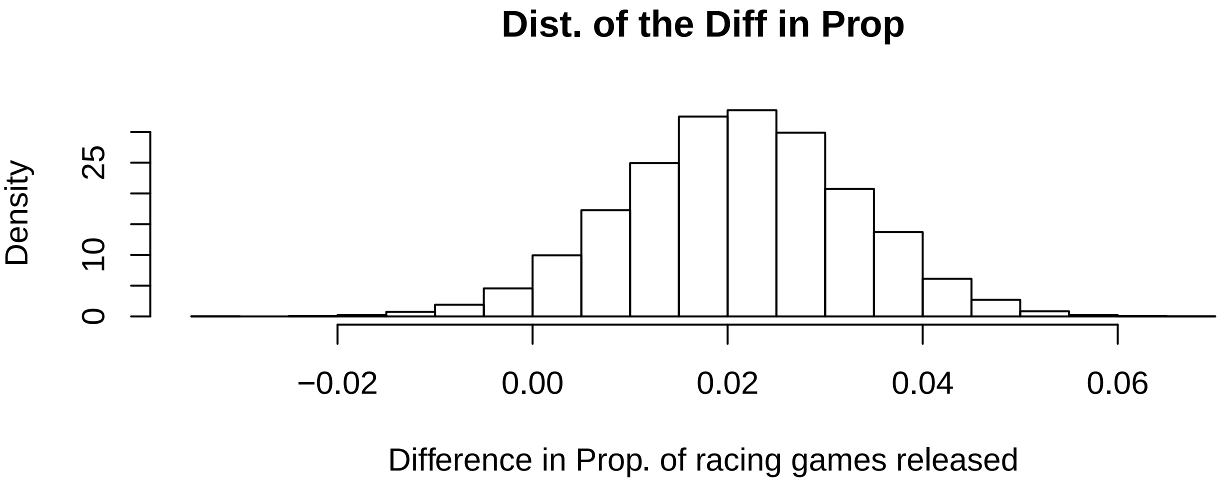
$$(\hat{p}_a - \hat{p}_n) \pm z(0.025)\sqrt{\frac{\hat{p}_a(1-\hat{p}_a)}{n_a} + \frac{\hat{p}_n(1-\hat{p}_n)}{n_n}}$$

The calculated confidence intervals are -0.0019348 and 0.0439353. Our 95% Confidence interval extends from approximately -0.0019348 to 0.0439353. We can see our null hypothesis value 0 is present in our calculated confidence intervals which strengthen our claim that we fail to reject the null hypothesis.

3.3.10 Bootstrap Approach

3.3.10.1 Bootstrap confidence intervals

For performing the bootstrap two sample proportion test a random samples from the two groups with the same size is to be taken from the original sample with replacement and the proportions are computed on these samples. The distribution of difference in proportions is to be plotted by performing this finate number of times.



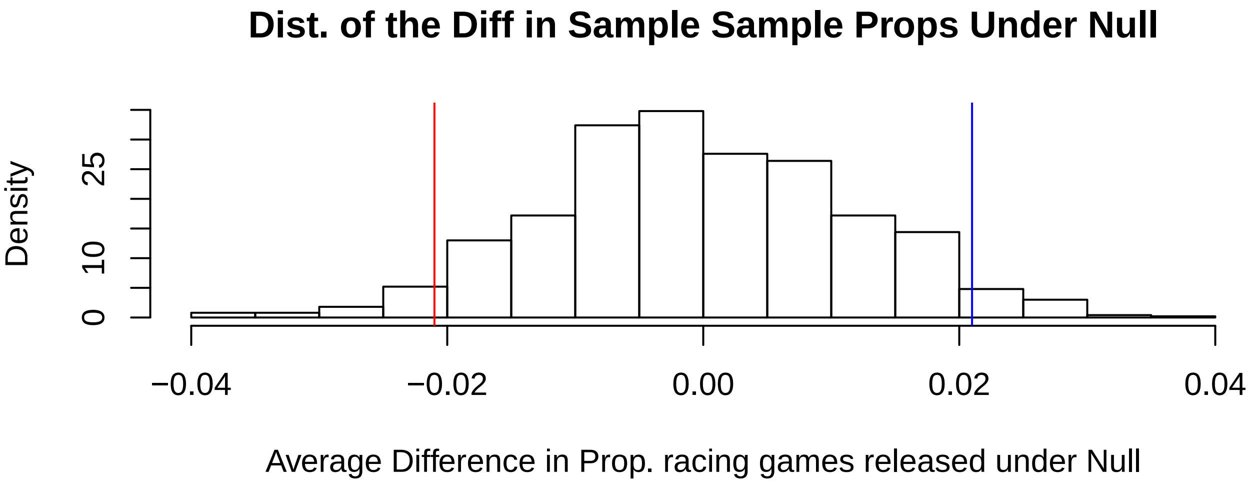
As we simulated the calculation of sample means for 1000 times we can see that the distribution has become more symmetric. The confidence intervals are calculated by finding the 0.025 and 0.975 quantiles on the sampling distribution of the sample mean which are calculated as −0.0024104 and 0.0435909.

3.3.10.2 Bootstrap p-value

From the definition of the p-value: The p-value is the probability that we observe a test statistic as or more unusual than the one we observed, given that the null hypothesis is true.

If the null hypothesis is true, then the true mean difference between the two groups is 0. We need to shift our distribution so that this is true or we can create a randomization distribution

In the case of randamization distribution, we have to simulate many samples assuming the null hypothesis is true. As the Null hypothesis is true, there is no relationship between two variables.
1. Create many samples were the treatment group is shuffled (randomized).
2. Compute the difference in groups for each of the samples.
3. Create a histogram of the randomized statistics. This is an approximation of the null distribution.



The lines in the graphs represent the difference in the proportions of the racing games calculated from the original sample. The p-value is calculated as the counts of values more extreme than the test statistic in our original sample, given the null hypothesis is true, which is calculated as 0.085. As the bootstrap p-value of the test is known, which is greater than our significance level (0.05), so we fail to reject the null hypothesis.

3.3.11 Interpretation

There is no evidence (p-value = 0.0727135) to suggest that there is a difference between the true proportion of racing games released by Activision compared to Nintendo between 1980 and 2016. We fail to reject the null hypothesis that the true proportion of racing games released by Activision is equal to the true proportion of the racing games released by Nintendo. With 95% confidence, the true population proportion difference is between -0.0019348 less racing games to 0.0439353 more racing games released by the Activision than Nintendo. The null hypothesized difference of 0 is in the confidence interval which agrees with our failure to reject the null hypothesis.

3.4 Games released by Electronic Arts

From the exploratory data analysis, it is found that almost half of the games released by Electronic Arts between 1976 and 2017 are sports genre. To make further inferences to this claim with statistical evidence we have to conduct a hypothesis test.

3.4.1 Selection of statistical test

Whether half of the games released by Electronic Arts publishers between 1976 and 2017 are sports genre? From the question, the goal is to compare the number of games with sports theme released by the Electronic Arts with other games. As the comparison involves only one sample and deals with the proportion of sports theme released, the test that needs to be selected is the one-sample test of proportions.

3.4.2 Population parameter

As we are comparing the proportion of genres released by one publisher and it is clear from the sample data that genre is a categorical variable, so the population parameter should be taken as the proportion of games by Electronic arts with sports genre.

$$p_s$$

Where p_s is the population proportion of games with sports theme released by the Electronic Arts.

3.4.3 Null Hypothesis

Assumed null hypothesis as the proportion of games released by the Electronic Arts is 0.5

$$H_0 : p_s = 0.5$$

3.4.4 Alternate Hypothesis

In contrast to the null hypothesis, the alternate hypothesis is assumed as the proportion of games released by the Electronic Arts is less than half.

$$H_a : p_s < 0.5$$

3.4.5 Conditions for applying the test

For applying the one-sided one-sample test of proportions the sample data should satisfy the following conditions.
i. Exact Binomial Test
Yes, genre contains exactly two categories: Sports and others, sports refers to success trail whereas the other refers to failure trail.
ii. $np \geq 10$ and $n(1-p) \geq 10$ for both the populations
Yes, the sample had a sufficient number of records to satisfy this condition.

3.4.6 Sample Statistic

With respect to the population parameter the sample statistic would be,

$$\hat{p}_s = \frac{\textbf{number of games with sports theme}}{\textbf{Total number of games}}$$

3.4.7 Test statistic

Adding to the testing statergy selected, the alternate hypothesis states that the prpportion of games released with sports theme is less than half, so the one sided testing statergy is to be applied. Which mean, the test statistic should be

$$z = \frac{(\hat{p}_s - p_s)}{\sqrt{\frac{p_s(1-p_s)}{n}}} \sim N(0,1)$$

3.4.8 Test Results

Before we conduct the test we have to choose the test significance values α. In this case, we considered the test significance $\alpha = 0.05$, so the confidence intervals for the test would be 95%. From the sample proportion $\hat{p}_a$ and the sample size n we can easily compute the test statistic. The z is calculated as -15.0686206. Under the null hypothesis, the test statistic has a normal distribution. To complete the test, we need to obtain the p-value of the test. Since this is a one-sided test with not less than the alternative, we need the area under the left tail of the normal curve.

$$P(z < -15.0686206)$$

The p-value is calculated as $1.3024765 \times 10^{-51}$. As the p-value of the test known, which is less than our significance level (0.05), so we reject the null hypothesis.

3.4.9 Building confidence intervals

The 95% Confidence Interval for the proportion of games released by Electronic Arts with sport theme is calculated by the following formula,

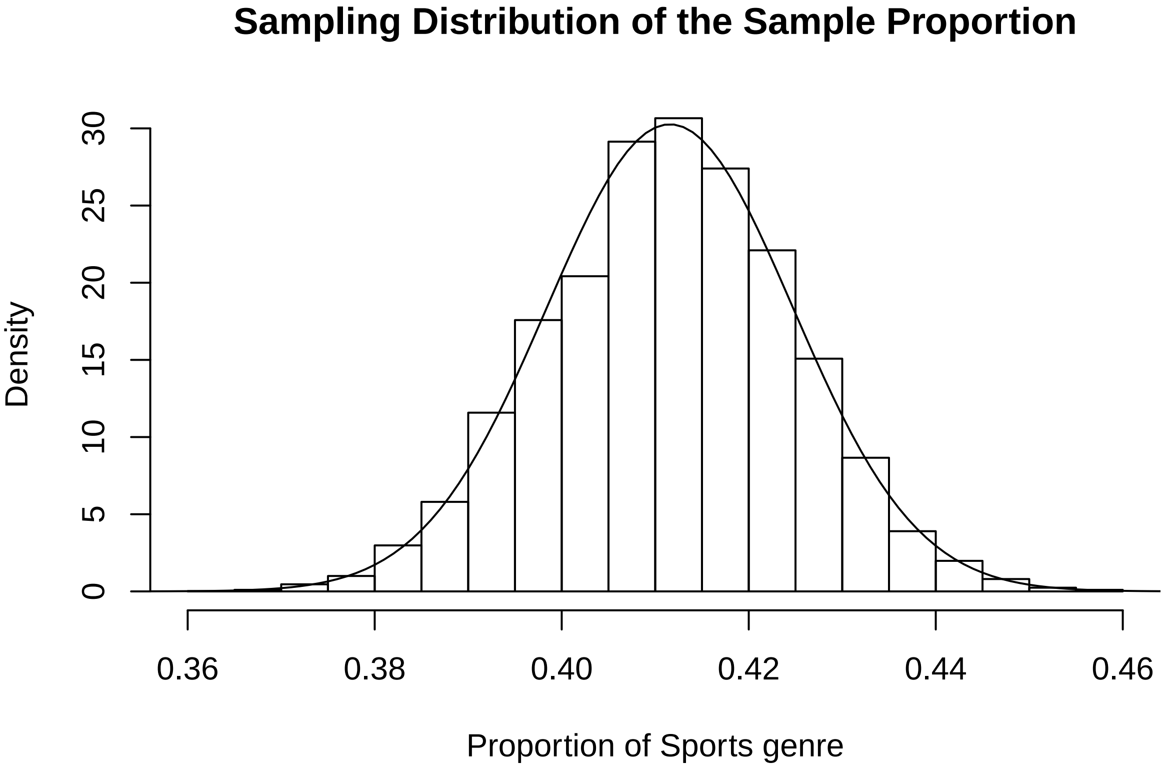
$$\hat{p}_s \pm z(0.05)\sqrt{\frac{\hat{p}_s(1-\hat{p}_s)}{n}}$$

The calculated confidence intervals are 0 and 0.4333844.As, our 95% Confidence interval extends from approximately 0 to 0.4333844. it can br observed that the null hypothesis value 0.5 is not present in our calculated confidence intervals which strengthen our claim that we reject the null hypothesis.

3.4.10 Bootstrap Approach

3.4.10.1 Bootstrap confidence intervals

For performing the bootstrap one-sample proportion test a random sample with the same size is to be taken from the original sample with replacement and the proportion is computed on these samples. The distribution of proportions is to be plotted by performing this finite number of times.



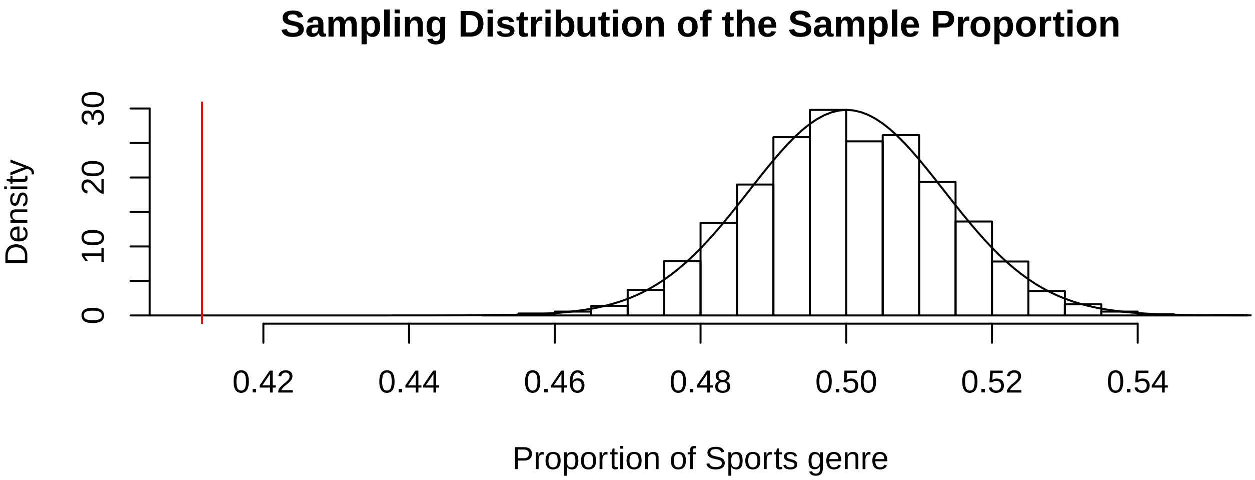
As we simulated the calculation of proportions for 1000 times we can see that the distribution has become more symmetric. The confidence intervals are calculated by finding the 0 and 0.95 quantiles on the sampling distribution of the sample mean which are calculated as 0.3601449 and 0.4333333.

Notice that the one-sided confidence interval from the bootstrap stops at 0.3601449 because, in our empirical distribution, no values go below 0.3601449 however theoretically they could. Our empirical results are also giving a smaller confidence interval compared to our exact and normal approximations in this case.

3.4.10.2 Bootstrap p-value

From the definition of the p-value: The p-value is the probability that we observe a test statistic as or more unusual than the one we observed, given that the null hypothesis is true.

If the null hypothesis is true, then the true proportion of the sports genre released by Electronic Arts would be 0.5. We need to shift our distribution so that this is true.



The line in the graphs represents the proportion of the sports genre from the sample. The p-value is calculated as the counts of values more extreme than the test statistic in our original sample, given the null hypothesis is true, which is calculated as 0. As the bootstrap p-value of the test is known, which is less than our significance level (0.05), so we reject the null hypothesis.

3.4.11 Interpretation

There is very strong evidence (p-value=$1.3024765 \times 10^{-51}$) that the true proportion of games with spots theme released by Electronic Arts is less than 0.5. We can reject the null hypothesis that the true proportion of games with sports theme released by Electronic arts is equal to .5. With 95% confidence, the true proportion of games with sports theme released by Electronic arts is between 0 and 0.4333844.

3.5 Games released in 2008

This part of statistical analysis makes inference on the number of games released in 2008. From the Exploratory data analysis, we found that in the year 2008 the number of games released with Platform, Fighting, Strategy genres are almost equal. To make further inferences to this with statistical evidence we have to conduct a hypothesis test.

3.5.1 Selection of statistical test

As the analysis is performed on the genre variable which consists of more than two categories the test to be chosen is chi-squared goodness of fit.

3.5.2 Population parameter

As the comparison is made among the proportion of Platform, Fighting, Strategy genres released in 2008 and it is clear from the sample data that genre is a categorical variable, so the population parameter should be taken as the proportion of Platform, Fighting, Strategy genres.

$$p_p, p_f, p_s$$

Where p_p, p_f, p_s are the population proportions of games with Platform, Fighting, Strategy genres respectively.

3.5.3 Null Hypothesis

Assumed null hypothesis as the proportion of all the three games released are equal. There are 179 observations in our sample. If each of the solution categories had the same frequency, then each solution category would have a count of 59.

$$H_0 : p_p = p_f = p_s = 0.33\bar{3}$$

3.5.4 Alternate Hypothesis

In contrast to the null hypothesis, the alternate hypothesis is assumed as the proportion of all the three games released are not equal. i.e., at least one of the proportions is not equaled to 0.33

This test does not tell us which proportion is not equal to 0.333. All this test tells us is that at least one of the proportions are significantly different than 0.333.

$$H_a : p_i \neq 0.333\bar{3}$$

3.5.5 Conditions for applying the test

For applying the chi-square goodness of fit test the sample data should satisfy the following conditions.
i. Single categorical variable with more than 2 categories
Yes, genre contains exactly three categories: Platform, Fighting, Strategy.
ii. The expected count of each category is at least 5
Yes, the count of each of three variable is grater than 50.

3.5.6 Sample Statistic

With respect to the population parameter the sample statistic would be,

$$\hat{p}_p, \hat{p}_f, \hat{p}_s$$

3.5.7 Test statistic

Adding to the testing statergy selected, the alternate hypothesis states that the proportion of games released are not equal. So, the test statistic should be

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E)^2}{E} \sim \chi^2_{k-1}$$

3.5.8 Test Results

From the sample proportion $\hat{p}_a$ and the sample size n we can easily compute test statistic. The chi-square is calculated as 0.4134078. Under the null hypothesis, the test statistic has a chi-square distribution with degree of freedom 2. To complete the test, we need to obtain the p-value of the test. Since the chi-squared is one-sided distribution, we need the area under the upper tail of the curve.

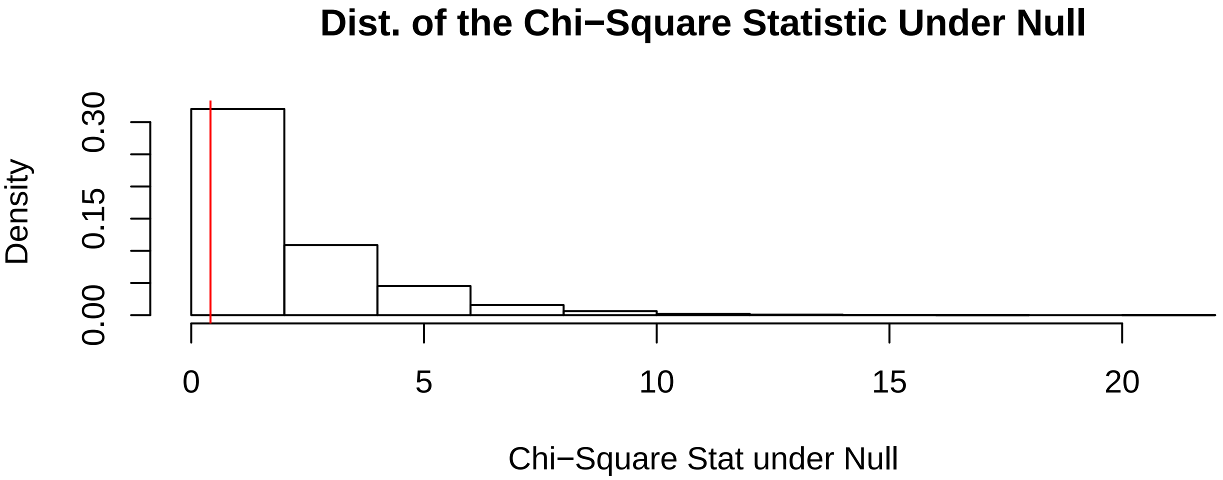
$$P(\chi > 0.4134078)$$

The calculated p-value is 0.8132604. As the p-value of the test is known, which is greater than our significance level (0.05), so we fail to reject the null hypothesis.

3.5.9 Randomization Approach

Let's build our intuition for the chi-squared test using randomization. This approach creates a distribution by assuming that the null hypothesis is true.

As per the null hypothesis, the proportions of games released with themes Platform, Fighting, Strategy genres are equal. The distribution is plotted by creating a dummy data and sampling 1000 times using this condition which exactly looks like Chi-squared distribution.



The line form the above graph represents our chi-squared test statistic which is present in the denser region of the graph. This is the indication that our hypothesis is true.

The p-value from this approach is calculated as 0.8269. As it is observed that p-value for the randomized approach is very close to the p-value from a normalized approach. As the p-value is greater than our significance level (0.05), so we fail to reject the null hypothesis.

3.5.10 Interpretation

There is evidence (p-value=0.8269) that the proportion of games released with themes Platform, Fighting, Strategy genres in 2008 are equal. We failed to reject our null hypothesis that proportions of games released with themes Platform, Fighting, Strategy are equal at the $\alpha = 0.05$ significance level.

4 Conclusion

4.1 Summary

This report provides an analysis of the global sales, genres, critic scores and publishers of the video games. Methods of analysis include two-sample t-test, one-sample t-test, the two-sample test of proportions, one sample test of proportions and chi-square goodness of fit test. Results of the data analysed shows that regardless of the platform the sales patterns of PS2 and Wii games are equal, the average critic score of video games differs from theoretically said value, major publishers Activision and Nintendo released equal proportions of racing games between 1980 and 2008, and the proportion of games released with Platform, Fighting, Strategy themes are equal in the year 2008

4.2 Implications

1. By conducting two-sample t-test it is concluded that the global sales patterns of the games released on PS2 and Wii platforms are equal. With 95% confidence, this test tells that regardless of the platform the sales patterns of PS2 and Wii games are equal.

2. As said in many articles and blogs the average critic score of the video games is not equalled to 70. With 95% confidence, one-sample t-test concludes that the average critic score attained by the video games is less than 70.

3. It is not true the Electronic Arts publishers primarily focus on sports genre games. With 95% confidence, one sample test of proportions concludes that the proportion of sports genre games released by Electronic Arts is not equalled to half.

4. It is true the Activision and Nintendo had released equal proportions of racing games between 1980 and 2008. With 95% confidence, one sample test of proportions concludes that equal proportions of racing games are released by these two publishers.

5. By conducting chi-square goodness of fit test it is concluded that equal proportions of games with Platform, Fighting, Strategy genres released in 2008. With 95% confidence, this test tells that the proportion of games released with Platform, Fighting, Strategy themes are equal in 2008.

4.3 Extensions and Limitations

4.3.1 Extensions

This study just focussed on analyzing the global sales of certain platforms, it can be extended to study the regional-wise sales patterns of different platforms and how these patterns vary with genres, ratings and publishers.

4.3.2 Limitations

1. Scraping algorithm used to collect this data focused on the sales of the game units and also this data does not represent the modern gaming industry.

2. The data used to perform the tests is not symmetrically distributed, which makes the test inconsistent. So, we can’t say that the test results are completely reliable.

4.4 Further questions, next steps

This report outlined the analysis of the sales pattern of the video games based on different genres, publishers and platforms. For further questions, I want to consider what are the other factors that affect the sales of video games, such as the adult population in the region. Secondly, I want to consider other populations, like PC gaming and mobile gaming.

4.5 Appendix

4.5.1 Dataset

Dataset - https://www.kaggle.com/kendallgillies/video-game-sales-and-ratings

4.6 R Script

```
knitr::opts_chunk$set(echo = FALSE)
#Loading Packages
library("tidyverse")
library("readxl")
library(plotrix)
library(dplyr)
library("ggplot2")
library(cowplot)
library(lattice)
library(readr)
# Loading dataset
dataset <- read_csv(file="/cloud/project/Dataset/final_proj_data.csv")
#Two sample t-test
## Data Cleaning
dataset_filtered_platform_globalsales<-dataset[complete.cases(dataset["Platform"]),]
dataset_filtered_platform_globalsales<-dataset_filtered_platform_globalsales[complete.cases(dataset_fil
## Data processing
dataset_filtered_platform_globalsales_grouped<-dataset_filtered_platform_globalsales %>%
  dplyr::group_by(Platform) %>%
  dplyr::summarise(mean_sales = mean(Global_Sales))
#Two sample t-test
## Disribution of global sales by platform
ggplot(data=dataset_filtered_platform_globalsales_grouped, aes(x=Platform, y=mean_sales)) +
  geom_bar(stat="identity",position = position_dodge(width=0.5),fill = "#d9b886")+
  theme(axis.text.x = element_text(angle = 90, hjust = 1))+
  xlab("Platform") +
  ylab("Mean global sales")+
  ggtitle("Disribution of global sales by platform")
## Data Filtering
dataset_subset_platform_globalsales <- dataset_filtered_platform_globalsales[dataset_filtered_platform_g


dataset_grouped_year_platform<-dataset_subset_platform_globalsales[complete.cases(dataset_subset_platfo

dataset_grouped_year_platform<-dataset_grouped_year_platform %>%
  dplyr::group_by(Platform,Year_of_Release) %>%
  dplyr::summarise(mean_sales = mean(Global_Sales))

dataset_grouped_year_platform$Year_of_Release<-as.factor(dataset_grouped_year_platform$Year_of_Release)
##Global sales PS2 and Wii games
ggplot(data=dataset_grouped_year_platform, aes(x=Year_of_Release, y=mean_sales, fill=Platform)) +
geom_bar(stat="identity")+
  theme(axis.text.x = element_text(angle = 90, hjust = 1))+
  xlab("Year Of Release") +
  ylab("Mean global sales")+
  ggtitle("Global sales PS2 and Wii games")
## Data Filtering
dataset_subset_platform_globalsales<-dataset_subset_platform_globalsales[dataset_subset_platform_globals
## Frequency distribution of Global sales of PS2 and Wii games
ggplot(dataset_subset_platform_globalsales, aes(x=Global_Sales, fill=Platform)) +
  geom_histogram(color="black",alpha=0.8)+
  xlab("Global Sales") +
  ylab("Frequency")+
  ggtitle("Frequency distribution of Global sales of PS2 and Wii games")
# Box plot to understand the spread of global sales
boxplot(dataset_subset_platform_globalsales$Global_Sales[dataset_subset_platform_globalsales$Platform==
        dataset_subset_platform_globalsales$Global_Sales[dataset_subset_platform_globalsales$Platform==
        names = c("PS2", "Wii"),
        col = c("blue","red"),
        notch = TRUE,
        ylab="Global Sales",
        main='Box plot for Global sales of PS2 and Wii games')
#One sample t-test
dataset_filtered_critic_score<-dataset[complete.cases(dataset["Critic_Score"]),]
## Frequency distribution Critic Scores
ggplot(dataset_filtered_critic_score, aes(x=Critic_Score)) +
  geom_density(fill="#4065a1",color="#4065a1")+
  xlab("Critic Scores") +
  ylab("Frequency")+
  ggtitle("Frequency distribution Critic Scores")+
  geom_vline(aes(xintercept=mean(Critic_Score)),
             linetype="dashed",color="red")
# Box plot to understand the spread Critic Score
boxplot(dataset_filtered_critic_score$Critic_Score,
        notch = TRUE,
        ylab="Critic Score",
        col = c("blue"),
        horizontal=T,
        main='Box plot for critic scores')
## Two proportion test

dataset_filtered_publisher_genre<-dataset[complete.cases(dataset["Publisher"]),]

dataset_filtered_publisher_genre<-dataset[complete.cases(dataset["Genre"]),]

dataset_filtered_publisher_genre<-dataset_filtered_publisher_genre[dataset_filtered_publisher_genre$Publ

dataset_filtered_publisher_genre$Genre <- ifelse(dataset_filtered_publisher_genre$Genre == 'Racing', da

racing_activision <- t(replicate(length(which(dataset_filtered_publisher_genre$Publisher == "Activision
                        c("Racing", "Activision")))
racing_nintendo <- t(replicate(length(which(dataset_filtered_publisher_genre$Publisher == "Nintendo" & d
                        c("Racing", "Nintendo")))
other_activision <- t(replicate(length(which(dataset_filtered_publisher_genre$Publisher == "Activision"
                        c("Other", "Activision")))
other_nintendo <- t(replicate(length(which(dataset_filtered_publisher_genre$Publisher == "Nintendo" & da
                        c("Other", "Nintendo")))
```

```r
dataset_2 <- data.frame(rbind(racing_activision,
                              racing_nintendo,
                              other_activision,
                              other_nintendo))
names(dataset_2) <- c("Genre", "Publisher")
ds<-table(dataset_2)
ds<-addmargins(ds)
#One proportion test
dataset_filtered_electronic_arts<-dataset[complete.cases(dataset["Publisher"]),]
dataset_filtered_electronic_arts<-dataset_filtered_electronic_arts$Genre[dataset_filtered_electronic_ar
## Frequency distribution genres by Electronic Arts
ggplot(data.frame(dataset_filtered_electronic_arts), aes(x=dataset_filtered_electronic_arts)) +
  geom_bar()+
  xlab("genres") +
  ylab("Frequency")+
  ggtitle("Frequency distribution genres by Electronic Arts")+
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
## Data Filtering
dataset_filtered_electronic_arts <- ifelse(dataset_filtered_electronic_arts == 'Sports', dataset_filtere
## Frequency distribution genres by Electronic Arts
ggplot(data.frame(dataset_filtered_electronic_arts), aes(x=dataset_filtered_electronic_arts)) +
  geom_bar()+
  xlab("genres") +
  ylab("Frequency")+
  ggtitle("Frequency distribution genres by Electronic Arts")+
 coord_flip()
#chi_squared
## Data Filtering
dataset_filtered_year_genre<-dataset[complete.cases(dataset["Year_of_Release"]),]
dataset_filtered_year_genre<-dataset_filtered_year_genre[complete.cases(dataset_filtered_year_genre["Ge

  dataset_graph <- dataset_filtered_year_genre[dataset_filtered_year_genre$Year_of_Release %in% c('2008

##Frequency distribution of genres in 2008

  ggplot(data.frame(dataset_graph$Genre), aes(x=dataset_graph$Genre)) +
  geom_bar(fill='#99c4a4')+
  xlab("genres") +
  ylab("Frequency")+
  ggtitle("Frequency distribution of genres in 2008")+
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
 ## Data Filtering
  dataset_filtered_year_genre <- dataset_filtered_year_genre[dataset_filtered_year_genre$Genre %in% c('

  dataset_filtered_year_genre<-dataset_filtered_year_genre$Genre[dataset_filtered_year_genre$Year_of_Rel
  pie(table(dataset_filtered_year_genre))
dataset_filtered_year_genre_unique<-unique(dataset_filtered_year_genre)
# Checking the normality of the data.
par(mfrow=c(2,2))
#Normality check for Global sales of Wii Games
qqnorm(dataset_subset_platform_globalsales$Global_Sales[dataset_subset_platform_globalsales$Platform=="
qqline(dataset_subset_platform_globalsales$Global_Sales[dataset_subset_platform_globalsales$Platform=="
#Normality check for Global sales of Wii Games
```

```
qqnorm(dataset_subset_platform_globalsales$Global_Sales[dataset_subset_platform_globalsales$Platform=="W
qqline(dataset_subset_platform_globalsales$Global_Sales[dataset_subset_platform_globalsales$Platform=="W

#Histogram showing the Global sales of PS2 Games
hist(dataset_subset_platform_globalsales$Global_Sales[dataset_subset_platform_globalsales$Platform=="PS2
     xlab="Global sales in Millions",
     main="Density plot PS2 games")
#Histogram showing the Global sales of Wii Games
hist(dataset_subset_platform_globalsales$Global_Sales[dataset_subset_platform_globalsales$Platform=="Wi
     xlab="Global sales in Millions",
     main="Density plot Wii Games")
#Histogram to show the distribution of  sample statistic
hist(dataset_subset_platform_globalsales$Global_Sales[dataset_subset_platform_globalsales$Platform=="PS2
     xlab="Sample Statistic",
     ylab="frequency",
     main="Sample Statistic Distribution")
# the parts of the test statistic
# sample means
x_bar_p <- mean(dataset_subset_platform_globalsales$Global_Sales[dataset_subset_platform_globalsales$Pla
x_bar_w <- mean(dataset_subset_platform_globalsales$Global_Sales[dataset_subset_platform_globalsales$Pla
# null hypothesized population mean difference between the two groups
mu_0 <- 0
# sample variances
s_p_sq <- sd(dataset_subset_platform_globalsales$Global_Sales[dataset_subset_platform_globalsales$Platf
s_w_sq <- sd(dataset_subset_platform_globalsales$Global_Sales[dataset_subset_platform_globalsales$Platf
# sample size
n_p <- length(dataset_subset_platform_globalsales$Global_Sales[dataset_subset_platform_globalsales$Plat
n_w <- length(dataset_subset_platform_globalsales$Global_Sales[dataset_subset_platform_globalsales$Plat
# t-test test statistic
t <- (x_bar_p - x_bar_w - mu_0)/sqrt((s_p_sq/n_p) + (s_w_sq/n_w))
#Calculation of P Value
two_sided_diff_t_pval <- pt(q = t, df = min(n_p, n_w)-1, lower.tail = FALSE)*2

#Critical value
t_critcal_1<-qt(0.025,min(n_p,n_w)-1)
##Graphical representation of test statistic
pts<-seq(-5,5,0.1)
plot(pts,dt(pts,df=1200),
     col='red',
     type='l',
     xlab="t-stat",
     ylab="density",
     main="Graphical representation of test statistic")
 abline(v=qt(0.025,df=1200), col="black")
 abline(v=qt(.975,df=1200), col="black")
 text(qt(0.025,df=1200),0.3,"LB")
 text(qt(0.975,df=1200),0.3,"UB")
 abline(v=t, col="blue")
 text(t,0.3,"tstat")
##Building confidence intervals
#lower bound
conf_lower_1<-(x_bar_p-x_bar_w)+(qt(0.025,min(n_p,n_w)-1)*sqrt((s_p_sq/n_p)+(s_w_sq/n_w)))
```

```r
#higher bound
conf_upper_1<-(x_bar_p-x_bar_w)-(qt(0.025,min(n_p,n_w)-1)*sqrt((s_p_sq/n_p)+(s_w_sq/n_w)))

num_sims <- 10000
# A vector to store my results
results <- rep(NA, num_sims)
# A loop for completing the simulation
for(i in 1:num_sims){
 mean_ps2 <- mean(sample(x = dataset_subset_platform_globalsales$Global_Sales[dataset_subset_platform_gl
 size = n_p,
 replace = TRUE))
 mean_wii <- mean(sample(dataset_subset_platform_globalsales$Global_Sales[dataset_subset_platform_global
 size = n_w,
 replace = TRUE))
 results[i] <- mean_ps2 - mean_wii
}
# Finally plot the results
hist(results, freq = FALSE, main='Sampling Distribution of the Sample Mean', xlab = 'Average Difference
## Bootstrap lower confidence interval
lowerbound_bootstrap<-c(quantile(results, .025))
## Bootstrap upper confidence interval
upperbound_bootstrap<-c(quantile(results, .975))
set.seed(0)
num_sims <- 1000
# A vector to store my results
results_given_H0_true <- rep(NA, num_sims)
# A loop for completing the simulation
for(i in 1:num_sims){
 # idea here is if there is no relationshipm we should be able to shuffle the groups
 shuffled_global_sales <- transform(dataset_subset_platform_globalsales, Global_Sales=sample(Global_Sal
 mean_ps2 <- mean(shuffled_global_sales$Global_Sales[shuffled_global_sales$Platform=="PS2"])
 mean_wii <- mean(shuffled_global_sales$Global_Sales[shuffled_global_sales$Platform=="Wii"])
 results_given_H0_true[i] <- mean_ps2 - mean_wii
}
# Finally plot the results
hist(results_given_H0_true, freq = FALSE,
 main='Dist. of the Diff in Sample Means Under Null',
 xlab = 'Average Difference Global Sales under Null',
 ylab = 'Density')
diff_in_sample_means <- mean(dataset_subset_platform_globalsales$Global_Sales[dataset_subset_platform_gl
abline(v=diff_in_sample_means, col = "blue")
abline(v=abs(diff_in_sample_means), col = "red")
## Bootstrap P-value
count_of_more_extreme_lower_tail <- sum(results_given_H0_true <= diff_in_sample_means)
count_of_more_extreme_upper_tail <- sum(results_given_H0_true >= abs(diff_in_sample_means))
bootstrap_pvalue <- (count_of_more_extreme_lower_tail + count_of_more_extreme_upper_tail)/num_sims
# Checking the normality of the data.
#Normality check for the critic score variable
par(mfrow=c(1,2))
qqnorm(dataset_filtered_critic_score$Critic_Score)
qqline(dataset_filtered_critic_score$Critic_Score)

#Histogram showing the frequency distribution of critic score
```

```
hist(dataset_filtered_critic_score$Critic_Score,
     xlab="Critic score",
     main="Density plot of critic score")
#Histogram to show the distribution of  sample statistic
hist(dataset_filtered_critic_score$Critic_Score,
     xlab="Sample Statistic",
     ylab="frequency",
     main="Sample Statistic Distribution")
# the parts of the test statistic
# sample mean
x_bar <- mean(dataset_filtered_critic_score$Critic_Score)
# null hypothesized population mean
mu_0 <- 70
# sample st. dev
s <- sd(dataset_filtered_critic_score$Critic_Score)
# sample size
n <- length(dataset_filtered_critic_score$Critic_Score)
# t-test test statistic
t <- (x_bar - mu_0)/(s/sqrt(n))
# two-sided p-value so multiply by 2
two_sided_t_pval <- pt(q = t, df = n-1,lower.tail = TRUE)*2
#Critical value
t_critcal_1<-qt(0.025,n-1)
##Graphical representation of test statistic
pts<-seq(-8,8,0.1)
plot(pts,dt(pts,df=8335),
     col='red',
     type='l',
     xlab="t-stat",
     ylab="density",
     main="Graphical representation of test statistic")
 abline(v=qt(0.025,df=8335), col="black")
 abline(v=qt(.975,df=8335), col="black")
 text(qt(0.025,df=8335),0.3,"LB")
 text(qt(0.975,df=8335),0.3,"UB")
 abline(v=t, col="blue")
 text(t,0.3,"tstat")
##Building confidence intervals
#lower bound
conf_lower_1<-(x_bar)+(qt(0.025,n-1)*s/sqrt(n))
#higher bound
conf_upper_1<-(x_bar)-(qt(0.025,n-1)*s/sqrt(n))
num_sims <- 1000
# A vector to store my results
results <- rep(NA, num_sims)
# A loop for completing the simulation
for(i in 1:num_sims){
 results[i] <- mean(sample(x = dataset_filtered_critic_score$Critic_Score,
 size = n,
 replace = TRUE))
}
# Finally plot the results
hist(results, freq = FALSE, main='Sampling Distribution of the Sample Mean', xlab = 'Average Critic Sco
```

```r
# estimate a normal curve over it - this looks pretty good!
# Shift the sample so that the null hypothesis is true
critic_score_given_H0_true <- dataset_filtered_critic_score$Critic_Score - mean(dataset_filtered_critic_

num_sims <- 1000
# A vector to store my results
results_given_H0_true <- rep(NA, num_sims)
# A loop for completing the simulation
for(i in 1:num_sims){
 results_given_H0_true[i] <- mean(sample(x = critic_score_given_H0_true,
 size = n,
 replace = TRUE))
}
# add line to show values more extreme on lower end
low_end_extreme <-mean(results_given_H0_true)+(mean(results_given_H0_true)-x_bar)
bootstrap_SE_X_bar <- sd(results)
# an estimate is to use the formula statistic +/- 2*SE
bootstrap_lower<-x_bar - 2*bootstrap_SE_X_bar
bootstrap_upper<-x_bar + 2*bootstrap_SE_X_bar
# Shift the sample so that the null hypothesis is true
# Finally plot the results
hist(results_given_H0_true, freq = FALSE,xlim=c(68.8,72.3), main='Sampling Distribution of the Sample M
# add line to show values more extreme on upper end
abline(v=x_bar, col = "red")
# add line to show values more extreme on lower end
low_end_extreme <-mean(results_given_H0_true)+(mean(results_given_H0_true)-x_bar)
abline(v=low_end_extreme, col="red")
# counts of values more extreme than the test statistic in our original sample, given H0is true
# two sided given the alternate hypothesis
count_of_more_extreme_lower_tail <- sum(results_given_H0_true >= low_end_extreme)
count_of_more_extreme_upper_tail <- sum(results_given_H0_true <= x_bar)
bootstrap_pvalue <- (count_of_more_extreme_lower_tail + count_of_more_extreme_upper_tail)/num_sims
# the parts of the test statistic
# sample props

p_hat_a <- length(racing_activision)/(length(racing_activision)+length(other_activision))

p_hat_n <- length(racing_nintendo)/(length(racing_nintendo)+length(other_nintendo))

# null hypothesized population prop difference between the two groups
p_0 <- 0
# sample size
n_a <- (length(racing_activision)+length(other_activision))/2
n_n <- (length(racing_nintendo)+length(other_nintendo))/2
# sample variances
den_p_a <- (p_hat_a*(1-p_hat_a))/n_a
den_p_n <- (p_hat_n*(1-p_hat_n))/n_n
# z-test test statistic
z <- (p_hat_a - p_hat_n - p_0)/sqrt(den_p_a + den_p_n)
# two-sided p-value
two_sided_diff_prop_pval <- pnorm(q = z, lower.tail = FALSE)*2
#Critical value
z_critcal<-qnorm(0.025)
```

```r
##Graphical representation of test statistic
pts<-seq(-5,5,0.1)
plot(pts,dnorm(pts),
     col='red',
     type='l',
     xlab="t-stat",
     ylab="density",
     main="Graphical representation of test statistic")
 abline(v=qnorm(0.025), col="black")
 abline(v=qnorm(.975), col="black")
 text(qnorm(0.025),0.3,"LB")
 text(qnorm(0.975),0.3,"UB")
 abline(v=z, col="blue")
 text(z,0.2,"test-stat")
##Building confidence intervals
# lower bound
twoprop_lower_bound<-(p_hat_a - p_hat_n)+(qnorm(0.025)*sqrt(den_p_a + den_p_n))
# upper bound
twoprop_upper_bound<-(p_hat_a - p_hat_n)-(qnorm(0.025)*sqrt(den_p_a + den_p_n))

# Make the data
activision <- rep(c(1, 0), c((length(racing_activision)/2), n_a - (length(racing_activision)/2)))
nintendo <- rep(c(1, 0), c((length(racing_nintendo)/2), n_n - (length(racing_nintendo)/2)))
num_sims <- 10000
# A vector to store my results
results <- rep(NA, num_sims)
# A loop for completing the simulation
for(i in 1:num_sims){
 prop_act <- mean(sample(activision,
 size = n_a,
 replace = TRUE))
 prop_nin <- mean(sample(x = nintendo,
 size = n_n,
 replace = TRUE))
 results[i] <- prop_act - prop_nin
}
# Finally plot the results
hist(results, freq = FALSE, main='Dist. of the Diff in Prop', xlab = 'Difference in Prop. of racing game
lowerbound_bootstrap<-c(quantile(results, .025))
upperbound_bootstrap<-c(quantile(results, .975))
# Make the data
df_combined <- data.frame("racing_games" = c(activision, nintendo),
 "publisher" = rep(c("activision", "nintendo"), c(n_a, n_n)))

set.seed(0)
num_sims <- 1000
# A vector to store my results
results_given_H0_true <- rep(NA, num_sims)
# A loop for completing the simulation
for(i in 1:num_sims){
 # idea here is if there is no relationshipm we should be able to shuffle the groups
 shuffled_groups <- transform(df_combined, publisher=sample(publisher))
 prop_activision <- mean(shuffled_groups$racing_games[shuffled_groups$publisher=="activision"
])
 prop_nintendo <- mean(shuffled_groups$racing_games[shuffled_groups$publisher=="nintendo"])
 results_given_H0_true[i] <- prop_activision - prop_nintendo
}
# Finally plot the results
hist(results_given_H0_true, freq = FALSE,
 main='Dist. of the Diff in Sample Sample Props Under Null',
 xlab = 'Average Difference in Prop. racing games released under Null',
 ylab = 'Density')
diff_in_sample_props <- p_hat_a - p_hat_n
abline(v=diff_in_sample_props, col = "blue")
abline(v=-diff_in_sample_props, col = "red")
# counts of values more extreme than the test statistic in our original sample, given H0, is true
# two sided given the alternate hypothesis

count_of_more_extreme_lower_tail <- sum(results_given_H0_true <= -diff_in_sample_props)
count_of_more_extreme_upper_tail <- sum(results_given_H0_true >= diff_in_sample_props)
bootstrap_pvalue <- (count_of_more_extreme_lower_tail + count_of_more_extreme_upper_tail)/num_sims
# the parts of the test statistic

p_hat<-0.4115942
z <- (p_hat - .5)/ sqrt((.05*(1-.05)) / length(dataset_filtered_electronic_arts))

# two-sided p-value
one_sided_diff_prop_pval <- pnorm(z, lower.tail = TRUE)
##Building confidence intervals
# lower bound
oneprop_lower_bound<-0
# upper bound
oneprop_upper_bound<-p_hat - (qnorm(0.05))*sqrt(((p_hat)*(1-p_hat))/length(dataset_filtered_electronic_a

Sports <- rep(c(1, 0), c(568, 1380-568))

num_sims <- 10000
# A vector to store my results
results <- rep(NA, num_sims)
# A loop for completing the simulation
for(i in 1:num_sims){
  results[i] <- mean(sample(x = Sports,
                            size = 1380,
                            replace = TRUE))
}
# Finally plot the results
hist(results, freq = FALSE,
     main='Sampling Distribution of the Sample Proportion',
     xlab = 'Proportion of Sports genre', ylab = 'Density')
# estimate a normal curve over it - this looks pretty good!
lines(x = seq(.35, .75, .001),
      dnorm(seq(.35, .75, .001),
      mean = mean(results), sd = sd(results)))
## Bootstrap confidence intervals
bootstrap_lower<-quantile(results,0)
bootstrap_upper<-quantile(results, .95)
```

```
Sports <- rep(c(1, 0), c(690, 1380-690))


num_sims <- 10000
# A vector to store my results
results <- rep(NA, num_sims)
# A loop for completing the simulation
for(i in 1:num_sims){
  results[i] <- mean(sample(x = Sports,
                            size = 1380,
                            replace = TRUE))
}
# Finally plot the results
hist(results, freq = FALSE,xlim=c(0.41,0.55),
     main='Sampling Distribution of the Sample Proportion',
     xlab = 'Proportion of Sports genre', ylab = 'Density')
# estimate a normal curve over it - this looks pretty good!
lines(x = seq(.35, .75, .001),
      dnorm(seq(.35, .75, .001),
      mean = mean(results), sd = sd(results)))
abline(v=0.41159420, col="red")
## Bootstrap P-value
count_of_more_extreme_upper_tail <- sum(results <= 0.41159420)
bootstrap_pvalue <- count_of_more_extreme_upper_tail/num_sims
# the parts of the test statistic
equal_prop<-length(dataset_filtered_year_genre)/length(dataset_filtered_year_genre_unique)
chi_square<-sum(((table(dataset_filtered_year_genre) - equal_prop)^2)/equal_prop)
df <- length(dataset_filtered_year_genre_unique)-1
## Calculation of p-value
chisquared_pval <- pchisq(chi_square,df=df , lower.tail = FALSE)
#chi-squared with randamization
solutions_under_H_0<-rep(dataset_filtered_year_genre_unique,equal_prop)
num_sims <- 10000
# A vector to store my results
chisq_stats_under_H0 <- rep(NA, num_sims)
# A loop for completing the simulation
for(i in 1:num_sims){
 new_samp <- sample(solutions_under_H_0, length(dataset_filtered_year_genre), replace = T)
 chisq_stats_under_H0[i] <- sum(((table(new_samp) - equal_prop)^2)/equal_prop)
}

## Dist. of the Chi-Square Statistic Under Null
hist(chisq_stats_under_H0, freq = FALSE,
 main='Dist. of the Chi-Square Statistic Under Null',
 xlab = 'Chi-Square Stat under Null',
 ylab = 'Density')
abline(v=sum(((table(dataset_filtered_year_genre) - equal_prop)^2)/equal_prop), col="red")
# Randomized p-value
randamized_p<-sum(chisq_stats_under_H0 >= sum(((table(dataset_filtered_year_genre) - equal_prop)^2)/equa
```