Statistical Inferences Report

The original purpose of this capstone project was to see if there was a relationship between the Metacritic rating of a video game and how well it sold globally. Many would think that video games that were rated highly should have done well in sales. If many people liked a game, more of it would be bought, right? Through this part of the capstone project, the data was looked at to see how well it represented that statement. On top of looking at just the Metacritic scores and the global sales of video games, many other video game variables were also analyzed to see how they were related to global sales.

Metacritic is a media rating website that gives two different types of scores to a video game; critic scores and user scores. The critic score is derived from professional critics and the user score is derived from the users of the site. A curiosity that came to mind was whether or not there is a difference between the two scores? Plotting the histogram of these two scores showed that there was a higher frequency of critic scores below 75 and a higher frequency of user scores above 75. For this dataset, The average of the critic scores and the average of the user scores is 70.45 and 74.34, respectively. This gives a difference of average scores of 3.88, in favor of user scores. On average, user scores appeared to score higher than critic scores. But how can we know that this difference of averages wasn't just due to chance? By performing a hypothesis test!

For the hypothesis test, the null hypothesis was as follows: The average of critic scores and user scores is the same. The alternative hypothesis was: There is a difference between the average scores of critics and users. The significance level for this test was 0.05. After simulating a dataset where the means between critic and user scores were the same, it was found that the probability, the p-value, of observing our actual difference of 3.75 was practically zero. The p-value was below our significance level of 0.05, allowing for the rejection of the null hypothesis and acceptance of the alternative that there is a statistically significant difference between the average scores of critics and users.

Next, the original question of how Metacritic scores related to global sales was looked at. For the analysis, the critic scores were used to observe their effects on a video game's sales. It was found that in order to visualize the data better and to make a more significant correlation, the log of global sales had to be used. A linear regression model was fitted to this data and a hypothesis test with a significance level of 0.05 was performed. The null hypothesis was that there is no correlation between log of global sales and critic scores. The alternative hypothesis is that there is a correlation between them. Using the python module, scipy stats, the p-value and R^2 value were calculated and a line was fitted to the data. The p-value was found to be practically zero, allowing for the rejection of the null hypothesis and acceptance of the alternative hypothesis of there being a correlation. The R^2 was calculated to be 0.17, or 17%. This means that 17% of the log of global sales variation could be explained by critic scores in the linear regression. A quick look at the R^2 for user scores and global sales revealed it to be 5%. These R^2 values indicate that the linear regression did not fit the data very well and that there may be other variables that could also contribute to the global sales.

The next relationship that was explored was the linear regression of the total count of video games for each publisher with their respective total sales. It was found that in order to represent the data appropriately, the log of both global sales and video game counts had to be taken. The hypotheses for this test was the same as the previous linear regression test. The p-value was found to be practically zero and the null hypothesis was rejected. The hypothesis of a correlation existing was accepted. The linear regression line was a decent fit for this analysis. The R^2 value between the log of video game counts per publisher and the log of global sales was calculated to be 69%.

Another video game variable that was analyzed was the genre. There are 12 possible genres that the video games in the dataset could be categorized into. The video games in the dataset were grouped up by their genres and aggregated by the sum of their global sales. Action games were seen to have the highest global sales by nearly 400 millions units sold over the next 2 highest selling genres. It was found that about 20% of all video games in the dataset were action games. Through bootstrap resampling the 95% confidence interval of the true action game proportion was calculated to be 18.6%-21% . A hypothesis test with a significance level of 0.05 was also performed to see if this observed proportion occurred by chance. The null hypothesis was: The proportion of all the genres are equal, action games have a proportion of 1/12, or 8%. The alternative hypothesis was: There is a difference in the proportions of video game genres, action game proportion does not equal 1/12, or 8%. The p-value was calculated to be 0.00, so the null hypothesis was rejected. The large proportion of video games being action could explain why the total global sales of action video games was so high, relative to the other genres. Looking at the average, rather than the sum total, global sales of video game genres should be a better representation of the relationship.

Looking at the top two highest average sales of genres, shooter and sports, the standard deviation of each were calculated to see if there was a difference between the dispersion of the two genres. The standard deviations of global sales for shooter and sports were 3.52 and 2.70, respectively, with a difference of 0.82. For the hypothesis test the significance level will be 0.05. The null hypothesis is: There is no difference between the standard deviations of shooter and sports global sales. The alternative hypothesis is: There is a difference between the standard deviations of shooter and sports global sales. Through permutation resampling, the p-value was calculated to be 0.11. This value is larger than the significance level, therefore the null hypothesis can be accepted. There is no statistically significant difference between the standard deviation of global sales for shooter and sports video games.

Throughout the statistical analysis of the video game dataset, many inferences and correlations were found. It was seen that while there was a correlation between Metacritic scores and global sales of video game genres, there were also other factors that affected a video game's sales. A significant correlation in the amount of video games a publisher makes and their total global sales was found. Initially, it seemed that action games affected global sales the most. But it was determined that the reason for the large value of global sales was because a large proportion of the video games in the dataset were action games. Looking at the average global sales, instead, showed that shooters and sports games performed the best in sales. It was also shown that the standard deviation of global sales for shooters and sports were not statistically significantly different.