[redacted]

Intro to Probability and Statistics

MATH-1223_01

Dr. Jacob

Global Sales of Top 25 Video Games

## Part One

The data set that I have provided was obtained from a data website called kaggle. This data set was last updated 5 years ago in 2016 by the creator of the set, Gregory Smith. Some variables that are seen in this set include categorical, such as different companies that these games were published from, the genre of the games, and how well they sold across North America, Europe, Japan, as well as some extra sales from other locations. Another example of variable is ordinal, which would be the year the games were released. The platforms that the games were released on are also categorical, more specifically nominal. I am interested in the data mainly because I want to see how many of these games are from Nintendo since I grew up playing Nintendo games and they seem to be so loved among video game fans. I am hoping to compare the differences in how well the platforms did as well as their publishers, and to see if one region enjoyed some games more than others. To make things clear, the sales are by the millions.

| Rank | Name | Platform | Year | Genre | Publisher | NA_Sales | EU_Sales | JP_Sales | Other_Sales | Global_Sales |
|---|---|---|---|---|---|---|---|---|---|---|
| **1** | Wii Sports | Wii | 2006 | Sports | Nintendo | 41.49 | 29.02 | 3.77 | 8.46 | 82.74 |
| **2** | Super Mario Bros. | NES | 1985 | Platform | Nintendo | 29.08 | 3.58 | 6.81 | 0.77 | 40.24 |
| **3** | Mario Kart Wii | Wii | 2008 | Racing | Nintendo | 15.85 | 12.88 | 3.79 | 3.31 | 35.82 |
| **4** | Wii Sports Resort | Wii | 2009 | Sports | Nintendo | 15.75 | 11.01 | 3.28 | 2.96 | 33 |
| **5** | Pokemon Red/Pokemon Blue | GB | 1996 | Role-Playing | Nintendo | 11.27 | 8.89 | 10.22 | 1 | 31.37 |
| **6** | Tetris | GB | 1989 | Puzzle | Nintendo | 23.2 | 2.26 | 4.22 | 0.58 | 30.26 |
| **7** | New Super Mario Bros. | DS | 2006 | Platform | Nintendo | 11.38 | 9.23 | 6.5 | 2.9 | 30.01 |
| **8** | Wii Play | Wii | 2006 | Misc | Nintendo | 14.03 | 9.2 | 2.93 | 2.85 | 29.02 |
| **9** | New Super Mario Bros. Wii | Wii | 2009 | Platform | Nintendo | 14.59 | 7.06 | 4.7 | 2.26 | 28.62 |
| **10** | Duck Hunt | NES | 1984 | Shooter | Nintendo | 26.93 | 0.63 | 0.28 | 0.47 | 28.31 |
| **11** | Nintendogs | DS | 2005 | Simulation | Nintendo | 9.07 | 11 | 1.93 | 2.75 | 24.76 |
| **12** | Mario Kart DS | DS | 2005 | Racing | Nintendo | 9.81 | 7.57 | 4.13 | 1.92 | 23.42 |
| **13** | Pokemon Gold/Pokemon Silver | GB | 1999 | Role-Playing | Nintendo | 9 | 6.18 | 7.2 | 0.71 | 23.1 |
| **14** | Wii Fit | Wii | 2007 | Sports | Nintendo | 8.94 | 8.03 | 3.6 | 2.15 | 22.72 |
| **15** | Wii Fit Plus | Wii | 2009 | Sports | Nintendo | 9.09 | 8.59 | 2.53 | 1.79 | 22 |
| **16** | Kinect Adventures! | X360 | 2010 | Misc | Microsoft Game Studios | 14.97 | 4.94 | 0.24 | 1.67 | 21.82 |
| **17** | Grand Theft Auto V | PS3 | 2013 | Action | Take-Two Interactive | 7.01 | 9.27 | 0.97 | 4.14 | 21.4 |
| **18** | Grand Theft Auto: San Andreas | PS2 | 2004 | Action | Take-Two Interactive | 9.43 | 0.4 | 0.41 | 10.57 | 20.81 |
| **19** | Super Mario World | SNES | 1990 | Platform | Nintendo | 12.78 | 3.75 | 3.54 | 0.55 | 20.61 |
| **20** | Brain Age: Train Your Brain in Minutes a Day | DS | 2005 | Misc | Nintendo | 4.75 | 9.26 | 4.16 | 2.05 | 20.22 |
| **21** | Pokemon Diamond/Pokemon Pearl | DS | 2006 | Role-Playing | Nintendo | 6.42 | 4.52 | 6.04 | 1.37 | 18.36 |
| **22** | Super Mario Land | GB | 1989 | Platform | Nintendo | 10.83 | 2.71 | 4.18 | 0.42 | 18.14 |
| **23** | Super Mario Bros. 3 | NES | 1988 | Platform | Nintendo | 9.54 | 3.44 | 3.84 | 0.46 | 17.28 |
| **24** | Grand Theft Auto V | X360 | 2013 | Action | Take-Two Interactive | 9.63 | 5.31 | 0.06 | 1.38 | 16.38 |
| **25** | Grand Theft Auto: Vice City | PS2 | 2002 | Action | Take-Two Interactive | 8.41 | 5.49 | 0.47 | 1.78 | 16.15 |

Link to data set:
https://www.kaggle.com/gregorut/videogamesales

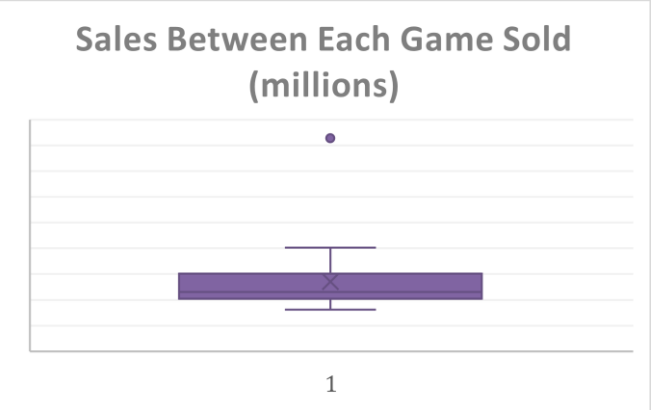
## Part Two

For the second part of this project, I will include a frequency and relative frequency table for one of my nominal categorical variables, which will be the number of 25 top games released relative to the publisher.

| Publisher | Frequency | Relative Frequency |
|---|---|---|
| **Microsoft Game Studios** | 1 | 0.04 |
| **Nintendo** | 20 | 0.8 |
| **Take-Two Interactive** | 4 | 0.16 |
| **Total** | **25** | **1** |

Following this table is a two-way table for two other categorical variables from this set; I will look at the genre of these games as well as the publisher and how often they show up on the table.

| | Nintendo | Microsoft Game Studios | Take-Two Interactive | Total |
|---|---|---|---|---|
| **Sports** | 4 | 0 | 0 | 4 |
| **Platform** | 6 | 0 | 0 | 6 |
| **Racing** | 2 | 0 | 0 | 2 |
| **Role-playing** | 3 | 0 | 0 | 3 |
| **Action** | 0 | 0 | 4 | 4 |
| **Misc.** | 2 | 1 | 0 | 3 |
| **Shooter** | 1 | 0 | 0 | 1 |
| **Simulation** | 1 | 0 | 0 | 1 |
| **Puzzle** | 1 | 0 | 0 | 1 |
| **Total** | **20** | **1** | **4** | **25** |



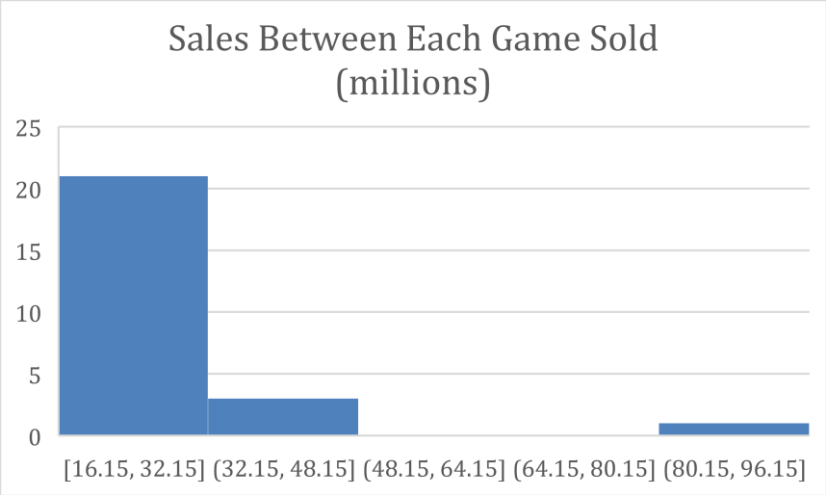

It is very apparent from this table that Nintendo had the most variety among genres compared to the other two publishers, but that should not come as a surprise considering Nintendo also has the most games sold.

## Part Three

In this next section I will include both a box plot and a histogram for some quantitative data, which is the amount of global sales per game.

The outlier of this box-and-whisker plot is *Wii Sports* for the Nintendo Wii, selling almost double the amount of the game in second place, *Super Mario Bros.* for the Nintendo Entertainment System.

The following is a histogram of the same data:




The five number summary for this data includes the following:

- Minimum: **16.15**
- Q1: **20.415**
- Median: **23.1**
- Q3: **30.135**
- Maximum: **82.74**

The mean is **27.0624** and the standard deviation according to Excel is **12.93206**. Keep in mind that these are by the millions. There is also a lone outlier that is a lot more than the rest of the games. The histogram appears to be skewed right, but it is difficult to tell since the histogram grouped together the sales the way it did.

## Part Four

For the next part of this project, I will set up hypothesis tests for both quantitative and categorical variables. As for my quantitative alternative hypothesis, I will say that the average sales of Japan is 3.77 million units of Wii Sports. My quantitative null hypothesis will be that the average sales of Japan was not 3.5 million units for the top 25 video games.

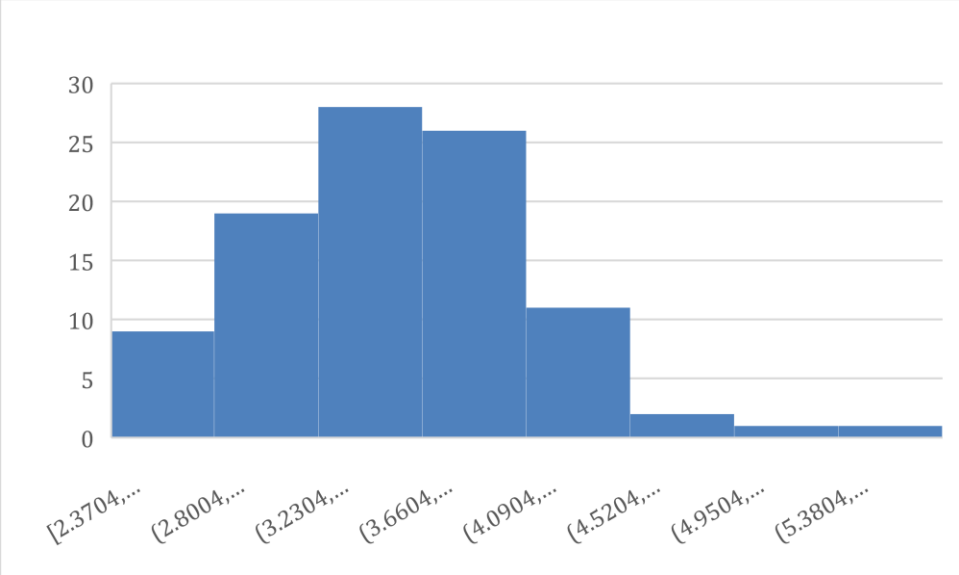
$H_0$: $\mu = 3.5$
$H_a$: $\mu \neq 3.5$

My categorical null hypothesis is that the proportion of games published by Nintendo is 0.65 or more. Out of the 25 games on the list, 20 of them are from Nintendo while the other 5 are from the two other publishers. My categorical alternative hypothesis is that the proportion of games published by Nintendo is less than 0.65.

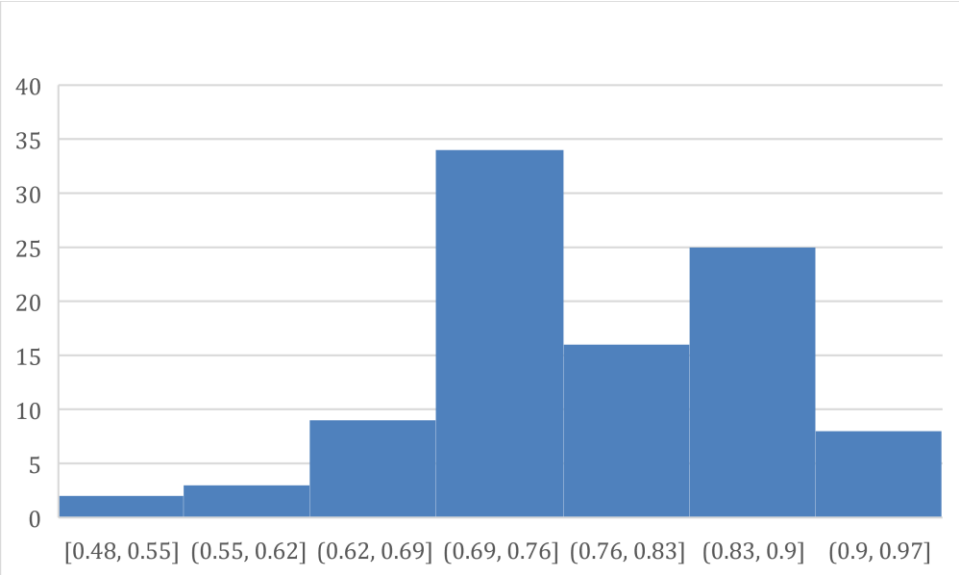
$H_0$: $p = 0.65$
$H_a$: $p < 0.65$

## Part Five

In order to test my null hypotheses, I will use a method called bootstrapping to find the standard error of the data. The following is a histogram which was made using my quantitative null hypothesis, which is that the average sales of video games in Japan is not 3.592 million. I gathered around 100 bootstrapping samples from the average sales of all video game sales in Japan and returned a standard error of 0.493131. The confidence interval for a 95% confidence level was 2.423276-4.655512. I can reject my quantitative null hypothesis.




My categorical null hypothesis is that the proportion of games published by Nintendo is 0.65. Here I did essentially the same thing as my quantitative null hypothesis; I gathered 100 bootstrapping samples from all of the 25 video games sales and determined whether or not the game was from Nintendo. Standard error was 0.09185. Confidence interval for 95% confidence level was 0.595269-0.962669. The p-value is less than the lowest point of the confidence level, so the null hypothesis is rejected.



## Part Six

For part six of this project, I will be repeating the categorical hypothesis test from the previous part while utilizing the appropriate formulas. My categorical null hypothesis is that the proportion of games published by Nintendo is 0.65. The two most important formulas I will be using is standard error, which is SE = sqrt{p(1-p)/n}and Z = (p hat-p)/SE. My sample has 25 observations, which means n=25. The p=0.65. The p-hat is 0.80. The formula for the SE would be SE= SQRT(0.65*(1-0.65)/25) which equals 0.095394. I can now input this into the Z score formula, Z=(0.80-0.65)/0.095394. The result I got was 1.572427. Using Excel, I was able to find the P-value from the Z score, which was 0.942074. Since this is higher than the significance level (0.05), I failed to reject the null hypothesis.

## Part Seven

I will be repeating this same process for part seven, only now I will be using the quantitative null hypothesis. My quantitative null hypothesis will be that the average sales of video games in Japan was not 3.5 million units.

$H_0$; $\mu = 3.5$
$H_a$; $\mu \neq 3.5$

The first step here is to find the standard error. The standard error can be found using SE = $\sigma \sqrt{n}$. Since we are using a 95% confidence level for this formula, the **$\sigma = 0.05$**. Therefore, SE=0.05/SQRT 25. This equals **0.01**. Then we can find the 95% confidence interval, which is the point estimate ± z*SE. The point estimate would be the same as the sample mean, which for my data would be the average of all the game units sold in Japan, 3.952 million.

3.952 ± 1.96*0.01 = 3.9324-3.9716

I'm going to go ahead and calculate the degrees of freedom, which would be 24 in this case since there are 25 samples. Using a t table, 2.06 is the critical value. I was also able to find the standard deviation of the sample using Excel, which is 2.50267. I can now calculate the test statistic using this information: t= (x-bar / mu)/SD/SQRT(n).

(3.952-3.5)/2.50267/SQRT(25)= 0.903035558

Since the test statistic is less than the critical value, I failed to reject the null hypothesis that the average sales of video games in Japan was not 3.5 million units since the t value was not more extreme than the critical t value. I was able to reject the null hypothesis with my bootstrapping results unlike I was here.

## Part Eight

Using the two-way table from part two, I will create 2 conditional probabilities and interpret their meanings and how they were computed. The formula for conditional probability is $P(A|B) = [P(A \cap B)] / [P(B)]$

I will include the same table for ease of viewing:

| | Nintendo | Microsoft Game Studios | Take-Two Interactive | Total |
|---|---|---|---|---|
| **Sports** | 4 | 0 | 0 | **4** |
| **Platform** | 6 | 0 | 0 | **6** |
| **Racing** | 2 | 0 | 0 | **2** |
| **Role-playing** | 3 | 0 | 0 | **3** |
| **Action** | 0 | 0 | 4 | **4** |
| **Misc.** | 2 | 1 | 0 | **3** |
| **Shooter** | 1 | 0 | 0 | **1** |
| **Simulation** | 1 | 0 | 0 | **1** |
| **Puzzle** | 1 | 0 | 0 | **1** |
| **Total** | **20** | **1** | **4** | **25** |

For my first conditional probability, I would like to find the probability of obtaining a platform game that was published by Nintendo selected completely at random. I can use this information to create the equation needed to find the probability:

$$\mathbf{P}(\text{Platform} \mid \text{Nintendo}) = \mathbf{P}(\text{Platform} \cap \text{Nintendo}) / \mathbf{P}(\text{Nintendo}) = 6/20 = 0.3 = 30\%$$

My second conditional probability will be the probability of obtaining a game published by Take-Two Interactive that is an action game.

$$4/4 = 1 = 100\%$$

Unfortunately for the other two publishers, they only seem to excel at one genre each, so the odds of picking up an action Take-Two Interactive game is 100% since they only have four games in the list and all four are action games.