[redacted]

Intro to Probability and Stats

Statistics Project

Dr. Jacobs

Video Game Sales in North America and Japan

The data I will be covering is over the top 100 bestselling videogames as of 2017. I will be referencing GitHub account nurfnick as all of data will be collected from the spreadsheet linked below. My categorical variables will be the game publisher and game console while the quantitative data is the North American sales and Japanese sales. For example, Pokémon Platinum Version is published by Nintendo, the console was the DS, it sold 2.82 million copies in North America and sold 2.69 million copies in Japan. I chose this data set because I find it interesting and nostalgic to look back on some of the most popular games in the past, and it's also something I actually keep up with every now and then. What I hope to learn and want to look deeper into is comparing sales of different games in North America and Japan. It would be interesting to see how games developed in North America sell in Japan and how games developed in Japan sell in North America and compare the two. Here is some data for the categorical variables, showing how many games were published by each company as well as how many games per console for the top 100 bestselling games. Here is a sample of the statistics showcasing the top 10 bestselling games.

Rank	Name	Platform	Publisher	NA_Sales	JP_Sales
1	Wii Sports	Wii	Nintendo	41.49	3.77
2	Super Mario Bros.	NES	Nintendo	29.08	6.81
3	Mario Kart Wii	Wii	Nintendo	15.85	3.79
4	Wii Sports Resort	Wii	Nintendo	15.75	3.28
5	Pokémon Red/Pokémon Blue	GB	Nintendo	11.27	10.22
6	Tetris	GB	Nintendo	23.2	4.22
7	New Super Mario Bros.	DS	Nintendo	11.38	6.5
8	Wii Play	Wii	Nintendo	14.03	2.93
9	New Super Mario Bros. Wii	Wii	Nintendo	14.59	4.7
10	Duck Hunt	NES	Nintendo	26.93	0.28

Here is some data for the categorical variables, showing how many games were published by each company as well as how many games per console for the top 100 bestselling games. I also made a two-way table to show how some of the top publisher sells on some of the top console. Obviously, publishers like Nintendo will sell well on their own consoles (Nintendo publishes games and sells their own consoles as well which primarily features games published by them, and while third-party games do exist on Nintendo consoles, they rarely sell that much which is reflected in the data sample below). Consoles like the Xbox and PS2 have less first-party games so there will be more diversity when it comes to publishers.

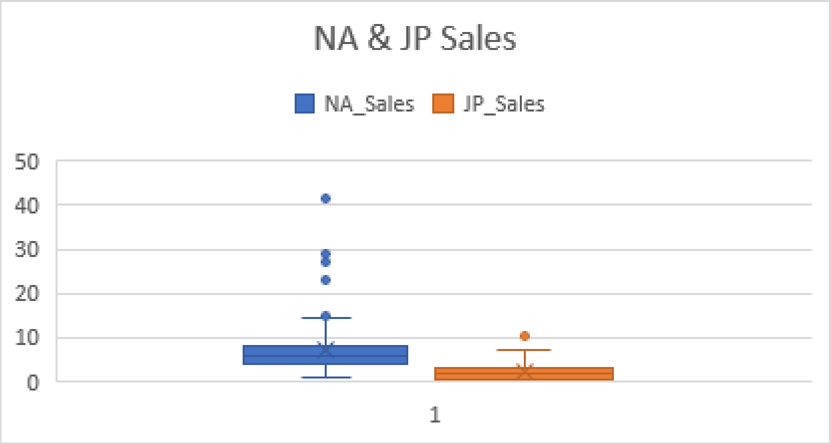
Publisher	Frequency	Relative Frequency
Nintendo	52	0.52
Microsoft Game Studios	6	0.06
Take-Two Interactive	9	0.09
Activision	14	0.14
Sony Computer Entertainment	8	0.08
Ubisoft	2	0.02
Electronic Arts	5	0.05
Atari	1	0.01
Sega	1	0.01
Bethesda Softworks	1	0.01
SquareSoft	1	0.01
Total	100	1

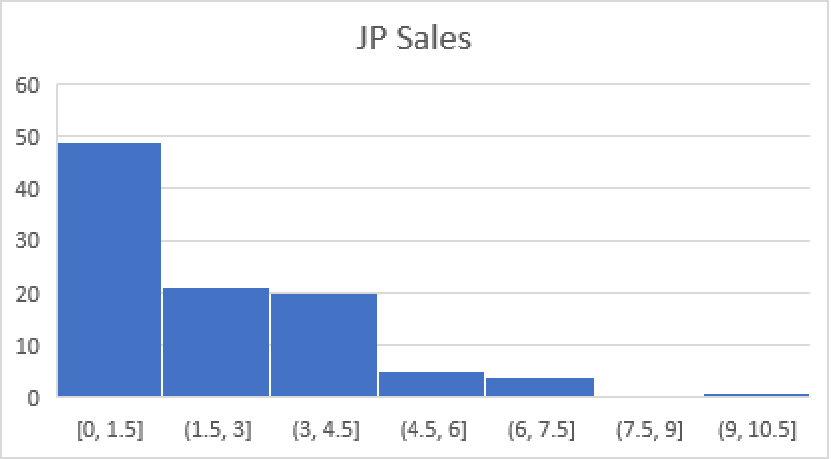
Console	Frequency	Relative Frequency
Wii	15	0.15
NES	4	0.04
GB	6	0.06
3DS	7	0.07
GBA	2	0.02
DS	13	0.13
X360	16	0.16
PS	5	0.05
PS2	6	0.06
PS3	9	0.09
PC	1	0.01
PS4	5	0.05
PSP	1	0.01
2600	1	0.01
N64	4	0.04

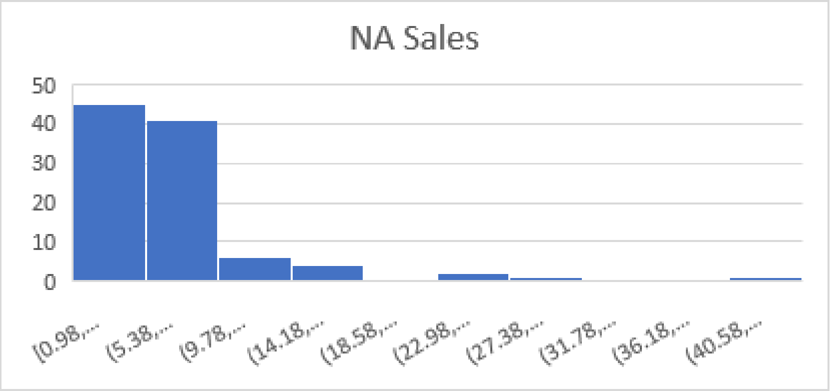
SNES	4	0.04
XB	1	0.01
Total	100	1

	Wii	NES	X360	PS3	PS2
Nintendo	12	4	0	0	0
Microsoft Game Studios	0	0	5	0	0
Take-Two Interactive	0	0	2	2	3
Sony Computer Entertainment	0	0	0	1	3
Activision	0	0	7	5	0

Here are my quantitative variables. The box plot below compares NA and JP sales for the top 100 best selling games. There is are two histograms below show the sales of NA and JP sales individually (measured in the millions). According to the box plot, there seems to be 6 outliers. The histograms are both skewed to the right. From the looks of it, North American sales seem to be much higher than Japanese sales for the most part. North America is represented by “NA” while Japan is represented by “JP”.







Here is the 5-number summary, mean, median, mode, and standard deviation. If you look at the minimum for Japan, you'll notice that it's 0. That's because some games like Just Dance 3 didn't sell at all in Japan for whatever reason.

Column1	**NA**	**JP**
Mean	7.0499	2.0462
Median	5.675	1.635
Mode	4.99	0.13
Standard Deviation	5.87323	2.048607
Range	40.51	10.22
Minimum	0.98	0
Q1	3.8625	0.23
Q3	8.29	3.1475
Maximum	41.49	10.22

For my quantitative null hypothesis, I believe the average amount of copies sold for the top 10 best-selling games in North America is 20 million. My alternative hypothesis is that the average amount of copies sold for the 10 best-selling games in North America is not in 20 million.

H_0: $\mu = 20$

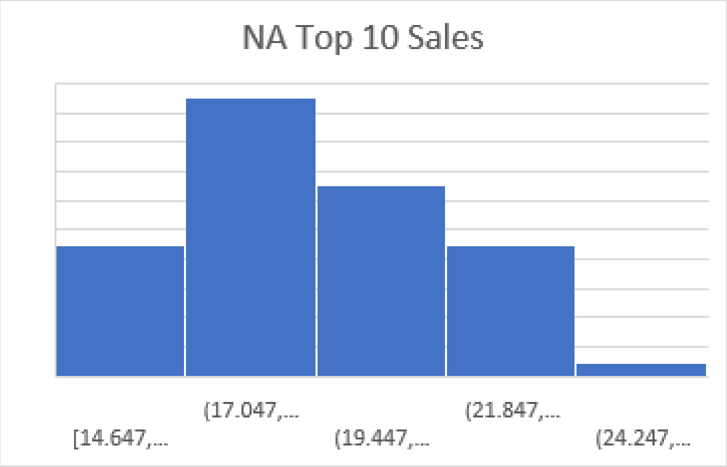
Ha: $\mu \neq 20$

For my categorical null hypothesis, I believe more than 5 of the top 10 best-selling games in North America are from Nintendo despite the fact that they come from a different country. If my hypothesis is correct, then that would mean that there are more people buying foreign games in North America rather than games published in North America. It would also mean that US sales are imperative for Nintendo as North America has a much higher population. My alternative hypothesis is that there are less than 5 Nintendo games in North America's top 10 bestselling games.

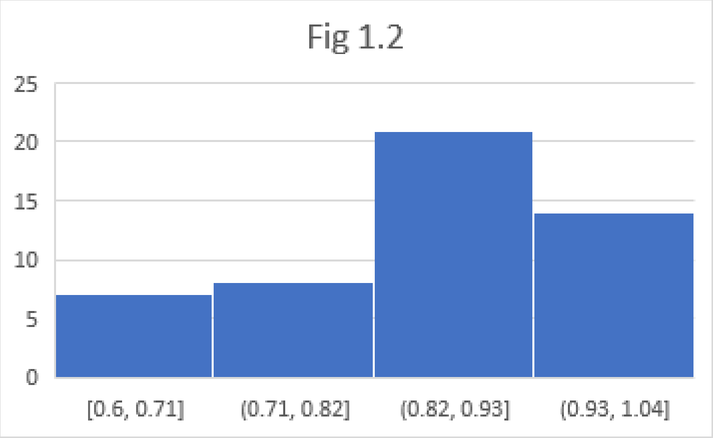
H_0: $p > 0.5$

Ha: $p < 0.5$

To test my quantitative null hypothesis, I created 100 bootstrap samples containing the number of copies sold for the top 10 best-selling games in North America. The standard error was 2.44883 and the 95% confidence interval was 15.209-25.0043, so I think we can conclude my null hypothesis of the average being 20 million is correct.



To test my categorical null hypothesis, I made 100 bootstrap samples containing the top 10 best-selling games in North America, with games published by Nintendo being represented with a “1” and other games being represented by a “0”. The standard error was 0.082 and the 95% confidence interval was 73%-100%, so yes, we can conclude that my null hypothesis is correct and more than half of the top 10 best-selling games in North America were published by Nintendo.



My categorical null hypothesis stated that at least more than half of top 10 best-selling games in North America were published by Nintendo. This time, rather than using a bootstrap sample, I'll be using equations in Excel to find standard error, z-score, 95% confidence interval, p, z*, and p-hat. The n will be 10 since it's the top 10 best-selling games. My predicted 0.5 of the top ten games being from Nintendo will be represented by "p". The p-hat is found by counting the number of Nintendo games in the top ten (9) and dividing it by the total (10) which 0.9. The standard error was found using the formula SQRT(p(1-)/n), which gave me an SE of 0.158114. Using the formula (p-hat - p)/SE) I was able to find the z-score, which was 2.529822. Next, I found the p-value using excel, which was 0.807021. The 95% confidence interval is found with the formula CI= p – (z*SE) and p+(z*SE), with the lower bound interval being 0.183772 and the upper bound being 0.8162278. We can conclude that my null hypothesis is correct, just like in the bootstrapping results.

Column1	**Column2**	**Column3**	**Column4**	**Column5**
n	p	p-hat		95% CI
10	0.5	0.9		0.1837722
				0.8162278
SE		z		p
SQRT(p(1-)/n)		(p-hat - p)/SE)		
0.1581139		2.5298221		0.994294
				0.005706

My quantitative null hypothesis stated that the average amount of copies sold for the top 10 best-selling games in North America is 20 million. I will test my hypothesis by using formulas to attain the x-bar, t-test statistic, standard error, T*, 95% confidence interval, and x. Each number will be represented in millions, so for example, 10 will equal 10,000,000. First, I got my x-bar by getting the average for the top 10 games which was 20.357. I got the standard deviation with the STDEV.S formula in Excel and got 9.7294982. I found the standard error with the formula (SD/sqrt(n)) and got 3.0767375. I used the formula (x-bar - mu)/se to find the t statistic, which was 0.357. To find the t* I used the formula T.INV(0.975, n-1) and got 2.2621572. Now I'm ready to find the confidence interval, which is found with the formula (x-bar – (t*)*se) for the lower bound, which was 13.396936, and (x-bar + (t*)*se) for the upper bound, which was 27.317064. So, we can conclude my null hypothesis is correct, just like in the bootstrapping results.

	x-bar	**SD**	**n**	**mu**	**a**
	20.357	9.7294982	10	20	0.05
SE	3.0767375				
t	0.357				
t*	2.2621572				
CI upper	27.317064				
CI lower	13.396936				

I will create 2 conditional probabilities using my two-way table. Again, the data used in this table comes from the top 100 best-selling games overall.

	Wii	NES	X360	PS3	PS2
Nintendo	12	4	0	0	0
Microsoft Game Studios	0	0	5	0	0
Take-Two Interactive	0	0	2	2	3
Sony Computer Entertainment	0	0	0	1	3
Activision	0	0	7	5	0

In order to find the conditional probability, I must use the formula: P(A|B) = P(A & B)/P(B). What I'm going to be looking for first is the probability of getting a PS2 game given that it was published by Take-Two Interactive if I were to randomly select it. The equation would be P(PS2|Take-Two Interactive) = P(PS2 & Take-Two Interactive)/P(Take-Two Interactive) which equals 0.43 or 3/7. This means there is a 43% chance I would pick a PS2 game given it was a game published by Take-Two Interactive. The second conditional probability I will be looking for is the probability of getting a game published by Activision given that it was an X360 game if I were to randomly select it. The equation would be P(Activision|X360) = P(Activision & X360)/P(X360) which equals 0.50 or 7/14. This means there is a 50% chance I would pick a game published by Activision given it was an X360 game.

References-

https://github.com/nurfnick/Data_Sets_For_Stats/blob/7509aa499ae7f0232fb0975dd8587ca8d023ac20/videogames.xlsx