Video Game Sales

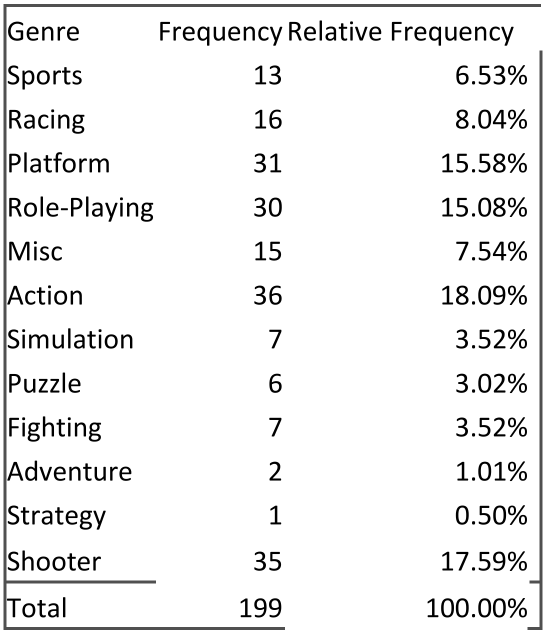
[redacted]

Intro to Probability and Statistics

Nicholas Jacob

I got the data from Dr. Jacob's GitHub and the data is video game sales. The categorical variables in the dataset are platform, genre, publisher, rankings, and the year it was released. The quantitative variables are North America sales, Europe sales, Japan sales, other sales, and global sales. The questions I would like to answer from this data set are which platform had the highest global sales, which genre had the highest global sales, which publisher had the highest global sale, and which genre sells the most in North America.

Below there are four tables. the first three tables are frequency and relative frequency for genre, platform , and publisher The last table is a two way table. The data in the two way table is platform and publisher. I can use this data with sales to determine which publisher has the highest sales and I can determine with sales which platform has the highest sales.



Genre	Frequency	Relative Frequency
Sports	13	6.53%
Racing	16	8.04%
Platform	31	15.58%
Role-Playing	30	15.08%
Misc	15	7.54%
Action	36	18.09%
Simulation	7	3.52%
Puzzle	6	3.02%
Fighting	7	3.52%
Adventure	2	1.01%
Strategy	1	0.50%
Shooter	35	17.59%
Total	199	100.00%

Platform	Frequency	Relative frequency
Wii	21	10.24%
NES	6	2.93%
GB	11	5.37%
DS	19	9.27%
X360	28	13.66%
PS3	27	13.17%
PS2	18	8.78%
SNES	6	2.93%
GBA	7	3.41%
3DS	8	3.90%
PS4	8	3.90%
N64	7	3.41%
PS	14	6.83%
3DS	8	3.90%
XB	2	0.98%
PC	4	1.95%
PSP	3	1.46%
Xone	3	1.46%
GC	3	1.46%
WiiU	2	0.98%
Total	205	100.00%

Publisher	Frequency	Relative Frequency
Nintendo	80	42.11%
Microsoft Game Studios	15	7.89%
Take-Two Interactive	13	6.84%
Sony Computer Entertainment	17	8.95%
Activision	21	11.05%
Ubisoft	9	4.74%
Bethesda Softworks	3	1.58%
Electronic Arts	16	8.42%
Sega	3	1.58%
Atari	1	0.53%
505 Games	2	1.05%
Capcom	3	1.58%

LucasArts	1	0.53%
Konami Digital Entertainment	3	1.58%
Square Enix	3	1.58%
Total	190	1

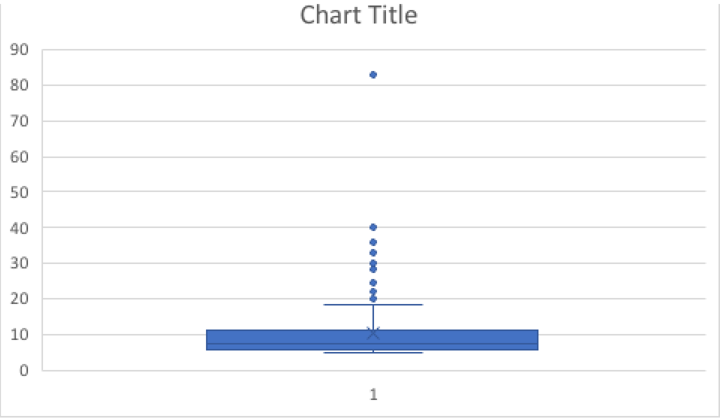
	Nintendo	Ubisoft	Activision	Electronic arts	Sega
Wii	14	4	0	0	1
NES	6	0	0	0	0
GB	11	0	0	0	0
DS	17	0	0	0	1
X360	0	3	7	2	0
PS3	0	2	7	4	0
PS2	0	0	0	6	0
SNES	5	0	0	0	0
GBA	7	0	0	0	0
3DS	8	0	0	0	0
PS4	0	0	3	3	0
N64	7	0	0	0	0
PS	0	0	0	0	0
3DS	8	0	0	0	0
XB	0	0	0	0	0
PC	0	0	2	1	0
PSP	0	0	0	0	0
Xone	0	0	2	0	0
GC	3	0	0	0	0
WiiU	2	0	0	0	0

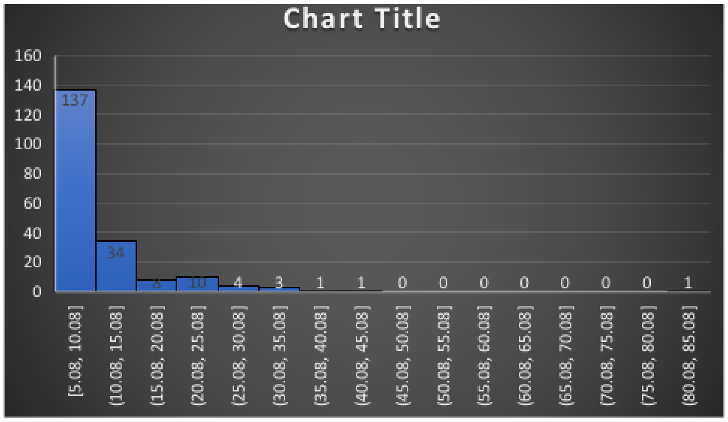
Below is the data for the top 200 global sales, which is mean, standard deviations, and the five number summary and two graphs showing the data.

Mean	43.91
Standard deviations	54.91391263
Min	5.08
Q1	5.88
Median	43.91
Q3	11.255

Max	82.74

I couldn’t change the tables to better represent the data because I’m using excel on my iPad Pro and it doesn’t allow me to change the range of numbers.





Quantitative variable

Ho μ NAS = 8 million

Hp μ NAS $\neq$ 8 millions

My hypothesis is that the average sales in North America is 8 million. My alternative hypothesis would then be that it does not equal 8 million. For this I calculated the mean of North America sales and I got 4.9 million sold. My hypothesis would be wrong .

Categorical Variable

Ho: proportion, p, is the proportion of video games sold on Nintendo's platforms p =.5

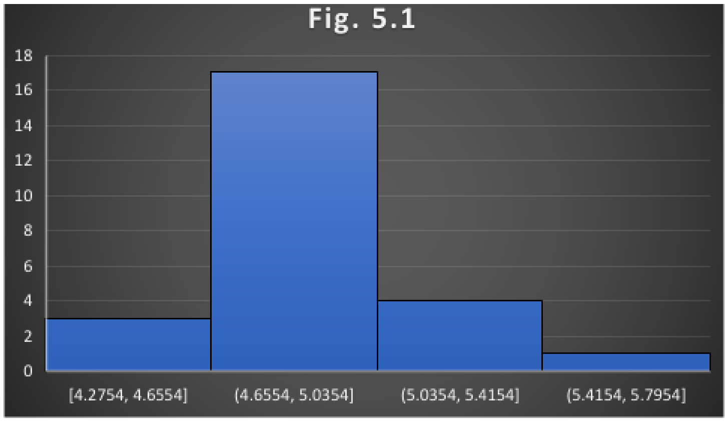
Ha: proportion, p, is the proportion of video games sold on Nintendo's platforms p $<$ .5

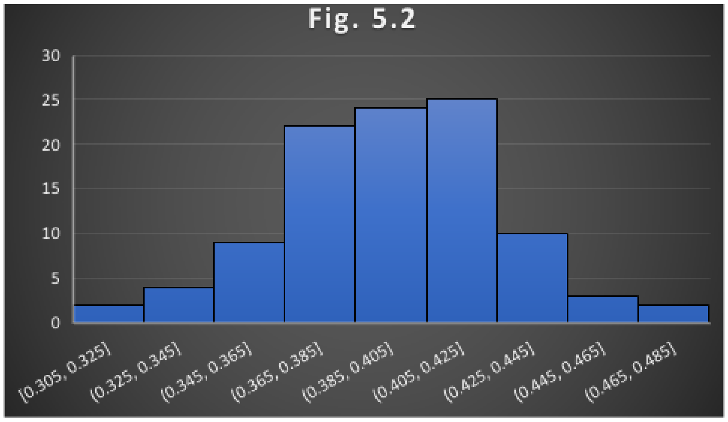
My hypothesis is that the proportion of games sold on a Nintendo's platforms is equal to 50% of the total. My alternative hypothesis would then be that the proportion of games sold on a Nintendo's platforms is less than 50% of the total . To do this I will add up Nintendo's platforms. The platforms included are Wii, NES, GB, GBA, SNES, DS, 3DS, N64, GC, and Wii U. The answer that I got is 40% of the total. My hypothesis would be wrong.

Bootstrap Sampling

Going back and testing the quantitative hypothesis using bootstrap sampling I found the standard error for the sample to be .32014. I then calculated the 95% confidence interval for the mean is between 4.24 and 5.52. My conclusion is with this data we can reject the null hypothesis because the estimated mean is outside the confidence interval. **Figure 5.1** below is the histogram for the quantitative variable bootstrap distribution.

I also tested the categorical hypothesis using bootstrapping. I found the standard error to be .035. I then calculated The 95% confidence interval which was between .335 and .46. My conclusion is with this data we can reject the null hypothesis, because the estimated mean is outside the confidence interval. **Figure 5.2** below is the histogram for the categorical variable bootstrap distribution.





Project Part 6

Im retesting my categorical variable hypothesis using formulas in excel. My hypothesis was that the proportion of games sold on Nintendo’s platforms is equal to 50% of the total. My alternative hypothesis was that the proportion of games sold on Nintendo’s platforms is less than 50% of the total. My standard error from using the formulas is .03536. My 95% confidence interval is .32929 at the lower end to .47071 at the higher end. I also calculated p and it was .00234. Comparing.00234 to the alpha which is .05 I can reject the null hypothesis, because.00234 is less than .05. Comparing the formulas test to the bootstrapping test results the 95% confidence intervals are slightly different. The 95% confidence interval for the bootstrapping was .335 for lower end to .46 at the higher end. As you can see the confidence intervals are very similar and I think one of the reasons for the slight variation is that I used 2 in my calculations rather then 1.96. Using either formulas test or the bootstrapping test I can reject my null hypothesis, because .5 is not in the confidence interval. The formulas I used are

SE = SQRT(p*(1-p)/n), Z= (p-hat – p)/SE, P= Norm.S.Dist(Z, true), P = 1-P

CI = p – (Z*SE) and p+(Z*SE)

Project part 7

For part 7 of the project I am retesting my quantitative variable using a t test. My hypothesis was that the average sales in North America is 8 million. My alternative hypothesis was that it does not equal 8 million. My assumed mean was 8 million and my sample mean was 4.8916. My t- test was -9.2985 when I calculated it. My standard error was .33429 and my standard deviation was 4.72758. My 95 % confidence interval was 4.23 at the low end and 5.55 at the higher end. Comparing that to my bootstrap 95% confidence interval, which was 4.24 at the lower end and 5.52 at the higher end, they are almost identical. Comparing the bootstrap standard error, which was .32014, they are very close. Whichever test I use I can reject my null hypothesis, because it is not in the confidence interval. The equations I used are

T-test = (x – u/(SD/ SQRT(n)),   SE = (SD/SQRT(n), T* = T.INV(alpha, n-1), 95% CI = x + (t*) * SE, and x- (t*) * SE

Project part 8

For this part I'm going to create two conditional probabilities using my two way table.

	Nintendo	Ubisoft	Activision	Electronic arts	Sega
Wii	14	4	0	0	1
NES	6	0	0	0	0
GB	11	0	0	0	0
DS	17	0	0	0	1
X360	0	3	7	2	0
PS3	0	2	7	4	0
PS2	0	0	0	6	0
SNES	5	0	0	0	0
GBA	7	0	0	0	0
3DS	8	0	0	0	0
PS4	0	0	3	3	0
N64	7	0	0	0	0

PS	0	0	0	0	0
3DS	8	0	0	0	0
XB	0	0	0	0	0
PC	0	0	2	1	0
PSP	0	0	0	0	0
Xone	0	0	2	0	0
GC	3	0	0	0	0
WiiU	2	0	0	0	0

The formula I'm going to use is **P(A|B) = P(A and B)/ P(B).** For my first conditional probability I'm going to find P(Wii|Nintendo) = P(Wii and Nintendo)/ P(Nintendo) which is P(Wii|Nintendo) = 14/88. 14/88 is about 16% chance. That means there is a 16% chance that a wii is chosen given that is came from Nintendo. The second conditional probability I'm going to find is P(Nintendo|Wii) = P(Nintendo and Will)/P(Wii) which is P(Nintendo|Wii) = 14/19. 14/ 19 is about 73.7% chance that means there is a 73.7% chance that Nintendo is chosen given it came from Wii.