[redacted]

Intro to Probability and Statistics

Dr. Nicholas Jacob

January 14, 2022

Marvel Cinematic Universe

Part 1:

The data that was collected on Marvel movies and was found on a website provided by Dr. Jacob: https://www.kaggle.com/promptcloud/all-marvel-cinematic-universe-movies-on-imdb. This data was collected into an excel spreadsheet containing 24 movies, not up to date obviously. The spreadsheet provides variables on these movies, which include: the year the movie was produced, the genre, the release date, movie rating, the review rating, the run time (in minutes), the run time integer, the plot, the director, the cast, the language in which it was released, the filming locations, the collections, and the date of collection. Not all of these will be included on the spreadsheet example (Table 1.1) due to the length, but they will be defined into categories. An example of this spreadsheet will be shown below. The variables will also be defined in a separate table beneath as quantitative or categorical in Table 1.2. I took interest in Marvel movies because I'm kind of a nerd. They also have some amazing one liners, and it was a franchise I could support all behind Stan Lee, may he rest in peace. Though he gets all the credits and more, he still deserves recognition and support. We all know how popular these films are; however, I am curious to know which movie was the most popular among the ones listed. Each of these were such a hit in theatres, but which one ranked most popular after the initial watch?

Title	Year	Genres	Release Date	Movie Rating	Review Rating	Movie Run Time
Guardians of the Galaxy Vol. 3 (2021) - IMDb	2023					
Iron Man 2 (2010) - IMDb	2010	Action|Adventure|Sci-Fi	7 May 2010 (USA)	PG-13	7	124 min
Guardians of the Galaxy (2014) - IMDb	2014	Action|Adventure|Comedy|Sci-Fi	1 August 2014 (USA)	PG-13	8.1	121 min
Black Panther (2018) - IMDb	2018	Action|Adventure|Sci-Fi	16 February 2018 (USA)	PG-13	7.3	134 min
Iron Man 3 (2013) - IMDb	2013	Action|Adventure|Sci-Fi	3 May 2013 (USA)	PG-13	7.2	130 min
The Incredible Hulk (2008) - IMDb	2008	Action|Adventure|Sci-Fi	13 June 2008 (USA)	PG-13	6.7	112 min

Table 1.1 This is a small portion of the spreadsheet providing 6 of the 24 movies.

Here is a link to the full spreadsheet: https://eastcentraluniversity-my.sharepoint.com/:x:/g/personal/eminhen1_email_ecok_edu/EYJHkGcqKBhBiheczAIuuBwBS2LDoaIf9HQ6lpyf4gwamA?e=rkdpCD

Variable	Type of Data
Year of Production	Quantitative
Genres	Categorical
Release Date	Quantitative
Movie Rating	Quantitative
Review Rating	Quantitative
Movie Run Time (in minutes)	Quantitative
Run Time Integer	Quantitative

Plot	Categorical
Director	Categorical
Cast	Categorical
Language	Categorical
Filming Locations	Categorical
Collection	Quantitative
Collection Date	Quantitative

Table 1.2 This describes each variable in a category.

Part 2:

With some trial and error, I chose to use the variables “language” and “filming location.” The reason I decided on these specific two is because I feel that the filming location may be a deciding factor in what language the movie is produced. After calculating the frequency(s) of the language variable, it has concluded that though all the movies were produced in English, so I took that factor out and used all the other languages. After that, it comes down to French, Russian, Spanish, and Xhosa. Each of those hold 3 tallies a piece. As individual factors, they were about 11%, so together that totals 44%. The information shows the total amount of frequencies to be 27, and the reason for that is because several of the movies were produced is multiple languages. This can be seen on the full excel sheet when opened (it would be in column L).

Language	**Frequency**	**Relative**	**Cumulative**
French	3	0.111111111	0.111111111
Russian	3	0.111111111	0.222222222
Swahili	1	0.037037037	0.259259259
Nama	1	0.037037037	0.296296296
Xhosa	3	0.111111111	0.407407407

Korean	2	0.074074074	0.481481481
Portuguese	1	0.037037037	0.518518519
Spanish	3	0.111111111	0.62962963
Persian	1	0.037037037	0.666666667
Urdu	1	0.037037037	0.703703704
Arabic	1	0.037037037	0.740740741
Hungarian	1	0.037037037	0.777777778
German	1	0.037037037	0.814814815
Romanian	1	0.037037037	0.851851852
Hindi	2	0.074074074	0.925925926
Norwegian	1	0.037037037	0.962962963
Japanese	1	0.037037037	1
	27	1	

Table 2.1 This is the language table, extra column included for cumulative.

On the second categorical variable, filming location, I decided to break down in two ways, country and state. I chose to break it down this way because a lot of the UK and USA locations are in the same "state," so I thought two tables would be beneficial. I found that the filming location in the USA is the highest, ranking at 70%. On the state scale for the USA, Georgia is the highest rank, standing at 35% while the UK has a majority in Surrey, which is 13%.

Location-Country	**Frequency**	**Relative**	**Cumulative**
USA	16	70%	0.695652174
England	5	22%	0.913043478
Canada	1	4%	0.956521739
Australia	1	4%	1
	23	1	

Table 2.2 This shows the country frequencies, also includes cumulative.

Location-"State"	Frequency	Relative	Cumulative
California	4	17%	0.173913043
Georgia	8	35%	0.52173913
North Carolina	1	4%	0.565217391
New Mexico	1	4%	0.608695652
New York	1	4%	0.652173913
Pennsylvania	1	4%	0.695652174
Surrey	3	13%	0.826086957
Queensland	1	4%	0.869565217
Ontario	1	4%	0.913043478
Hertfordshire	1	4%	0.956521739
Merseyside	1	4%	1
	23	1	

Table 2.3 This shows the state frequencies, also includes cumulative.

On the two-way table, I decided to compare the country vs the language in which the movie was produced. Example: the USA's main language is English (and Spanish), so I would mark the 24 under USA (because all the movies were produced in English).

Two-way Table	**USA**	**England**	**Canada**	**Australia**
English	24	24	24	24
French	0	0	3	0
Russian	0	0	0	0
Swahili	0	0	0	0
Nama	0	0	0	0
Xhosa	0	0	0	0
Korean	0	0	0	0
Portuguese	0	0	0	0
Spanish	3	0	0	0
Persian	0	0	0	0
Urdu	0	0	0	0
Arabic	0	0	0	0
Hungarian	0	0	0	0
German	0	0	0	0

Romanian	0	0	0	0
Hindi	0	0	0	0
Norwegian	0	0	0	0
Japanese	0	0	0	0

Table 2.4 This shows the coloration between places filmed vs the language.

On a side note, for this table above, the information seems redundant, but it is not. The relation that I felt would work is the primary language in the country in which it was filmed. Therefore, all of these countries' primary language is English. That is why every movie is counted under each one. I threw in Spanish for the USA because that is technically our second language.

By seeing this, there is hardly any relation between the places filmed and the languages the movies were produced. Also, Australia has no "official language," so it doesn't make a large effect.

Part 3:

By using Excel, I was able to figure out most of the information that was needed. The summary statistics include: mean (the average), standard deviation, mode (most often), and the five number summary, which contain the minimum, quartile 1, the median, quartile 3, and the maximum. I chose the quantitative variable of "review ratings." I felt like it was a close-range set that would allow more details rather than a broader spectrum. It is calculated that the mean is 7.518. The standard deviation is 0.549143371 (specifically). The mode is 7.3. Lastly, the five number summary is min: 6.7, Q1: 7.1, med: 7.4, Q3: 7.875, and max: 8.8. The skewness seems to be right skewed but in a symmetrical way. I claim it to be symmetrical. There are also no outliers. The number set was relatively close together as I previously stated. I would also like to state that due to one of the movies not being produced yet and another one doesn't have a rating, there is 22 instead of 24.

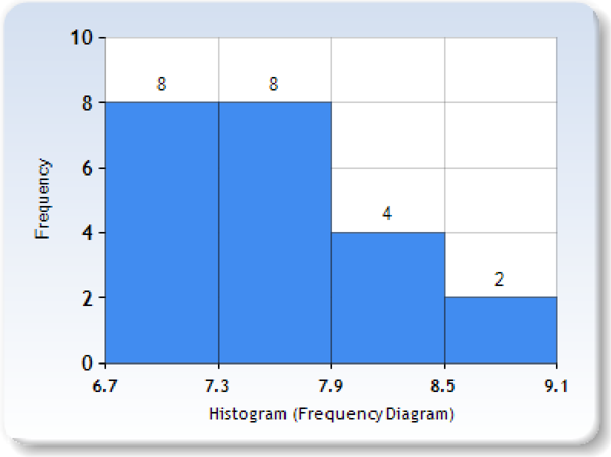


Table 3.1 This is a histogram graph.

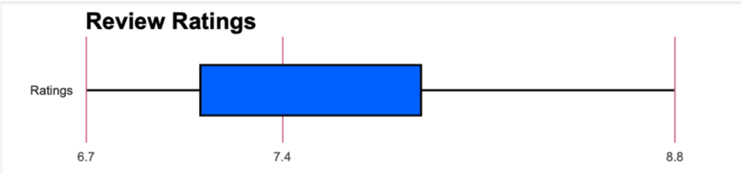


Table 3.2 This is a box plot graph.

Part 4:

I have decided to observe and create a hypothesis on two variables. The first one is a categorical variable: genre. The second variable is quantitative: movie run time.

Categorical hypothesis:

My hypothesis is that 70% of the movies are of the Sci-Fi genre. My null hypothesis is that of the 24 movies 70% is not Sci-Fi.

H_0: p = 70%

Ha: p ≠ 70%

Quantitative hypothesis:

The average run time for 22, out of the 24, movies is two hours, or 120 minutes. My hypothesis is that the average run time for Marvel films is 120 minutes.

H_0: $\mu = 120$

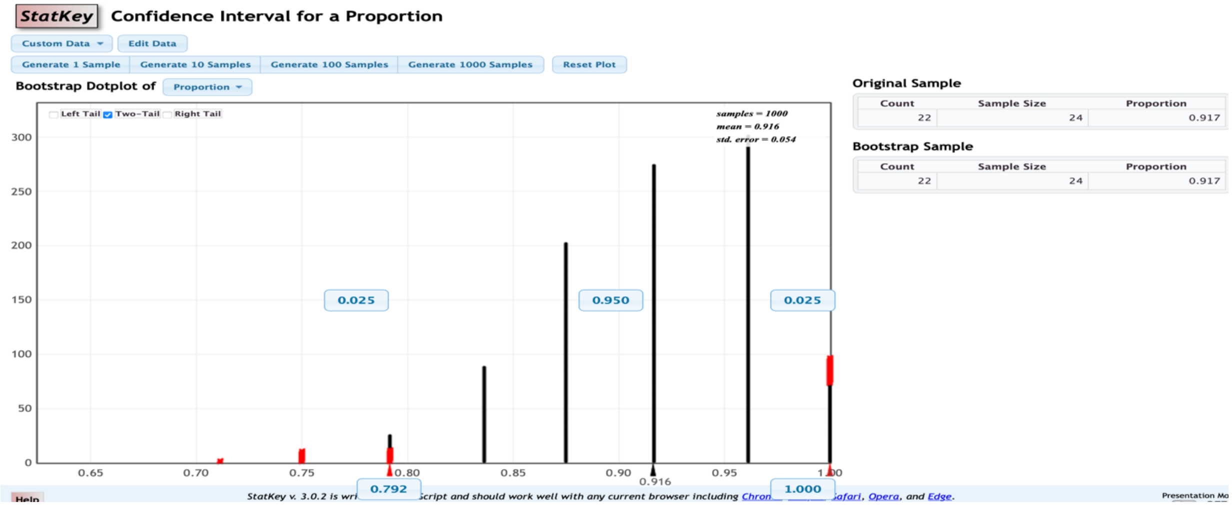
Ha: $\mu \neq 120$

Part 5:

Using the information from the previous part, we will look at the standard error, mean, median, confidence interval, and standard deviation (if all are applicable).

Categorical:

After entering the data into StatKey, I was given the information. I will note that though there are 24 movies, only 22 have been produced. Instead of making the sample 22, I just threw in a couple zeros, making the sample 24. The categorical data provides SE: 0.054 (5.4%) and the mean: 0.916 (91.6%). The confidence interval of 95% is {0.792, 1.000} (79.2%, 100%). The centered point is 0.916 (91.6%), which is about 21% higher than estimated. This proves that my H_0 hypothesis was rejected because my original guess was outside of the confidence interval.


StatKey Confidence Interval for a Proportion

Bootstrap Dotplot of Proportion

□ Left Tail ☑ Two-Tail □ Right Tail

samples = 1000
mean = 0.916
std. error = 0.054

Original Sample

Count	Sample Size	Proportion
22	24	0.917

Bootstrap Sample

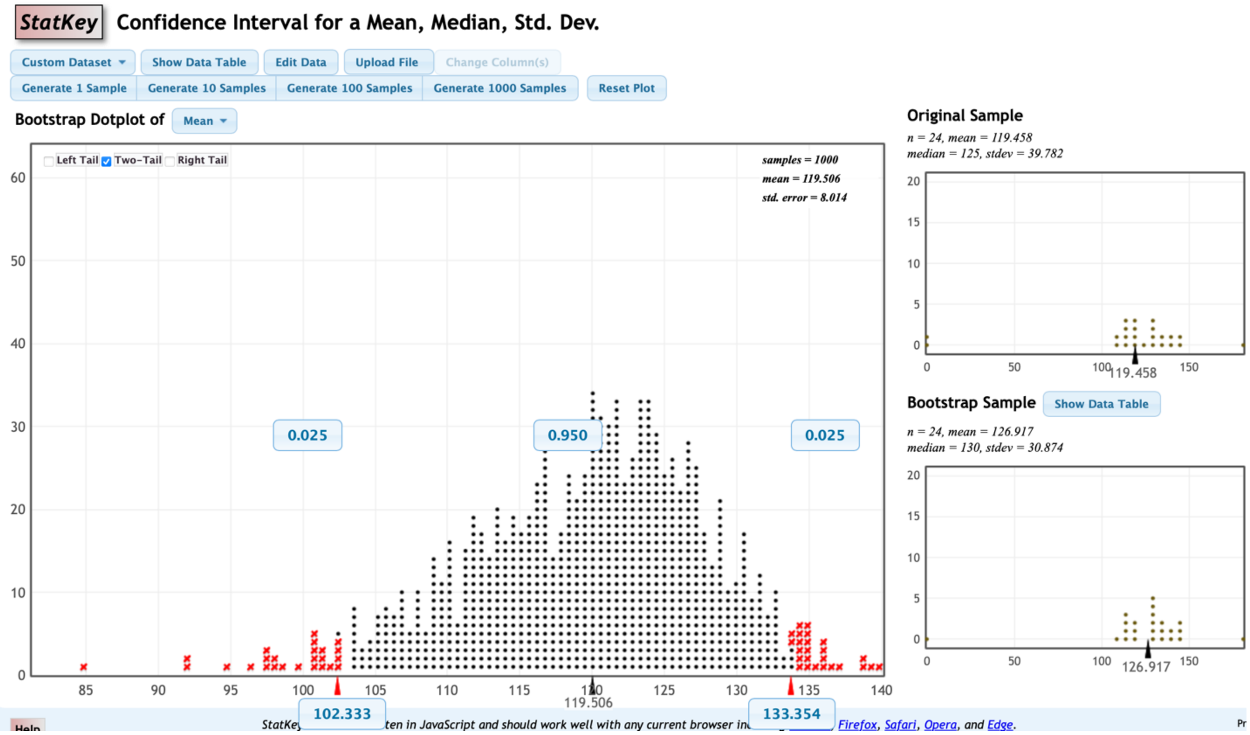
Count	Sample Size	Proportion
22	24	0.917

StatKey v. 3.0.2 is wri… …cript and should work well with any current browser including Chron… …afari, Opera, and Edge.

Table 5.1

Quantitative:

After entering the data into StatKey, the website provided me with information. As I previously noted, though there are 24 movies, only 22 have been produced. Instead of making the sample 22, I just threw in a couple zeros, making the sample 24. The quantitative data provides, SE: 8.014, mean: 126.917, median: 130, standard deviation: 39.782. The confidence interval for the 95% is {102.333, 133.354}. The centered point is around 119.506, which is near the 120 that was estimated. This proves that my H_0 hypothesis was fail rejected due to the result being within the confidence interval.



StatKey Confidence Interval for a Mean, Median, Std. Dev.

Bootstrap Dotplot of Mean

□ Left Tail ☑ Two-Tail □ Right Tail

samples = 1000
mean = 119.506
std. error = 8.014

Original Sample
n = 24, mean = 119.458
median = 125, stdev = 39.782

Bootstrap Sample
n = 24, mean = 126.917
median = 130, stdev = 30.874

Table 5.2

Part 6:

On this portion, I will be retesting my categorical hypothesis and compare the differences with bootstrapping.

The formulas I used were provided by previous projects posted on BlackBoard, YouTube videos posted on BlackBoard, and Google. I will also provide a table below explaining those.

Inquired information (x)	number of occurrences
Sample size (n)	entire size of group
Proportion (p)	the hypothesis %
P-Hat	=x/n
z-value	=Phat-P/SE
z-critical or z*	=NORM.S.INV(1-0.05/2)
p-value	https://www.omnicalculator.com/statistics/p-value
Standard Error (SE)	=SQRT(P*(1-P)/n)
Low 95%	=Phat-z critical*SE
High 95%	=Phat+z critical*SE

Table 6.1

Using the calculations provided above, I have filled in the blanks with my own data. I found each of these values listed, calculated through excel. I will show the results below.

Inquired information (x)	22
Sample size (n)	24
Proportion (p)	0.7
P-Hat	0.916666667
z-value	-6.566648107
z-critical or z*	1.959963985
p-value	0
Standard Error (SE)	0.093541435
Low 95%	0.733328824
High 95%	1.10000451

Table 6.2

My null hypothesis was that 70% of the Marvel movies were under the Sci-Fi genre (H_0: p = 70%). My alternative hypothesis stated that 70% were not under the Sci-Fi category (Ha: p ≠ 70%). After using the formulas, the 95% confidence intervals show {0.733328824, 1.10000451}. Therefore, I will reject my null hypothesis because 0.70 is outside of the confidence interval.

Comparison:

The 95% confidence intervals from the bootstrap test show {0.792, 1.000}. Whereas the inference 95% confidence intervals show {0.733328824, 1.10000451}. The differences are quite separate. Because the bootstrap high 95% is right at 1.00, it seems that it is more accurate. Looking back at the facts, the movies that are under the Sci-Fi genre are 100%. By knowing this, the low 95% interval of the bootstrap test seems more correct. For this variable, I would say that bootstrap is more accurate.

Part 7:

This portion will retest my quantitative hypothesis.

As I stated in the previous part, the formulas I used were provided by previous projects posted on BlackBoard, YouTube videos posted on BlackBoard, and Google. I will also provide a table below explaining those.

MU	Hypothesis
X-Bar	The average =(sum/n)
Standard Deviation (Sigma)	=STDEV.S(column of numbers)
Sample Size (n)	entire size of group
Standard Error (SE)	=Sigma/SQUT(n)
T-score	=(Xbar-MU)/SE
Alpha	For 95% CI (1-0.95)
T-critical or T*	=T.INV (alpha, n-1)
Low 95%	=(Xbar-Tcritical)*SE
High 95%	=(Xbar+Tcritical)*SE

Table 7.1

Using the calculations provided above, I have filled in the blanks with my own data. I found each of these values listed, calculated through excel. I will show the results below.

MU	120
X-Bar	119.4583333
Standard Deviation (Sigma)	39.78199196
Sample Size (n)	24
Standard Error (SE)	8.120465104
T-score	104.6808546
Alpha	0.05
T-critical or T*	-1.713871528
Low 95%	105.5408994
High 95%	133.3757673

Table 7.2

My null hypothesis states that the average run time is 120 minutes (H_0: $\mu = 120$). My alternative hypothesis is that the average run time is not 120 minutes (Ha: $\mu \neq 120$). After using the formulas in Excel, the 95% confidence intervals show to be {105.5408994,133.3757673}. Therefore, I will fail to reject my hypothesis because 120 falls within the 95% confidence intervals.

Comparison:

The 95% confidence intervals on the bootstrap test show {102.333, 133.354}. Whereas the inference 95% confidence intervals show {105.5408994, 133.3757673}. These results are much closer than the categorical results were. Due to the inference being calculated by formulas rather than computer generated estimated, it seems more accurate.

Part 8:

On the final part of the project, I will create two probabilities by using a two-way table. Though the two-way table from part 2 should have used for this part, my information wasn't good enough to reuse. I created a new two-way table using the release date data.

Two-Way	Jan	Feb	Mar	April	May	June	July	Aug	Sept	Oct	Nov	Dec	**Total**
2008	0	0	0	0	1	1	0	0	0	0	0	0	**2**
2009	0	0	0	0	0	0	0	0	0	0	0	0	**0**
2010	0	0	0	0	1	0	0	0	0	0	0	0	**1**
2011	0	0	0	0	1	0	1	0	0	0	0	0	**2**
2012	0	0	0	0	1	0	0	0	0	0	0	0	**1**
2013	0	0	0	0	1	0	0	0	0	0	1	0	**2**
2014	0	0	0	1	0	0	0	1	0	0	0	0	**2**
2015	0	0	0	0	1	0	1	0	0	0	0	0	**2**
2016	0	0	0	0	1	0	0	0	0	0	1	0	**2**
2017	0	0	0	0	1	0	1	0	0	0	1	0	**3**
2018	0	1	0	1	0	0	1	0	0	0	0	0	**3**
2019	0	0	1	1	0	0	1	0	0	0	0	0	**3**
Total	**0**	**1**	**1**	**3**	**8**	**1**	**5**	**1**	**0**	**0**	3	**0**	23

Table 8.1

The first probability I will make is if one month is randomly selected. What is the probability that the month is May if the given year is 2010? To find the probability, I will use the formula:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

I will enter my own variable inside the equation to match my data. This is just to visualize what I am doing. I will make M for May and Y for year, in this case 2010. Then the formula becomes: **P(M|Y) = P(M∩Y)/P(Y)**. The find **P(M∩Y)**, one can look at Table 8.1. The junction of May and 2010 is 1. Figuring out **P(Y)** from the table, it also seems to be 1. So, **P(M|Y)** = 1/1, which equals 1. This means there is a 100% chance that May will be selected if the given year is 2010.

The second probability I will make is if one year is randomly selected. What is the probability that the month is February if the given year is 2018? Using the same steps as I did above, I replaced the example formula with my own data. I will use F for February and still Y for the year. The new formula would be **P(F|Y) = P(F∩Y)/P(Y).** Finding **P(F∩Y)** in the table, it is seen that the intersection at 2018 and February is 1. The **P(Y)** is 3. Meaning, **P(F|Y)** =1/3, which is 0.33. Therefore, there is 33% chance that February will be selected if the given year is 2018.