Research Data on Anime Dataset Regarding Statistics

Melvin Vazquez Andujar

April 26, 2024

Table of Contents

[Abstract: 1](#_Toc165129829)

[Chapter 2: Events, Conditional Probability, Permutation, Combination 2](#_Toc165129830)

[Chapter 3: Binomial Probability Distribution, Geometric Distribution, Negative Binomial Distribution, Hypergeometric Distribution, Poisson Distribution, and Tchebysheff’s Theorem 4](#_Toc165129831)

[Binomial Distribution: 4](#_Toc165129832)

[Geometric Probability Distribution: 4](#_Toc165129833)

[Negative Binomial Distribution: 5](#_Toc165129834)

[Hypergeometric Distribution 5](#_Toc165129835)

[Conclusion: 5](#_Toc165129836)

[Bibliography: 6](#_Toc165129837)

# Abstract:

This report will be based on a data set found from a website called Kaggle on Anime. The data found in the CSV will be used to get results from inputting that data into formulas from my Probability and Statistics class. The results will give a deeper understanding of the data set and will help better understand the dataset.

# Chapter 2: Events, Conditional Probability, Permutation, Combination

**For chapter 2 there is not many equations that can be used for the data set, however, when looking at the data we see that each category has multiple options when looking at each anime. For example, when we look at a single anime, we also see the source that it came from, the studio that animated it, and which company licensed the anime. We can make a sample space that every anime has three options when it comes to source, either it is anime original or it is directly based on the source material, which would be the manga, or it is based on the light novel. So, an example would be**   
  
S = {O}, S = {M}, S = {L} – O = Original M = Manga L = Light Novel

From those three options we know we have a one-third chance of it being one of those. Additionally, this list can become even more complicated as you add more options and more categories like genre, type of anime, studios, etc.

For Theorem 2.2 that deals with Permutation, we can use example 2.8 as a guide on how to come up with a similar problem for this dataset. A way we can do it is putting 100 anime inside of a hat and drawing 6 different anime shows from that hat. By doing that we can use the permutation formula because we now have the n, which is 100 and we have the r, which is the amount drawn.

That is the total number of sample points from that problem alone.

Another formula we can use is the Combinations formula and this can be used in correlation with chances of picking 3 anime shows from a set of 30. We can use this to find the number of combinations that can occur when choosing 3 anime shows from a set of 30 books.

There are 4060 combinations that can occur when picking 3 anime shows from a set of 30.

When using conditional probability, we can say that P(A) = .33 since there is a 1/3 chance of an anime coming from one of 3 different sources, P(B) can be .33 that there is only 1/3 chance of the anime being either a movie, tv show, or an ova. Since these two events are independent of each other, the chances of an anime being both ova and manga is 1/9 since P(A) is independent from P(B). This can be found in definition 2.10.

From this formula a conclusion can be made that the events are independent of each other.

With this information we can also do the additive law of probability which is just

Which basically reads as there being a chance that either A or B will occur. So, a chance of being a source or type of anime show, movie, or OVA.

The user wants to select two different anime to watch, He doesn’t know what to watch so he puts all three types of anime equally into the list. There are a total of 9 different anime to pick from. So this problem is a probability distribution function in which you can find total sample points from doing:

Since we know that you can get any of the three types equally as likely you can do

That is the same as finding the probability for the other types since they all can be chosen equally, then we divide the result by the sample points to get the probability of selecting two different anime to watch.

# Chapter 3: Binomial Probability Distribution, Geometric Distribution, Negative Binomial Distribution, Hypergeometric Distribution, Poisson Distribution, and Tchebysheff’s Theorem

**Chapter 3 has more formulas better suited for this data, so there will be a lot more formulas in this section of the report. The formula used will be given and a description of where the information is gotten from and sample problem, I came up with using textbook as a guide to how it fits, and it comes up with a story.**

Binomial Distribution: With the same problem above we can use it to find binomial distribution from chapter 3 depending on if we want to select exactly one tv show. So, the formula for binomial distribution is:

p = probability of success, q = probability of failure, n = trials, y = success. So we know that p =

Now lets plug into the formula:

This translates to: The probability of picking exactly one tv show when picking two different anime is .

Geometric Probability Distribution: Another interesting thing about this example is that it is versatile in that it can be used for multiple formulas from chapter 3. One of those formulas is Geometric Probability Distribution. One thing to add to this is that we want to check after 5 trials.

Formula:

Known info:

Plug in:

Translates to probability it takes exactly 5 trials to select a tv show anime is .

Negative Binomial Distribution: Using the same problem above we will look for the probability of selecting exactly a certain amount of TV shows when picking an anime randomly from list. So we will look at finding a second tv show anime occurring on the sixth anime picked.

Formula:

Known info:

Plug in: :

Result:

Translates to the probability of getting a second tv show being found after the sixth anime is picked is

Hypergeometric Distribution: For this formula we will use a new problem, Similar to example 3.16 of the textbook we will look at 20 anime, 8 are randomly selected to be animated by a studio. What is the possibility that 8 of those anime are 5 of the best rated anime in the group of 40.

Formula:

Known info: N = 40, n=8, r=5, Y=5

Plug in:

Result:

Translates to there being a .361% chance of the top 5 rated anime from the list of 20 being in the randomly selected list.

# Conclusion:

In conclusion this research project helped me gain a better understanding of how statistics can be applied to real life data and how depending on what you are looking for, there can be drastic changes to the data. I did not put any chapter 4 and 5 as I am not as proficient in understanding how I would go about inputting that data into integrals with limits and finding a solution. I also did not put the mean, median, or standard deviation in this project as the dataset is over 2000 values and each category has something different from each other. Instead, I put in some chapter 2 and most of chapter 3 that I could use my dataset for. I did enjoy making this research project as it made me understand better how to read my dataset and creating a story of finding a tv show anime through different formulas and how it differs from other formulas because of differences each formula looks for. Some look for first instance, some look for third instance, and some look for there being a chance of finding a third tv from looking at your 7th anime.

# Bibliography:

Uddin, Sajid. “Anime Dataset 2023.” 2023, Accessed 2024.

Wackerly, Dennis  D, et al. *Mathematical Statistics with Applications, Seventh Edition*. seventh edition ed., Cengage Learning, 2008.

[..\..\CSVs\anime-dataset-2023.csv](../../CSVs/anime-dataset-2023.csv)