# 

# 

# 

# 

# A Statistical Look At Steam Games

# By:

# Darwyn Abreu

## 

[Conditional Probability](#_am62ljncd5y6) 3

[Independence and Bayes Theorem](#_4rk9t4mn6dth) 4

[Probability Distribution](#_myv019262aya) 5

[Binomial Distribution](#_mvp1grei3fk2) 6

[Negative Binomial Distribution](#_lcwl0ehf71g1) 7

[Geometric Distribution](#_gf66f3fifum7) 8

[Hypergeometric Probability](#_t9f5t27mz0q3) 9

[Possession DIstribution](#_uyae23ecb4hj) 10

[Tchebysheff Theorem](#_bthl634w9idn) 11

[Uniform Distribution](#_an9bo692ntnz) 12

[Marginal Probability](#_409r7ughz7n5) 12

[Conditional Probability (based on Marginal)](#_fc2tg8k026jc) 13

[Takeaways](#_n9dkiz9cjhpq) 14

## 

## Conditional Probability

When looking at the data set of games we can see that we are able to view a game’s final price as well as the type of review it has at the time of collecting the data. With this being said an interesting point that can be made about most consumers is that they are more accepting of a low quality product as long as the price of said product is equally as low. To test this thought process we can look at the probability that a game has some sort of positive review (Overwhelmingly Positive, Very Positive, Positive) given that the game is less than or equal to $5 **P(Positive Rating | Price<=5).** In order to find the percentage that a game will be positive given that it has a price of 5 dollars there are some steps we can follow to solve this problem:

Fulfill equation p(a |b)= anb / b

1. Filter all games down to those that are under $5 and find the number of games let this be called **B=26693**
2. Calculate the Probability of a game being less that $5 given by **( B / 50872) =.5247**
3. From **B** find the number of games that have a positive rating let this be called **A =18848**
4. Now find the chances of a games being positive and being below doing **(A/50872)** and let this be called **C=.3705**
5. Now to calculate the chances of a positive review given that a game is under $5 can be calculated from **C**/**B** our final answer works out to be **.7061**

After following through with calculations we can see that there is a **.7061** probability or a **70.61%** chance of a game being reviewed as positive given the fact that the **original** price of the game is $5 or less.

### Independence and Bayes Theorem

An important part to figuring out how much these stats can be relied on is figuring out if the relationship between the price and reviews are independent or not. There are many ways to test for this such as the rules of P(A |B) =P(A) P(B |A) =P(B) P(AnB)=P(A)P(B) which tells that the set is independent if one of the rules holds true.

To determine this we already have p(A|B) where A=Positive Rating and B=Price<5 so to verify each equation. First we must obtain P(A) which is **.7172** and because we already obtain P(B) and P(AnB) from the original problem and plug in values which yields **(.7172)(.5247)=.3705** which simplifies to **.3763=.3705** which is not true as well as when looking at P(A|B)=P(A) you obtain **.7061=**.**7172**  which is also not true. This means that we must check the final equation of p(B|A)=P(B). Due to the circumstance of having a data set available It may be easier to write code to find the inverse of a relationship P(B|A) however the method can be solved using Bayes Theorem which is given by:P(B|A)=P(A|B)P(B)/P(A|B)P(B)+P(A|B)P(‘B)

When solving with bayes theorem we get **.5165=.5165** which means that the data is dependent. This doesn’t necessarily mean that a cheap game will be good, however it does mean that there is some trend between the price of a game and the review it receives.

## Probability Distribution

Sometimes during large sales there are third party websites that sell game bundles that will give away a random amount of keys for a given price. A common example seen is something along the lines of 5 games for 5 dollars. Now given that is such a large market of games the user may be weary of paying 5 dollars for 5 games that they won’t end up playing. In order to help display the likelihood of getting a good game a user can look at the probability distribution of getting a good combination of games. To do this first we need to find the probability of getting 0 good games, then 1, then 2, and so on until we find the chances of getting 5 good games. This can be represented as such (combo good games)(combo bad games)/(total games). First we need to obtain the number of games total which is 50872 and the number of good games which is 36489, this means the remaining games are mix-ed or negative which leaves us with 14383 games in the bad collection: Solving for the probability distribution is as follows:

What this distribution shows is that a user is most likely to get 4 out of 5 good games with a 37.42% chance of getting exactly 4 good games.

## Binomial Distribution

When choosing games a majority of players are on windows. With that said there are players who play on linux or macos who are customers of these bundle deals so an important part of the deals would be to make sure that they are compatible with your system. Although most websites should say what system is compatible with it’s software, in the case of bundling games there may be a chance not all games are compatible. With that said a user may want to find out what are the chances that all 5 games are compatible with macOS given that 74.41% of the dataset is macOS compatible. To solve we use the Binomial Distribution equation (ny)p^y q^n-y

* The rest of of the distribution can also be solved

Overall the user has the greatest chance at having 4 of 5 games be macOS compatible with a 39.25% of having exactly 4 games be macOS compatible

## Negative Binomial Distribution

Some users may look at the data and see that around 75% of the games are macOS compatible and may assume that their whole cart is compatible with their OS. Although this assumption may not be true so to demonstrate to users that they should always check the specs we find the chances of getting your 3rd non-compatible macOS game on your 9 purchase.

To calculate this we use the negative binomial probability equation which is (y-1 r-1) p^r q^y-r

This solves out to be:

This means that there is a 0.17% chance of getting a 3rd game that isn’t compatible on the 9th purchase. Furthermore this means that if choosing 9 random games from the dataset that the user can feel fairly confident in their choice of games working.

## Geometric Distribution

To look even further at the chances of finding a game that doesn’t suit the system we can use Geometric DIstribution to find the probability of getting your first non compatible game at different points in the entire purchase. For a list of 9 purchases the disruption is as follows

* %

by the end of everything we can see the p(9) is 2.26% which means that there is about a 2% chance of getting 8 compatible games in a row followed by one that doesn’t work. This along with the trend of the data tells us that by the time a user has reached 9 games there is a great chance that the user has bought at least one game that isn’t compatible with their system if they are using macOS.

## Hypergeometric Probability

Another common issue seen between users is when playing with friends or swapping computers they would like to know how many games they are able to play. In the case of somebody who has 144 games it might be reasonable to expect at least a quarter of the games to work on other operating systems. With that said, from a selection of 144 games, what are the chances that 36 games are compatible with mac and the rest are not. We know that there are macOS compatible games and mac incompatible games. With that information we cause the Hypergeometric Probability Distribution to find the answer. This is done with the following equation:

* with N=50872 r=37854 n=144 y=36
* =0 which means that there is a 0 percent chance when selecting from all the cards however if we have a smaller set of gams with the same ratio of compatibility we can have a library of 144 games with 109 mac compatible games, out of that subset if we wanted to choose 36 games and find that chances that half of them are macOS compatible we get the answer of == which means there is about a 1% chance that exactly 18 games are macOs compatible. Also given the chance we want to find only 1 game is macOS compatible we get a result of **0.00968** which is **0.968%** which means that overall finding at least one game that is macOS compatible should be relatively easy to find.

## Possession DIstribution

On average for every there are macOS compatible games, with that said there are about 37 mac os compatible games for every 50 games. With that said if a user purchases a bundle of 50 games and a possession diction is followed then the chances of a random game library also having 37 games out of 50 is given by equation: which is 0.0637

## Tchebysheff Theorem

GIven the prices of games a developer may want to make sure they are pricing their game within a relatively close proximity to other games to avoid coming off as money hungry or cheap. With that said the average price of a game from the data set is at $8.62 with a standard deviation of 11.51. What can be said about games that are listed at the typical AAA price of $60.

If all prices are determined based upon our numbers the chances of a game getting listed at $60 follows the equation P(60-mean)>1(1/k^2)

p=60-8.62=51.38 so therefore k is 4 because 60 is within 4 standard deviations of the mean. This means that overall the probability of a game being listed randomly for $60 is 1/k^2 which is 1/16 chance. In other words a new game being listed for $60 has a 1/16th chance of occurring so most games that are new are most likely listed closer to the mean price.

## Uniform Distribution

If we have game pages with listings of 15 games per page and the user only looks at 5 consecutive listings before the next listing what are the chances that it is the first 5 entries before they click off. This can be found from the equation

= so there is a ⅔ chance that a person only looks at the 5 first listings before moving on

## Marginal Probability

Given that we have 36489 good games and 12157 mixed games and 2226 bad games a user may want to find the chances of getting exactly 2 good games and 1 bad game out of 3. The probability of finding this distribution of games can be found by using the following equation.

This means that the chances of obtaining 2 good games, 0 mixed games and 1 bad game is 6.75%. To find the rest of distribution use the same equation adjusting p(x,y) until all combinations of good and bad games are covered respectfully which results in the following values: x=# of good games y=#of bad games

p(0,0)=.0136=1.36%

p(0,1)=.007496=.75%

p(0,2)=.00137=.14%

P(0,3)=.0000836=.00% rounded

p(1,0)=.12288=12.3%

p(1,1)=.045=4.5%

p(1,2)=.004118=.41%

p(2,0)=.3689=36.89%

p(2,1)=.06753=6.75%

P(3,0)=.3690=36.9%

## Conditional Probability (based on Marginal)

Given the distribution that was just found, users should be able to find the probability of getting a certain distribution of games given a certain amount of good or bad games. For example, if a player has picked one bad game already, what are the chances that the other two games are good? To find this we can use the following formula

This means that there is a 56.25% chance of getting 2 good games if we know we already have one bad game.

## Takeaways

Overall when approaching this project I knew I felt like the harder part was coming up with the actual questions. With that said thought I was able to look at a couple aspects of games that I really didn’t think about prior to doing this project. After this project one thing that has stuck with me was the concept that people are more forgiving of a cheap game than of an expensive game. I also realized how rare a $60 game is. Most games seen today in stores are usually at this higher price tag and are usually created by bigger studios however as somebody who may want to get into game development it is nice to know that there is an advantage to having a lower price. If the data had the info to show the number of purchases I would have definitely looked at that data as well to see if purchases correlate to price as I think it might. Overall though I think I gained a better perspective of game marketplaces like Steam,Origin, Epic Games, etc. When it comes to a large platform like steam having games that are compatible with other Operating systems it seems like users have plenty of options and would explain the popularity of the platform compared to others. Another fact I took note of was the likelihood of receiving “good games” from bundles of games. At the time the data was collected the chances of getting a mostly good bundle had the highest percent and still very positive results for receiving smaller qualities of good games. In other words when games are on sale and there are super bundles offered it seems like that as long as I actually end up playing the games buying those bundles are a good deal. I feel like my data set lacked some of the player interaction with the game that I would want in order to perform a full analysis in the way that I want; however This data set was also the most complete, clean, and easy to work with data set that I was able to find. Other data sets that gave me player information simply lacked the game data I needed or had large amounts of missing data. Overall this data set fit my needs best.

One major point of learning for this project was setting expectations for how long things will take me to get down. At first I was planning on using an MMA data set and approach this project with a more sports betting type of aspect along with a large extra credit game portion that may have been based on the stats gathered. I had to pivot from my original idea however because I simply fell behind my own self imposed deadlines. Falling back on my own personal deadlines really did show me that I at least have reasonable expectations for myself and the amount of work I can get done. Even though I fell back on my work, my deadlines for myself were made to give myself a safe cushion of free time so it just so happened that the need to start time really was the time that I needed to start. Realistically the project was not super difficult and actually did give a decent look at how to view things. If I wanted to eventually become a developer or run a marketplace I think I have a slightly better idea of how to go about things in terms of offering games and what not.

At the end of the day I would say that this project was a light data mining project that I think I was able to handle with relative ease. When it came to gathering the information and numbers I needed from the data set, writing the code to find those numbers was not too difficult. The hardest part for me was making sure I was using statistics properly as seeing as the answers I’ve gained make logical sense to me I feel confident in my answer. I think this project was a very good showcase at why Prob and Stats is even a required course for Comp Sci in the first place.