STATISTICAL ANALYSIS OF ANIME DATA

MyAnimeList

MAL

NAME: MARTA FILIPA RAMALHO

GROUP: E

INTRODUCTION

DESCRIPTION OF THE DATASET

In this assignment we are going to work with the following dataset: MyAnimeList dataset. This dataset contained 14477 observations and 31 variables originally. Each observation represents an anime with its corresponding variables such as title, title synonyms, genre, studio, licensor, producer, duration, rating, score, airing date, episodes, source (manga, light novel etc.) and many other important data about individual anime providing sufficient information about trends in time about important aspects of anime.

This information was gathered from a website which is dedicated to store all the information about anime and where users can score them. This website is MyAnimeList and it has over 5 million users. As we can see in the pictures bellow, in this website we can find all the important information about an anime.

Information

Type: TV
Episodes: 148
Status: Finished Airing
Aired: Oct 2, 2011 to Sep 24, 2014
Premiered: Fall 2011
Broadcast: Sundays at 10:55 (JST)
Producers: VAP, Nippon Television Network, Shueisha
Licensors: VIZ Media
Studios: Madhouse
Source: Manga
Genres: Action, Adventure, Fantasy, Shounen, Super Power
Duration: 23 min. per ep.
Rating: PG-13 - Teens 13 or older

Statistics

Score: 9.08[1] (scored by 1,040,943 users)
Ranked: #6[2]
Popularity: #12
Members: 1,804,195
Favorites: 157,545

As for the data used in the assignment, I used the following criteria:

	How many were used?	Criteria followed
Observations	3509	I only considered animes with more than 1000 members, since less than that means that that anime is either barely known and it will have missing data, or that that anime is very recent and there is not enough information about it. I also deleted all the anime tagged as OVA, Music or Specials since Music is a category that does not represent anime series, it just represents music related to them; and OVA and Specials because they are not isolated animes, they are always related to another anime series or movie. Since they are not independent and I wanted to analyze isolated animes, I did not consider them for this analysis.
Variables	6	I chose 6 of the 31 variables in the original dataset because they were the ones needed to complete this assignment. I picked 2 qualitative variables and 4 numeric variables. Since the dataset had too many qualitative variables, I picked the 2 with less than 10 variants. From the 4 numeric variables, I picked the only continuous variable in the dataset and 3 discrete variables with a high range of values.

As for the variables chosen, this is what each one represents:

Type (F_1)	This variable represents the type of the anime, that is: • **TV:** The anime is a series broadcasted on Japanese TV. • **Movie:** The anime is a movie released on Japanese cinemas. • **ONA:** Meaning Original Net Animation, is an anime which is not broadcasted on TV and can only be watched online. As said before, I did not consider Music, Specials and OVAs for the reasons already mentioned.
Source (F_2)	This variable represents the original source on which the anime is based. Some of the variants have been modified, for example, I have included the variable book in novel and some variables which only appeared once have been included in "Others". Then, the anime can be based on: • **4-koma manga**: Read as "Yonkoma Manga" in Japanese, it is a comic format that consists of four panels of equal size ordered from top to bottom. • **Game**: It can be a videogame, a card game, or any other type of tabletop game. • **Light novel**: Novel format that includes drawings. • **Manga**: Japanese comic. • **Novel** • **Visual novel**: Type of novel which is read and played as a videogame, getting your own story based on your choices. • **Web Manga**: Type of manga that is only published online. • **Original**: The anime is not based on anything and the story is original. • **Other**: The anime is based on any other type of media.
Score (X_1)	It represents the score of the anime, which can go from 0 to 10, based on the score each user gave that specific anime.
Members (X_2)	It represent the number of users that have added the anime to their lists.
Popularity (X_3)	It represents how popular an anime is, that is, how many people know about it. The more popular it is the smaller the value for this variable.
Favorites (X_4)	It represents the number of users that have added this anime to their favorites'

PARTICULAR OBJECTIVES

Given that we have the source and type of each anime, it would be interesting to know which are the most frequent variants of each variable and try to think why that happens, and also discuss how the two variables are related and the most frequent source for each type of anime. It would also be interesting to see the kind of distribution that each numeric variable follows and see if any of them follow a normal distribution.

DISCUSSION ABOUT THE SAMPLE AND THE POPULATION

The population in this case is the set of all the animes stored in MyAnimeList. As for the sample used in the assignment, only 3509 of all those animes are being used. Also, the population is always increasing since new animes come out and are added to MyAnimeList every year. That being said, I consider my sample to be more or less representative of the population since I am using animes which contain all of the information for the variables even after reducing the set of 14477 animes the dataset contained originally. Also, I am using all of the possible variants that represent anime for the qualitative variables. (Except Music, OVA and Specials as said before).

DESCRIPTIVE STATISTICS

1) COMPUTE A FREQUENCY TABLE WITH THE VARIABLE TYPE AND DISCUSS THE MOST RELEVANT INFORMATION DEDUCED FROM THE TABLE.

Table 1: Frequency table for "Type"

Class	Value	Frequency	Relative Frequency	Cumulative Frequency	Cum. Rel. Frequency
1	ONA	229	0,0653	229	0,0653
2	Movie	766	0,2183	995	0,2836
3	TV	2514	0,7164	3509	1,0000

This table shows the number of times each value of Type occurred, that is, its absolute frequency, as well as percentages and cumulative statistics. For example, in 229 rows of the data file, Type equaled ONA. This represents 6,52608% of the 3509 values in the file. Therefore, we can deduce that the most frequent value of Type is TV, which represents the 71,64% of the animes in the data file, and the less frequent is ONA. This makes sense since most animes want to reach a wider audience by being broadcasted on TV, and the ones that are only released online (ONA) are a minority. It also makes sense that there are more Movies than ONAs and less than TV, since Movies require a higher quality of production than anime series, but they also facture more money by being released on cinemas. Also, there are more TV animes than ONA and Movies together.

2) PLACE A BAR CHART AND A PIE CHART OF THE VARIABLE SOURCE.

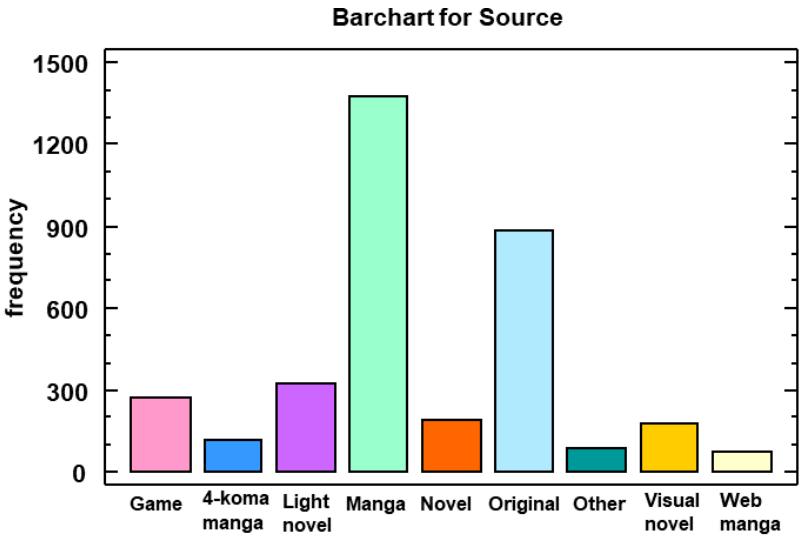

Figure 2: Barchart for "Source"

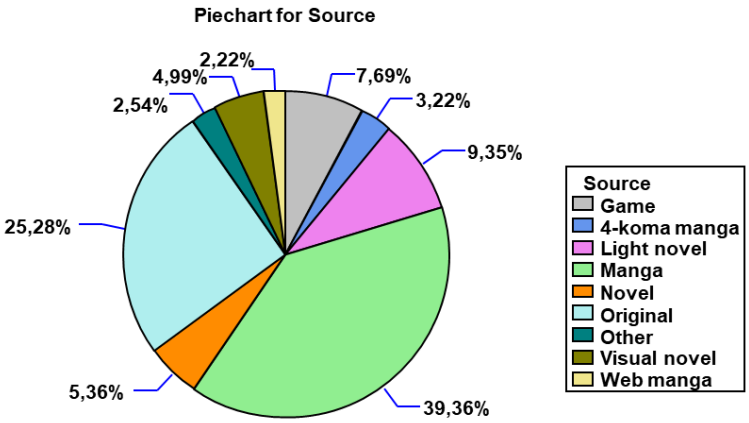

Figure 1: Piechart for "Source

WHAT DOES THE CONTINOUS SCALE ON THE BAR CHART REPRESENT?

The continuous scale of the bar chart represents the absolute frequency of each variant of the variable Source, that is, the number of times each variant appears in the data list.

WHAT DOES A PIE CHART CONSIST OF?

The pie chart consists of a circular graphic which is divided into slices to illustrate the numerical proportion of each instance of the variable, that is, to represent the relative frequency of each value. Each slice takes the same percentage of the circle than the corresponding variant takes of the data set. Each variant is represented in a different color to avoid mistakes.

DISCUSS THE MOST IMPORTANT RESULTS DEDUCED FROM THE CHARTS

From these charts we can see that the most frequent variant of the variable Source is "Manga" with 39,36% and the least frequent is "Web Manga" with 2,22%. The second most frequent variant is "Original" with 25,28%. This makes sense since in Japan, Manga is a product consumed by everyone and it is obvious that they will produce animes based on it. It is surprising though, that "Original" is the second most frequent one, since it does not have source material, it is riskier for the companies when they produce the anime since they do not yet have an established fanbase. As for Web Manga, it makes sense that it is the least frequent, since those are usually mangas that are less popular and facture less money. We can also say that the barchart is not ideal to represent exactly how many instances of one variant exist in a data set, since we cannot pinpoint the exact value just by looking at it.

3) COMPUTE A TABLE OF CROSSED FREQUENCIES WITH THE VARIABLES "TYPE" AND "SOURCE"

Table 2: Crossed frequencies of "Type" and "Source"

	Manga	Original	Light novel	4-koma manga	Game	Novel	Other	Visual novel	Web manga	Row Total
TV	1033	563	267	97	172	114	62	148	58	2514
	41,09%	22,39%	10,62%	3,86%	6,84%	4,53%	2,47%	5,89%	2,31%	71,64%
Movie	285	256	47	5	70	64	20	14	5	766
	37,21%	33,42%	6,14%	0,65%	9,14%	8,36%	2,61%	1,83%	0,65%	21,83%
ONA	63	68	14	11	28	10	7	13	15	229
	27,51%	29,69%	6,11%	4,80%	12,23%	4,37%	3,06%	5,68%	6,55%	6,53%
Column Total	1381	887	328	113	270	188	89	175	78	3509
	39,36%	25,28%	9,35%	3,22%	7,69%	5,36%	2,54%	4,99%	2,22%	100,00%

"ROW PERCENTAGES" OR "COLUMN PERCENTAGES"? WHY?

In this case, I chose row percentages because the variable "Type" has less variants than "Source" and I was interested in knowing what is the percentage that each source occupies in TV, Movie and ONA.

EXPLAIN THE DIFFERENCE BETWEEN ABSOLUTE AND RELATIVE FREQUENCIES

Absolute frequency represents the number of times a value appears, and the sum of all absolute frequencies is the total number of observations. On the other hand, relative frequency is the absolute frequency divided by the total number of values in the data set, obtaining a number from 0 to 1. The sum of all the relative frequencies is equal to 1.

EXPLAIN THE DIFFERENCE BETWEEN MARGINAL AND CONDITIONAL FREQUENCIES

Marginal frequencies are the ones that represent the total frequencies of each value of the variable. Conditional frequencies are the ones that are computed depending on the values of one variable. For example, in the table above, we can see that in red we represent the conditional frequency of Manga with

respect to TV, that is, 41,09% of the data that have Type TV have Manga as source. However, in "Row Total" and "Column Total" we represent the marginal frequencies of each variant with respect to the total amount of observations.

DISCUSS THE MOST IMPORTANT RESULTS DEDUCED FROM THE TABLE

In the case of TV, the most frequent source is "Manga" and the least frequent one is "Web Manga". This happens because, as said before, Manga is something highly consumed by Japanese people, while Web Manga has a very narrow number of fans since they are only published online.

In the case of Movie the most frequent source is "Manga" and the least frequent ones are "4-koma manga" and "Web Manga", both representing 0,65% of the observations with "Movie". The same explanation given for TV can be applied here, taking into account that "4-koma manga" are just short comics with very little story, it makes them a bad choice to adapt an anime from.

In the case of ONA, the most frequent source is "Original", unlike with the previous two, and the least frequent one is "Other". Since ONA is the least frequent type, it makes sense that the most frequent source is "Original" since most famous mangas and novels will get a TV adaptation because they are sure to facture a grand amount of money. Therefore, if a director that is not too famous wants to create his original story, he will have to resort to "ONA" since they are less expensive. Also, since "ONA" is not that popular, it means that the sources can't be that diverse either, then it makes sense that "Others" is the least frequent one.

4) COMPUTE A TABLE WITH THE MAIN STATISTICS FOR EACH OF THE 4 CONTINOUS VARIABLES

Table 3: Main Statistics for each continuous variable

Parameters	Type	Score	Members	Popularity	Favorites
Minimum	Position	2,97	1006	1	0
Maximum	Position	9,63	1,45638E6	8267	106895
Range	Dispersion	6,66	1,45637E6	8266	106895
Interquartile Range	Dispersion	1,39	81279	2982	472
Average	Position	6,5038	79829,7	2626,89	1229,24
Median	Position	6,54	28443	2179	88
Variance	Dispersion	1,044484	1,8237E10	4,089E6	2,695E7
Standard Deviation	Dispersion	1,022	135045	2022,12	5191,5
Standard Skewness Coefficient	Shape	-1,69509	91,4452	18,1393	248,831
Standard Kurtosis Coefficient	Shape	-0,760781	237,404	-4,21739	1721,59

5) PLACE A HISTOGRAM AND A BOX-WHISKER PLOT OF THE VARIABLE SCORE

HISTOGRAM

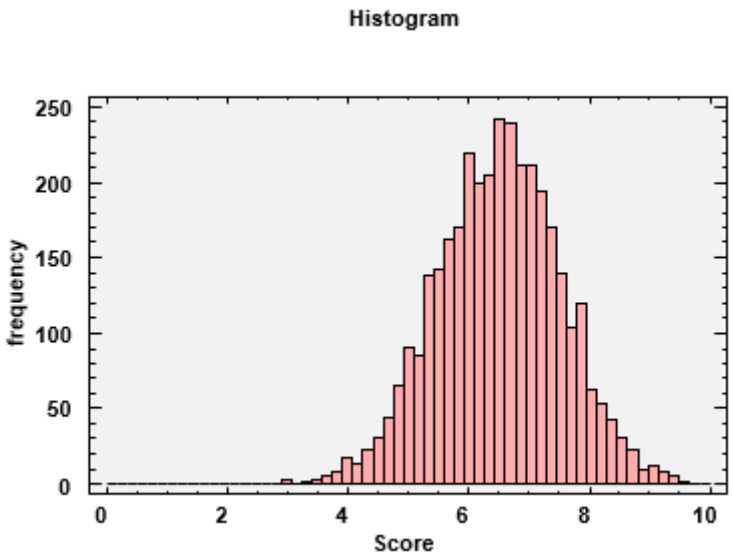

Figure 3: Histogram for "Score"

WHAT DOES THE VERTICAL SCALE OF THE HISTOGRAM REPRESENT?

It represents the absolute frequency of each value, that is, the number of times the values in that interval appear in all of the observations.

DO YOU CONSIDER THE NUMBER OF INTERVALS OF THE HISTOGRAM TO BE THE MOST SUITABLE?

Yes, it is. For it to be the most suitable, I computed the square root of the total amount of observations (3509) and obtained a value of 59,23. Then, changed the number of intervals of the histogram to 59.

INDICATE AN ADVANTAGE AND A DISADVANTAGE OF THIS GRAPHIC AS A DESCRIPTIVE TECHNIQUE

The histogram is one of the most used graphics to represent data. An advantage of the histogram would be that it is very useful to represent a large number of values, making it easy to compare the data. However, they are less useful and offer less information when dealing with a small amount of data. That is because it is harder to find the number of intervals that would show the information that we need, since using too many blocks can make analysis difficult, while too few can leave out important data.

BOX WHISKER PLOT

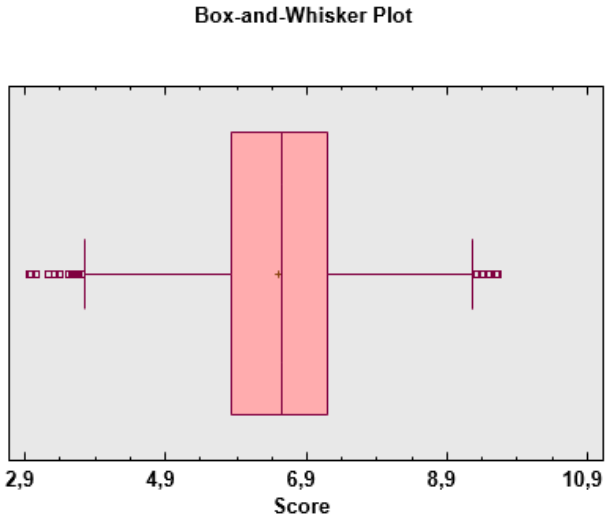

Figure 4: Box-Whisker Plot for "Score"

WHAT DOES THIS GRAPHIC PRETEND TO REPRESENT?

The Box-Whisker plot is a very useful graphic for representing small quantities of data. The box represents 50% of the values, that is, those that are comprised between the first and third quartiles. The whiskers represent the rest of the values, going from the minimum value to the maximum. It also represents some points that differ from the rest of the data and that could be considered abnormal.

INDICATE AN ADVANTAGE AND A DISADVANTAGE OF THIS GRAPHIC AS A DESCRIPTIVE TECHNIQUE

Box plots are useful as they provide a visual summary of the data making it easy to quickly identify mean values, the dispersion of the data set, and signs of skewness. Moreover, they are very useful to identify outliers, that is, extreme values that are considered abnormal. However, this plot only represents a summary of the data and lacks some information, for example the exact number of observations we are working with.

WHICH ONE GIVES A MORE USEFULL INFORMATION IN THIS CASE?

In this case, I consider the histogram to be more useful because the amount of data we are working with is very large and the extreme values of the box whisker plot are very close to the whiskers, which means that there are no outliers since the distribution is normal, so the box whisker plot does not give us any new information. The histogram is also better to observe that this variable follows a normal distribution.

6) STUDY AND DISCUSS THE PATTERN OF VARIATION OF THE VARIABLE SCORE ACCORDING TO THE PREVIOUS GRAPHICS

DISCUSS THE RANGE OF THE DATA, THE MEDIAN, THE AVERAGE AND THE MOST RELEVANT INFORMATION.

According to the Box-Whisker plot we know that the maximum value is approximately 9’6 and the minimum value is around 2’9, which are quite close to the real values. This means that the range of the values would be 6’7 and the interquartile range would be 1’4 approximately. As for the median, it is represented by the

line inside the box, which gives us a value of 6,6 approximately. The average is the point located inside the box and it represents the value 6,5. These values are very similar to those in the table of exercise 4, but they are not as accurate. As for the shape of the histogram we can see that it is more or less symmetric, and it resembles the shape of a normal distribution. There is no mix of populations.

TAKING INTO ACCOUNT THE STANDARD SKEWNESS AND KURTOSIS COEFFICIENTS, IS IT REASONABLE TO ASSUME A NORMAL DISTRIBUTION FOR THE DATA?

Yes, it is. As seen in the table of exercise 4, both of these coefficients belong to the interval [-2,2] which means that the distribution is normal.

ARE THERE ANY ABNORMAAL VALUES THAT SHOULD BE ERASED FROM THE STUDY?

No, there are not. In a normal distribution, those extreme values that are very close to the end of the whiskers should not be considered outliers and erased from the study. In this case, that is exactly what happens as seen in the Box-Whisker plot. All the extreme values are close to the end of the whiskers and they should not be erased. However, this only happens because the distribution is normal.

7) PLACE A HISTOGRAM AND A NORMAL PROBABILITY PLOT OF THE VARIABLE MEMBERS

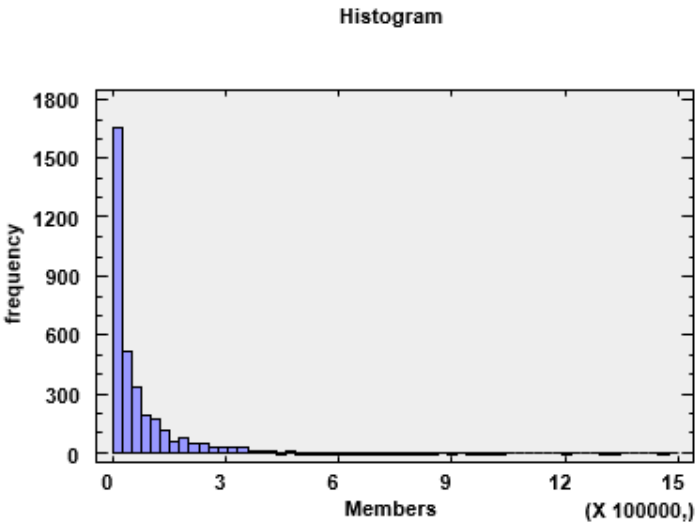

Figure 6: Histogram for "Members"

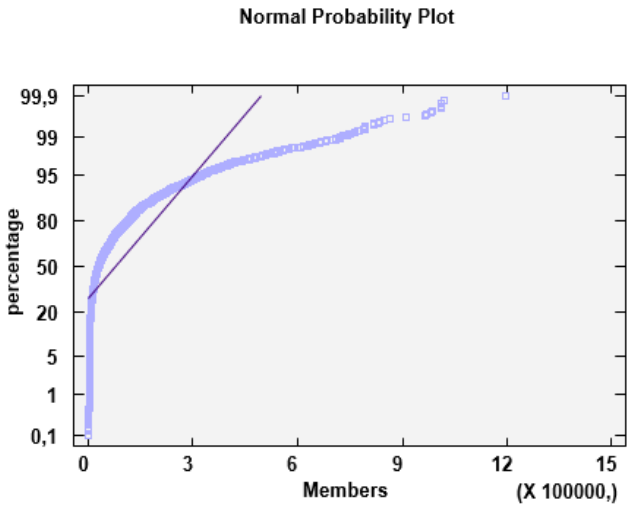

Figure 5: Normal Probability Plot for "Members"

WHAT DOES THE NORMAL PROBABILITY PLOT CONSIST OFF?

The normal probability plot is used to see if a set of data follows a normal distribution or not. Each point of this plot represents one observation, and by representing all the points we obtain its normal probability plot. The horizontal scale represents the value of each observation and the vertical scale represents the percentage of values that are less or equal to that of each observation, that is, the cumulative frequency's percentages of each value. By representing the cumulative percentage in the vertical scale, it makes it easy to find percentiles. Then, if the points clearly follow a straight line, it is normal. If not, then it does not follow a normal distribution.

INDICATE AN ADVANTAGE AND A DISADVANTAGE OF THE NORMAL PROBABILITY PLOT AS A DESCRIPTIVE TECHNIQUE

The normal probability plot is very useful to quickly see if a distribution is normal or if it is skewed. However, this plot can be very confusing if the distribution is not normal, making it hard for us to make assumptions based on it.

WHICH GRAPH DO YOU THINK GIVES US THE MOST USEFUL INFORMATION IN THIS CASE?

Since this distribution is strongly skewed, the normal distribution does not give us the best insight. The histogram is more useful since it gives us exactly the same thing as the normal probability plot (we can clearly see that it is strongly skewed in both of them) and more, because with it we can know which range of values have the most observations.

8) STUDY AND DISCUSS THE PATTERN OF VARIATION OF THE VARIABLE MEMBERS ACCORDING TO THE PREVIOUS GRAPHICS

DISCUSS THE RANGE OF THE DATA, THE MEDIAN, THE AVERAGE AND THE MOST RELEVANT INFORMATION.

Looking at the histogram and the normal probability plot we cannot know the exact values of minimum and maximum but let us assume that the minimum is 0 and the maximum is around 15E5. This would mean that the range is 15E5, which is not too accurate. As for the average, there is no way for us to know the exact value just by looking at the plots. The median, however, can be known by looking at the normal probability plot. If we look at the point which has the value 50 in the vertical scale we can more or less know its value on the horizontal scale, which would be around 28000 according to the tool “locate” of statgraphics. By looking at both plots we can clearly see that the distribution is asymmetric.

IF THE DISTRIBUTION IS NOT SYMMETRIC, INDICATE ITS INTENSITY AND SIGN

This distribution is clearly not symmetric. It is strongly and positively skewed since CA>>>2.

ARE THERE ANY ABNORMAL VALUES THAT SHOULD BE ERASED FROM THE STUDY?

By computing the normal probability plot of the logarithm of the variable “Members” we observe that there are not any points that are clearly abnormal. Instead, all points follow the same curve, which means that we should not erase any of them, even if they do not form a straight line.

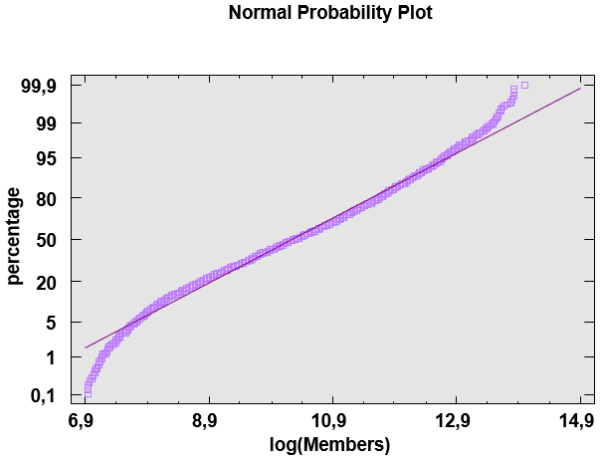

Figure 7: Normal Probability Plot of the logarithm of "Members"

9) PLACE A HISTOGRAM OF THE VARIABLE POPULARITY AND A MULTIPLE BOX-WHISKER PLOT IN FUNCTION OF THE VARIABLE TYPE

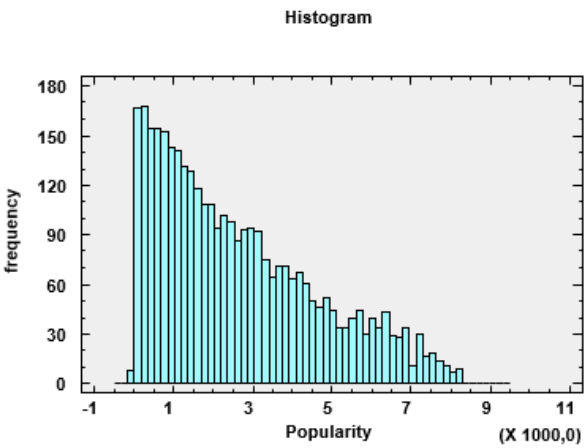

Figure 9: Histogram of "Popularity"

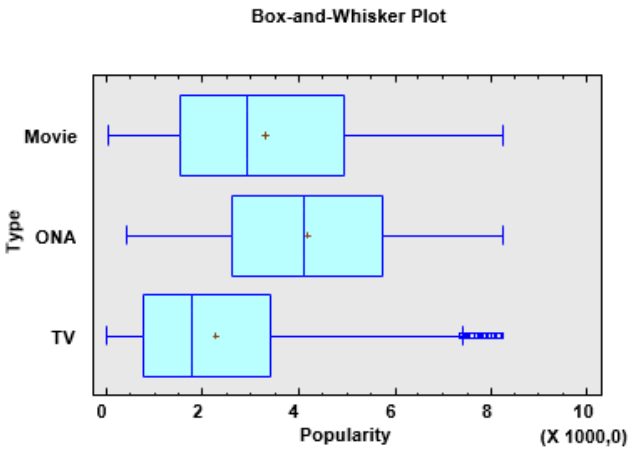

Figure 8: Multiple Box-Whisker plot of "Popularity" in function of "Type"

DISCUSS ABOUT WHAT THE MULTIPLE BOX-WHISKER PLOT CONSISTS OF

This plot consists of computing a box whisker plot of one variable, but instead of doing it with all the values, it computes one box whisker plot for each variant of a continuous variable, taking only the data which have that specific variant for each plot. In this case, since "Type" has three variants, it computed three box whisker plots.

COMPUTE A TABLE WITH THE STANDARD SKEWNESS AND KURTOSIS COEFFICIENTS OF THE VARIABLE "POPULARITY" IN FUNCTION OF THE DIFFERENT VARIANTS OF "TYPE"

By taking the values of "Popularity" and separating them according to which type they belong to, you can use Statgraphics to compute the Standard Skewness and Kurtosis Coefficients of Popularity according to each "Type". Doing this we obtain the following:

Table 4: Parameters of Shape of "Popularity" in function of the variable "Type"

	Movie	ONA	Type
Standard Skewness Coefficient	5,76599	0,626506	18,1092
Standard Kurtosis Coefficient	-4,82539	-2,97784	0,0595737

10) STUDY AND DISCUSS THE PATTERN OF VARIATION OF THE VARIABLE POPULARITY FOR EACH VARIANT OF "TYPE"

The variable "Popularity" has the following coefficients: CA_{std}= 18,1393 and CC_{std}= -4,21739, meaning that it is positively and moderately skewed and it is also a platikurtic distribution. This information will be used in the following discussion.

Looking at the Box-Whisker plots and the table of coefficients we can deduce the following information:

- For **Movie**, the median is not in the center of the box, instead it is located to the left. Moreover, it has the right whisker longer than the left one. These are all signs of a positively skewed distribution, meaning that it is asymmetric. We can also see that the Box-Whisker has no extreme values and therefore no outliers, which means that there are no data that should be erased from the study. The Standard Coefficients also indicate that the distribution is not normal, since none of them belong to the interval [-2,2]. This means that the distribution is moderately and positively skewed since CA_{std}>2. As for the CC_{std}, it is lower than -2 which means that we have platykurtic data. This does not differ from the original distribution of the variable "Popularity" since both the histogram and the coefficients match this information. Therefore, we have that for "Movie", the variable "Population" follows a similar distribution as the whole variable "Popularity" itself.

- For **ONA**, the median is almost at the center of the box and both whiskers present the same length. Like in the previous case, there are no extreme values that could be considered outliers and therefore, no data should be erased. The plot presents a symmetric shape, unlike the previous case. By looking at the standard coefficients we can also see that the CA_{std} belongs to the interval [-2, 2] which means that it is symmetric. However, we cannot say that it is normal since the CC_{std} does not belong to said interval. Instead, it is lower than -2, meaning that we have platykurtic data. In this case, the distribution is not that similar to that of the whole variable "Population" since it is symmetric.

- Lastly, for **TV**, we have that the median is not in the center of the box, but to the right. We also have that the right whisker is way longer than the left one and there are some extreme values close to the end of the right whisker. These are all signs of a positively skewed distribution, which we can also see by noticing that the CA_{std} is bigger than 2. However, it cannot be considered to be heavily skewed, so let us say that it is moderately skewed. Furthermore, all the extreme values are very close to the right whisker and are simply a sign of skewness, which means that they should not be considered outliers and should not be erased from the study. Also, by looking at the CC_{std}, we see that it is very close to 0, and it is inside the interval [-2,2]. This could mean that the distribution is normal, however, the CA_{std} does not belong to this interval. Therefore, this distribution is asymmetric and moderately and positively skewed, like the distribution of the whole variable "Popularity". However, it is not a platykurtic distribution.

11) TO DESCRIBE GRAPHICALLY THE PATTERN OF VARIATION OF FAVOURITES, CHOOSE THE GRAPHIC THAT GIVES THE MOST INFORMATION

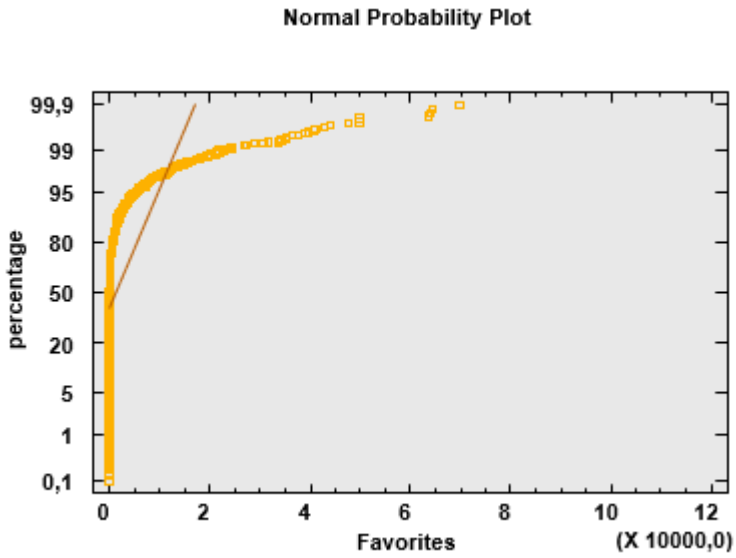

Figure 10: Normal Probability Plot of "Favorites"

WHY DID YOU CHOOSE THIS GRAPHIC?

I chose this graphic because the distribution of the data is highly skewed and there are some values that differ from the others, meaning that with the histogram, some of the bars were so small that you could not see them, since there were only a couple of values in that interval. As for the Box-Whisker plot, the box was so narrow you could barely see it and it was much more confusing.

DISCUSS THE MOST IMPORTANT INFORMATION DEDUCED FROM THE GRAPHIC

From the graphic we can see that the distribution is highly and positively skewed since it follows that kind of curve. The distribution is clearly not normal or symmetric, it is instead asymmetric. Moreover, we see that most values are very close to zero, being the ones further from it points that are isolated. However, since the points that are isolated still follow the same curve as the rest of the distribution, we can assume that they are not abnormal values and should not be removed from the study. However, to be sure let us represent the logarithm of the variable "Favorites". By doing it we can see that there are no points that clearly diverge from the rest of the values and therefore we do not have any outliers in the distribution.

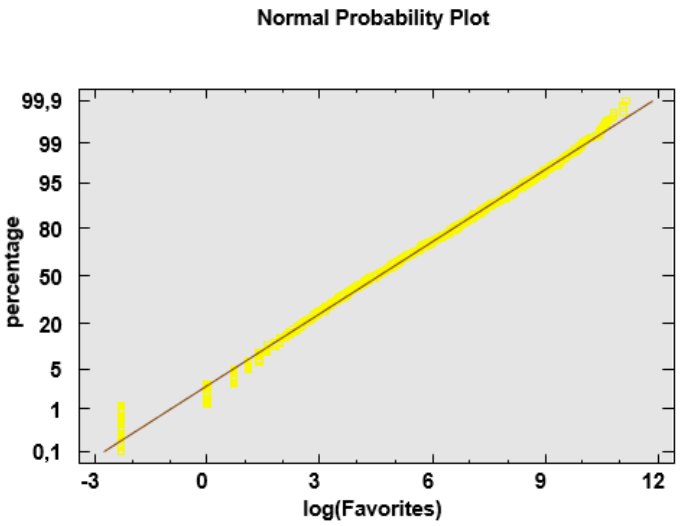

Figure 11: Normal Probability Plot of the logarithm of "Favorites"

12) WITH THE VARIABLE "SCORE", CHOOSE RANDOMLY ONE HALF OF THE VALUES AND MULTIPLY THEM BY 2. PLOT A HISTOGRAM WITH ALL OF THE DATA. DISCUSS THE RESULTS

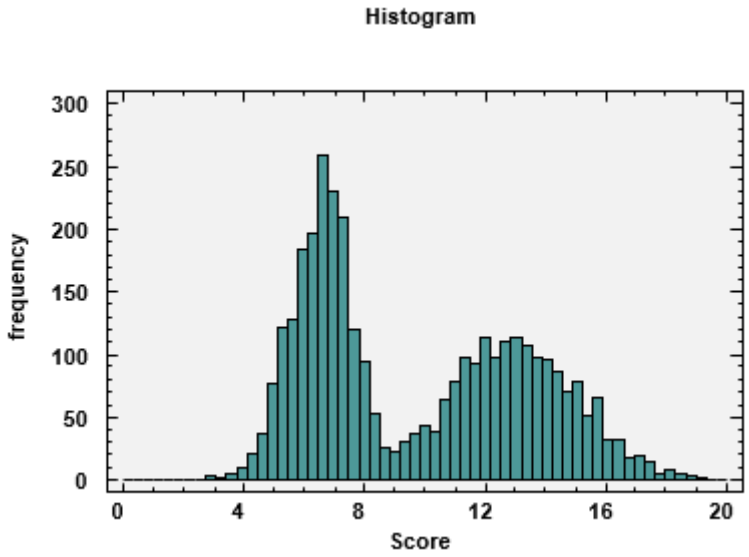

Figure 12: Histogram of mixed populations

After multiplying half of the values by 2 and plotting the histogram we see that we do not have a normal distribution anymore. Instead, we have a shape that resembles two normal distributions together. This means that we do indeed have a bimodal distribution, as if we have two different populations. This happens because we multiplied half of the values by 2, creating a "new population" that differs from the previous one. By computing the histogram of all the values, we are mixing the two populations and obtaining a bimodal distribution. That is also why we have a peak much higher than the other.

DISCRETE AND CONTINUOUS DISTRIBUTIONS

13) SIMULATE A POISSON VARIABLE THAT HAS THE SAME AVERAGE AS "SCORE". PLACE THE GRAPHIC OF THE DENSITY FUNCTION OF THE NEW VARIABLE AND A HISTOGRAM OF "SCORE". WHAT COULD BE DEDUCED FROM BOTH GRAPHICS?

*Since the only discrete variables that I have considered have a really big range of values and they behave as if they were continuous variables, I decided to do the simulation of the Poisson variable for this exercise. Moreover, in exercise 14 I have to do the same thing that I was asked to do in this exercise, so it would have been repeated. For this reason, I decided to do this option of the exercise and not the other one.

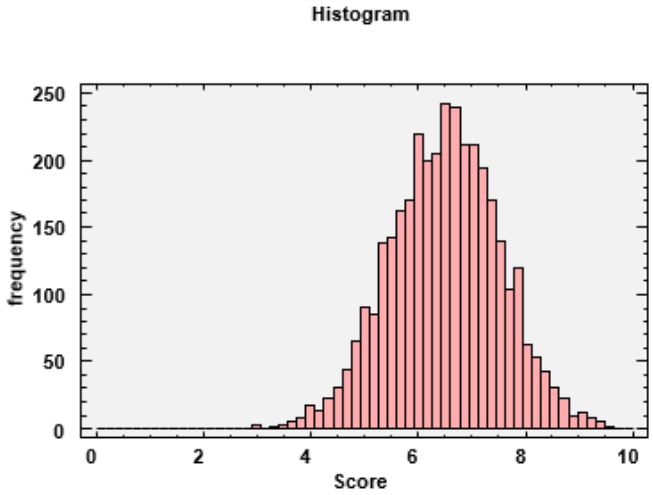

Figure 13: Histogram of "Score"

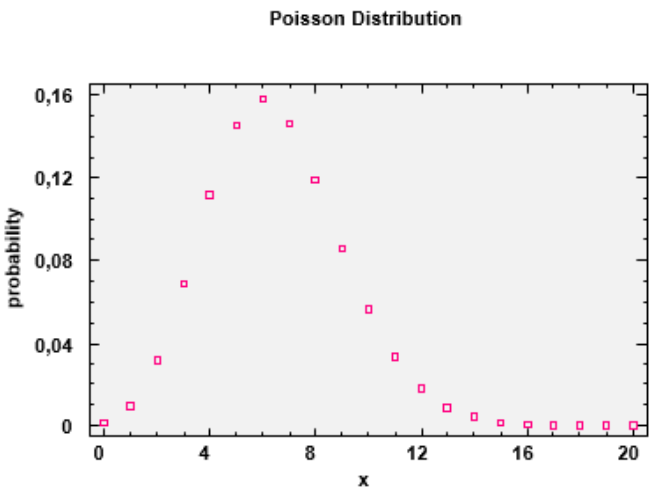

Figure 14: Density function of the Poisson distribution

We know that the variable Score follows a normal distribution. If we compute a Poisson distribution with its average equal to the average of "Score", then we will get a similar shape to that of the histogram of Score like we can see in the pictures. This happens because a Poisson distribution can be approximated to a normal distribution with its average and variation equal to the average of the Poisson distribution. That is, since we computed a Poisson distribution with λ=6'5038, this distribution can be approximated to a Normal distribution with parameters m=6,5038 and σ=2'55 (square root of λ). Taking into account that Score follows a distribution N(m=6,5038, σ= 1,022), these distributions are pretty similar and will have a very similar shape to each other, as we can see in the pictures.

14) FOR EACH CONTINUOUS VARIABLE WHOSE DISTRIBUTION IS NOT NORMAL, STUDY THE GOODNESS OF FIT FOR DIFFERENT MODELS OF CONTINUOUS DISTRIBUTIONS: UNIFORM, EXPONENTIAL OR LOG-NORMAL. PLACE THE GRAPHIC OF THE BEST FIT. DISCUSS THE CONCLUSIONS.

MEMBERS

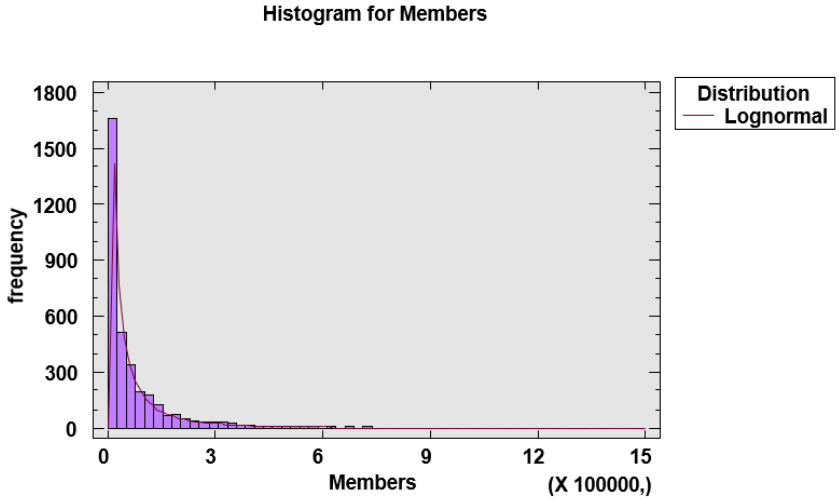

Figure 15: Fit of a lognormal distribution to "Members"

After fitting the variable Members to the three distributions mentioned, we can conclude that this variable does not come from any of the three distributions. We can say that because Statgraphics vives us a P-value that is less than 0,05 in the three distributions. However, the one that gives us a value closer to 0,05 in the log-normal distribution, with a value of 0,000244837. Both of the other two gave us a value of 0. We can see in the graphic that the shape of the histogram is kind of similar to that of the lognormal distribution, but it doesn't exactly fit to be considered a lognormal distribution.

POPULARITY

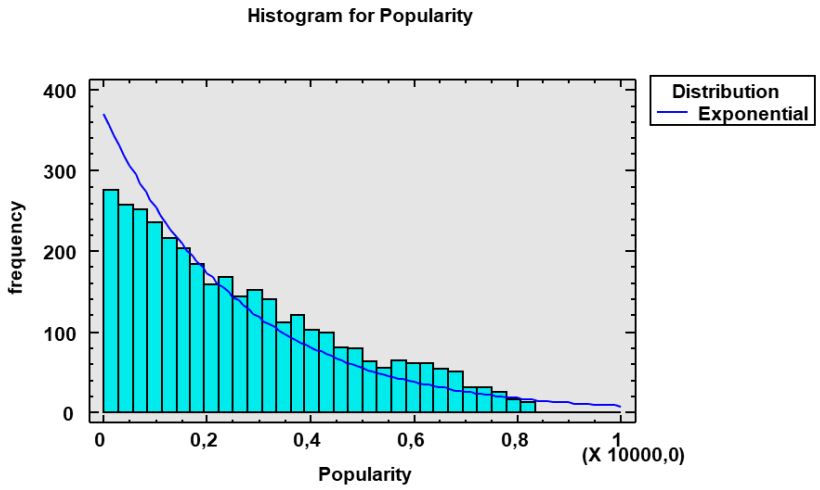

Figure 16: Fit of an exponential distribution to "Popularity"

For the variable Popularity none of the three distributions fit correctly. All of them have a P-Value of 0, meaning that Popularity does not come from either of the three, making it difficult to choose the best one.

Just by looking at the graphics, I decided that the exponential is the best fit, simply because of how the shape is kind of similar, but still not good enough. The only information that we get from this analysis is that popularity does not follow a lognormal distribution, an exponential distribution nor a uniform distribution.

FAVORITES

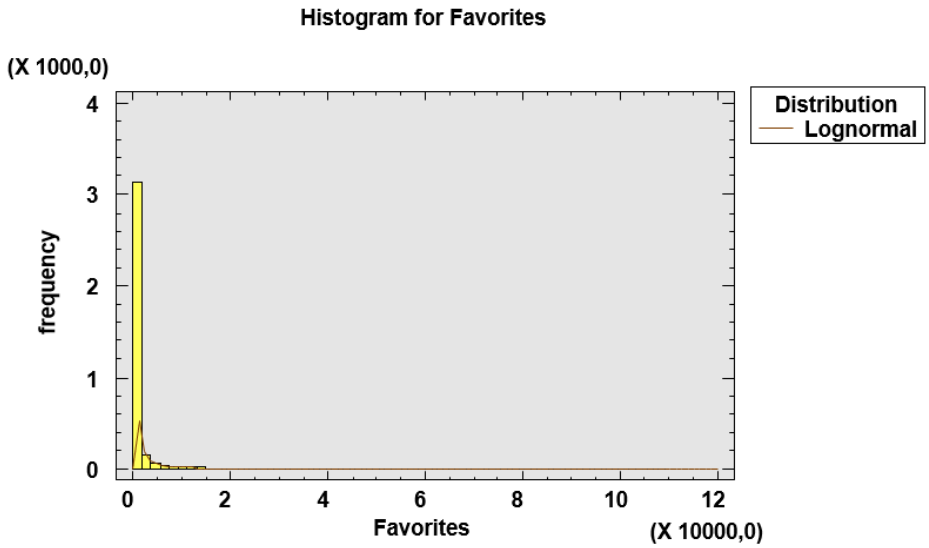

Figure 17: Fit of a lognormal distribution to "Favorites"

For the variable Favorites, both the exponential and the uniform distributions had a P-Value of 0, meaning that we can reject the possibility that this variable comes from any of those two distributions. However, with the lognormal distribution it had a P-Value of 0,0479515, which is almost 0,05. It still does not mean that it comes from a lognormal distribution since it is still less than 0,05, but we cannot say that it is not a lognormal distribution with 100% confidence. Therefore, we will use the next exercise to check if this distribution is in fact lognormal or not.

15) FOR EACH CONTINUOUS VARIABLE WITH AN ASSYMMETRIC POSITIVE DISTRIBUTION, STUDY IF ANY TRANSFORMATION IS ABLE TO "NORMALIZE" THE DATA. REPRESENT THE NORMAL PROBABILITY PLOT OF EACH VARIABLE, OF ITS SQUARE ROOT, ITS FOURTH ROOT AND ITS LOGARITHM.

OF ALL OF THE TRANSFORMATIONS, DISCUSS THE ONE THAT BEST NORMALIZES THE DATA AND PLACE ITS NORMAL PROBABILITY PLOT. JUSTIFY YOUR ANSWER.

MEMBERS

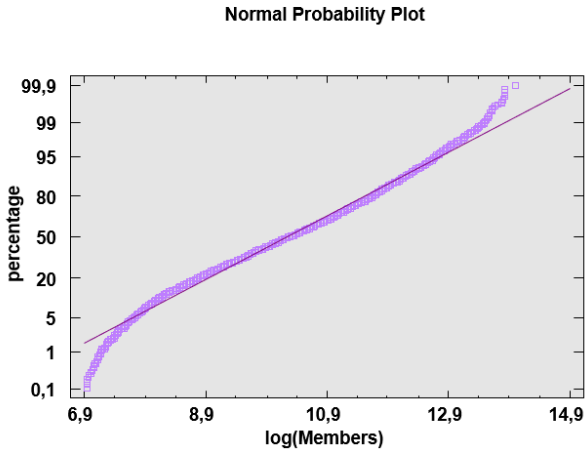

Figure 18: Normal Probability plot of the logarithm of "Members"

For the variable "Members" the transformation that best normalizes the data is the one of the logarithm. That is because its standard skewness is within the interval [-2,2], with a value of -0,468447. However, the standard kurtosis is not withing this interval, having a value of -9,37759. This means that this distribution is not totally normalized since only one of the two coefficients is between -2 and 2. For both of the other transformations, none of the coefficients were within the interval [-2,2], that is the reason I chose this as the most normalized.

POPULARITY

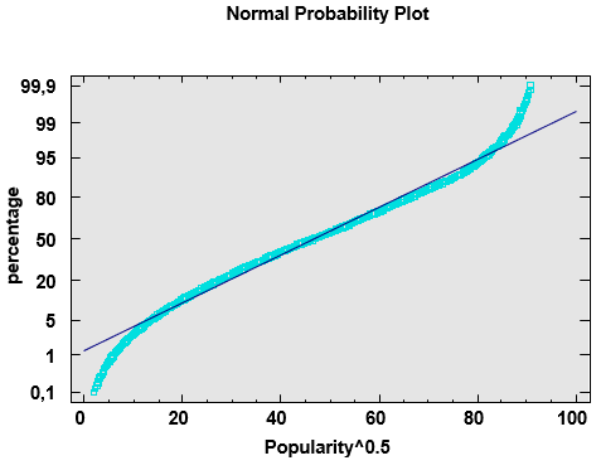

Figure 19: Normal Probability Plot of the square root of "Popularity"

For the variable popularity, the best transformation is the square root since it has a standard skewness of 0,989697 and a standard kurtosis of -10,2817. For all the other transformations, none of the coefficients was

within the interval [-2,2]. For this reason, the square root is the best option since at least the standard skewness is between -2 and 2. However, it does not normalize the data completely, because for that the standard kurtosis should have a value between -2 and 2 as well.

FAVORITES

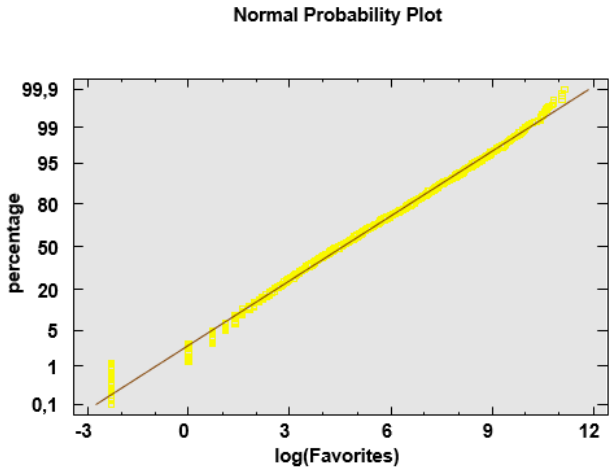

Figure 20: Normal Probability plot of the logarithm of "Favorites"

For the variable Favorites, the best transformation is definitely the logarithm since both its standard skewness (0,566484) and kurtosis (-0,130617) are within [-2,2], This means that this transformation is able to completely normalize the data that we have, unlike the other two transformations. This is the only variable of the three that has been normalized by one of the transformations applied to it.

ACCORDING TO THE RESULTS, CAN WE SAY THAT ANY OF THE VARIABLES FOLLOWS A LOG-NORMAL DISTRIBUTION?

Yes, the variable “Favorites” follows a lognormal distribution. In exercise 14 we arrived at the conclusion that Favorites could follow a lognormal distribution because its P-Value was very close to 0’05, but since it wasn’t more than 0’05, we could not say for sure that it was a lognormal distribution. However, now we have checked that by applying this transformation to the variable “Favorites” we are able to normalize the data. Then, we can say for sure that “Favorites” follows a lognormal distribution, that is, that by applying logarithm to all of the values we get a normal distribution.

16) WHAT VARIABLE OF THE STUDY FOLLOWS A NORMAL DISTRIBUTION? WHAT ARE ITS PARAMETERS? JUSTIFY YOUR ANSWER

The variable “Score” follows a normal distribution with average 6’5038 and standard deviation 1,022. We can say that this variable follows a normal distribution with 100% confidence because both the standard skewness and kurtosis are within the interval [-2,2]. For this reason and by looking at its normal probability plot, we know that it follows this kind of distribution N(m=6’5038, σ=1’022).

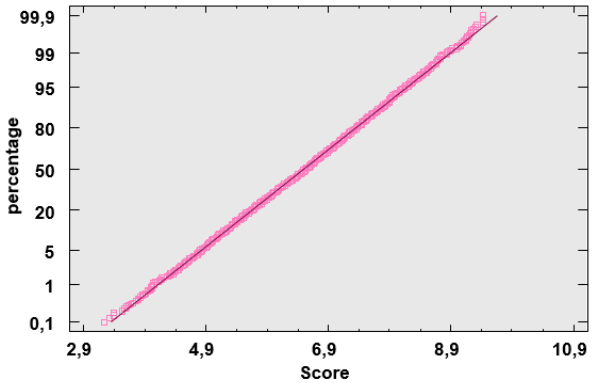

Figure 21: Normal Probability Plot of "Score"

17) GENERATE 100 RANDOM VALUES OF A NORMAL DISTRIBUTION WITH AVERAGE 10 AND STANDARD DEVIATION 4, AND 100 MORE VALUES FOLLOWING A NORMAL DISTRIBUTION WITH AVERAGE 2 AND STANDARD DEVIATION 3. GENERATE A NEW VARIABLE BY SUMMING THE VALUES BY PAIRS.

FROM THE VALUES OF THE STANDARD SKEWNESS AND KURTOSIS COEFFICIENTS, STUDY IF THE VARIABLE SUM FOLLOWS A NORMAL DISTRIBUTION. IF NOT, EXPLAIN WHY.

Yes, it does. The standard skewness has a value of -0,157718 and the standard kurtosis has a value of 0,986538, both within the interval [-2,2], meaning that the distribution is normal. This was to be expected, since by summing two normal distributions you always get another normal distribution.

COMPUTE THE AVERAGE AND THE STANDARD DEVIATION OF THE VARIABLE SUM WITH STATGRAPHICS.

According to Statgraphics, this variable, obtained by summing the two normal distributions, has an average of 12,294 and a standard deviation of 5,22703.

COMPUTE THEORETICALLY THE AVERAGE AND THE STANDARD DEVIATION THAT WOULD BE EXPECTED FROM THE VARIABLE SUM.

Since the variable “Sum” is obtained by summing two normal distributions, the average and the standard deviation that should be expected are:

- Average: m = m1 + m2 = 10 + 2 = 12
- Standard Deviation: σ = $\sqrt{\sigma^2} = \sqrt{\sigma 1^2 + \sigma 2^2} = \sqrt{16 + 9} = \sqrt{25} = 5$

WHY DO THE THEORETICAL PARAMETERS NOT COINCIDE WITH THE OBSERVED ONES?

That is because we are generating random numbers for both of the normal distributions and saving them. Since we are only taking 100 random values that follow that kind of distribution, the average and the standard deviation are never going to be totally accurate. For instance, for the first normal distribution with average 10 and standard deviation 4 we obtain values of 10,3828 and 3,78979 respectively. That is because 100 values are not enough to obtain an accurate distribution. If we increase the number of values from 100 to 10000, then we obtain an average of 10,0319 and a standard deviation of 4,01758 for the first distribution, which are much more accurate. The parameters of the variable sum computed theoretically do not take into account the random values that we are generating, but Statgraphics does. That is why they do not match how we would expect them to. However, the more random values you generate, the more accurate they become. Since in this case we only have 100 values it is normal to get an average and standard deviation that are only an approximation of the theoretical ones.