STATISTICAL ANALYSIS OF ANIME DATA

MyAnimeList

NAME: MARTA FILIPA RAMALHO

GROUP: E

SUMMARY

In this assignment I have discussed different aspects about anime and relations between its variables. To discuss the relations and see what distributions the variables follow, I used the software Statgraphics which provides us with enough tools to make all the possible statistical analysis that we would be interested in making.
In the assignment we have discussed problems and obtained conclusions about the following topics: Descriptive Statistics, Discrete and Continuous Distributions, Inference, ANOVA and Regression. The main **conclusions** obtained from the assignment are the following:

The most predominant variant of Source is Manga and the least predominant one is Web Manga. By making a table of crossed frequencies of Type and Source we concluded that the most predominant combination is TV and Manga. By analyzing the qualitative variables, we obtained that the only variable that followed a normal distribution was Score, while the other three had a positively skewed distribution.

We also studied which kind of transformation was the best option to **normalize** the positively skewed variables, obtaining that the only one that could be normalized was the variable Favorites by applying the logarithm. The best option for Popularity was the square root and for Members was the logarithm, even though they were not completely normal. We also studied the **goodness of fit** of these variables to different kinds of distributions, obtaining that the best fit for Favorites and Members was the Log-Normal distribution while the best fit for Popularity was an exponential distribution.

Then, we obtained different intervals and probabilities for the samples that could be taken from the normal variable Score, both by hand and with Statgraphics. We also concluded that there were no statistically significant differences between the standard deviations of the variable Score by separating it into the three variants of Type.

After that, we made **ANOVA** analysis to study the effect of the qualitative variables in the quantitative variables, obtaining that the effect of Score and Type, as well as their interaction, was statistically significant for both Score, Members and Popularity. This was concluded after analysing the LSD intervals and the P-Values associated. By studying the distribution of the residuals we checked that there were no outliers that should be erased.

Finally, we checked that the couple of quantitative variables with higher correlation coefficient were Score and Favorites. By plotting the dispersion graphic, we observed that the relation was linear, moderate and positive. Then we identified a cause effect relation between Members and Popularity. We applied regression to it, obtaining a very strong and negative correlation between both variables. By analyzing the residuals, we found one abnormal value that was erased from the study. We also identified a quadratic effect between the residuals of the regression model and the independent variable Members.

DISTRIBUTIONS IN SAMPLING – INFERENCE ABOUT ONE POPULATION

1) ASSUME THAT THE VARIABLE SCORE FOLLOWS A NORMAL DISTRIBUTION WITH AVERAGE EQUAL TO THE SAMPLE AVERAGE AND STANDARD DEVIATION EQUAL TO THE SAMPLE VALUE. IF WE TAKE A RANDOM SAMPLE WITH 5 VALUES AND WE COMPUTE ITS AVERAGE, COMPUTE THE CONFIDENCE INTERVAL THAT WOULD COMPRISE THE 99% OF THESE VALUES.

Assuming that the variable Score follows a normal distribution N(m=6,5038; σ=1,022) and taking 5 random numbers, we have that N=5.

We know that $\frac{\bar{x}-m}{\sigma/\sqrt{N}} \sim N(0,1)$, that is, we know that $\frac{\bar{x}-m}{\sigma/\sqrt{N}}$ follows a normal distribution with m=0 and σ=1, being $\bar{x}$ the average of the sample of 5 values. Now, since we want the interval that would comprise the 99% of the values, we need to compute this interval for N(0,1) with Statgraphics. Doing so, we obtain the interval [-2,5758; 2,5758], which comprises the 99% of the values of a normal distribution N(0,1).

Now, we know that $\frac{\bar{x}-m}{\sigma/\sqrt{N}}$ belongs to the interval [-2,5758; 2,5758], so let us compute the interval to which $\bar{x}$ belongs.

$$\bar{x} \in \left[-2{,}5758\frac{\sigma}{\sqrt{N}}+m;\ 2{,}5758\frac{\sigma}{\sqrt{N}}+m\right] = \left[-2{,}5758\frac{1{,}022}{\sqrt{5}}+6{,}5038;\ 2{,}5758\frac{1{,}022}{\sqrt{5}}+6{,}5038\right] = [5{,}3265; 7{,}6811]$$

Therefore, the confidence interval that we are looking for is [5,3265; 7,8911].

2) COMPUTE WITH STATGRAPHICS THE PERCENTILE 48 OF THE VARIABLE SCORE.

According to Statgraphics, the percentile 48 of the variable Score is equal to 6,49.

SOLVE WITH THE APPROPIATE FORMULAS THE CONTRAST OF HYPOTHSEIS H_0: M = Z_{48} AND H_1: M ≠ Z_{48}. CONSIDER AS SIZE OF THE SAMPLE THE NUMBER OF OBSERVATIONS OF THE DATASET AND A SIGNIFICANCE LEVEL OF 1%.

Null hypothesis: H_0: m = Z_{48}

Alternative hypothesis: H_1: m ≠ Z_{48}

We have computed with Statgraphics that Z_{48}= 6,49. The average of the sample is 6,5038 and the standard deviation of the sample (which is equal to all of the observations) is 1,022.

We need to obtain the value of t_{N-1} from the tables. By doing so, we obtain that t_{3508}=2’581 with α=0,01.

Now, let us compute the tcalc $= \frac{\bar{x}-m}{s/\sqrt{N}} = \frac{6{,}5038-6{,}49}{1{,}022/\sqrt{3509}} = 0{,}799871$

Since |tcalc|<t_{3508} we do not have enough evidence to reject H_0 and therefore we accept the hypothesis that m=6,49.

SOLVE THE SAME CONTRAST OF HYPOTHESIS WITH STATGRAPHICS AND VERIFY THAT WE OBTAIN THE SAME VALUE AND THE SAME CONCLUSIONS

With Statgraphics we put the following data:

Sample mean = 6,5038
Sample standard deviation = 1,022
Sample size = 3509
Alpha=0,01
Null Hypothesis: mean = 6,49
Alternative: not equal

And we obtain the following parameters:

Computed t statistic = 0,799871 (the same value that we obtained previously)
P-Value = 0,423784

Therefore, since the P-value (0,423784) is bigger than alpha (0,01), we cannot reject the null hypothesis and we obtain the same conclusion that we did by making the computations by hand.

3) ASSUME THAT THE VARIABLE SCORE FOLLOWS A NORMAL DISTRIBUTION WITH AVERAGE EQUAL TO THE SAMPLE AVERAGE AND STANDARD DEVIATION EQUAL TO THE SAMPLE VALUE. IF WE TAKE A RANDOM SAMPLE WITH 5 VALUES AND WE COMPUTE ITS VARIANCE, COMPUTE THE CONFIDENCE INTERVAL THAT WOULD COMPRISE THE 90% OF THESE VALUES.

Assuming that the variable Score follows a normal distribution N(m=6,5038; σ=1,022) and taking 5 random numbers, we have that N=5.

We know that $(N-1)\frac{s^2}{\sigma^2}$ follows a Chi Square distribution with N-1 degrees of freedom, in this case it follows a Chi Square distribution with 4 degrees of freedom and being s^2 the variance of the sample. We have to compute the interval that comprises the 90% of the values of a Chi Square distribution X^2_4. We compute this interval with Statgraphics and obtain [0,7107; 9,4877].

Now we know that $(N-1)\frac{s^2}{\sigma^2}$ belongs to the interval [0,7107; 9,4877]. We need to compute the interval to which s^2 belongs: $s^2 \in \left[0{,}7107\frac{\sigma^2}{N-1}; 9{,}4877\frac{\sigma^2}{N-1}\right] = \left[0{,}7107\frac{1{,}022^2}{4}; 9{,}4877\frac{1{,}022^2}{4}\right] = [0{,}1856; 2{,}4774]$

Therefore, the confidence interval for the variance in 90% of the cases is $[0{,}1856; 2{,}4774]$.

4) ASSUME THAT THE VARIABLE SCORE FOLLOWS A NORMAL DISTRIBUTION WITH AVERAGE EQUAL TO THE SAMPLE AVERAGE AND STANDARD DEVIATION EQUAL TO THE SAMPLE VALUE. IF WE TAKE A RANDOM SAMPLE WITH 8 VALUES AND WE COMPUTE ITS VARIANCE, WHAT IS THE PROBABILITY OF IT BEING BIGGER THAN 2*STD DEVIATION.

We assume that the variable Score follows a normal distribution N(m=6,5038; σ=1,022). Let us consider that we take 8 random values, which would mean that N=8.

Knowing that $(N-1)\frac{s^2}{\sigma^2}$ follows a Chi Square distribution with N-1 degrees of freedom, we can compute the probability we are asked to in the following way:

P(s^2>2σ)=$P\left(s^2\frac{N-1}{\sigma^2} > 2\sigma\frac{N-1}{\sigma^2}\right) = P(X_7^2 > 13{,}69863)$. This probability can be computed by looking at the table of the Chi Square distribution, however this would not give us an exact value, so the best option is to compute it using Statgraphics. By doing so, we obtain a value of 0,0568.

Therefore, the probability that the variance of a sample of 8 values exceeds the value of 2σ is very low.

5) IF WE TAKE TWO SAMPLES OF 15 VALUES OF THE VARIABLE SCORE, WHAT IS THE PROBABILITY THAT THE VARIANCE OF THE SECOND SAMPLE IS MORE THAN TWICE THE VARIANCE OF THE FIRST ONE?

If we take two samples of 15 values, that will mean that N=15 for both samples. Let us call s_1^2 to the variance of the first sample and s_2^2 to the variance of the second one.

We know that $\frac{s_1^2/\sigma^2}{s_2^2/\sigma^2}$ follows a Fisher distribution, to be precise, it follows a $F_{N-1,\,N-1}$. We know that N is 15 for both cases, then $\frac{s_1^2/\sigma^2}{s_2^2/\sigma^2} \sim F_{14,14}$. Since we know that both samples come from the same population, then σ^2 will be the same in both cases. This means that we can simplify it, getting $\frac{s_1^2}{s_2^2} \sim F_{14,14}$.

We want to compute the probability that $s_2^2 > 2s_1^2$. We can compute this probability in the following way:

P(s_2^2>2s_1^2)=$P\left(\frac{s_2^2}{s_1^2} > 2\right)$. Since we know that $\frac{s_1^2}{s_2^2} \sim F_{14,14}$ then we can say that $P\left(\frac{s_2^2}{s_1^2} > 2\right) = P(F_{14,14} > 2) = 0{,}103539$. This probability was computed with Statgraphics.

Therefore, we can conclude that the probability of s_2^2 being greater than 2 s_1^2 is 0,103539.

6) OBTAIN WITH STATGRAPHICS A CONFIDENCE INTERVAL FOR THE AVERAGE OF SCORE AT A POPULATION LEVEL, WITH A CONFIDENCE LEVEL OF 99%.

With m=6.5038, σ=1.022, N=3509 and α=0.01, we obtained a confidence interval of [6,45936;6,54824] with Statgraphics.

WHAT WOULD HAPPEN IF SCORE DID NOT ADJUST TO A NORMAL DISTRIBUTION?

These kinds of inference techniques assume that the data we are working with follows a normal distribution. That is why it is important to make sure that the data are normally distributed before making these kinds of assumptions. If Score did not adjust to a normal distribution, then the confidence interval that we obtained would not be useful and the study would lack any type of sense. These techniques only apply to normally distributed data.

COMMENT YOUR OPINION ABOUT THE FOLLOWIND STATEMENT: "IF WE TAKE ANY VALUE INSIDE THE OBTAINED INTERVAL AND WE MAKE A HYPOTHESIS TEST ABOUT THE MEAN, THE CONCLUSION OF SAID TEST WILL ALWAYS BE THE SAME CONSIDERING α=1%"

If we take any value inside the obtained interval and do a hypothesis test about the mean, we will always obtain the same conclusion since being inside the interval means that we cannot reject the hypothesis and we have to accept that m=x being x any value inside the interval and m the average of the population.

7) OBTAIN WITH STATGRAPHICS A CONFIDENCE INTERVAL FOR THE STANDARD DEVIATION OF SCORE AT A POPULATION LEVEL, WITH A CONFIDENCE LEVEL OF 95%. COMPUTE THE INTERVAL WITH A CONFIDENCE LEVEL OF 99% AS WELL.

For a confidence level of 95%, that is, with α=0.05, we obtain an interval of [0,998637;1,04649]. On the other hand, for a confidence level of 99%, that is, with α=0.01, we obtain an interval of [0,99145;1,05436].

WHAT INTERPRETATION DOES THE INTERVAL HAVE IN PRACTICE?

There exists a 1-α probability that the interval obtained comprises the real standard deviation of the population. This does not mean that there is a 1-α probability that the standard deviation is inside the interval, but that the interval contains the standard deviation. If we take any of the values inside the interval as a hypothesis for the standard deviation, it will always be accepted for that specific α.

WHICH INTERVAL DO YOU FIND MORE SUITABLE? WHAT DOES THIS DECISION DEPEND ON?

I find the interval [0,99145;1,05436] more suitable, that is, the interval for a confidence level of 99%. This decision depends only on the value of α, or the value of the confidence level, which would mean the same since confidence level=1-α. Since we are looking for the best interval which has the most probability of containing the real standard deviation, the one with the highest confidence level would be the best option, even if with a higher confidence level the risk of making a type II error also increases. This way, we have an interval that has a 0.99 probability of containing the standard deviation of the whole population.

8) INDICATE IN A TABLE THE VARIANCE OF SCORE AND THE NUMBER OF DATA FOR EACH ONE OF THE VARIANTS OF TYPE.

Table 1: Variance of Score for variants of Type

Type	Score	N
Movie	$S^2 = 1.00832$	766
ONA	$S^2 = 0.91272$	229
TV	$S^2 = 1.00912$	2514

IF WE TAKE THE HIGHEST AND THE SMALLEST VARIANCES, CAN WE SAY THAT THE DIFFERENCES OBSERVED ARE STATISTICALLY SIGNIFICANT?

Let us call S_1^2 to the highest variance and S_2^2 to the smallest variance.

Then S_1^2= 1.00912 and S_2^2=0.91272.

To check if the differences are statistically significant, we need to compute the ratio S_1^2/ S_2^2 and check if it is "close to 1". To do this we will compute the confidence interval for S_1^2/ S_2^2 and check if 1 is comprised in this interval. By using Statgraphics and going to Compare->Two Samples->Hypothesis Tests and choosing the option Normal Sigmas, we can input both standard deviations and we will obtain the confidence interval for the variance. In this case, we obtain the interval [0,905311;1,32941]. Since 1 does belong to this interval, we cannot say that there are statistically significant differences between both variances.

ANALYSIS OF VARIANCE

9) MAKE AN ANALYSIS OF VARIANCE (ANOVA) TO STUDY THE EFFECT OF SOURCE ON THE VARIABLE SCORE.

In this case, we will use a either a significance level of 0,05 or 0,01. I decided to do it with a significance level of 0,05 because it provides a better balance between making type I and type II errors, since 0,01 increases the possibility of accepting the null hypothesis (that there are no statistical differences between the averages) when it is false.

Table 2: ANOVA table for Score by Source

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Between groups	338,325	8	42,2906	44,51	0,0000
Within groups	3325,7	3500	0,950201		
Total (Corr.)	3664,03	3508			

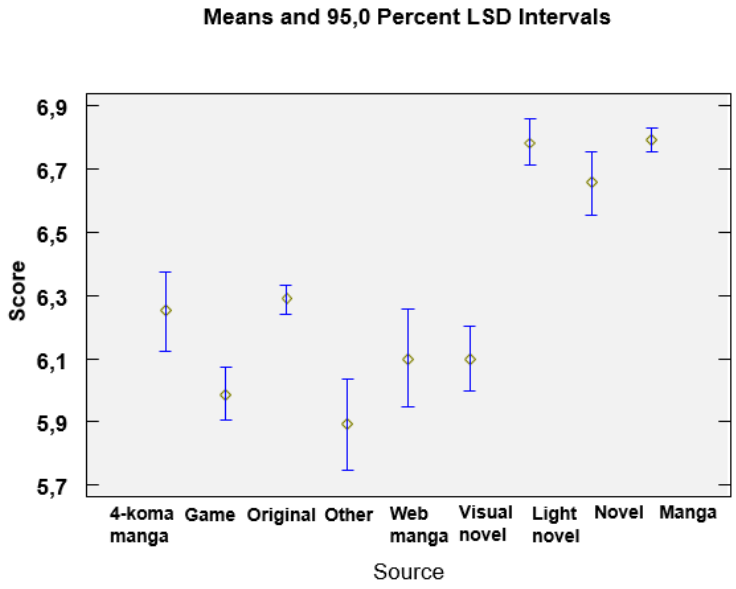

Figure 1: LSD intervals for Score by Source

WHY DON'T ALL THE INTERVALS HAVE THE SAME AMPLITUDE?

Because these intervals are the LSD (Least Significant Differences) intervals, which are computed following this formula: $\overline{x_i} \pm \frac{\sqrt{2}}{2} t_{d.f.resid.}^{\alpha/2} \sqrt{\frac{MS_{resid}}{K}}$ being $\overline{x_i}$ the sample average corresponding to the ith variant of the factor and K the total number of data used to compute $\overline{x_i}$. Since this formula depends on both the average of the different variants and the data used, it is obvious that the intervals will be different for each variant.

ARE THE CONCLUSIONS THAT DERIVATE FROM THE GRAPHIC COHERENT WITH THE ANOVA TABLE?

Yes, they are. From the ANOVA table we know that there are statistically significant differences between the averages because the P-Value is lower than α=0,05. By looking at the LSD intervals, we can clearly see which variants overlap and which ones do not. If they overlap it means that there are no statistically significant differences between them, but if they do not then there are significant differences.

INTERPRET THE MAIN CONCLUSIONS DERIVATED FROM THE ANALYSIS.

As said before, since α=0.05 and the P-Value is lower than α, then it means that there are statistically significant differences between the mean of all the variants. But just by looking at the table we cannot know which variants' means differ from the rest. That is why it is very important to look at the LSD intervals. In the graphic we can clearly see which intervals overlap each other, meaning that there are no statistically significant differences between the means. So, if we want to choose the variant whose score is lower on average, we can either pick Other, Web Manga, Visual Novel or Game because, since the LSD interval for Other is the lowest one, all the variants whose

LSD intervals overlap with it will have the same mean. On the other hand, if we want to pick the variant whose score is higher on average, we can either pick Light Novel, the highest one; or Novel or Manga, since they overlap with Light Novel.

STUDY IF THERE ARE ANY ABNORMAL RESIDUALS THAT SHOULD BE ERASED FROM THE MODEL.

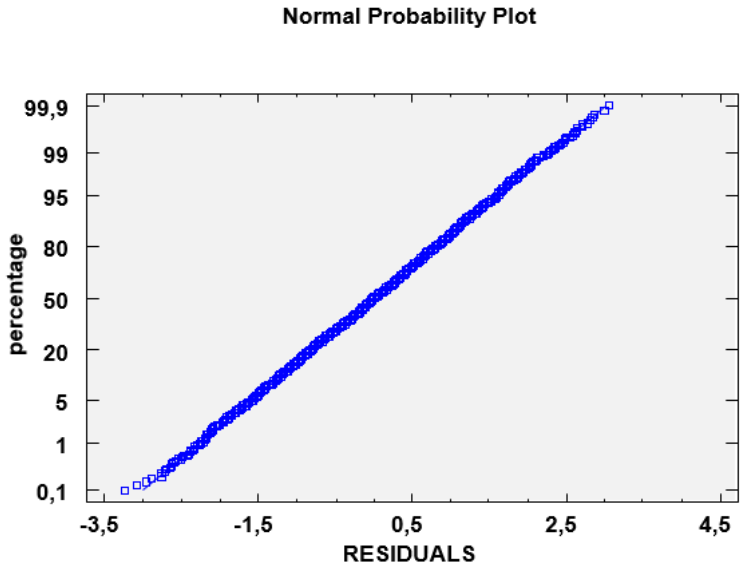

Figure 2: Normal probability plot of the residuals of the previous ANOVA

By saving all of the residuals and representing the data on a normal probability plot, we can clearly see that there are no outliers and therefore all of the residuals follow a normal distribution. This means that there are no abnormal values, and nothing should be erased from the model.

10) INCORPORATE TO THE PREVIOUS MODEL THE FACTOR TYPE AND THE DOUBLE INTERACTION.

Table 3: ANOVA table for Score by Type with Double Interaction.

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
MAIN EFFECTS					
A:Type	38,6407	2	19,3203	21,29	0,0000
B:Source	106,664	8	13,333	14,69	0,0000
INTERACTION: AB	45,0877	16	2,81798	3,11	0,0000
RESIDUAL	3160,06	3482	0,907543		
TOTAL (CORRECTED)	3664,03	3508			

To check if any of the factors are non-significant we need to compute $F_{df.\ factor,\ df.\ residual}$ for each of the factors and check if it is bigger than the F-Ratio obtained in the table. For all of the computations we will use the same α as in the previous exercise (0,05).

For the factor Type, $F_{2;3482}$ = 2,998311125<F-Ratio=21,29. Type is significant.

For the factor Source, $F_{8;3482}$ = 1,941061673<F-Ratio=14,69. Source is significant.

For the interaction, $F_{16;3482}$ = 1,646416971<F-Ratio=3,11. The interaction is significant.

Since none of the factors are non-significant, we cannot erase any of them from the model.

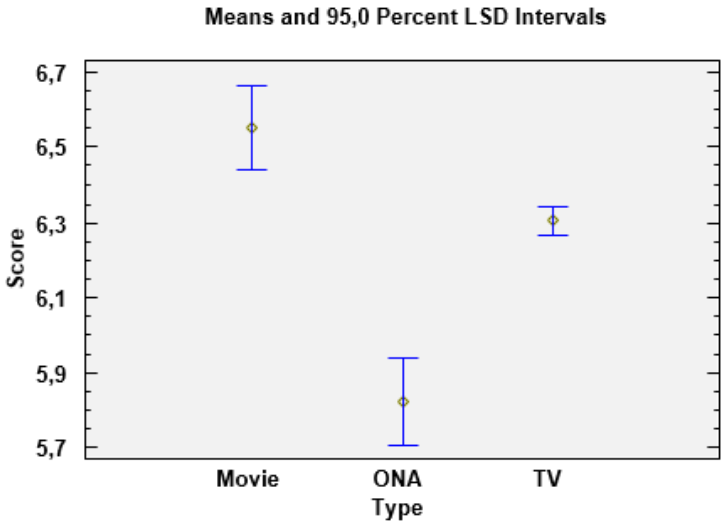

Figure 3: LSD intervals for Score by Type

ARE THE CONCLUSIONS DERIVATED FROM THE GRAPHIC COHERENT WITH THE ANOVA TABLE?

Yes, they are, since the P-Value from the ANOVA table for Type is equal to 0 and less than α=0.05, then it means that there are significant differences between the means of the variants. By looking at the graphic we see that none of the intervals overlap each other, which means that there are differences between the three of them and none of their means are “similar”.

DOES IT MAKE SENSE TO STUDY THE GRAPHIC OF THE DOUBLE INTERACTION? IF THE ANSWER IS YES, INTERPRET THE MAIN CONCLUSIONS DERIVATED FROM THE GRAPHIC

Yes, it does. Since the double interaction turned out to be significant as we have said before, it does make sense to study the graphic of the double interaction.

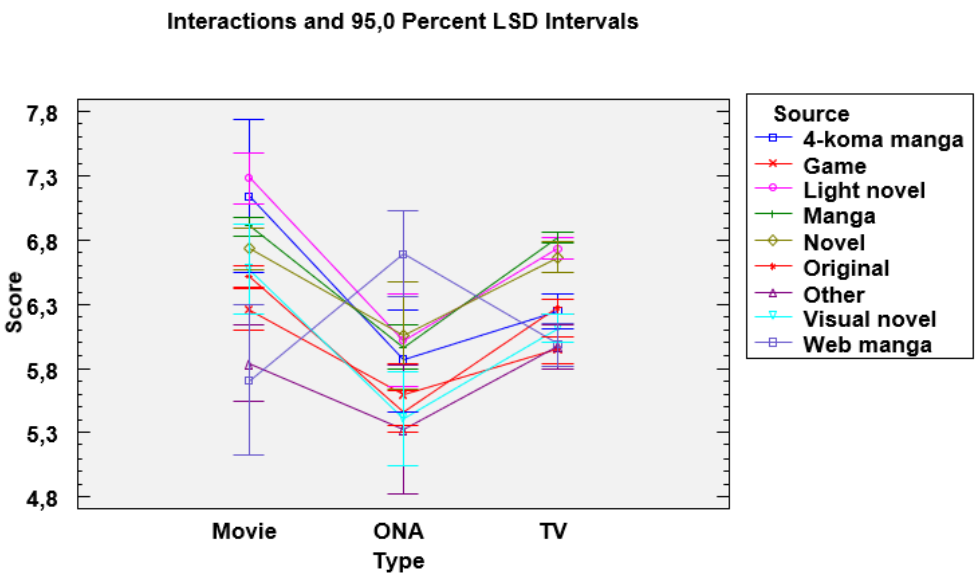

Figure 4: Graphic of interaction of Type and Source for Score with LSD intervals

In this graphic we can clearly see that the interaction is significant since not all the lines are parallel. We can also determine which means are equal since the points are very close together. For instance, we can say that the means for Game, Other and Web Manga in the case of TV are equal since the points practically coincide. On the other hand, the line corresponding to Web Manga is not parallel with any of the other lines and it crosses all of them, meaning that the interactions between them are statistically significant. That is the aspect of the plot that breaks the parallelism the most.

FROM THE MODEL IN WHICH ALL THE EFFECTS INCLUDED ARE STATISTICALLY SIGNIFICANT, SAVE THE RESIDUALS AND REPRESENT THEM ON A NORMAL PROBABILITY PLOT. WHAT CAN BE DEDUCED?

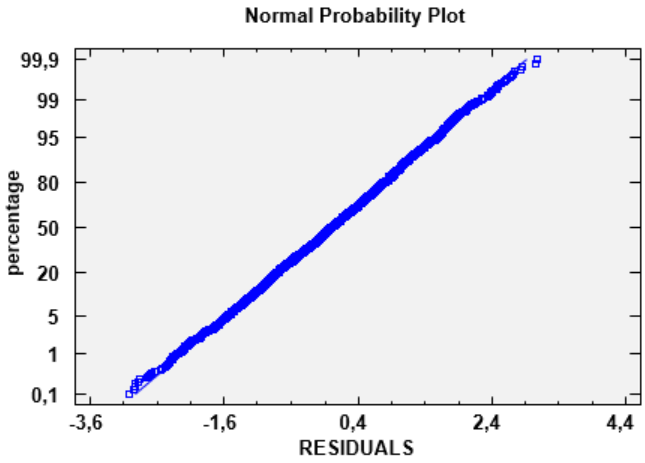

Figure 5: Normal Probability Plot of residuals of previous model

By representing all of the residuals on a normal probability plot we can clearly see that they all follow a straight line, meaning that they do indeed follow a normal distribution. This means that there are no abnormal values in the original variable and no data should be erased from the study.

11) MAKE AN ANOVA TO STUDY THE EFFECT OF THE FACTORS TYPE AND SCORE ON THE VARIABLE MEMBERS.

In this case, we will use a either a significance level of 0,05 or 0,01. I decided to do it with a significance level of 0,05 because it provides a better balance between making type I and type II errors. Also, since the P-Value is 0, it will be less than 0.05 and 0.01, so it is not really relevant which one we choose.

Table 4: Multifactor ANOVA table for Members

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
MAIN EFFECTS					
A:Type	1,90718E12	2	9,53592E11	57,15	0,0000
B:Source	3,07307E12	8	3,84134E11	23,02	0,0000
RESIDUAL	5,83689E13	3498	1,66864E10		
TOTAL (CORRECTED)	6,39758E13	3508			

To check if any of the factors are non-significant we need to compute $F_{df.\ factor,\ df.\ residual}$ for each of the factors and check if it is bigger than the F-Ratio obtained in the table. For all the computations we will use the same α (0,05).

For the factor Type, $F_{2;3498}$ = 2,998299323<F-Ratio=57,15. Type is significant.

For the factor Source, $F_{8;3498}$ = 1,941049556<F-Ratio=23,01. Source is significant.

Since none of the factors are non-significant, we cannot erase any of them from the model.

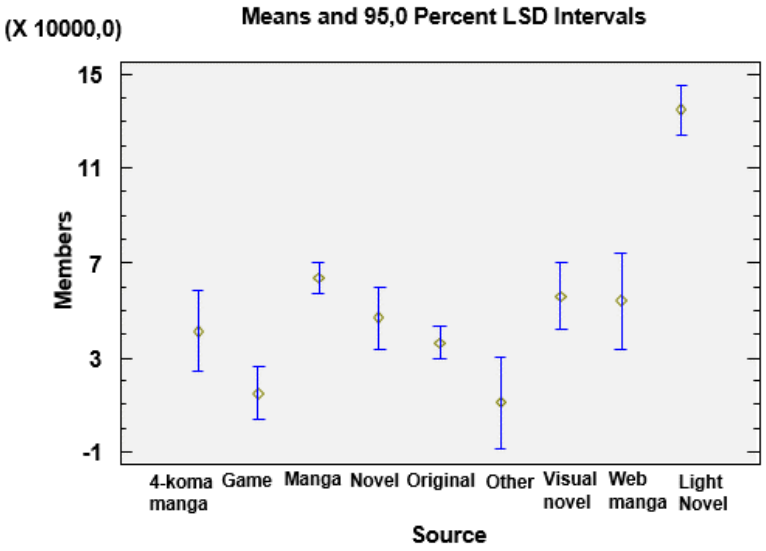

Figure 7: LSD intervals of Source for Members

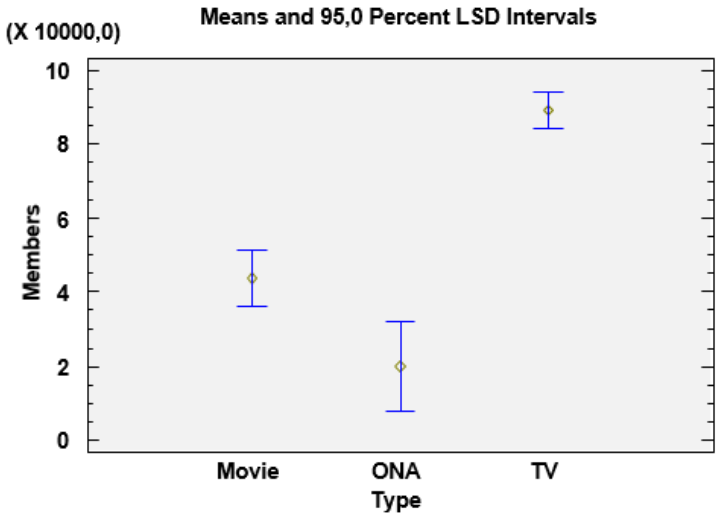

Figure 6: LSD intervals of Type for Members

INTERPRET THE MAIN CONCLUSIONS DERIVATED FROM THE ANALYSIS

As we can clearly see, in the table both P-Values are lower than α, which means that the means are not equal for all the variants and that we should see which ones differ from the rest. In the case of Source, we can clearly see that the variants with less members on average are Other, Original, Game and 4-koma manga. On the other hand, if we are looking for an anime which has many members, it is recommended to pick one whose source is Light Novel. In the case of Type, however, none of the intervals overlap, which means that none of the means are similar in that case and that there are statistically significant differences between the means of the three variants. Then, the variant with less members on average is ONA and the one with most members is TV.

12) INCORPORATE TO THE PREVIOUS MODEL THE DOUBLE INTERACTION

Table 5: Multifactor ANOVA for Members with double interaction

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
MAIN EFFECTS					
A:Type	7,07212E11	2	3,53606E11	21,31	0,0000
B:Source	3,87963E11	8	4,84953E10	2,92	0,0030
INTERACTION: AB	5,97372E11	16	3,73358E10	2,25	0,0030
RESIDUAL	5,77715E13	3482	1,65915E10		
TOTAL (CORRECTED)	6,39758E13	3508			

To check if any of the factors are non-significant we need to compute $F_{df.\ factor,\ df.\ residual}$ for each of the factors and check if it is bigger than the F-Ratio obtained in the table. For all of the computations we will use the same α as in the previous exercise (0,05).

For the factor Type, $F_{2;3482}$ = 2,998311125<F-Ratio=21,31. Type is significant.

For the factor Source, $F_{8;3482}$ = 1,941061673<F-Ratio=2,92. Source is significant.

For the interaction, $F_{16;3482}$ = 1,646416971<F-Ratio=2,25. The interaction is significant.

Since none of the factors are non-significant, we cannot erase any of them from the model.

CAN WE CONCLUDE THAT THE INTERPRETATION OF THE RESULTS IS THE SAME AS IN THE PREVIOUS EXERCISE, OR SHOULD WE POINT OUT SOME ASPECTS?

The interpretation is the same since there are still differences between the means of the variants for both Type and Source, because the P-Value is still lower than α. However, now we have to add that the double interaction is also statistically significant, because as we can see the P-Value of AB is still lower than α.

DOES IT MAKE SENSE TO STUDY THE GRAPHIC OF THE DOUBLE INTERACTION IN THIS CASE?

Yes it does because as we said before, the interaction is significant, so it would be useful to study the conclusions derived from the graphic of the double interaction.

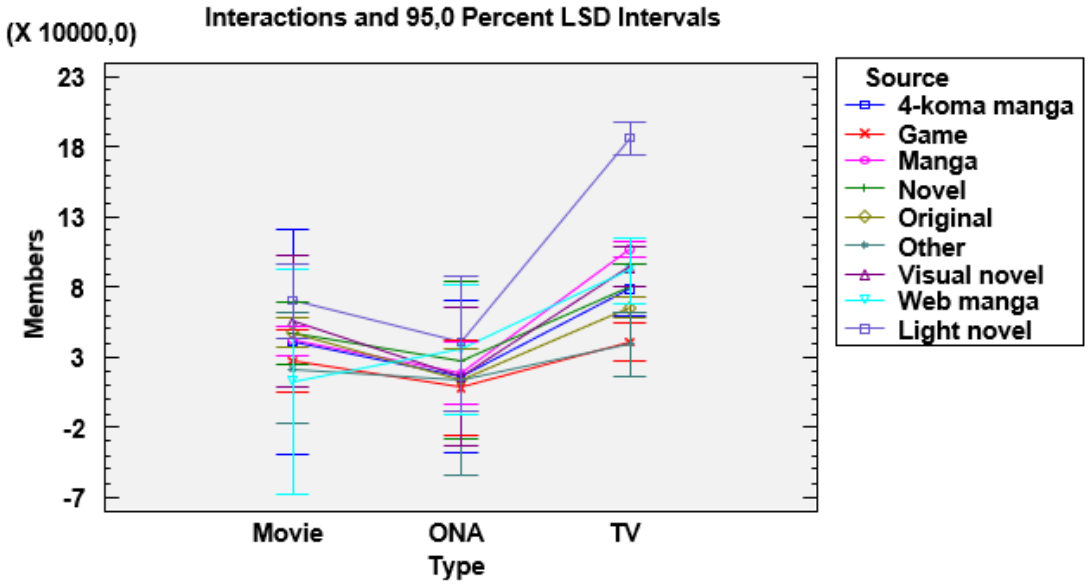

Figure 8: Graphic of double interaction for Members

WHAT ADDITIONAL INNFORMATION DOES THE STUDY OF THE DOUBLE INTERACTION GIVE IN THIS CASE?

By adding the double interaction to the study, we are modifying the LSD intervals for both Type and Source because we are changing the degrees of freedom of the residuals, which are used to compute the $t_{df.residuals}{}^{\alpha}$ used to compute the exact values of the intervals. Then, the conclusions can change and then the variants whose means differ from the rest can be different. For example, in this case we can see that all the LSD intervals overlap with some of the others except for one: the LSD interval of Light Novel for the case of TV. This conclusion is the same one that we got in the previous exercise.

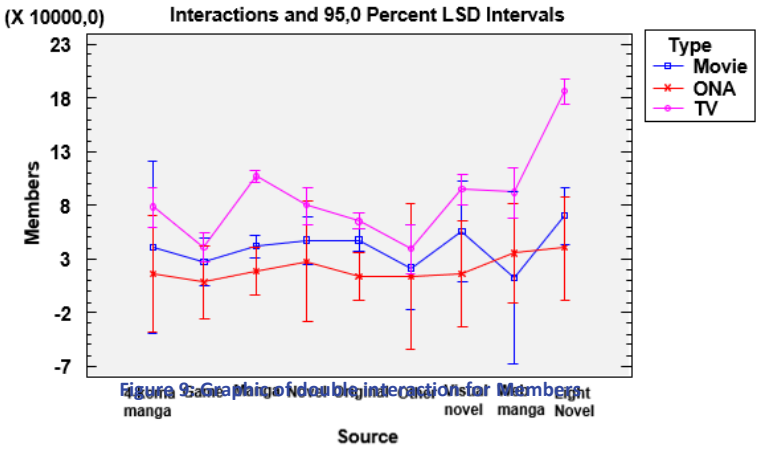

Figure 9: Graphic of double interaction for Members

However, if we look at the same plot but with the variants of Source in the horizontal axis, we see that the only variant of Type that has LSD intervals that do not overlap with any other is TV, meaning that Movie and ONA have the same average, but differ from the average of TV. This is clearly different from the conclusion we obtained in the previous exercise.

Then, since there are LSD intervals that do not overlap in both graphics, it means that the interaction is indeed significant and should be studied more deeply.

FROM THE MODEL IN WHICH ALL THE EFFECTS INCLUDED ARE STATISTICALLY SIGNIFICANT, SAVE THE RESIDUALS AND REPRESENT THEM ON A NORMAL PROBABILITY PLOT. IN THE CASE OF OBSERVING A POSITIVE ASSYMETRY, WHAT WOULD BE RECOMMENDED?

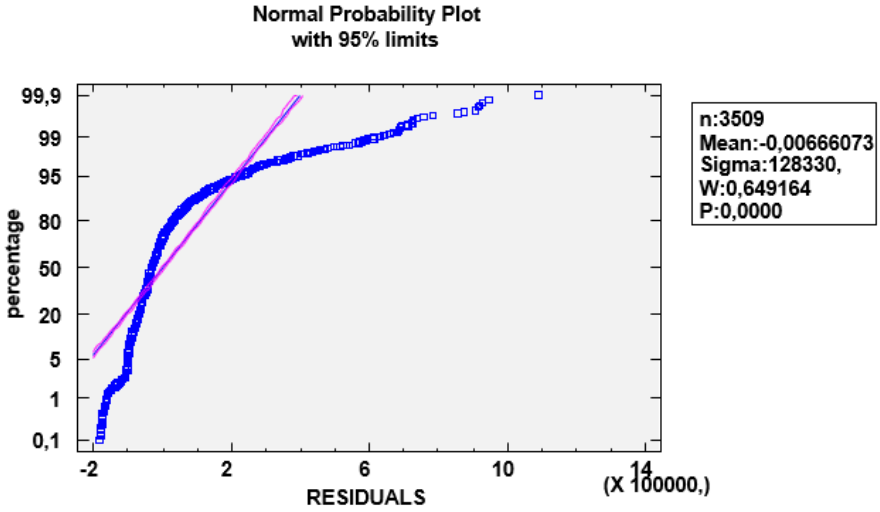

Figure 10: Normal probability plot of the residuals of previous model

As we can see, the residuals in this case do not follow a normal distribution, instead they follow a positively skewed distribution, which means that we cannot know yet if there are any values that are abnormal and should be erased. Then, what we should do in this case, is to apply different transformations to the original variable Members, not to the residuals, and check which transformation normalizes the residuals to check if we can find any outliers.

13) STUDY WITH ANOVA THE EFFECT OF THE FACTORS TYPE AND SOURCE AND OF THEIR DOUBLE INTERACTION IN THE VARIABLE POPULARITY.

In this case, we will use either a significance level of 0,05 or 0,01. I decided to do it with a significance level of 0,05 because it provides a better balance between making type I and type II errors. Also, since the P-Value is very small, it will be less than 0.05 and 0.01, so it is not relevant which one we choose.

Table 6: Multifactor ANOVA for Popularity with double interaction.

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
MAIN EFFECTS					
A:Type	3,83145E8	2	1,91572E8	55,75	0,0000
B:Source	3,42112E8	8	4,2764E7	12,45	0,0000
INTERACTION: AB	1,13092E8	16	7,06827E6	2,06	0,0078
RESIDUAL	1,19646E10	3482	3,43613E6		
TOTAL (CORRECTED)	1,43442E10	3508			

IS THE EFFECT OF THE DOUBLE INTERACTION STATISTICALLY SIGNIFICANT?

To check if the interaction is significant, we need to compute $F_{df.\ interaction,\ df.\ residual}$ and check if it is bigger than the F-Ratio obtained in the table. We will use the same α (0,05).

$F_{16;3482}$ = 1,646416971<F-Ratio=2,06. The interaction is significant.

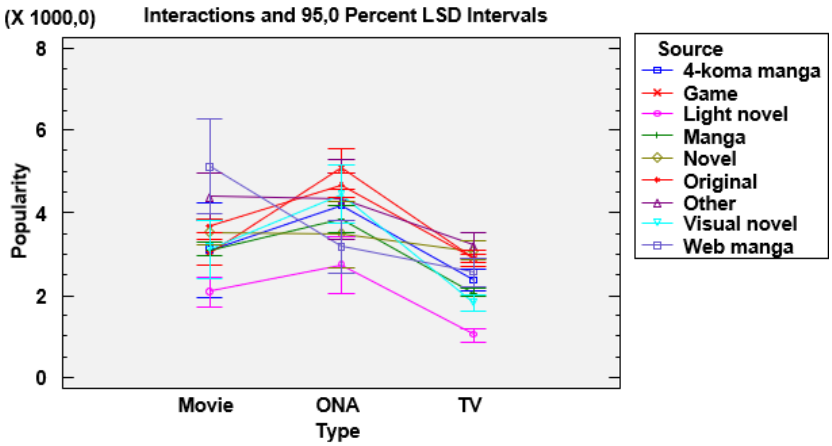

Figure 11: Graphic of double interaction for Popularity

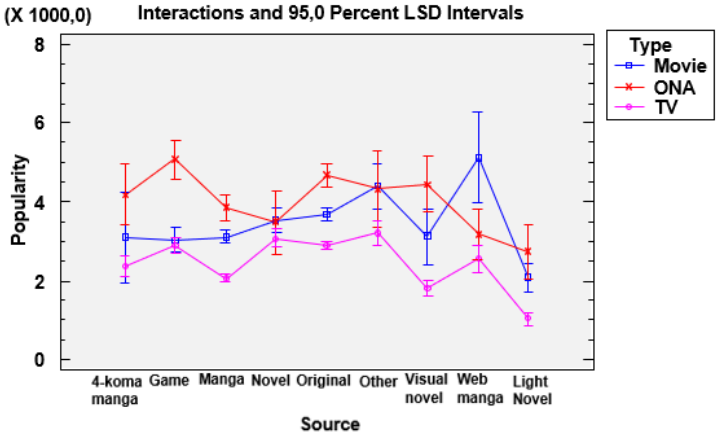

Figure 12: Graphic of double interaction for Popularity

FROM WHICH GRAPHIC DO WE OBTAIN A CLEARER INFORMATION? INTERPRET THE INFORMATION GIVEN BY BOTH GRAPHICS.

We can see in both graphics that the interaction is significant since the lines are not parallel and some of them cross each other. None of the graphics offer a very clear information, but in the first one we can barely see the lines since they are all together because of the many variants of Source. On the other hand, for the second graphic there are less lines but more points in which the lines change their slope, but this one still gives a clearer image than the first one.

We need to take into account that, for this variable, the lower the number the more popular it is, and not the other way around. That being said, if we want to pick an anime which is very popular it would be recommended to pick an anime of type TV and source Light Novel. This information can be deduced from any of the graphics since the LSD interval for TV and Light Novel does not overlap with any of the other intervals and it is also the one whose average is lower, which means that it is more popular.

14) STUDY OF THE RESIDUALS OF THE PREVIOUS MODEL:

FROM THE MODEL IN WHICH ALL THE EFFECTS INCLUDED ARE STATISTICALLY SIGNIFICANT, SAVE THE RESIDUALS AND REPRESENT THEM ON A NORMAL PROBABILITY PLOT. WHAT CAN BE DEDUCED?

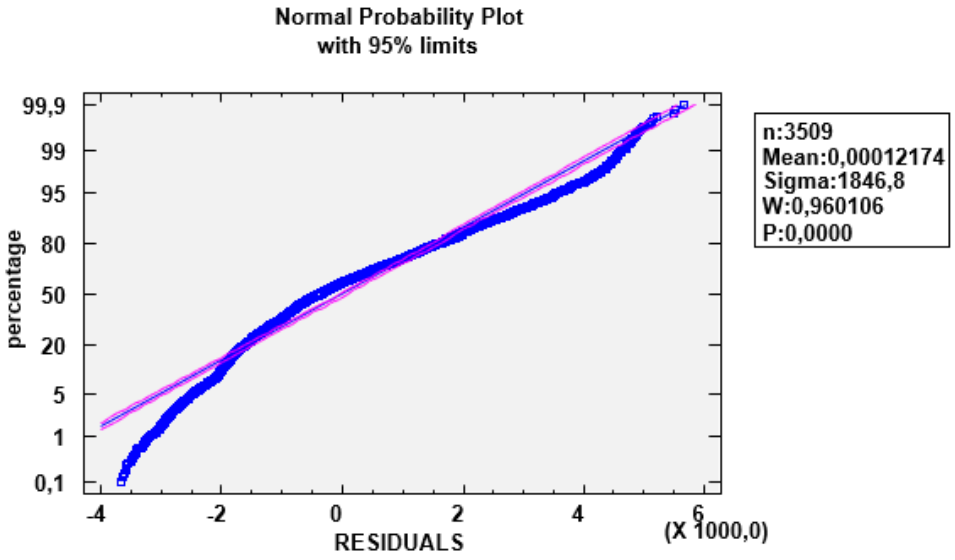

Figure 13: Normal probability plot of the residuals

As we can see, the residuals do not follow a normal distribution, so we need to study which transformation is the best to normalize the residuals of the model, and then check if there are any outliers.

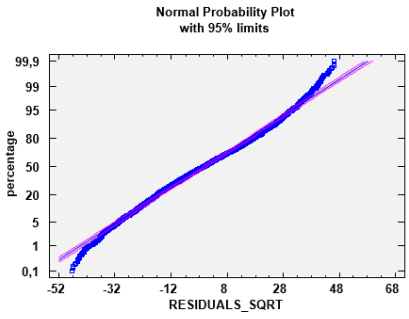

Figure 14: Normal Probability Plot of the Residuals^0.5

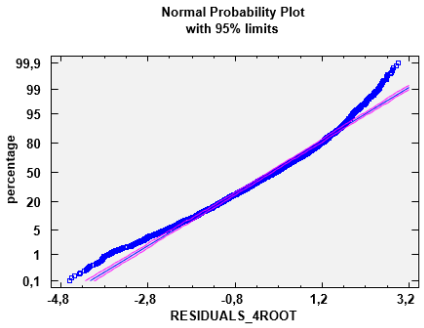

Figure 16: Normal Probability Plot of the Residuals^0.25

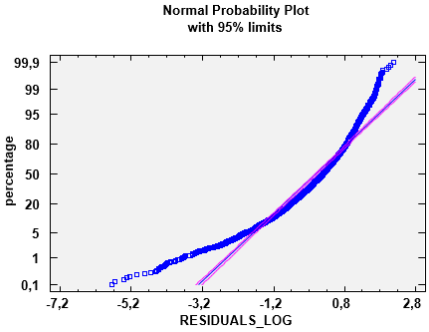

Figure 15: Normal Probability Plot of log(Residuals)

As we can see, the one that best fits into a normal distribution and that still maintains the shape of the original distribution of the residuals is the first one, the one where we applied the square root to the variable Popularity.

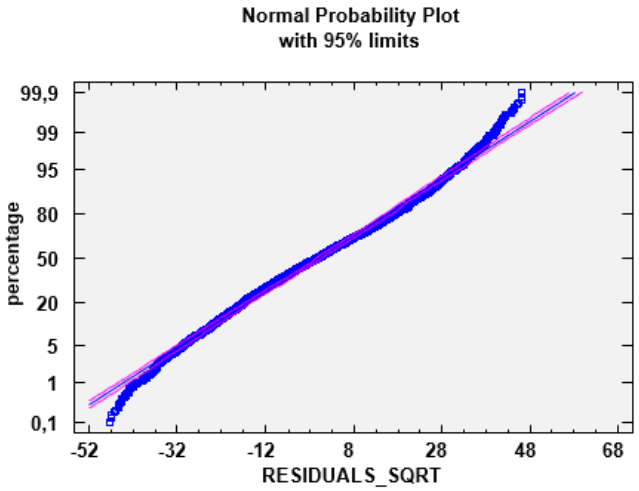

Figure 17: Normal Probability Plot of the Residuals^0.5

There are no outliers that should be erased from the study because all the points follow the same curve and there are no values that clearly differ from the rest. Therefore, there are no abnormal values in the original variable Popularity that should be erased because they are affecting the study negatively. Therefore, the conclusions that we got in the previous exercise are valid and we do not need to change any of them.

INDICATE WHICH OF THE FOLLOWING AFFIRMATIONS IS TRUE:

1- IN THE CASE THAT THE DEPENDENT VARIABLE IS POSITIVE AND ASSYMETRIC, IT IS CONVENIENT TO NORMALIZE IT TO GET THE RESIDUALS OF THE ANOVA TO ADJUST TO A NORMAL MODEL.

2- IN THE CASE THAT THE DEPENDENT VARIABLE IS POSITIVELY ASSYMETRIC, IT IF PREFERIBLE TO ADJUST THE ANOVA MODEL AND STUDY THE DISTRIBUTION OF THE RESIDUALS; IN CASE THEY FOLLOW AN ASSYMETRIC DISTRIBUTION, IT IS CONVENIENT TO TRY DIFFERENT TRANSFORMSTIONS IN THE DEPENDENT VARIABLE UNTIL WE GET THAT THE RESIDUALS ADJUST TO A NORMAL MODEL.

The correct affirmation is the second one. Even if the dependent variable does not follow a normal distribution, the residuals can still be normal. That means that we should not modify the variable before we adjust the ANOVA model. After we make the ANOVA analysis, we should save the residuals and check if they follow a normal distribution or not. If they do not, then we should apply the transformations to the dependent variable and save the residuals again for each transformation to check which one best normalizes the residuals. It is important to remember that the transformations are not applied to the residuals themselves.

LINEAR REGRESSION

15) OBTAIN THE MATRIX OF VARIANCES-COVARIANCES OF THE VARIABLES SCORE, MEMBERS, POPULARITY AND FAVORITES. WHAT USEFUL INFORMATION DOES THIS MATRIX GIVE? WHY IS IT SYMMETRIC?

Table 7: Matrix of Variances-Covariances

	Score	Members	Popularity	Favorites
Score	1,04448	63727,8	-1060,21	1842,59
	(3509)	(3509)	(3509)	(3509)
Members	63727,8	1,82371E10	-1,556E8	5,53143E8
	(3509)	(3509)	(3509)	(3509)
Popularity	-1060,21	-1,556E8	4,08899E6	-2,87604E6
	(3509)	(3509)	(3509)	(3509)
Favorites	1842,59	5,53143E8	-2,87604E6	2,69517E7
	(3509)	(3509)	(3509)	(3509)

This matrix gives us the information about the covariances of the variables with respect to one another, but it also provides us with the variances for each variable. This is because in the main diagonal of the matrix, what appears are the variances of each variable since the row and column coincide. This matrix is symmetric because we are representing the variables on the same order in the columns and rows and the covariance of X1 with respect to X2 is the same as the covariance of X2 with respect to X1, then the matrix will be symmetric because when we compute the covariance of two variables, it will not matter which one is in the row and which one is in the column, the result will be the same.

16) OBTAIN THE MATRIX OF CORRELATION OF THESE VARIABLES. IN THE CASE OF POSITIVE ASSYMETRY NORMALIZE THE VARIABLES.

I tried to normalize the data, but the variables Popularity and Members cannot be normalized with any of the transformations, so I decided to choose the best transformation for them. In the case of Score it is already normal and for Favorites the logarithm does normalize the data.

Table 8: Matrix of correlation

	Score	log(Members)	Popularity^0.5	log(Favorites)
Score		-0,0031	0,0133	0,5239
		(3509)	(3509)	(3509)
		0,8540	0,4307	0,0000
log(Members)	-0,0031		-0,9976	0,1345
	(3509)		(3509)	(3509)
	0,8540		0,0000	0,0000
Popularity^0.5	0,0133	-0,9976		0,0746
	(3509)	(3509)		(3509)
	0,4307	0,0000		0,0000
log(Favorites)	0,5239	0,1345	0,0746	
	(3509)	(3509)	(3509)	
	0,0000	0,0000	0,0000	

WHAT CAN BE DEDUCED FROM THIS MATRIX? COMMENT THE CORRELATION BETWEEN THE VARIABLES.

This matrix shows partial correlation coefficients between each pair of variables. For each pair, there are three values, the first one being the correlation coefficient, the second one the number of data and the third one the P-Value that is useful to know if the correlation is statistically significant. In this case, it is used a confidence level of 95% which means that if the P-Value for one pair is less than 0,05 then it means that the correlation is not significant.

For example, the correlations between Score and Popularity and Score and Members are not statistically significant because the P-Value is greater than 0,05. On the other hand we have that the correlation between Members and popularity is very strong since the correlation coefficient is almost 0.

EXPLAIN WHAT IS THE VALUE OF THE ELEMENTS OF THE MAIN DIAGONAL

The main diagonal is empty because it would not give us any useful information. This is because the values of the correlation coefficients of the main diagonal are always one, since the correlation coefficient of one variable with itself will always be one. This is why Statgraphics does not show any values on the main diagonal of this matrix.

17) FROM THE PREVIOUS MATRIX, IDENTIFY THE COUPLE OF VARIABLES WITH A GREATER DEGREE OF CORRELATION (BUT LESS THAN 0.95). PLOT A DISPERSSION GRAPHIC BETWEEN BOTH.

The two variables with a greater degree of correlation that is less than 0.95 are Score and Favorites, in this case Score already follows a normal distribution, and Favorites can be normalized by applying the logarithm.

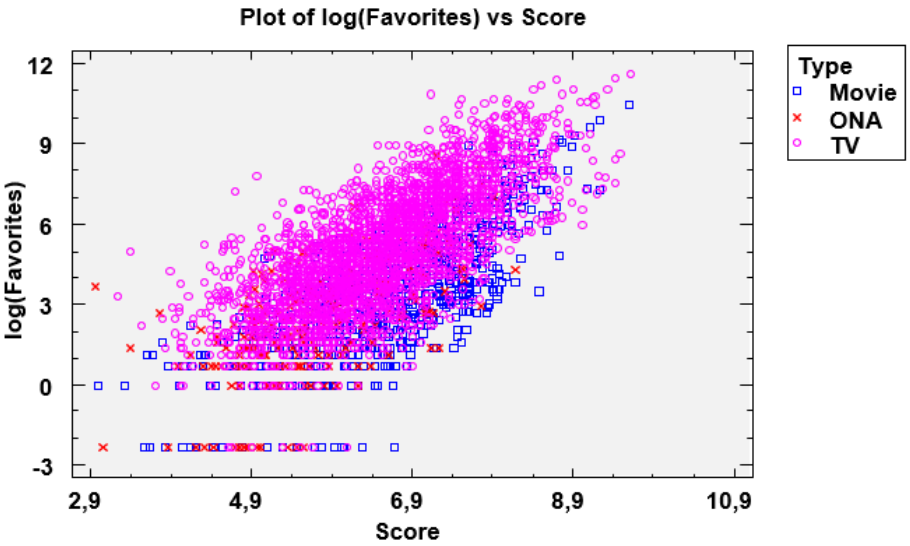

Figure 18: Disperssion Graphic between Score and log(Favorites)

WHAT CAN BE DEDUCED FROM THIS GRAPHIC?

From the graphic we can see that both variables are correlated since, when the score is higher, the number of Favorites also increases. This makes a lot of sense since if an anime has a higher score, meaning that it is very good, more users will have this anime added to their favorites' list. We can also see that most of the animes with higher score and more favorites are of type TV. We can see that depending on which Type we look at, the correlation could be stronger or weaker. For example, the points for TV and Movie appear to be closer to each other while ONAs are more dispersed.

DESCRIBE THE RELATION BETWEEN BOTH VARIABLES.

These variables have a positively linear relation, since we can fit a straight line to the plot, and this line would have positive slope. This relation appears to be moderate since the points could be closer, but since there are so many observations, it is normal that they appear more dispersed, but not dispersed enough to be considered weak. So, we will say that the relation is linear, positive and moderate.

18) BETWEEN THE FOUR CUANTITATIVE VARIABLES, CHOSE THE ONE (Y) THAT COULD BE CONSIDERED TO BE A RESPONSE VARIABLE. FROM THE MATRIX OF CORRELATION IDENTIFY THE VARIABLE (X) WITH MORE CORRELATION WITH Y. MAKE AN ANALYSIS OF SIMPLE LINEAR REGRESSION THAT ALLOWS US TO PREDICT THE VALUES OF Y IN FUNCTION OF X.

The variable that could be considered a response variable is Popularity, since the popularity of an anime is computed from the number of members that they have. By looking at the matrix, we can clearly see that the variable which is more correlated to it is Members, with a correlation coefficient of almost 1. Then, we will say that the dependent variable is Popularity while the independent one is Members.

To normalize the data as much as possible I applied the square root to Popularity and the logarithm to Members.

PUT THE GRAPHIC OF DISPERSSION OF Y IN FUNTION OF X ALONG WITH THE FITTED REGRESSION LINE, AND THE PREDICTION INTERVAL (WITH A CONFIDENCE LEVEL OF 95%). COMMENT ABOUT THE UTILITY OF THE INTERVAL IN PRACTICE.

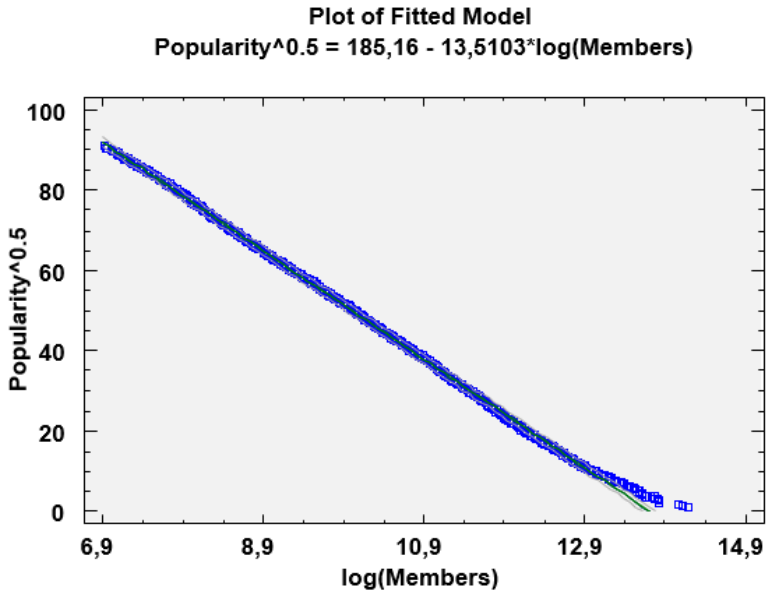

Figure 19: Plot of dispersion of log(Members) and Popularity^0.5

The prediction interval gives us an idea of how the regression line will be. Basically, we know that the points will be inside the green interval with 95% confidence. They are useful to know how disperse or how close to the line the points will be.

PUT THE SUMMARY TABLE OF THE MODEL OBTAINED WITH STATGRAPHICS. FROM THE INFORMATION OF THE TABLE, CAN WE SAY THAT THE OBSERVED CORRELATION IS STATISTICALLY SIGNIFICANT?

Table 9: Table of regression model

Parameter	Estimate	Starndard Error	T Statistic	P-Value
CONSTANT	185,134	0,0654718	2827,69	0,0000
log(Members)	-13,5076	0,00632339	-2136,13	0,0000

Yes, we can say that it is statistically significant since the P-Value of Members is 0. It doesn't really matter which α we chose since the P-Value will be less than α in all cases. This is coherent with the information we get from the graphic since it is almost a straight line.

WRITE THE MATHEMATICAL EQUATION OF THE MODEL Y=A+B*X. COMMENT ON THE STATISTICAL SIGNIFICANCE OF BOTH COEFFICIENTS, USING THE SIGNIFICANCE LEVEL THAT YOU CONSIDER TO BE THE BEST.

The mathematical equation of the model is:

Popularity^0.5 = 185,134 - 13,5076*log(Members)

In this case we have that A=185,134 and B=-13,5076. A is the point of the X axis at which the line intercepts it, while B is the slope of the line. Since the slope is negative, we get that the variables have a negative relation.

19) WITH RESPECT TO THE PREVIOUS EXERCISE:

WHAT IS THE PRACTICAL INTERPRETATION OF THE COEFFICIENTS A AND B FROM THE MODEL?

We can interpret the coefficients in the following way:

Interpretation of B: when log(Members) increases in 1, the square root of the number associated with their popularity reduces in 13,5076 on average. The number associated with the popularity reduces because the smaller the number the more popular the anime is, which actually means that it becomes more popular. Then, B is the average increase of Popularity^0.5 expected if log(Members) increases one unit.

Interpretation of A: When log(Members)=0, or the same, when Members is equal to 1, then the square root of the number associated with the popularity is equal to 185'134 on average, or the same, the number associated with the popularity will be $185'134^2$=34274,6 on average. This happens because if the anime only has one member then it won't be popular, which means that the Popularity coefficient will be very high, since, as said before, the smaller the number, the more popular it is.

COMMENT THE POSSIBLE CAUSALITY OF THE CORRELATION: FROM THE PHYSICAL INTERPRETATION OF THE VARIABLES, IS IT POSSIBLE TO SUSPECT THAT THE CORRELATION OBSERVED IS DUE TO A CAUSE-EFFECT RELATION, A PARTIAL DEPENDENCE, OR AN INTERDEPENDANCE BETWEEN THE VARIABLES?

The correlation observed is due to a cause-effect relation since the variable Popularity depends only on the number of members the anime has. Therefore, the cause is the number of members and the effect is the Popularity value, which will be smaller the more members it has.

20) SAVE THE RESIDUALS OF THE MODEL AND REPRESENT THEM ON A NORMAL PROBABILITY PLOT. WHAT CAN BE DEDUCED?

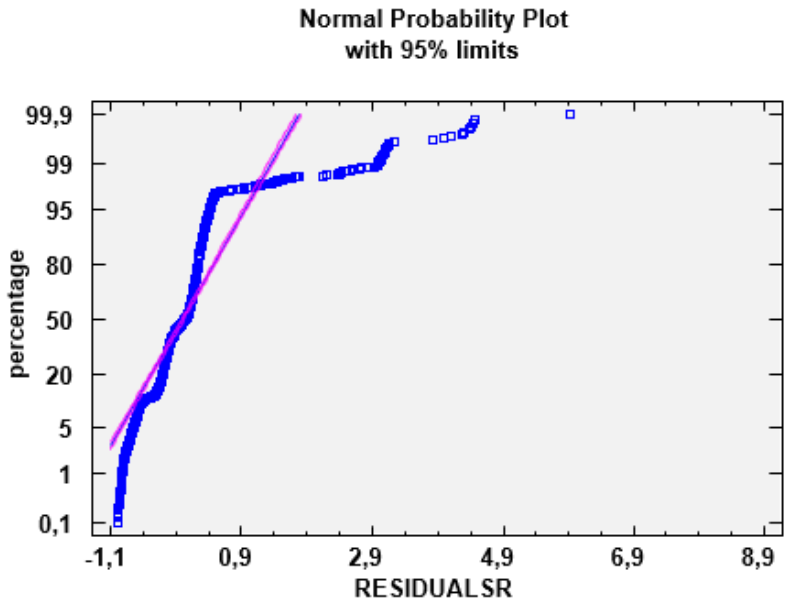

Figure 20: Normal Probability Plot of residuals of previous Regression model

By looking at the graphic we can see that the residuals do not follow a normal distribution, however we can more or less know which values are outliers because they differ from the rest of the points. In this case, the point for the highest value of the residuals does not follow the curve that the rest of the points follow, so we should remove it since it looks like an outlier.

REPRESENT THE RESIDUALS IN FUNCTION OF X. CAN THERE BE A QUADRATIC EFFECT?

After removing the outlier and representing the residuals in function of X we obtain the following plot:

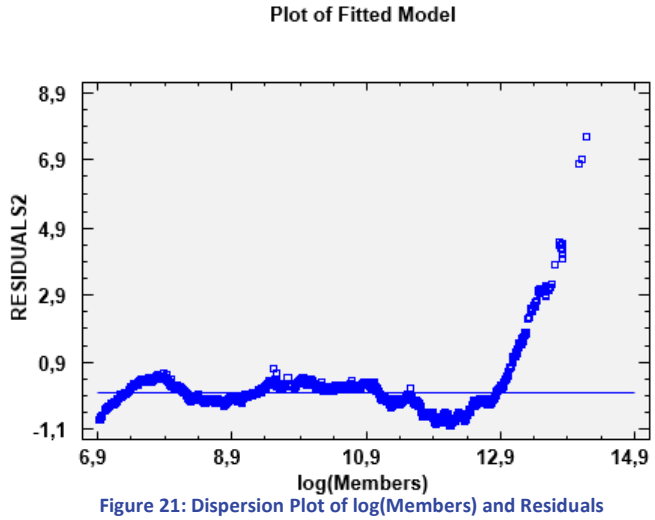

Figure 21: Dispersion Plot of log(Members) and Residuals

As we can see, the fitted line is horizontal which could mean that there is no correlation between log(Members) and the residuals. However, this is not the case because the points follow some kind of curve that cannot be fitted into a straight line. This means that there is no linear effect but there could be a quadratic effect between both variables that we need to check by adding a new independent variable which would be log(Members)^2.

EXPLAIN HOW WE CAN VERIFY IF SAID EFFECT IS STATITICALLY SIGNIFICANT. IN THE AFFIRMATIVE CASE, INTERPRET SAID EFFECT.

As we said before, to check this effect we have to add another independent variable to the model, log(Members)^2. Then we need to check if the P-Value is lower than alpha, which would mean that the effect is statistically significant. By doing so, we obtain the following table:

Table 10: Table of regression model between log(Members) and Residuals

Parameter	Estimate	Standard Error	T-Statistic	P-Value
CONSTANT	5,89605	0,366113	16,1045	0,0000
log(Members)^2	0,0577484	0,00353371	16,3421	0,0000
log(Members)	-1,1805	0,0724858	-16,2859	0,0000

We can clearly see that the P-Values are lower than α=0,5% which means that the effect is indeed statistically significant.

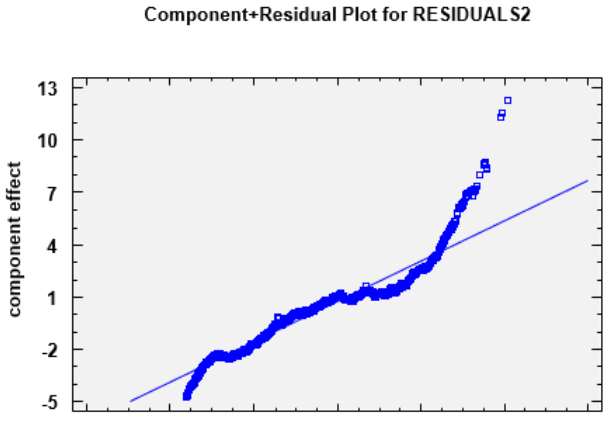

Figure 22: Dispersion Plot of log(Members)^2 and Residuals

By looking at the plot we can see that now the points fit better into a straight line since we added the new dependent variable. Then, we can say that the model has a quadratic effect, but maybe there is some other effect that would fit better this model. Since the effect is statistically significant, we cannot erase any of the variables and we do not have enough evidence to reject the quadratic effect and therefore we will accept the hypothesis.

USING THE FORMULAS NEEDED AND FROM THE INFORMATION SEEN IN THE TABLE OF THE MODEL, COMPUTE A PREDICTION INTERVAL OF Y WHEN X IS ON ITS FIRST QUARTILE (WITH A CONFIDENCE LEVEL OF 95%) JUSTIFY THE COMPUTATIONS. WHAT PRACTICAL INTERPRETATION DOES THIS RESULT HAVE?

First, we need to compute the mathematical equation of the model by looking at the table. Since this is a quadratic effect, the equation will be of the type Y=a+bx+cx2. The values of a, b and c can be seen in the table, in the "Estimate" column. Then, we have that a=5.89605, b=-1.1805 and c=0,0577484. Therefore, the equation of the curve is the following:

RESIDUALS = 5,89605 - 1,1805*log(Members) + 0,0577484*log(Members)^2

Now, let us compute the first quartile of X with Statgraphics: Z_{25}=9,04611.

We know that RESIDUALS/(log(Members)=Z25) follows a normal distribution with the following average and standard deviation:

Average: is computed by replacing log(Members) by Z_{25} in the equation of the curve.

m = 5,89605 - 1,1805*9,04611 + 0,0577484*9,046112= -0.05721

Standard Deviation: in Statgraphics it corresponds with the Standard Error of Est.=0,547989

Now, we need to check with Statgraphics the values for percentiles 2,5 and 97,5 of a normal distribution N(m=-0,05721; σ= 0,547989) in order to get the interval. These values are -1,131250759 and 1,016830759, so the prediction interval is [-1,131250759; 1,016830759].
This result tells us that the 95% of the values of the residuals when log(Members) is equal to 9,04611 will be comprised in the interval [-1,131250759; 1,016830759].