

Data analysis plan

Question 1: Is there a relationship between the type/genre and score?

Type:

To address this question, we have decided to make a hypothesis test. First of all, we should define our null and alternative hypothesis, which is that the overall score of Documentaries is higher than the overall score of Films:

H0: $\mu_D = \mu_F$ (REWRITE)

H1: $\mu_D > \mu_F$ (REWRITE)

The next step would be to determine the value of the sample statistic. We chose to use the DBM (Difference Between Means) because the size of the population is much larger than the sample size and samples are independent (or at least we consider so).

— · · —
 $x_D - x_F$ (REWRITE)

After applying the statistic of choice to the set of randomized samples, we would compute the p-value, which would tell us the proportion of samples on the distribution with an statistic as extreme or more than the observed sample.

If the p-value is lower than a 5% significance level ($\alpha=0.05$ is used as a default level), the conclusion would be that the results are statistically relevant, thus rejecting H0. If the p-value is not lower than the significance level, it would mean that we cannot reject H0, and the results would be not statistically relevant.

Genre:

This question address a similar problem as the previous one, but with many different values for the qualitative explanatory variable. Despite not being sure on how to address this problem, our guess would be to perform the same analysis using an one-versus-all approach, computing the p-value of each of the different genres and comparing all of them individually against the significance level.

Question 2: Is there any difference on the audience/critics/imdb score by genre?

We could perform a Principal Component Analysis to measure the percentage of variance explained of each of the three variables (audience_score, critics_score and imdb_rating) to see which of them is the highest.

Before doing so, we should standardize the explanatory variables to the same scale. We are not sure about how to do this, as the variables do not seem to be distributed equally between variables.

This would tell us which of the three variables should be taken into account the most.

As graphs suggest that critics tend to evaluate the lowest, we could perform an hypothesis test on each of the genres under the null and alternative hypothesis of the critics score being respectively equal or lower than the rest of the scores.

H0: $\mu_C = \mu_O$ (REWRITE)

H1: $\mu_C > \mu_O$ (REWRITE)

Question 3: Are oscar-awarded films more liked?

We find the procedure done at the first question to adapt well to this question too, so we would make the same

test, with our null and alternative hypothesis:

H0: $\mu_O = \mu_N$ (REWRITE)

H1: $\mu_O > \mu_N$ (REWRITE)

Where the null hypothesis states that being nominated to the best picture award does not affect the overall score, while the alternative hypothesis states that this has a positive impact on it.

Question 4: What are the trends over the years?

Which is the referred month for releases? Over the years? Does it affect the score?

Without much analysis, we could say that the most films are released on Christmas, Summer and Easter.

Does the genre change over the years?

no idea

Do older films tend to have higher score/number of votes?

how to measure if the regression tends to go lower?

Question 5: Do actors and directors have a higher score once they won an Oscar?

comparing time series of the overall score before the award against time series of the overall score after the award?