Project Report

Hypothesis testing to infer a population mean for movie data- G13

Member 1
Rohan Arava
S20210010028
UG-3 CSE

Member 2
PRS Pramod
S20210020304
UG-3 ECE

Member 3
Pamidi Mohammad Ashraf
S20210020303
UG-3 ECE

Member 4
Sivasanath Kumar Medavarapu
S20210020322
UG-3 ECE

Member 5
Hruday Chowdary Gurijala
S20210020278
UG-3 ECE

November 15, 2023

Contents

1 Objective

Our project's goal is to conduct hypothesis testing to determine the population mean.

1. Data pre-processing:Prepare the data for analysis by cleaning and transforming it.
2. Calculate population mean: Using imdb_score, get the population mean of all the films released through 2016. This will function as the reference point when comparing it to the 2017 sample data.
3. Acquire sample data: Gather a selection of all the films released in 2017. We will test the hypothesis that "Popularity of films increases" using this dataset.
4. Test the hypothesis: Apply the next two techniques to test the hypothesis: The population standard deviation is known: To verify the hypothesis, apply the Z-test. Compute test statistic and p-value to decide if the null hypothesis should be rejected or not. Population standard deviation is unknown: To test the hypothesis, apply the t-test. Compute test statistic and p-value to decide if the null hypothesis should be rejected or not.
5. Conclusion: Provide an overview of the hypothesis tests' outcomes and make inferences from them.

2 Implementation

2.1 Methods and materials

Methods:

We used libraries like

1. magrittr - For forward pipe operator,
2. tidyr - Tidy messy data,
3. dplyr - A Grammar of data manipulation,
4. modeest - Mode estimation.

We have programmed functions like "eda" - for exploratary data analysis of the data frame taken as parameter, "delete_nans" - to delete nan values and replace with a statistic according to the data, "boxplots" - to plot boxplots for every feature in the dataset, "remove_outliers" - to remove outliers from each feature in the dataset.

Materials:

The datasets we have used are based on the movie data in 2016 and 2017 from the official IMDb database.

The "movie_metadata.csv" of 2016 contains the following features:

"color" "director_name" "num_critic_for_reviews" "duration" "director_facebook_likes" "actor_3_facebook_likes" "actor_2_name" "actor_1_facebook_likes" "gross" "genres" "actor_1_name" "movie_title" "num_voted_users" "cast_total_facebook_likes" "actor_3_name "facenumber_in_poster" "plot_keywords" "movie_imdb_link" "num_user_for_reviews" "languag "country" "content_rating" "budget" "title_year" "actor_2_facebook_likes" "imdb_score" "aspect_ratio" "movie_facebook_likes"

The "2017 Movie List.csv" contains the following features:

"Position" "Const" "Title" "URL" "Title.Type" "IMDb.Rating" "Runtime..mins." "Year" "Genres" "Num.Votes" "Directors"

But out of all the features we use only the "imdb_score" from movie_metadata.csv and sample from "IMDb.Rating" from 2017 Movie List.csv for our hypothesis testing.

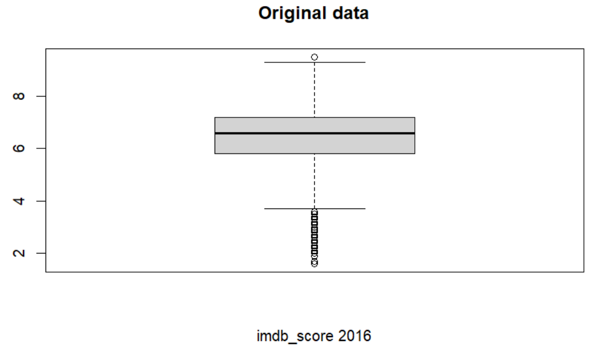
2.2 Data pre-processing and EDA

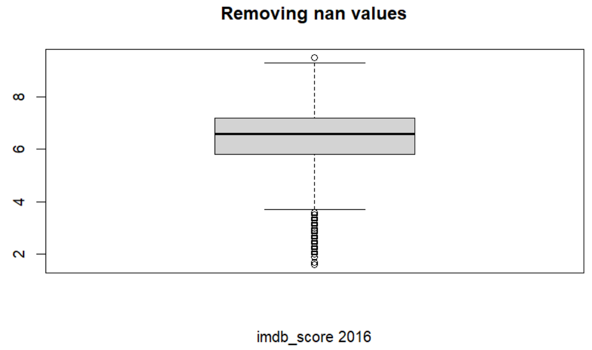
In EDA , we identify whether data in each column is numerical or categorical and print the summaries and plot histograms if the data is numerical and barplots if the data is categorical and we also print the percentage of nan values in the entire data.

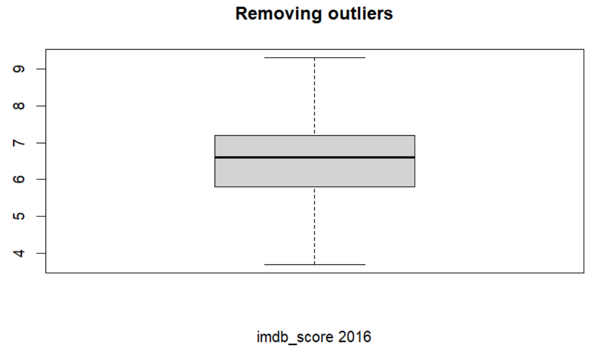
In the data-preprocessing,

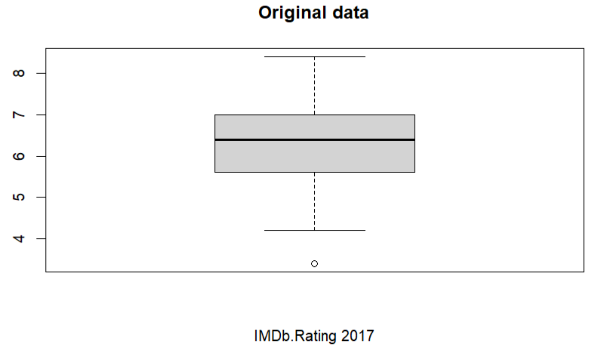
Removal of nan values: the nan values are identified and replaced by median value if the data is numerical and mode if the data is categorical for each feature in the dataset.

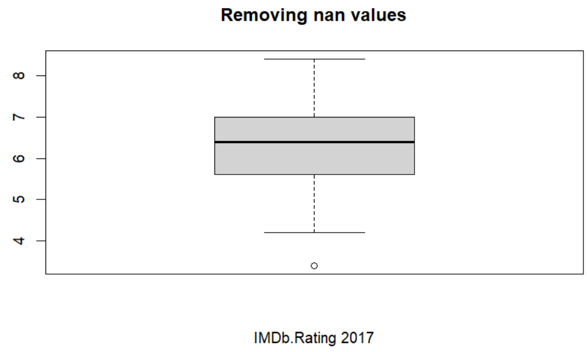
Removal of outliers:the outliers can be removed to make the data normally distributed and it can be done by identifying the IQR's and excluding the outliers for each feature in the dataset.

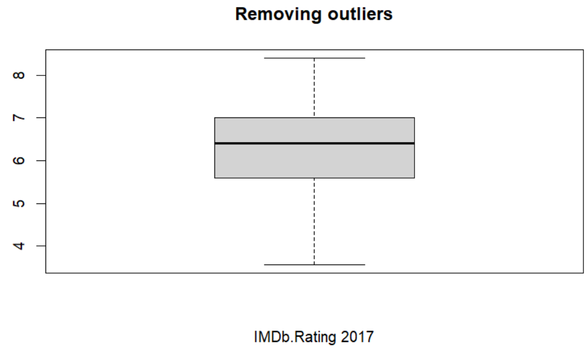

Original data
8
6
4
2
imdb_score 2016


Removing nan values
8
6
4
2
imdb_score 2016


Removing outliers
9
8
7
6
5
4
imdb_score 2016


Original data
8
7
6
5
4
IMDb.Rating 2017





2.3 Hypothesis Testing

For Hypothesis testing, we calculate the population mean and standard deviation and same for the sample from 2017 dataset. Hypothesis testing is done based on these values given the 2 conditions i. Population standard deviation is known.
ii. Population standard deviation is unknown.

We perform Z-test for this comes under parametric testing.(normally distributed data and comparing the mean of a sample to some hypothesized mean for the population.

3 Theory

Z-Test: This is most frequently test in statistical analysis.

- It is based on the normal probability distribution.
- Used for judging the significance of several statistical measures particularly the mean.
- It is used even when binomial distribution or t distribution is applicable with a condition that such a distribution tends to normal distribution when n becomes large.
- Typically it is used for comparing the mean of a sample to some hypothesized mean for the population in case of large sample, or when population variance is known.

$$z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

4 Conclusion

i. Population standard deviation is known.

```
Null Hypothesis:  Mu=6.45746579417014"
"Alternate Hypothesis:  Mu>=6.45746579417014"
"P value:  0.127903146833744"
"At 90% Confidence level"
"We cannot reject the null hypothesis"
"We cannot say that popularity increases"
```

ii. Population standard deviation is unknown.

```
"Null Hypothesis:  Mu=6.45746579417014"
"Alternate Hypothesis:  Mu>=6.45746579417014"
"P value:  0.0982440216526019"
"At 90% Confidence level"
"We can reject the null hypothesis"
"We can say that popularity increases"
```

5 Contributions

Rohan Arava - Exploratary data analysis

PRS Pramod - Data preprocessing(removing nan values)

Pamidi Mohammad Ashraf - Data preprocessing(removing outliers)

Sivasanath Kumar Medavarapu - Hypothesis Testing(Population Standard Deviation is known)

Hruday Chowdary Gurijala - Hypothesis Testing (Population Standard deviation is unknown)