Academic Year 2019 - 2023

Introduction to Data Analytics

Project Report

Topic 1 : Hypothesis testing to infer a population mean
Date of Submission : 30-11-2021

Group: 16
Members:
Preethi G - S20190020241
Varun K - S20190020222
Kajal - S20190020215
Amruth P - S20190020242
Hardik Sharma - S20190010062

1 Problem Statement

Reference: MOVIE data
a) Calculate population mean from all the movies up to 2016 on imdb_score.
b) Collect a sample of all the movies in the year 2017.
c) Test the hypothesis that "Popularity of films increases".
To test the hypothesis consider following:
i. Population standard deviation is known.
ii. Population standard deviation is unknown

2 Understanding the theory

- Given a dataset with 5000 movies over the years 1916 to 2016, we are supposed to predict whether the popularity of films in general are increasing or not.

- We could use the imdb score a.k.a ratings and apply z-test if the population standard deviation is known or apply t-test if the population standard deviation is unknown.

- The null hypothesis will be
  H0: Mean rating of 2017 <= Population Mean
  H1: Mean rating of 2017 > Population Mean

- This is in fact just an one tailed test with null hypothesis
  H0: Mean rating of 2017 = Population Mean
  H1: Mean rating of 2017 > Population Mean

3 Implementation of the project

- **Step 1: Collecting the data**
  The movie dataset has been downloaded from the given kaggle metadata as a .csv file.
  The sample of 100 movies in the year 2017 was collected using web scrapping with the columns Movie Title, Release Year and Rattings.

- **Step 2: Data Pre-processing**
  Only the required columns are selected from the metadata and the rows are removed if null values are found in any of the selected columns.

- **Step 3: Finding the mean and standard deviation of imdb scores**
  The mean and standard deviation of ratings in both the metadata and movies of the year 2017 are calculated.

- **Step 4: Testing the Hypothesis** – Population standard deviation is known
  Since the population standard deviation is known we could use z-test to test the hypothesis.

  H0:Popularity of films has not increased
  H1:Popularity of films has increased

Level of significance is taken to be 5% i.e. $\alpha = 0.05$
The critical value is 1.645 from the z-test statistical table. We reject the Hypothesis if z is less than the critical value.

$z = (\overline{X} - \mu)/(\sigma/\sqrt{n})$

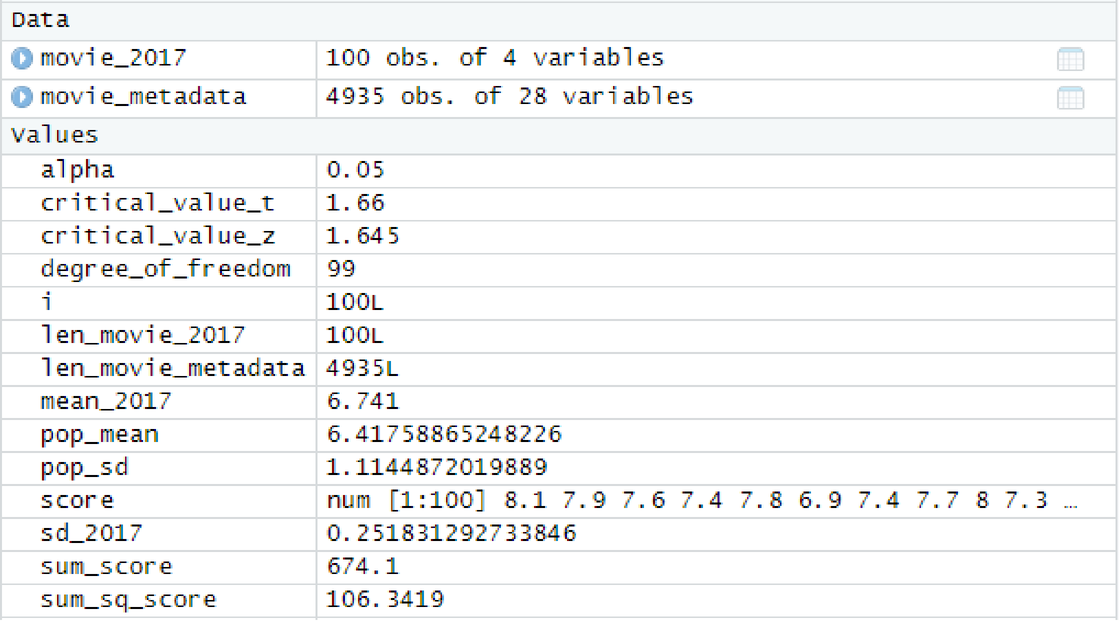
- **Step 5: Testing Hypothesis** – Population standard deviation is unknown
  Since the population standard deviation is unknown we could use t-test to test the hypothesis.

  H0:Popularity of films has not increased
  H1:Popularity of films has increased

  Level of significance is taken to be 5% i.e. $\alpha = 0.05$
  Degree of Freedom is 100 - 1 = 99 (sample size, n=100)
  The critical value is 1.660 from the t-test statistical table. We reject the Hypothesis if t is less than the critical value.

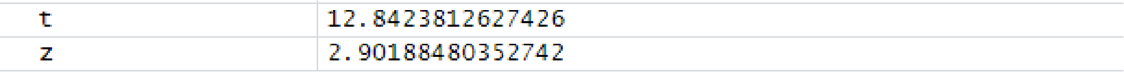
  $t = (\overline{X} - \mu)/(s/\sqrt{n})$

4 Data and Values



Data	
movie_2017	100 obs. of 4 variables
movie_metadata	4935 obs. of 28 variables
Values	
alpha	0.05
critical_value_t	1.66
critical_value_z	1.645
degree_of_freedom	99
i	100L
len_movie_2017	100L
len_movie_metadata	4935L
mean_2017	6.741
pop_mean	6.41758865248226
pop_sd	1.1144872019889
score	num [1:100] 8.1 7.9 7.6 7.4 7.8 6.9 7.4 7.7 8 7.3 …
sd_2017	0.251831292733846
sum_score	674.1
sum_sq_score	106.3419

5 Outputs



t	12.8423812627426
z	2.90188480352742

6 Conclusion

The values of both z and t are greater than their critical values respectively. Therefore, we reject the null hypothesis and conclude that the ”popularity of films are increasing”.