DATA ANALYTICS

PROJECT

Hypothesis testing to infer population mean

Group 12 :

- Cherukuri Nikhilesh - S20180010040
- Kore Nithish Kumar - S20180010086
- Puppala Sudheer - S20180010140
- Rajesh kumar - S20180010132
- Vasireddy Komal Kumar - S20180010189

Date of Submission : 14 April,2021

Understanding:

- The main objective of our project is to test the Hypothesis to infer a population mean on **IMDB movie dataset.**
- Steps to perform **Hypothesis testing**:
  - Specify H_0 and H1, the null and alternate hypothesis, and an acceptable level of alpha.
    - **H_0** : Popularity of films increases
    - **H_1** : Popularity of films does not increase.
  - Determine an appropriate sample-based test statistics and the rejection region for the specified H_0 .
  - The data follows Normal Population(Plot given below).The sample size is 3329(i.e. data belonging to year 2017) it is greater than 40. So the sample follows **Normal Distribution.**
    - Here, we have to use the following tests :
      - **Z-test** : It is a statistical test used when the population **standard deviations are known**. It can be used to test hypotheses in which the z-test **follows a normal distribution.**
      - **Z-test** : We perform this when sample **standard deviation is unknown.**
      - The rejection region/Significance level we are taking as **alpha(α) = 0.05**.
  - Collect the sample data and compute the test statistics.
    - **Z Test = $(\bar{x} - \mu) / (\sigma / \sqrt{n})$** (when standard deviation is known)
    - **Z Test = $(\bar{x} - \mu) / (S / \sqrt{n})$ (**when standard deviation is unknown)

- Now,make a decision either to reject the null hypothesis or accept the hypothesis.
  - We **accept** the hypothesis if **z known and z unknown variances are greater than z-alpha(critical region)** otherwise we reject the hypothesis.

Implementation:

Step 1: Reading and Analyzing dataset:

```
1
2  library(openxlsx)
3  DATA <- read.csv("IMDb movies.csv")
4  filter_data <- subset(DATA,(!is.na(DATA[,4])))
5  head(filter_data, n = c("title"))
6  filter_data <- subset(filter_data,(!is.na(filter_data[,15])))
7  head(filter_data, n = c("avg_vote"))
8
```

- First, we import the required libraries to read the IMDB movies dataset which consists of **85855 rows** and **22 columns**.Now, after reading the data we remove all the missing values(Null values) from the required columns that is **"year"** and **"avg_vote"**(Here avg_vote is considered as rating or imdb_score).

Step 2: Population mean from all the movies up to 2016 on imdb_Score(avg).

```
2  filter_data_2016<- subset(filter_data,(filter_data[,4]<=2016))
3  filter_data_2016[ , c("title",'year','avg_vote')]
4  populationMean <- mean(filter_data_2016[,15])
5  populationVariance <- var(filter_data_2016[,15])
6  print(populationVariance)
7  print(c("Population mean of all movies upto 2016 = ",populationMean))
8
```

- Now we filter the dataset of all the movies upto **2016** and we will calculate the population mean and variance of all the movies upto 2016 on Imdb_ score(i.e. “avg_vote”).
- **Output:**

```
[1] "The populationVariance  of all the movies upto 2016:"
[2] "1.48360500315701"
> print(c("The population mean of all the movies upto 2016:",populationMean))
[1] "The population mean of all the movies upto 2016:"
[2] "5.92409370951109"
```

❖ Step-3 : Collecting a sample of all the movies in the year 2017.

```
2  filter_data_2017 <- subset(filter_data,(filter_data[,4]==2017))
3  sampleMean <- mean(filter_data_2017[,15])
4  print(sampleMean)
5  sampleVariance <- var(filter_data_2017[,15])
6  print(sampleVariance)
7  n <- nrow(filter_data_2017)
```

- Here, we collect all the samples of movies in the year **‘2017’** and also we calculate the sample mean and variance of this sample in the dataset.
- **Output:**

```
> sampleMean <- mean(filter_data_2017[,15])
> print(c("sample mean:",sampleMean))
[1] "sample mean:"      "5.69678582156804"
> # calculates variance of avg_votes of all the movies in the year 2017
> sampleVariance <- var(filter_data_2017[,15])
> print(c("sample variance",sampleVariance))
[1] "sample variance"  "1.65659122845276"
>
```

❖ Step-4 : Testing the hypothesis that “Popularity of films increases”.

```
n <- nrow(filter_data2)
pop_mean_total = mean(filter_data[,15])
pop_var_total = var(filter_data[,15])
pop_mean = DATA[]
zknown <- (sampleMean-pop_mean_total)/(pop_var_total/sqrt(n))
alpha <- 0.01
z.alpha <- qnorm(1-alpha)
print(c("Z at alpha = 0.01 ",-z.alpha))
print(c("Z known variance = ",zknown))
zunknown <- (sampleMean-pop_mean_total)/(sampleVariance/sqrt(n))
print(c("Z unknown variance = ",zunknown))
```

- Here n denotes the number of rows of sample data(sample of all movies in the year 2017).We are assuming significance level(alpha) as **0.01.** Using z-test we perform this hypothesis and by observing in the z table we will get to find the value of P(z<0.01).
- we also calculate the total population mean and total population variance of “avg_vote” inorder to find the Z-known & Z-unknown
- After calculating the z-known and z-unknown, we compare it with the Z-alpha, to decide the rejecting or accepting the hypothesis.

Experimental results:

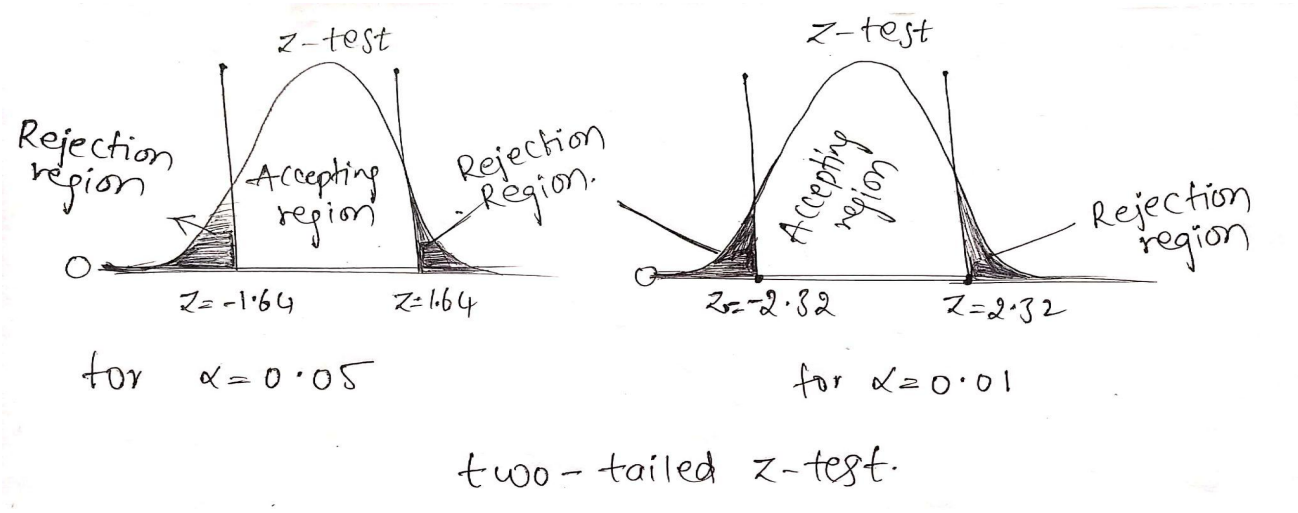
❖ For significance level(α) = 0.05

```
> alpha <- 0.05
> #value in the z-table for given significance level
> z.alpha <- qnorm(1-alpha)
> #critical value of z
> print(c("Z at alpha = 0.05 ",-z.alpha))
[1] "Z at alpha = 0.05 " "-1.64485362695147"
> #values of z when sample variance is known
> print(c("Z known variance = ",zknown))
[1] "Z known variance = " "-7.63666557987905"
> #when sample variance is unknown we use t-test that is z unknown
> zunknown <- (sampleMean-pop_mean_total)/(sampleVariance/sqrt(n))
> #printing z unknown
> print(c("Z unknown variance = ",zunknown))
[1] "Z unknown variance = " "-7.03094068910146"
```

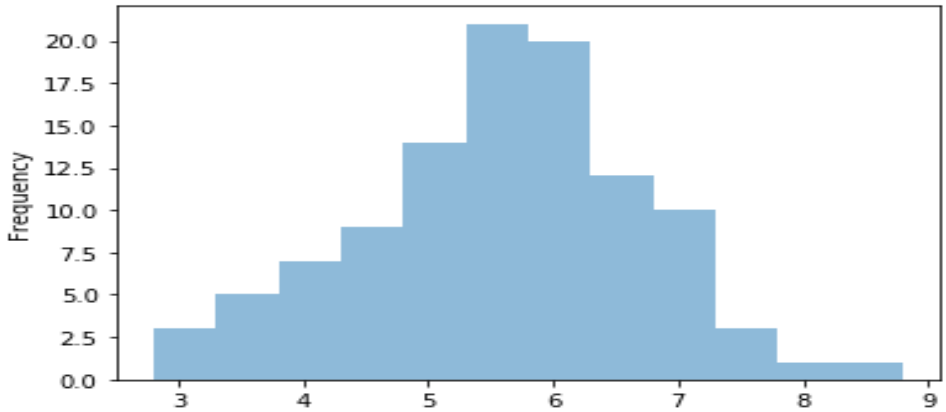
❖ For significance level(α) = 0.01

```
> alpha <- 0.01
> #value in the z-table for given significance level
> z.alpha <- qnorm(1-alpha)
> #critical value of z
> print(c("Z at alpha = 0.01 ",-z.alpha))
[1] "Z at alpha = 0.01 " "-2.32634787404084"
> #values of z when sample variance is known
> print(c("Z known variance = ",zknown))
[1] "Z known variance = " "-7.63666557987905"
> #when sample variance is unknown we use t-test that is z unknown
> zunknown <- (sampleMean-pop_mean_total)/(samplevariance/sqrt(n))
> #printing z unknown
> print(c("Z unknown variance = ",zunknown))
[1] "Z unknown variance = " "-7.03094068910146"
```

Graphical Interpretation of Hypothesis



Histogram for avg_vote(imdb_score)



From the above graph we can see that the bar graph resembles the bell shape so we can say that imdb_score follows Normal Distribution.

Observations :

By observing the above results, we can conclude that the **values of z-known and z-unknown are less than z-alpha for significance levels(0.05 &0.01).** Hence we **reject** the hypothesis that popularity of films does not increase with the year.