IT 350 Data Analytics
Lab 2: Inferential Data Analysis

P Akshara – 181IT132

Inferential analysis involves making prediction based on the patterns exhibited by the data supported by statistical justification. In this study, the following real-world datasets have been analyzed:

1. Spotify 1921-2020 music tracks
2. Covid-19 Indian lockdown data
3. Netflix movies & TV shows up to 2019.

The attached Jupyter notebooks contain the detailed analysis in Markdown cells corresponding to the tests done, and inferences drawn. Here only a summary is stated.

Tests carried out

1. Left tailed z-test on Spotify dataset on the mean acoustic value of songs.

Population: 2018-2020 songs. Sample: 300 random 2020 songs
H0: $\mu = \mu_0$
H1: $\mu < \mu_0$

Significance level: 0.05 (general application)
Critical value: -1.6448536269514729
Test statistic: -2.9416038396509134

As we have the variance of the population, we employ a left tailed z-test to suggest whether there is a significant reduction in the acoustic effect of songs in the year 2020. As the test statistic is less than critical value (it lies in the 5% of the leftmost region).

Hence H0 is rejected. Mean acoustic value in 2020 reduced compared to 2018-2020

2. 2 tailed z-test on Spotify dataset on the mean song title length

Population: 2018-2020 songs. Sample: 300 random 2020 songs
H0: $\mu = \mu_0$
H1: $\mu \mathrel{!=} \mu_0$

Significance level: 0.01 (critical application)
Critical value: 2.575829303548901
Test statistic: 2.1017781650237373

As we have the variance of the population, we employ a 2 tailed z-test with 0.01 significance level to suggest whether there is a significant difference in the title length of songs in the year 2020. Here we choose alpha = 0.01 critical significance value as the outcome of the results of this test would help artists to decide song names in future as 2020 songs were very popular. It directly affects the artists' earning and audience appeal. As the test statistic is less than negative critical value (it lies in the 2.5% of the leftmost region).

Hence H0 is NOT rejected. Mean song title length remained same in 2020

3. Right tailed t-test on Spotify dataset to study the mean popularity of songs

2 independent samples with 50 songs each
Sample 1: Before year 2000 songs. Sample 2: After year 2000 songs
H0: $\mu = \mu_0$
H1: $\mu > \mu_0$

Degrees of freedom: 98 (100 - 2)
Significance level: 0.05 (general application)
Pooled std dev: 16.62055525976365
Critical value: 1.6605512170440575
Test statistic: -8.170605486854903

As the 2 sample are independent and we don't know the variances, we employ a right tailed t-test to suggest whether there is a significant increase in the mean popularity of songs after year 2000 compared to before. Here the population is same, so pooled variance is used and DoF = n1 + n2 - 2. As the test statistic is less than critical value (it lies in the 95% region).

Hence H0 is NOT rejected. Mean song popularity did not significantly increase after year 2000

4. Chi-Squared test on Netflix dataset to study if there is association between show type and target audience

2 types of shows are there: TV and Movie shows.
The ratings of shows are TV-14, TV-PG, R, PG-13 and TV-MA. The 1st 4 indicate that parental guidance is required and it is meant for children, while TV-MA is for adults.
Accordingly, 2 categories Children and adult are created.

The aim of the test is to see if there exists some association between show type and the audience it is targeted at. Alpha = 0.05 (general application)
H0: There is no dependency between type of show (TV/movie) and target audience (children/adults).
H1: There is dependence between show type and audience targeted

Using Chi squared test on the contingency table we get,
Test statistic = 106.09479267108279
p-value = 7.02910957152232e-25
Degrees of freedom = 1

Since the 2 are categorical variables we employ the Chi-squared test for checking association/independence. As obtained p value is very low (< alpha)

Hence H0 is rejected. There is association between show type and target audience

5. 1-way ANOVA test to study if density of a region affects the count of COVID-19 positive cases

Based on the census of India in 2011, the population density of every state is found. Then 4 density groups are created based on the increasing density values. From the COVID dataset, the no. of positive cases state wise is estimated.

The aim of the test is to see if there exists some association between various density clusters and the no. of positive cases. Alpha = 0.01 (critical application)

H0: There is no effect of location density on COVID-19 cases
H1: Population location density affects count of COVID-19 cases

F-Statistic = 7.499
p = 0.000505

So, the F-statistic and p-values indicate that there is an overall significant effect of density groups on positive corona cases. Here alpha = 0.01 critical significance value as the outcome of the results of this test would affect the entire populations' life and way of living. As obtained p value is very low (< alpha)

Hence H0 is rejected. Location’s population density affects the no. of COVID-19 positive cases