511-FINAL-PROJECT-Hypothesis Test 2

HUITING SONG

2022-12-12

Import the cleaned dataset

```
df <- read.csv("spotify_cleaned.csv")
head(df,5)
```

```
##   X                                                name           artist
## 1 1 I Don't Care (with Justin Bieber) - Loud Luxury Remix       Ed Sheeran
## 2 2                        Memories - Dillon Francis Remix         Maroon 5
## 3 3                        All the Time - Don Diablo Remix     Zara Larsson
## 4 4                     Call You Mine - Keanu Silva Remix The Chainsmokers
## 5 5                Someone You Loved - Future Humans Remix    Lewis Capaldi
##   popularity year genre  subgenre danceability energy key loudness mode
## 1         66 2019   pop dance pop        0.748  0.916   6   -2.634    1
## 2         67 2019   pop dance pop        0.726  0.815  11   -4.969    1
## 3         70 2019   pop dance pop        0.675  0.931   1   -3.432    0
## 4         60 2019   pop dance pop        0.718  0.930   7   -3.778    1
## 5         69 2019   pop dance pop        0.650  0.833   1   -4.672    1
##   speechiness acousticness instrumantalness liveness valence   tempo duration
## 1      0.0583       0.1020         0.00e+00   0.0653   0.518 122.036   194754
## 2      0.0373       0.0724         4.21e-03   0.3570   0.693  99.972   162600
## 3      0.0742       0.0794         2.33e-05   0.1100   0.613 124.008   176616
## 4      0.1020       0.0287         9.43e-06   0.2040   0.277 121.956   169093
## 5      0.0359       0.0803         0.00e+00   0.0833   0.725 123.976   189052
```

From the daily observations, some music genres do dominate the front rank of the listener's music charts, such as pop music. There are many reasons for this phenomenon, such as the fact that these pop music are sung by favorite singers. Hence, with this complex relationship, whether the popularity of music is related to the genre of music itself needs to be studied. In the previous hypothesis test, the relationship between the average populative of rock music and of rap music has been addressed. However, the general relationship between the genres and the popularity cannot be explained by a test between two specific genres of musics.

In the second hypothsis test, whether the popularity is related to the genre will be examinated. Before implementing the test method, categorizing the popularity into three different groups - high,medium,low - will be helpful in understanding the relationship. Based on the 0 to 100 popularity score rules on spotify, three categories can be presented as: - High: popularity >= 66 - medium: 33 > popularity > 66 - low: popularity <= 33

The two-way table is generated to show the frequency of songs in the group of genre and popularity.

```
df$popularity <- as.factor(ifelse(df$popularity>=66, 'High',
                    ifelse(df$popularity<66 & df$popularity>33, 'Medium','Low')))
```

```
library(knitr)
t1 = table(df$popularity,df$genre)
kable(t1,align = "lccrr")
```

	edm	latin	pop	r&b	rap	rock
High	613	1378	1603	1096	921	1000
Low	2733	1343	1533	2043	1654	1748
Medium	2697	2432	2371	2292	3168	2203

In statistics, Chi-square test is commonly used in testing the independence of two variables. If two variables are independent, it means there is no relationship between two factors. Based on the question of whether the popularity is related to the genre, the null hypothesis (H0) can be set as the popularity and the genre is independent, and then correspondingly, the alternative hypothesis (Ha) will be the popularity and the genre is dependent.In this case, the specific hypotheses are:

- H0: There is no relationship between the popularity and the genre of musics.
- Ha: There is relationship between the popularity and the genre of musics.

```
t2 <- chisq.test(t1)
t2
```

```
##
##  Pearson's Chi-squared test
##
## data:  t1
## X-squared = 1270.6, df = 10, p-value < 2.2e-16
```

```
chisq.test(t1)$expected
```

```
##
##               edm    latin      pop      r&b      rap      rock
##   High   1216.957 1037.726 1109.016 1093.711 1156.542  997.0471
##   Low    2034.828 1735.143 1854.343 1828.752 1933.810 1667.1242
##   Medium 2791.215 2380.131 2543.641 2508.537 2652.647 2286.8287
```

From the output, the p-value is less than the significance level of 5%, which means the rejection of null hypothesis. In this context, rejecting the null hypothesis for the Chi-square test of independence means there is a significant relationship between the popularity and the genre of musics.