Billboard Top Song Popularity Score Comparison Project

Amy Lai

December 2023

1 Background

Spotify, the popular music streaming service, provides many statistics for songs. In this project, I will compare the popularity scores of the Billboard Top30 songs from 2018 and 2019. I would like to know if the popularity scores from 2018 and 2019 are indistinguishable through randomly mixing these populations.

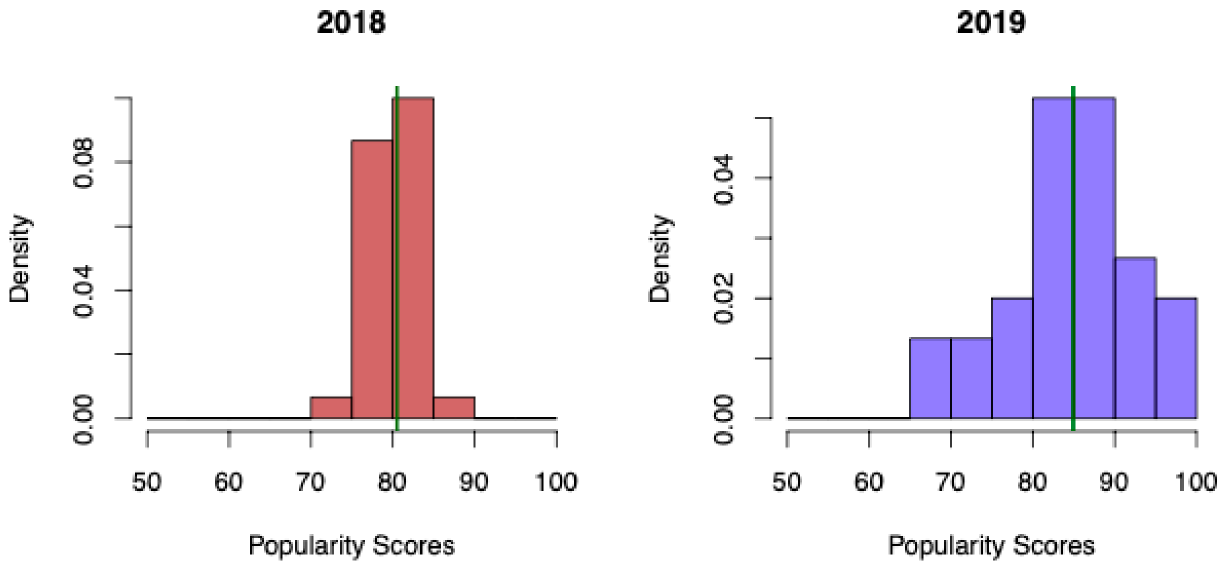
2 Data

Spotify popularity data for Billboard Top 30 songs in 2018 and 2019.

3 Analysis

3.1 Comparison by Mean

First we will examine the distribution of the the top 30 song's popularity in 2018 and 2019 respectively. A vertical line is superimposed to illustrate the mean.

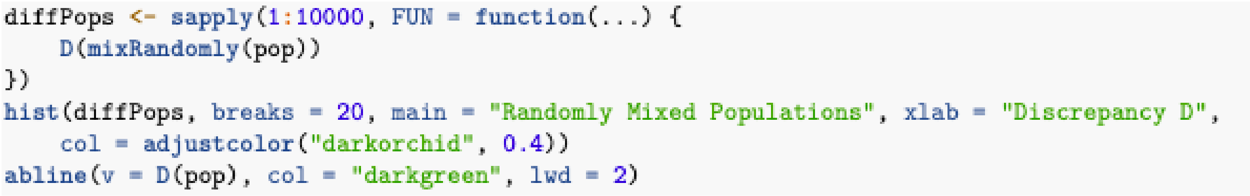


To test the null hypothesis that 2018 and 2019 popularity scores are indistinguishable, we will use the test statistic $|\bar{y}_{2018} - \bar{y}_{2019}|$.

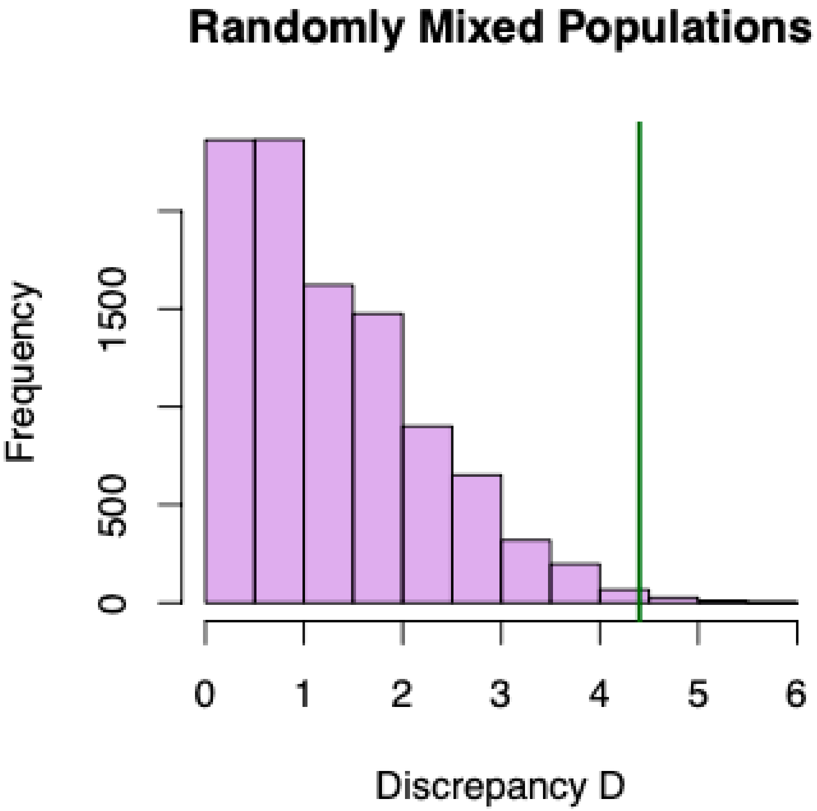
```
D <- function(pop) {
    ## First sub-population
    P1 <- pop[[1]]$popularity
    m1 <- mean(P1)

    ## Second sub-population
    P2 <- pop[[2]]$popularity
    m2 <- mean(P2)

    ## Calculate and return the Discrepancy
    abs(m1 - m2)
}
d_obs <- D(pop)
print(d_obs)
```

```
## [1] 4.4
```

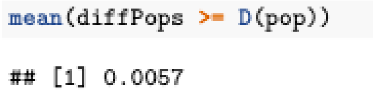
Now we'll mix the two populations 10000 times and plot a histogram of the 10000 values of the discrepancy.



```
diffPops <- sapply(1:10000, FUN = function(...) {
    D(mixRandomly(pop))
})
hist(diffPops, breaks = 20, main = "Randomly Mixed Populations", xlab = "Discrepancy D",
    col = adjustcolor("darkorchid", 0.4))
abline(v = D(pop), col = "darkgreen", lwd = 2)
```



The p-value is given by



```
mean(diffPops >= D(pop))
```

```
## [1] 0.0057
```

This p-value provides strong evidence against the null hypothesis that 2018 and 2019 popularity scores are indistinguishable based on a comparison of average.

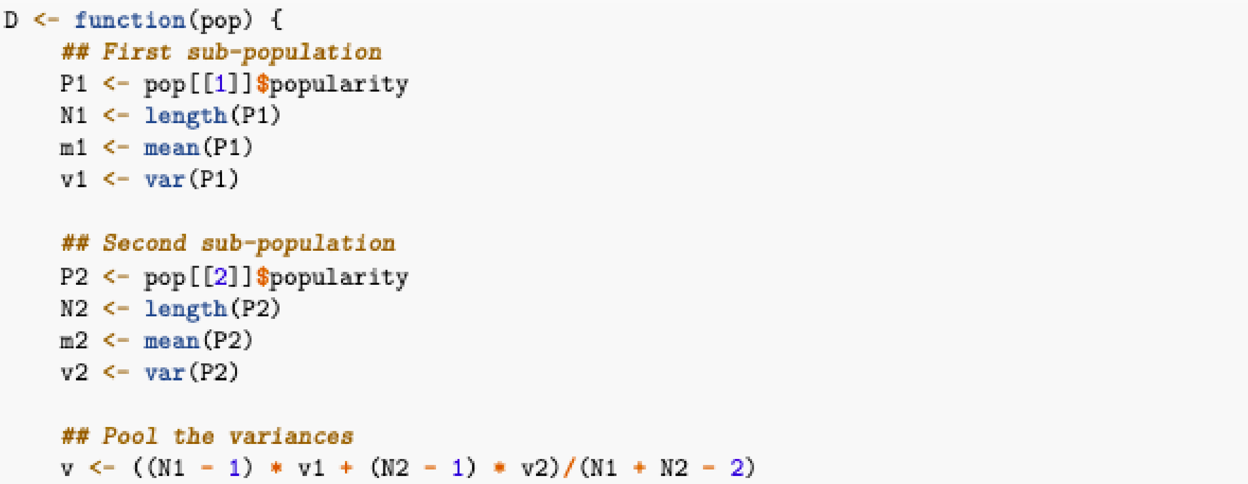
3.2 Comparison by Standard Deviation

To test the null hypothesis that 2018 and 2019 popularity scores are indistinguishable based on standard deviation, we will use the test statistic

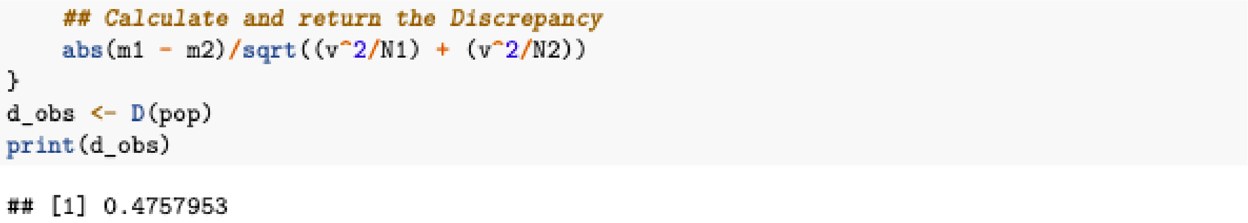
$$D(\mathcal{P}_{2018}, \mathcal{P}_{2019}) = \frac{|\overline{y}_{2018} - \overline{y}_{2019}|}{\sqrt{\frac{\tilde{\sigma}^2}{N_{2018}} + \frac{\tilde{\sigma}^2}{N_{2019}}}}$$

where

$$\tilde{\sigma}^2 = \frac{(N_{2018} - 1)\tilde{\sigma}^2_{2018} + (N_{2019} - 1)\tilde{\sigma}^2_{2019}}{(N_{2018} - 1) + (N_{2019} - 1)}$$



```
D <- function(pop) {
    ## First sub-population
    P1 <- pop[[1]]$popularity
    N1 <- length(P1)
    m1 <- mean(P1)
    v1 <- var(P1)

    ## Second sub-population
    P2 <- pop[[2]]$popularity
    N2 <- length(P2)
    m2 <- mean(P2)
    v2 <- var(P2)

    ## Pool the variances
    v <- ((N1 - 1) * v1 + (N2 - 1) * v2)/(N1 + N2 - 2)
```



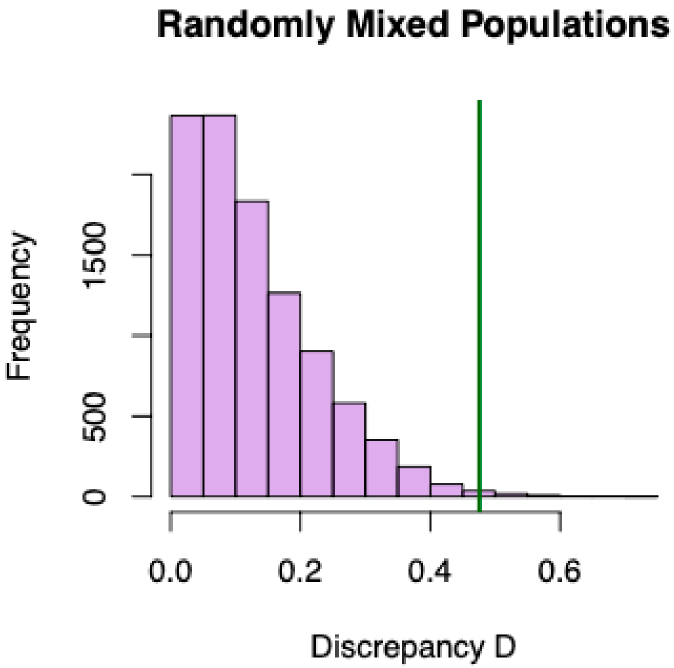
```
    ## Calculate and return the Discrepancy
    abs(m1 - m2)/sqrt((v^2/N1) + (v^2/N2))
}
d_obs <- D(pop)
print(d_obs)
```

```
## [1] 0.4757953
```

Using this formula, we obtain the observed discrepancy 0.4758.

Now we'll mix the two populations 10000 times and plot a histogram of the 10000 values of the discrepancy.

```
diffPopsT <- sapply(1:10000, FUN = function(...) {
    D(mixRandomly(pop))
})
hist(diffPopsT, breaks = 20, main = "Randomly Mixed Populations", xlab = "Discrepancy D",
    col = adjustcolor("darkorchid", 0.4))
abline(v = D(pop), col = "darkgreen", lwd = 2)
```



The p-value is given by

```
mean(diffPopsT >= D(pop))
```

```
## [1] 0.0053
```

This p-value provides strong evidence against the null hypothesis that 2018 and 2019 popularity scores are indistinguishable based on a comparison of standard deviation.

4 Conclusion

Based on the average and standard deviation, the popularity scores for Billboard Top 30 songs in 2018 and 2019 are indistinguishable.