data\_science\_final.R

matthewschultz

2022-11-28

The first step was to define my research question, which was to establish a correlation between certain variables and the popularity of the song. I ended up choosing dance ability in tempo as the two independent variables I’d be testing. To establish a correlation, I decided to perform multiple linear regression and ANOVA test to determine it.

H0: The F-value of Danceability and Tempo will be greater than 0.01 when compared to the popularity

H1: The F-value of Danceability and Tempo will be less than 0.01 when compared to the popularity

The next step in my data science project was just select a data set I was wanting to analyze. After looking through Kaggle, I ended up selecting an open Spotify dataset with 115,000 observations.

install.packages("data.table", repos = "http://cran.us.r-project.org")

##   
## The downloaded binary packages are in  
## /var/folders/tl/nvjxl2653rn0ttmj3t8c483h0000gn/T//RtmpUOMbQL/downloaded\_packages

install.packages("ggplot2", repos = "http://cran.us.r-project.org")

##   
## The downloaded binary packages are in  
## /var/folders/tl/nvjxl2653rn0ttmj3t8c483h0000gn/T//RtmpUOMbQL/downloaded\_packages

library(data.table)  
library(ggplot2)  
  
require(data.table)  
require(ggplot2)

# a tempo over 100 and a dancibility over .5 will cause the popularity of .60 or greater  
  
df <- read.csv("/Users/matthewschultz/Desktop/dataset.csv")  
  
# functions for the analysis  
# =================================================  
  
remove\_lower\_tail <- function(dataframe, low\_edge) {  
 i <- 1  
 temp <- dataframe[order(dataframe$scaled\_pop), ]  
 while (i < nrow(temp) && temp[i, "scaled\_pop"] <= low\_edge) {  
 i <- i + 1  
 }  
 return(temp[i:nrow(temp), ])  
}  
  
create\_prop <- function(data\_frame) {  
 count <- 0  
 for (row in 1:nrow(data\_frame)) {  
 if (data\_frame[row, "popularity"] >= 0.75 && data\_frame[row, "danceability"] >= 0.60 && data\_frame[row, "tempo"] >= 120) {  
 count <- count + 1  
 }  
 }  
 prop <- count / nrow(data\_frame)  
}  
  
find\_threshold <- function(data\_frame, mean\_, sd\_) {  
 two\_sds <- mean\_ + (sd\_ \* 2)  
 temp = data.frame()  
 for (row in 1:nrow(data\_frame)) {  
 if (data\_frame[row, "popularity"] >= two\_sds) {  
 temp <- rbind(temp, data\_frame[row,])  
 }  
 }  
 return(temp)  
}  
  
create\_sample\_df <- function(data\_frame, size=30) {  
 samp <- sample(1:nrow(data\_frame), size)  
 temp = data.frame()  
 for (i in samp) {  
 temp <- rbind(temp, data\_frame[i,])  
 }  
 return(temp)  
}  
  
simulation\_study <- function(data\_frame, iteration = 100, sample\_size = 50) {  
 # create a data frame to store the p-values for later analysis  
 temp\_df <- data.frame()  
   
 # used to iterate through the # of trials  
 for (i in 1:iteration) {  
 # creates a temporary sample  
 temp <- create\_sample\_df(data\_frame, sample\_size)  
 temp\_model <- lm(temp$scaled\_pop ~ temp$danceability + temp$tempo)  
 # adds p-value of danceability to data frame  
 temp\_df[i, 1] <- summary(temp\_model)$coefficients[2, 4]  
   
 # adds p-value of danceability to data frame  
 temp\_df[i, 2] <- summary(temp\_model)$coefficients[3, 4]  
 }  
 return(temp\_df)  
}  
  
# ==================================================

The first step in this process was to clean the data. I started off by performing a Z score standardization on the popularity data for the model to understand it better. This process included using the scale command and writing a function to remove any outliers of the data set. Approximately 20,000 observations were removed as outliers.

tem1 <- list(scale(df$popularity))  
temp2 <-rbindlist(Map(as.data.frame, tem1))  
  
df$scaled\_pop <- temp2$V1  
  
filtered\_df <- remove\_lower\_tail(df, -1.4)  
  
pop\_mean\_popularity <- mean(filtered\_df$popularity)  
pop\_sd\_popularity <- sd(filtered\_df$popularity)  
  
  
Next, I decided to run the multiple linear regression and ANOVA test on the filtered population. The MLR test reported that danceabilty had a positive correlation and tempo had a negative correlation. While the multiple linear regression test provided evidence for no statistical significance between the independent and dependent variables, the ANOVA test provided the opposite result, marking strong evidence to support the correlation.

model <- lm(filtered\_df$scaled\_pop ~ filtered\_df$danceability + filtered\_df$tempo)  
summary(model)$coefficient

## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 0.1353560499 1.445780e-02 9.362148 7.976660e-21  
## filtered\_df$danceability 0.3960354527 1.537278e-02 25.762122 7.519102e-146  
## filtered\_df$tempo -0.0004814305 8.849559e-05 -5.440164 5.336277e-08

# because of the t-value there is a level of signicants between the two variables   
# and popularity T-score of dancaebility: 25.76 / of temp: -5.44

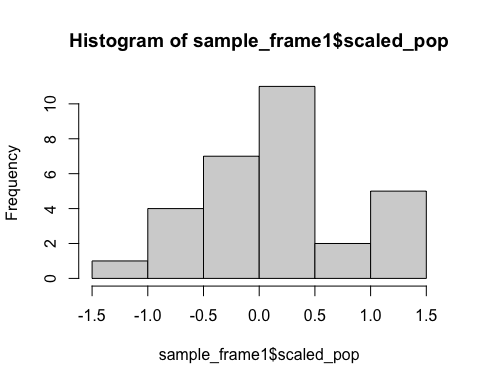
anova(model)

## Analysis of Variance Table  
##   
## Response: filtered\_df$scaled\_pop  
## Df Sum Sq Mean Sq F value Pr(>F)   
## filtered\_df$danceability 1 450 450.39 678.447 < 2.2e-16 \*\*\*  
## filtered\_df$tempo 1 20 19.65 29.595 5.336e-08 \*\*\*  
## Residuals 94801 62934 0.66   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

pop\_prop <- create\_prop(df)

Next, I decided to perform the same test on random samples with varying sample sizes, including 30, 50, and 75. Three random samples of size 30 were created and tested on. Each random sample reported no statistical significance through the multiple linear progression test and the ANOVA test. It is also important to note that the distribution of each sample is approximately normal. Only one random sample of size 50 was collected and tested on. The results of those test indicated no statistical significance between the dependent and independent variables for the linear regression test, however, marked a statistical significance for tempo on the ANOVA test. Finally, one Random sample of size 75 was collected and tested on. The results showed no statistical significance between the independent and dependent variables for both multiple linear regression and ANOVA test.

# small Samples  
# # ==================================  
  
# Sample 1  
# ==================================  
sample\_frame1 <- create\_sample\_df(filtered\_df)  
# sample\_prop1 <- create\_prop(sample\_frame)  
hist(sample\_frame1$scaled\_pop)



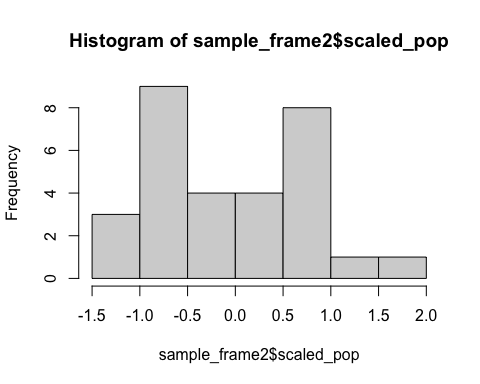
model1 <- lm(sample\_frame1$scaled\_pop ~ sample\_frame1$danceability + sample\_frame1$tempo)  
summary(model1)$coefficient

## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 0.409165639 0.827700194 0.4943404 0.6250641  
## sample\_frame1$danceability -1.218515666 0.848382796 -1.4362805 0.1624089  
## sample\_frame1$tempo 0.003831281 0.004936503 0.7761125 0.4444284

anova(model1)

## Analysis of Variance Table  
##   
## Response: sample\_frame1$scaled\_pop  
## Df Sum Sq Mean Sq F value Pr(>F)  
## sample\_frame1$danceability 1 1.1153 1.11531 2.2509 0.1451  
## sample\_frame1$tempo 1 0.2985 0.29846 0.6024 0.4444  
## Residuals 27 13.3784 0.49550

# Sample 2  
# ==================================  
sample\_frame2 <- create\_sample\_df(filtered\_df)  
hist(sample\_frame2$scaled\_pop)



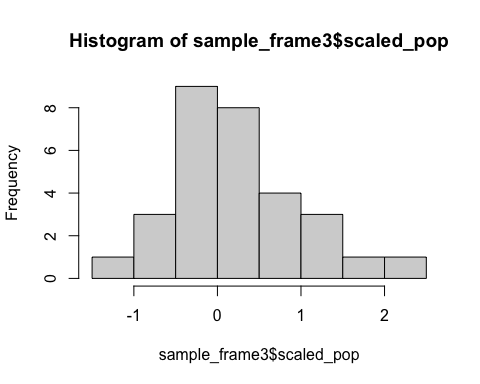
model2 <- lm(sample\_frame2$scaled\_pop ~ sample\_frame2$danceability + sample\_frame2$tempo)  
summary(model2)$coefficient

## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 0.0966016868 0.706707430 0.13669262 0.8922879  
## sample\_frame2$danceability -0.0448298132 0.824909623 -0.05434512 0.9570603  
## sample\_frame2$tempo -0.0008770255 0.004549976 -0.19275390 0.8485943

anova(model2)

## Analysis of Variance Table  
##   
## Response: sample\_frame2$scaled\_pop  
## Df Sum Sq Mean Sq F value Pr(>F)  
## sample\_frame2$danceability 1 0.0038 0.00381 0.0061 0.9382  
## sample\_frame2$tempo 1 0.0231 0.02312 0.0372 0.8486  
## Residuals 27 16.8031 0.62234

# Sample 3  
# ==================================  
sample\_frame3 <- create\_sample\_df(filtered\_df)  
hist(sample\_frame3$scaled\_pop)



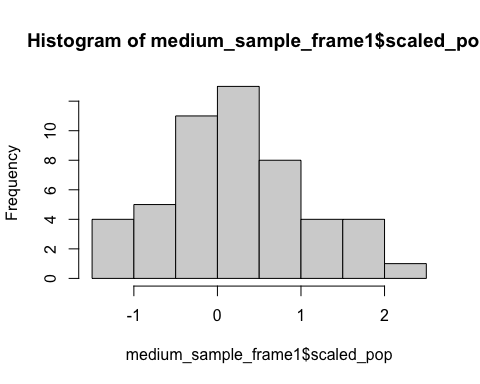
model3 <- lm(sample\_frame3$scaled\_pop ~ sample\_frame3$danceability + sample\_frame3$tempo)  
summary(model3)$coefficient

## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) -0.1048469088 0.608393063 -0.1723342 0.8644601  
## sample\_frame3$danceability 0.8964255138 0.606400880 1.4782721 0.1509070  
## sample\_frame3$tempo -0.0009621166 0.003959221 -0.2430066 0.8098370

anova(model3)

## Analysis of Variance Table  
##   
## Response: sample\_frame3$scaled\_pop  
## Df Sum Sq Mean Sq F value Pr(>F)  
## sample\_frame3$danceability 1 1.0945 1.09447 2.1513 0.1540  
## sample\_frame3$tempo 1 0.0300 0.03004 0.0591 0.8098  
## Residuals 27 13.7360 0.50874

# medium samples  
# ==================================  
  
# Sample 1  
# ==================================  
medium\_sample\_frame1 <- create\_sample\_df(filtered\_df, 50)  
# sample\_prop1 <- create\_prop(sample\_frame)  
hist(medium\_sample\_frame1$scaled\_pop)



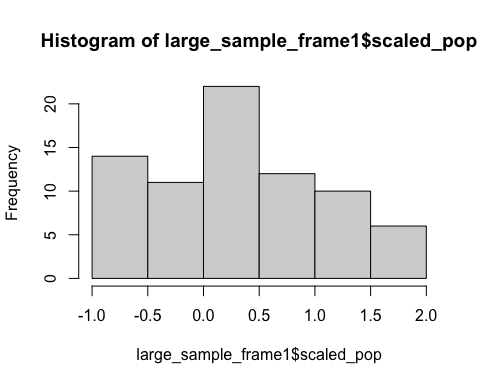
medium\_model1 <- lm(medium\_sample\_frame1$scaled\_pop ~ medium\_sample\_frame1$danceability + medium\_sample\_frame1$tempo)  
summary(medium\_model1)$coefficient

## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) -1.04292146 0.718577228 -1.4513700 0.15331935  
## medium\_sample\_frame1$danceability 0.13822851 0.657310448 0.2102941 0.83434717  
## medium\_sample\_frame1$tempo 0.00999534 0.004541364 2.2009553 0.03268683

anova(medium\_model1)

## Analysis of Variance Table  
##   
## Response: medium\_sample\_frame1$scaled\_pop  
## Df Sum Sq Mean Sq F value Pr(>F)   
## medium\_sample\_frame1$danceability 1 0.0013 0.00132 0.0021 0.96378   
## medium\_sample\_frame1$tempo 1 3.0738 3.07378 4.8442 0.03269 \*  
## Residuals 47 29.8227 0.63453   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# large samples  
# ==================================  
  
# Sample 1  
# ==================================  
large\_sample\_frame1 <- create\_sample\_df(filtered\_df, 75)  
# sample\_prop1 <- create\_prop(sample\_frame)  
hist(large\_sample\_frame1$scaled\_pop)



large\_model1 <- lm(large\_sample\_frame1$scaled\_pop ~ large\_sample\_frame1$danceability + large\_sample\_frame1$tempo)  
summary(large\_model1)$coefficient

## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 0.787796140 0.434655929 1.8124592 0.07408428  
## large\_sample\_frame1$danceability 0.293143714 0.460785593 0.6361825 0.52667474  
## large\_sample\_frame1$tempo -0.005189075 0.002741371 -1.8928758 0.06239393

anova(large\_model1)

## Analysis of Variance Table  
##   
## Response: large\_sample\_frame1$scaled\_pop  
## Df Sum Sq Mean Sq F value Pr(>F)   
## large\_sample\_frame1$danceability 1 0.214 0.21376 0.3906 0.53396   
## large\_sample\_frame1$tempo 1 1.961 1.96076 3.5830 0.06239 .  
## Residuals 72 39.401 0.54724   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

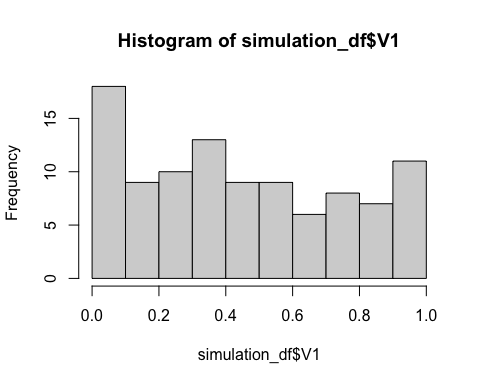
The discrepancies between the population ANOVA test and the different random samples ANOVA test led to further experimentation on the data set. Specifically, a simulation was conducted to find the variance of this sample over 100 iterations. A script was developed to log the F value for both tempo and danceability of 100 different samples.

The simulation established the sample size as 50 for the test. After running the script, the interquartile range and the distribution were then analyzed. The F value of danceability was analyzed first. The distribution returned approximately uniform. IQR returned a large range, with the minimum Being 0.0001564 and the maximum being 0.9898. Then, the F value of tempo was analyzed. The distribution of tempo returned approximately uniform, and the IQ are returned a large range, with the maximum being 0.9879 and the minimum being 0.009735.

simulation\_df <- simulation\_study(filtered\_df)  
  
summary(simulation\_df$V1)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0001564 0.1738043 0.3961578 0.4389218 0.7075489 0.9898030

hist(simulation\_df$V1)



summary(simulation\_df$V2)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.009735 0.239005 0.472561 0.473487 0.699406 0.987971

Due to the discrepancies between the population in sample test, there is enough evidence to support the null hypothesis. Even though the f-value for the population was less than 0.01 for both the population tempo and danceablity, the multiple random samples provide evidence to support the null.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Small Sample 1 | Small Sample 2 | Small Sample 3 | Medium Sample | Large Sample |
| Danceability | 0.1451 | 0.9570603 | 0.1540 | 0.96378 | 0.53396 |
| Tempo | 0.4444 | 0.8485943 | 0.8098370 | 0.03269 | 0.06239 |

Further analysis has led to an increasingly Complex rabbit hole with no trends to follow. As a result, one could speculate that there are lurking variables that are impacting the popularity of a song. On that note, it is important to acknowledge some limitations of the data set. For example, the data set does not include the author of the song’s follower counts or monthly listeners, which could tremendously impact the popularity of the song. Additionally, it is not known how the popularity of a song is calculated. Knowing this formula could become vital in determining a direct relationship on how a song can become popular. Finally, additional studies should be conducted on other variables in the data set and on variables outside of the data set to produce a correlation more effectively.