Homework 6

Caitlin Jagla

Problem 1 – Music Genres

For this homework we are going to see what type of music people seem to be listening to the most on Spotify. The observations for this dataset are at the track level, with each track belonging to a genre of music.

A. Understand the dataset

Load in the `spotify.csv` file and use a function to investigate the dataset.

```
library(tidyverse)


df <- read_csv("spotify.csv")
glimpse(df)
```

```
## Rows: 26,790
## Columns: 6
## $ track_name        <chr> "I Don't Care (with Justin Bieber) - Loud Luxury Rem~
## $ track_artist      <chr> "Ed Sheeran", "Maroon 5", "Zara Larsson", "The Chain~
## $ playlist_genre    <chr> "pop", "pop", "pop", "pop", "pop", "pop", "pop", "po~
## $ playlist_subgenre <chr> "dance pop", "dance pop", "dance pop", "dance pop", ~
## $ track_popularity  <dbl> 66, 67, 70, 60, 69, 67, 62, 69, 68, 67, 58, 67, 67, ~
## $ highly_popular    <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1~
```

B. Investigate the genre of each playlist. How many observations are there in each different playlist?

```
df %>% group_by(playlist_genre) %>% summarise(n = n())
```

```
## # A tibble: 5 x 2
##   playlist_genre     n
##   <chr>          <int>
## 1 latin           5155
## 2 pop             5507
## 3 r&b             5431
## 4 rap             5746
## 5 rock            4951
```

C. Let's say we are interested in whether the proportion of observations for each genre was equal. State your statistical hypotheses and alpha.

- H_0: Pr(latin) = Pr(pop) = Pr(r&b) = Pr(rap) = Pr(rock) = 1/5 = 0.2
  - aka there is no significant difference between the observed and expected proportion of observations.
- H_1: at least one genre has a different probability than listed in H_0
  - aka there is a significant difference between the observed and expected proportion of observations.
- $\alpha = 0.05$

D. What type of test do you think you should run? How do you know?

- You should run a Chi Square Goodness of Fit test because these are proportional data, and we want to compare their actual distribution to the theoretically equal theoretical distribution.

E. Check the assumptions of this test.

- a) random sampling is assumed
- b) expected values are all >5
  - 0.2 x 6 = 1.2

F. Run the test and interpret the output

```
obs_per_genre <- df %>%
  group_by(playlist_genre) %>%
  summarise(n_obs = n()) %>%
  mutate(p_exp = 0.2,
         n_exp = sum(n_obs)*0.2)

obs_per_genre_chi <- chisq.test(x = obs_per_genre$n_obs,
                                p = obs_per_genre$p_exp)

obs_per_genre_chi
```

```
##
##  Chi-squared test for given probabilities
##
## data:  obs_per_genre$n_obs
## X-squared = 71.842, df = 4, p-value = 9.266e-15
```

```
obs_per_genre
```

```
## # A tibble: 5 x 4
##   playlist_genre n_obs p_exp n_exp
##   <chr>          <int> <dbl> <dbl>
## 1 latin           5155   0.2  5358
## 2 pop             5507   0.2  5358
```

```
## 3 r&b             5431   0.2  5358
## 4 rap             5746   0.2  5358
## 5 rock            4951   0.2  5358
```

- The p-value ($p = 9.2657435 \times 10^{-15}$) is less than $\alpha = 0.05$, so I reject the null hypothesis and conclude that the proportion of observations for each genre is not equal. Comparing the observed proportion of observations to the expected proportions shows that there are more songs on the rap, pop, and r&b playlists than expected, while the latin and rock playlists have fewer songs than would be expected if the observations were equally proportioned across genres.

Problem 2 – Genre Popularity

Now I want to know how popular each of the genres are. This is indicated by the `highly_popular` column. I am curious if there is an association between `playlist_genre` and the proportion of highly popular songs.

A. Write out your statistical hypotheses and alpha

- H_0: Pr(popular) = Pr(not popular) = 1/2 = 0.5, for each genre
- H_1:
- $\alpha = 0.05$

B. Use the `xtabs()` function to create a cross tabulation of `highly_popular` and `playlist_genre`.

```
xt <- xtabs(~playlist_genre + highly_popular, data=df)
xt
```

```
##               highly_popular
## playlist_genre    0    1
##          latin 3309 1846
##          pop   3367 2140
##          r&b   3889 1542
##          rap   4234 1512
##          rock  3446 1505
```

```
# Add marginal totals
addmargins(xt)
```

```
##               highly_popular
## playlist_genre     0     1   Sum
##          latin  3309  1846  5155
##          pop    3367  2140  5507
##          r&b    3889  1542  5431
##          rap    4234  1512  5746
##          rock   3446  1505  4951
##          Sum   18245  8545 26790
```

```
# Get the proportional table
prop.table(xt, margin=1)
```

```
##               highly_popular
## playlist_genre         0         1
##          latin 0.6419011 0.3580989
##          pop   0.6114037 0.3885963
##          r&b   0.7160744 0.2839256
##          rap   0.7368604 0.2631396
##          rock  0.6960210 0.3039790
```

C. Check assumptions. Which type of test should you use?

- a) random sampling is assumed
- b) expected values are all >5 (see table above)
- c) You should use a Chi Square Contingency Table test because we want to test the association of two factors (playlist genre & highly popular songs)

D. Run the test

```
prop_popular_chi <- chisq.test(xt)
prop_popular_chi
```

```
##
##  Pearson's Chi-squared test
##
## data:  xt
## X-squared = 277.51, df = 4, p-value < 2.2e-16
```

E. Interpret the results. Are any genres more popular than expected by chance? If so, which ones?

```
chisq.test(xt)$expected #Expected > 5
```

```
##               highly_popular
## playlist_genre        0        1
##          latin 3510.749 1644.251
##          pop   3750.475 1756.525
##          r&b   3698.716 1732.284
##          rap   3913.243 1832.757
##          rock  3371.818 1579.182
```

```
xt
```

```
##               highly_popular
## playlist_genre    0    1
##          latin 3309 1846
##          pop   3367 2140
##          r&b   3889 1542
##          rap   4234 1512
##          rock  3446 1505
```

```
#Get difference between expected and observed values
xt - chisq.test(xt)$expected
```

```
##               highly_popular
## playlist_genre          0          1
##          latin -201.74935  201.74935
```

```
##          pop   -383.47462  383.47462
##          r&b    190.28425 -190.28425
##          rap    320.75737 -320.75737
##          rock    74.18234  -74.18234
```

- Since $p = 7.6629516 \times 10^{-59}$ is less than $\alpha = 0.05$, I reject the null hypothesis and conclude that highly popular songs are not equally distributed across genres. Comparing the expected and observed values shows that `latin` and `pop` playlists contain more highly popular songs than expected by chance.

Problem 3 – Table Creation

Create two publication worthy tables in RMarkdown that show:

a) The mean `track_popularity` by `playlist_genre`
b) The top 5 artists in terms of average `track_popularity`

```
library(ggpubr)
# mean `track_popularity` by `playlist_genre`
tbl1 <- df %>%
          group_by(playlist_genre) %>%
          summarise(`Mean Track Popularity` = mean(track_popularity)) %>%
          mutate(`Mean Track Popularity` = round(`Mean Track Popularity`, digits = 3)) %>%
          rename(`Playlist Genre` = playlist_genre)

tbl1_title <- paste0("Spotify playlists in the Latin and pop
                      genres contain more highly popular songs
                      than playlists of other genres.") %>%
                      strwrap(42) %>%
                      paste(collapse = "\n")

tbl1 %>%
  ggtexttable(rows = NULL, theme = ttheme("light")) %>%
 tab_add_title(text = tbl1_title, size = 11)
```

Spotify playlists in the Latin and pop genres contain more highly popular songs than playlists of other genres.

Playlist Genre	Mean Track Popularity
latin	47.027
pop	47.745
r&b	41.224
rap	43.215
rock	41.728

```r
#  The top 5 artists in terms of average `track_popularity`

tbl2 <- df %>%
          group_by(track_artist) %>%
          summarise(mean = mean(track_popularity)) %>%
          slice_max(mean, n = 5) %>%
          mutate(mean = round(mean, digits = 2)) %>%
          rename("Artist" = track_artist,
                 `Mean Track Popularity` = mean)



tbl2_title <- paste0("Top 5 artists by average track popularity.")

tbl2 %>% head(n=5) %>%
  ggtexttable(rows = NULL, theme = ttheme("light")) %>%
 tab_add_title(text = tbl2_title, size = 11)
```

Top 5 artists by average track popularity.

Artist	Mean Track Popularity
Trevor Daniel	97.00
Regard	94.00
Y2K	91.00
Don Toliver	90.50
Roddy Ricch	88.06