What Makes a Song Popular?

Nathan Zhang, Ira Bradie, Daniel Yan

Abstract

For this study we analyzed three different relationships between different aspects of a song (song tempo in beats per minute (BPM), song duration, and song time signature measured in beats per bar), and the song's popularity rating (measured from 0 to 1). We tested to see if there is a statistically significant correlation between song duration and song popularity using a bootstrap distribution. This 95% confidence interval did not contain 0, providing statistically significant but not practically significant evidence for an association between song duration and song popularity. To test for an association between song BPM and song popularity we used another bootstrap distribution to find a confidence interval, resulting in a 95% confidence interval that included 0, showing that there is not a statistically significant association between song BPM and song popularity. To test if there is a difference in mean song popularity between songs with different time signatures we conducted a one-way ANOVA test and found a statistically significant p-value for at least one of the means being different. Given our sample size of 4,212, however, this result is not practically significant.

Introduction:

We were interested in learning more about what aspects make a song popular. The results obtained could be used to provide guidelines for artists who want their songs to be as popular as possible. Our data was from the Million Song Dataset (MSD), a collection of one million "contemporary popular songs" gathered by the company Echo Nest, from which we took a random sample of 10,000 songs to analyze. Within these 10,000 randomly selected songs, we analyzed three relationships: the association between song duration and popularity, the association between song BPM and popularity, and the relationship between song time signature and popularity. Context for these analyses are given by Sharma et al. 's work on predicting music popularity through Spotify music metrics, and Vardo et al.'s study which uses prediction algorithms to provide insight on the most important factors (acoustics, tempo) that influence the popularity of a song .

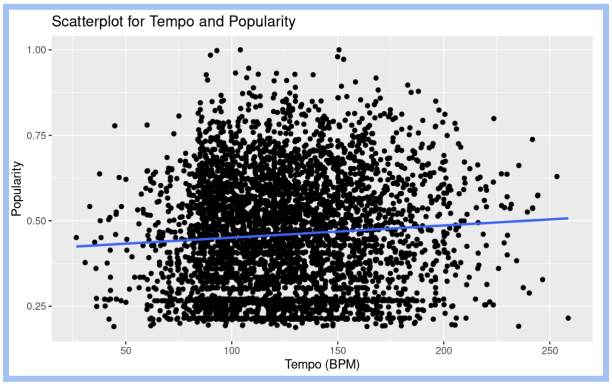
Data:

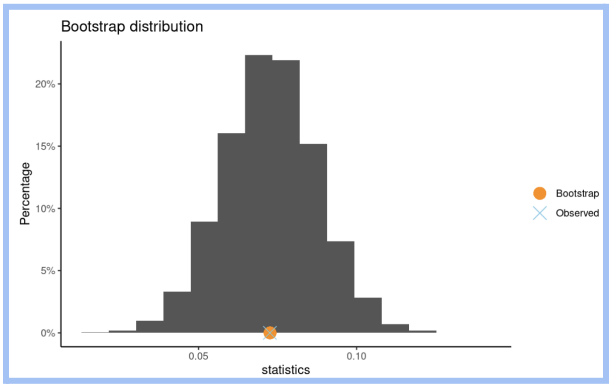
Our population of interest is the MSD and our subjects are the songs in the dataset. We used a random sample of 10,000 songs from the MSD provided by CORGIS (Whitcomb), and for each analysis we filtered out values that were less than or equal to 0 for our explanatory and response variables when applicable. This resulted in a sample size of 4,212 for measuring association between time signature and popularity, 4,214 for duration and popularity, 4,208 for tempo and popularity. Because our data was not found through an experiment, we are using information from an observational study. For all of our analyses, we were interested in the popularity of the song as our response variable. For our 3 analyses, our explanatory variable was the song duration, tempo (BPM), and time signature (beats per bar). Assuming the sample provided by CORGIS was selected randomly, we can use our analysis to answer the questions in the introduction, however, we are only able to generalize results to our population, the MSD, and not to all songs in general.

Results:

Is there an association between BPM and song popularity?

To find statistically significant evidence for an association between BPM and song popularity at the 0.05 significance level, the 95% confidence interval for the correlation should not contain 0. To answer this question, we used our filtered sample of 4208 songs to create a scatter plot and residual plot to observe whether the scatterplot has a linear relationship. There was no visible linear relationship in the plots, so we made attempts to transform the data using the log of tempo and/or song popularity; however, none of these transformations yielded a linear relationship. As a result, we used the data from our original scatter plot and found an r value of 0.07257. Next, we conducted a bootstrap analysis of the correlation which yielded a 95% confidence interval of (0.04398481, 0.10167353). This interval doesn't contain 0, which provides statistically significant evidence that there exists an association between BPM and song popularity. It is worth noting that none of the values in the confidence interval are practically significant in terms of application as the association is still low at best.





Is there an association between time signature and song popularity?

For the relationship between song time signature and song popularity, we were interested in determining whether songs with different time signatures led to different mean popularities or whether all time signatures were the same. We used the hypothesis test:

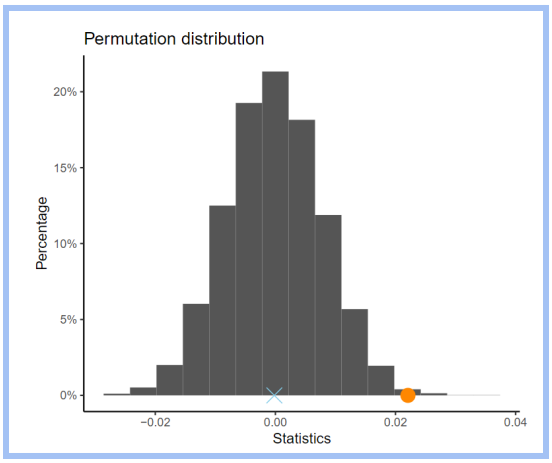
H_0: $\mu_1 = \mu_3 = \mu_4 = \mu_5 = \mu_7$

H_a: At least one mean is not equal to the others

Where μ_n is the mean popularity of songs with n beats per bar.

Using a one-way ANOVA test, we were able to determine that there is a statistically significant difference in at least one mean popularity of songs of time signature 1, 3, 4, 5, or 7. This is seen through the p-value of 0.000679 below. To determine which time signature was different from the others, we used a permutation test between the time signature with the highest mean popularity rating (4 beats per bar) and the time signature with the lowest mean popularity rating (3 beats per bar). Through this permutation test we found a p-value for our test statistic (the difference in mean popularities) to have a p-value of 0.0051, and a 95% confidence interval between 0.01402 and 0.03012. Therefore we know there is a statistically significant difference between the mean popularity of songs with time signature = 4 and songs with time signature = 3.

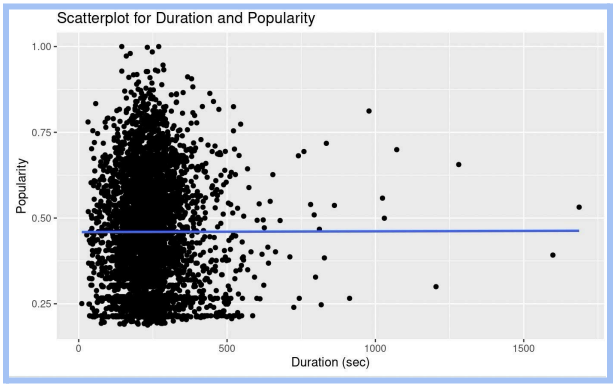
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
song.time2	4	0.55	0.1365	4.841	0.000679
Residuals	4207	118.63	0.0282		

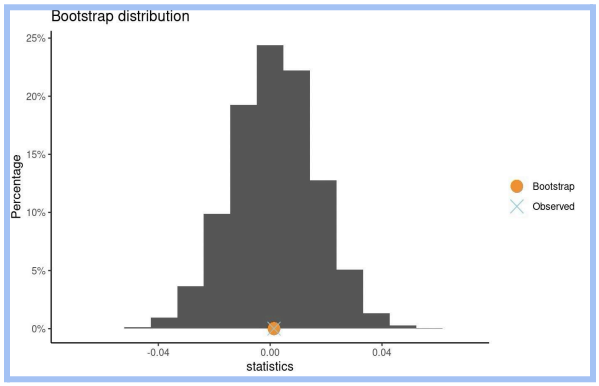


While we found there to be a statistically significant difference for there being a difference in mean popularity between songs with different time signatures (specifically time signatures 3 and 4), this does not mean that our results are practically significant. Analyzing our ANOVA test, we can see that our second degree of freedom is very large due to the sample size of 4,212, which in turn causes there to be a large F statistic and then a small p-value. Furthermore, looking at our confidence interval between songs with time signatures 3 and 4, while 0 is not contained in the interval the difference between the means is still quite small and in practice not significant.

<u>Is there an association between song duration and song popularity?</u>

To test the association between song duration and song popularity, we created a scatterplot to observe the relationship between the length of a song and its popularity using our sample size of 4,214 songs after filtering out null values. From the scatterplot, there did not appear to be any linear relationship between the two variables, even after attempting several transformations on both axes. We ended up using the correlation coefficient from the non-transformed scatterplot and found an extremely low r value of 0.001382954.





We then conducted a bootstrap analysis of the correlation which yielded a 95% confidence interval of (-0.02820780, 0.03098027). This interval does contain 0, which means that there is no significant correlation between the length of a song and its popularity.

Discussion:

We found statistically significant evidence for an association between song BPM and song popularity, no statistically significant evidence for an association between song duration and song popularity, and statistically significant evidence for at least one time signature having a different mean from the others. Although these results may have been practically significant, given the very large sample sizes for each of our questions (4,208 to 4,214), and the low values given by our 95% confidence intervals, we cannot conclude that this evidence is practically significant.

An ethical consideration is that the Million Song Dataset Website (Million) did not properly communicate how their songs were selected and how the data was collected, violating Principle B of the American Statistical Association's Ethical Guidelines for Statistical Practice. This also was a limitation to our findings as not knowing how Echo Nest chose their million songs, or the methods used to assign popularity ratings, makes it difficult to generalize our findings to all popular songs. This makes our population the MSD, which makes our findings less useful for any practical applications.

While we were unable to find a significant correlation between duration and popularity, we wondered if there was an ideal interval of song length that leads to higher popularity. We also wondered how representative the MSD was of all popular songs.

References:

L. Vardo, J. Jerkić and E. Žunić, "Predicting Song Success: Understanding Track Features and Predicting Popularity Using Spotify Data," *2023 22nd International Symposium INFOTEH-JAHORINA (INFOTEH)*, East Sarajevo, Bosnia and Herzegovina, 2023, pp. 1-6, doi: 10.1109/INFOTEH57020.2023.10094172.

Million Song Dataset. (2011, February 8). http://millionsongdataset.com/pages/getting-dataset/

Sharma, Dr. N., Pareek, Dr. P., Pathak, Mr. P., & Sakariya, Ms. N. (2022). Predicting music popularity using machine learning algorithm and music metrics available in Spotify. *JOURNAL OF DEVELOPMENT ECONOMICS AND MANAGEMENT RESEARCH STUDIES*, *09*(11), 10–19. https://doi.org/10.53422/jdms.2022.91102

Whitcomb, R. (2016, May 18). *Music CSV file*. CORGIS Datasets Project. https://corgis-edu.github.io/corgis/csv/music/

Appendix:

Song time signature	Mean song popularity
1	0.4415185
3	0.4399144
4	0.4668112
5	0.4613549
7	0.4449930

This table shows the different mean song popularities for the different time signatures 1, 3, 4, 5, and 7. Below are the boxplots of the distributions for each song popularity.