What makes a song "classic": an analysis of lyrics from top songs from 1960 to 2005?

Kuan-Jung Huang[1], Jiacheng Wang[2], Zhangqi Duan[3]

[1] Psychological and Brain Sciences, [2] Chemical Engineering, [3] Computer Science

Overview of the project and data collection

The project aims to reveal if there is a relationship between some attributes of the lyrics of a song and its popularity. We retrieved a dataset from Kaggle* that originally contained the top 10 songs of each year from 1960 to 2019, trimming it down to 1960-2005 (the advent of YouTube). The dataset has five columns: year, rank, artist, song title, and lyrics. We further scraped the YouTube website of each song and extracted view counts, likes, and comment counts using R.

* https://www.kaggle.com/datasets/stefancomanita/top-us-songs-from-1950-to-2019-w-lyrics

Objectives and interests

Music is an essential part of our life. Some songs are more popular (and more everlasting) than others. What makes it so? One potential factor is their lyrics. Here we analyze 445 songs' popularity (as defined by the view counts) as a function of the properties of the lyrics. While all songs were in the top 10 in their respective years, intuitively, some songs have a more long-lasting popularity than others. Our descriptive/nonparametric statistics concern "what the most common words in the lyrics among these top songs are" and "the extent of variation in view counts among these top songs", "skewness of view count distributions". For our main inferential statistical analysis, we will use *where the song title appears in the lyrics* (a: very first line of the song, b: in the chorus, and c: elsewhere) as a predictor for YouTube view counts. Our hypothesis is that songs with their title in the chorus will tend to be most popular because repeating the title feels punchy. Songs with their title in the very first line of the whole song should be the second most popular because of its explicitness (but the title usually doesn't repeat as much as in songs with the title in the chorus).

Class Material Used for the Analysis

web scraping and regex (extract view counts; analyze lyrics), ggplot (visualization), Monte Carlo (power analysis), bootstrapping (CI for coefficient of variation and for view counts), functions, control flows, & data transformation (generic purpose)

Responsibilities of Each Group Members

- Scrape YouTube data using the API to retrieve JSON output – Jiacheng.
- Extract view / like / comment count from JSON data – Galvin.
- Analyze most common content words in the lyrics – Kuan-Jung.
- Coefficient of variation – All members.
- Power analysis – All members.
- View count ~ song title position – All members.

'All members': we will do the analyses individually and cross-validate our results.

Conclusion

We would like to present the full conclusion on the presentation day.