Genre Bias in New Sounds

Sagnik Anupam

Formal Hypothesis

Let X1 be a distribution of songs played in the morning classified into genres. Let X2 be the corresponding distribution for the evening.

Null hypothesis H0: The distribution X1 follows the distribution of X2.

Alternative hypothesis: The difference in the distributions X1 and X2 is statistically significant.

Data Collection

- Use pandas for importing csv into Python.
- Compute a dictionary containing the number of songs of each genre played in the morning, and a similar one for the evening.
- There are some genres that have only been played in the morning, and some that have only been played in the evening.
- Total number of songs belonging to those genres is 8, vs 946 songs played in total in both mornings and evenings. As number of such songs <1% of total number of songs, so we only keep genres that were played in both mornings and evenings, and check their probability distributions against each other.

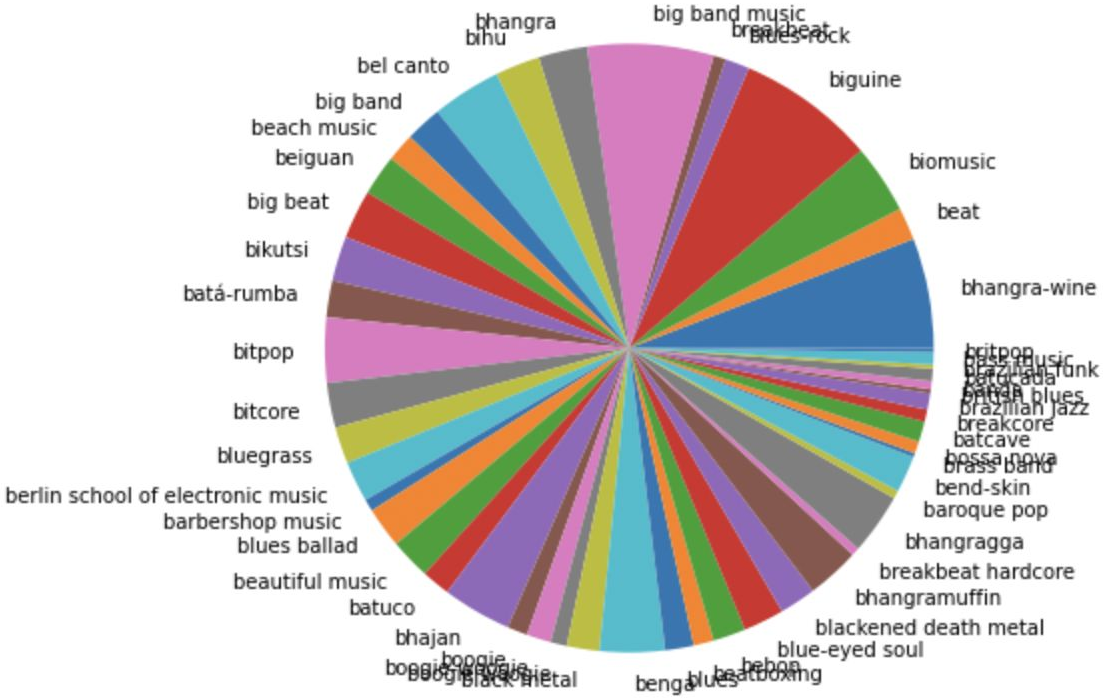
Chi-square Goodness of Fit Test

- Used to determine whether an observed distribution matches an expected distribution.
- Sum of observed and expected counts should both be the same, and for running test on Python, values should be non-zero (hence omitting genres played at only one point in time) vs setting their observed counts to 0.
- As we are comparing probability distributions, both sum to 1.
- The chi-square statistic is computed by summing (the squares of the differences in observed and expected values divided by the expected value).
- Checking the chi-square statistic against the table of values for degrees of freedom=N-1 returns a critical value which we can use to check against.

Accepting/Rejecting the Null Hypothesis

- Now, there are 463 songs that have been played in the morning, and 475 in the evening (not counting the 8 songs of genres played only in one timeframe).
- We take the mean number of songs (469) and compute the expected number of songs in each genre we would expect to hear based off the two probability distributions X1 and X2.
- When the evening songs are the expected distribution:
  - statistic=73.43111062635818, pvalue=0.0171101637044911
- When the morning songs are the expected distribution:
  - statistic=80.51012528371952, pvalue=0.004015387570516704
- In both cases, as $p < 0.05$, we can reject the null hypothesis with 95% confidence, hence, the difference in distributions is statistically significant.

Morning Distribution



Evening Distribution