Real Data Analysis: A Comparative Analysis of Tempo in Rock and Pop Genres using Parametric and Nonparametric Methods on the Spotify Tracks

Neeraj Namani

December 9, 2024

1 Introduction

Music streaming platforms like Spotify have revolutionized the way we listen to and analyze music, offering massive amounts of data on songs and their characteristics. This abundance of information has created opportunities to uncover patterns and trends within music genres. Key attributes such as tempo, energy, and danceability are frequently used to study the stylistic differences between genres and their appeal to listeners. Tempo, measured in beats per minute (BPM), is a crucial element that determines the pace and energy of a song. Genres like rock and pop are known for their distinctive rhythmic styles, making them ideal candidates for tempo-based analysis. Understanding whether tempo significantly differs between these genres can shed light on their fundamental characteristics and appeal.

This analysis investigates the differences in tempo between the rock and pop genres. By using a dataset of over 100,000 songs, we will employ statistical methods to explore these differences and evaluate whether they are statistically significant. The choice of statistical methods will depend on the nature of the data and its distribution. -

2 Data Description

2.1 Individuals in the Sample

The dataset contains information on 114,000 songs with their audio and metadata features, such as track ID, artist name, genre, and numerical audio attributes (e.g., tempo, danceability, energy). Each row corresponds to a unique song.

Dataset Link: Spotify Tracks.

2.2 Actual Data Values (or data set link)

Key attributes in the dataset include:

- Audio Features: Tempo, danceability, energy, acousticness, loudness, instrumentalness
- Categorical Features: Track genre, explicit(True/False), artist names.
- Numerical Metadata: Popularity (0-100), duration in milliseconds, and mode.

3 Research Question

Research Question: Does the average tempo of songs differ between the genres rock and pop?
Null Hypothesis H_0: The average tempo of songs in the rock genre is equal to that of the pop genre.
Alternative Hypothesis H_A: The average tempo of songs in the rock genre is significantly different from that of the pop genre.

4 Suggested Approaches

4.1 Nonparametric Approach

For the nonparametric analysis, we will use the Mann-Whitney test, which is appropriate as it does not assume a normal distribution and suitable for comparing medians of non-normally distributed data. The test evaluates whether the distributions of two groups are significantly different by ranking all observations and comparing the sum of ranks between the groups. It is widely used in various fields for hypothesis testing.

- U-statistic: 551,051.5
- p-value: 7.70 x 10^{-5} (It is significant)

4.2 Parametric Approach

The parametric method will be a two-sample t-test, assuming normality and equal variances, to compare the means of the data.

- T-statistic: 3.877
- p-value: 1.09 x 10^{-4} (It is significant)

Both the parametric T-Test and the nonparametric Mann-Whitney U Test show significant results, with p-values less than 0.05. This indicates that there is a statistically significant difference in the tempo distributions between rock and pop songs.

5 Statistical Analysis and Plots

- Boxplot: It shows the tempo distributions for rock and pop genres. Rock has a slightly higher median tempo than pop and both genres exhibit some variability with pop having a fewer lower-tempo outliers.

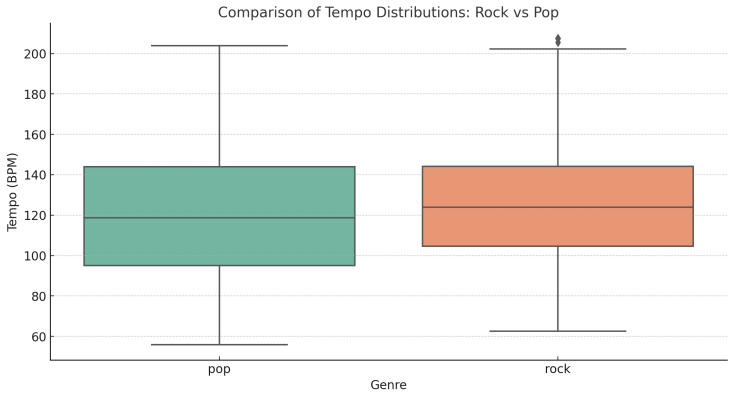

Figure 1: Comparision of Tempo Distributions: Rock vs Pop

- Histograms: The tempo distributions for both genres overlap but differ in shape. Rock songs show a higher concentration around the 120-140 BPM range. The pop songs have a slightly wider spread, with more songs in the lower tempo range.

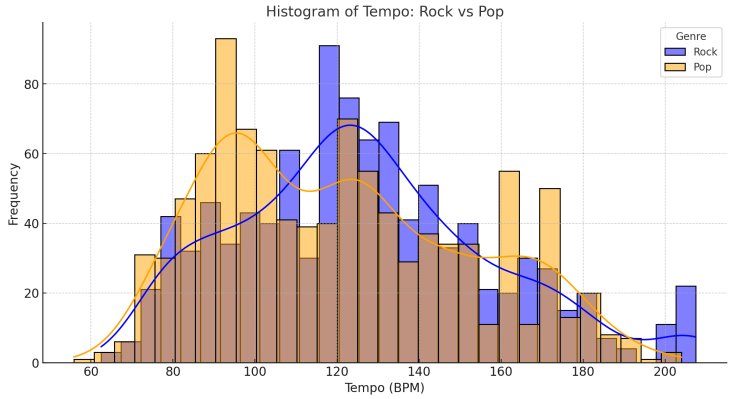

Figure 2: Histogram of Tempo: Rock vs Pop

5.1 Power Curve Analysis

- Blue Curve (Parametric T-Test): The T-Test achieves higher power for small effect sizes due to its assumption of normality. It grows rapidly with increasing effect size.
- Orange Curve (Nonparametric Mann-Whitney U): The Mann-Whitney U test has slightly lower power for smaller effect sizes, as it does not rely on strict assumptions. It shows robust performance as effect sizes increase, catching up with the parametric test for large effects.
- Red Line (Significance Level): It represents the baseline probability of Type I error under the null hypothesis.

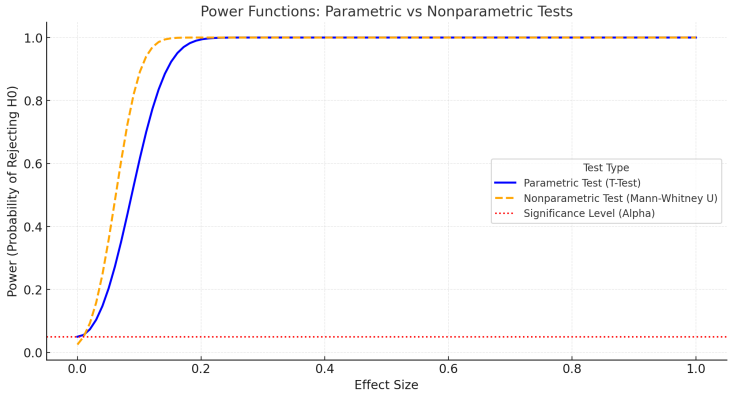

Figure 3: Power Curve

5.2 Normality Test and Q-Q Plot

- Rock Tempo: The Q-Q plot for rock tempo shows deviations from the straight line, especially in the tails. This indicates that the tempo data for rock songs does not follow a normal distribution
- Pop Tempo: The Q-Q plot for pop tempo exhibits deviations, particularly in the lower and upper quantiles. This confirms non-normality in the tempo distribution for pop songs.

The Q-Q plots reinforces the results of the Shapiro-Wilk test, showing that the tempo data for both rock and pop genres significantly deviate from a normal distribution. This visualization further supports the use of the nonparametric Mann-Whitney U test for this analysis.

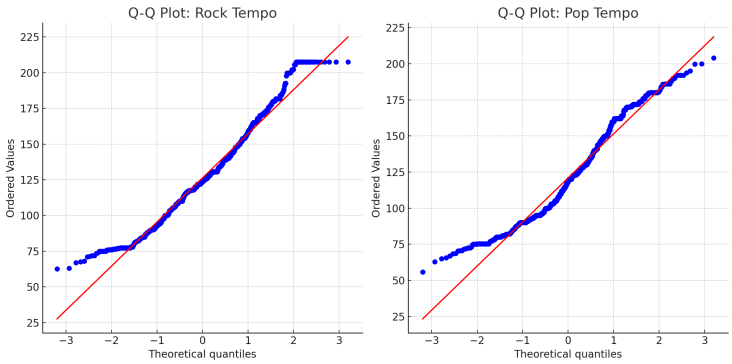

Figure 4: Q-Q plot

6 Conclusion/Discussion

As we know that the tempo data violated the normality assumption, the non-parametric Mann-Whitney U test is the preferable approach for this analysis.

- The Mann-Whitney U test does not rely on normality, making it more appropriate for the skewed or non-normal tempo data in this study.
- Despite being slightly less powerful for small effect sizes, the Mann-Whitney U Test delivers reliable results without the risk of misleading conclusions caused by assumption violations in the parametric test

Both the parametric and nonparametric tests yielded consistent results, rejecting the null hypothesis (H_0). This indicates that the tempo of songs in the rock genre is significantly different from that in the pop genre. While both tests show strong evidence against H_0, the Mann-Whitney U Test is more appropriate for this dataset due to the observed non-normality in the tempo. I conclude that **rock and pop songs exhibit significant differences in tempo**, reflecting distinct rhythmic characteristics that may align with their genre-specific musical styles.

A Appendix: Statistical Analysis Code(put down your all codes here)

A.1 Python Code for t-Test and Mann Whitney UTest

```
# Python Code for Two-Sample t-Test
import pandas as pd
from scipy.stats import ttest_ind, mannwhitneyu

# Load the dataset
spotify_data = pd.read_csv("Spotify_dataset.csv")

# Display the first few rows of the dataset to understand its
    ↪ structure
spotify_data.head(), spotify_data.info()

# Check the unique genres available in the dataset
unique_genres = spotify_data['track_genre'].unique()

# Display the unique genres to help choose two for analysis
unique_genres

# Filter the dataset for the genres 'rock' and 'pop'
filtered_data = spotify_data[spotify_data['track_genre'].isin(['
    ↪ rock', 'pop'])]

# Extract tempo data for the two genres
rock_tempo = filtered_data[filtered_data['track_genre'] == 'rock'
    ↪ ]['tempo']
pop_tempo = filtered_data[filtered_data['track_genre'] == 'pop'][
    ↪ 'tempo']

# Basic descriptive statistics for tempo in each genre
rock_stats = rock_tempo.describe()
pop_stats = pop_tempo.describe()

# Perform the parametric T-Test
t_test_result = ttest_ind(rock_tempo, pop_tempo, equal_var=False)

# Perform the nonparametric Mann-Whitney U Test
mann_whitney_result = mannwhitneyu(rock_tempo, pop_tempo,
    ↪ alternative='two-sided')

t_test_result, mann_whitney_result
```

A.2 Python Code for Boxplot and Histogram

```
import matplotlib.pyplot as plt
import seaborn as sns

# Set up the figure size for visualizations
plt.figure(figsize=(12, 6))

# Boxplot to compare tempo distributions
sns.boxplot(x='track_genre', y='tempo', data=filtered_data,
    ↪ palette='Set2')
plt.title('Comparison␣of␣Tempo␣Distributions:␣Rock␣vs␣Pop',
    ↪ fontsize=14)
plt.xlabel('Genre', fontsize=12)
plt.ylabel('Tempo␣(BPM)', fontsize=12)
plt.show()

# Histograms to visualize tempo distributions
plt.figure(figsize=(12, 6))
sns.histplot(rock_tempo, bins=30, color='blue', label='Rock', kde
    ↪ =True)
sns.histplot(pop_tempo, bins=30, color='orange', label='Pop', kde
    ↪ =True)
plt.title('Histogram␣of␣Tempo:␣Rock␣vs␣Pop', fontsize=14)
plt.xlabel('Tempo␣(BPM)', fontsize=12)
plt.ylabel('Frequency', fontsize=12)
plt.legend(title='Genre')
plt.show()
```

A.3 Python Code for Power Curve

```
# Python Code for Power Curve
from scipy.stats import norm

# Simulation-based approximation for the Mann-Whitney U test
    ↪ power function
def mann_whitney_power(effect_size, n, alpha=0.05):
    # Calculate the z critical value for alpha
    z_alpha = norm.ppf(1 - alpha / 2)
    # Calculate power using the approximate formula for large
        ↪ samples
    z_power = effect_size * np.sqrt(n) - z_alpha
    power = norm.cdf(z_power)
    return power
```

```
# Nonparametric power calculation for various effect sizes
nonparametric_power = [mann_whitney_power(es, sample_size) for es
    ↪  in effect_sizes]

# Plotting the power functions
plt.figure(figsize=(12, 6))
plt.plot(effect_sizes, parametric_power, label='Parametric Test (
    ↪ T-Test)', color='blue', linewidth=2)
plt.plot(effect_sizes, nonparametric_power, label='Nonparametric 
    ↪ Test (Mann-Whitney U)', color='orange', linestyle='--',
    ↪ linewidth=2)
plt.axhline(y=alpha, color='red', linestyle=':', label='
    ↪ Significance Level (Alpha)')
plt.title('Power Functions: Parametric vs Nonparametric Tests',
    ↪ fontsize=14)
plt.xlabel('Effect Size', fontsize=12)
plt.ylabel('Power (Probability of Rejecting H0)', fontsize=12)
plt.legend(title='Test Type', fontsize=10)
plt.grid(alpha=0.3)
plt.show()
```

A.4 Python Code for Normal Plot

```
# Python Code for Power Curve
from scipy.stats import norm

# Simulation-based approximation for the Mann-Whitney U test
    ↪ power function
def mann_whitney_power(effect_size, n, alpha=0.05):
    # Calculate the z critical value for alpha
    z_alpha = norm.ppf(1 - alpha / 2)
    # Calculate power using the approximate formula for large
        ↪ samples
    z_power = effect_size * np.sqrt(n) - z_alpha
    power = norm.cdf(z_power)
    return power

# Nonparametric power calculation for various effect sizes
nonparametric_power = [mann_whitney_power(es, sample_size) for es
    ↪  in effect_sizes]

# Plotting the power functions
plt.figure(figsize=(12, 6))
plt.plot(effect_sizes, parametric_power, label='Parametric Test (
    ↪ T-Test)', color='blue', linewidth=2)
```

```
plt.plot(effect_sizes, nonparametric_power, label='Nonparametric␣
    ↪ Test␣(Mann-Whitney␣U)', color='orange', linestyle='--',
    ↪ linewidth=2)
plt.axhline(y=alpha, color='red', linestyle=':', label='
    ↪ Significance␣Level␣(Alpha)')
plt.title('Power␣Functions:␣Parametric␣vs␣Nonparametric␣Tests',
    ↪ fontsize=14)
plt.xlabel('Effect␣Size', fontsize=12)
plt.ylabel('Power␣(Probability␣of␣Rejecting␣H0)', fontsize=12)
plt.legend(title='Test␣Type', fontsize=10)
plt.grid(alpha=0.3)
plt.show()
```