Statistical Inference

Lecturer: Abdol-Hossein Vahabie

Spring Semester 1401-1402

Writing Assignment V | Deadline 1402/03/24

1 Analyzing Factors

A music streaming platform wants to examine the factors that influence the popularity of songs on Spotify. They collected data on songs from four different genres: Pop, Hip-Hop, Rock, and Electronic. Please conduct analyses 1 to 5 based on the table provided below, which includes variables for each genre.

- **Popularity Score:** A measure of the overall popularity of a song on a scale of 1 to 100.
- **Danceability:** A score indicating how suitable a song is for dancing on a scale of 1 to 10.
- **Energy:** A measure of the intensity and activity level of a song on a scale of 1 to 10.

Genre	Popularity Score	Danceability	Energy
Pop	85, 88, 78, 80	7, 8, 6, 7	8, 7, 7, 6
Hip-Hop	75, 72, 80, 77	6, 5, 7, 6	9, 8, 7, 6
Rock	23, 11, 31, 8	5, 6, 4, 5	8, 7, 6, 5
Electronic	80, 82, 76, 75	8, 9, 7, 8	9, 9, 8, 7

1. Write the null and alternative hypotheses to test whether there is a significant difference in the mean popularity scores across the four genres.
2. Perform a one-way ANOVA to determine if there is a significant difference in the mean popularity scores across the genres. Use a significance level of 0.05, do the calculation and create the ANOVA table.
3. Is it possible to conduct post-hoc tests? For which variable? If it is possible conduct post-hoc tests (e.g., Tukey's HSD) to identify specific genre pairs that differ significantly in terms of mean for possible variables. Use a significance level of 0.05 for the post-hoc tests. If it is not possible, mention the reason.
4. Perform another ANOVA type test(type of test and variable is up to you but you should mention the reason that you think your test is appropriate).
5. Perform a multiple linear regression analysis to examine the relationship between popularity score (dependent variable) and danceability and energy (independent variables). Interpret the regression coefficients and assess the overall significance of the model. Use a significance level of 0.05.

2 Model Comparison

Suppose a Music Streaming platform is developing an automated system for curating personalized playlists of songs based on users' past preferences. The goal is to create playlists tailored to individual users by including songs highly similar to their favorite songs. This system is about to be a breakthrough in online music streaming platforms, and for the sake of publicity, the playlists must be precisely similar to what users want to hear. The system incorporates deep learning architectures trained for music genre classification to achieve this. Three models have been developed on various sets of songs in different genres. After training, they were tested on out-of-sample data, and accuracy, recall, and precision were recorded for each model in each test. As an expert, you have been asked to examine the results and determine which model is better suited for the task. The model's size is enormous, and because there are many music genres, it had to be tested on many samples so the ML Developers could only run the experiment 5 times for each model.

	Model A	**Model B**	**Model C**
Accuracy	0.85, 0.82, 0.87, 0.81, 0.88	0.80, 0.82, 0.78, 0.79, 0.81	0.88, 0.85, 0.89, 0.86, 0.90
Precision	0.95, 0.93, 0.96, 0.94, 0.97	0.85, 0.86, 0.84, 0.83, 0.86	0.90, 0.87, 0.92, 0.89, 0.93
Recall	0.76, 0.79, 0.82, 0.75, 0.81	0.72, 0.77, 0.74, 0.70, 0.75	0.85, 0.88, 0.90, 0.88, 0.87

1. Based on the given information and data, compare the model's performance, decide which model is better for this task, and defend it with a sufficient statistical explanation.

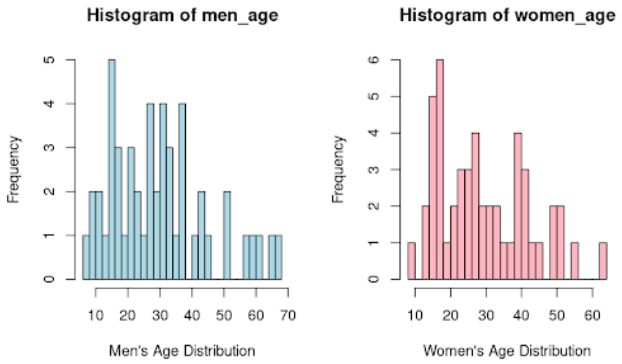
2. After you chose the winning model, the RD team made some adjustments and tested it again. The new results are given below. Is the latest model better? Why?

New Accuracy	0.88, 0.86, 0.90, 0.87, 0.89
New Precision	0.80, 0.79, 0.82, 0.80, 0.81
New Recall	0.92, 0.89, 0.91, 0.88, 0.90

3 Age Comparison of Fans

Imagine we want to analyze the age of male and female fans of a small new band. We want to test the hypothesis whether male fans are older than female fans. We collected a sample of 50 male fans and 50 female fans. Now, we want to analyze the age distribution to determine if there is a significant age difference between male and female fans.

Men's Age:	52 ,18, 27, 12 ,24 ,17, 68, 25 ,12 ,9 ,51 ,44 ,42 ,34, 44, 15, 21, 66, 61, 32, 31, 20, 6, 13, 34 ,38, 45, 17 ,16, 15, 36, 21, 29, 21, 29 , 9, 33, 15, 37, 27, 31 ,15, 57, 37, 27, 31, 38, 27, 60, 23
Women's Age:	36, 49, 20, 31, 51, 31, 15, 16, 39, 41, 52, 16, 39, 34, 18, 34 ,30, 18, 26, 18, 25, 16, 39, 49, 22, 37, 39, 21 ,16, 63, 45, 43 ,17, 28 ,29, 23, 42 ,23 ,28, 55 ,41 ,18, 23, 8 ,13 ,26, 13, 27, 28 ,18



1. Show whether the distribution is normal or not using statistical methods.

2. Can we use parametric tests in this problem? Examine all of the assumptions.

3. Can we transform the data to the normal distribution? Implement it and recheck the distribution.

4. If the parametric requirements are met, perform a parametric test and analyze the results.

5. Perform a non-parametric test on the original data and compare the results with the parametric one. Is there any difference? Compare the powers of two tests.

6. Which test is more appropriate in this situation? Why?

4 Artist Collaboration in the Spotify

Two given data-frames are the nodes and edges of the artist's collaboration graph in the Spotify platform. Nodes are the name of the artists plus some information related to their professional histories such as name, genre, number of followers, date of first and last releases, number of songs they released, and edges are the collaboration of the artists, hence the song they released together. The network is an undirected multigraph. (You are allowed to use any statistical test you want, but you should check its assumption beforehand)

1. Load the graph in R.

2. Read about PageRank's ranking algorithm and degree centrality, and run it over the graph.

3. Check out the nodes with the highest rank value; what do you think they have in common? What attributes do you think have the most significant impact on their rank? Form a hypothesis. Test the hypothesis using an appropriate statistical test (you don't need to use all the nodes in the graph. You can take a sample from the nodes)

4. Does collaboration have a relationship with the artist's popularity?

5. Is there a significant relationship between the work experience of the artist and their rank on the graph? What about the amount of activity?

6. Read about Label propagation algorithm (LBA) and cluster the graph using the algorithm.

7. Select the two biggest clusters, and analyze the results. What do you think the nodes in each cluster have in common? Form a hypothesis and test it.

8. Take a sample from the artists in three music genres: "pop", "hip hop" and "Classical" In which of these genres it's harder to become famous? (based on the number of songs, years of experience, or the number of collaborations in which genre do you get fewer followers or popularity?)

9. Are the collaborations between the artists in these genres random, or do they have some pattern? Find a way to prove this with statistics (HINT: read about Random Graphs)

Required Document

- Please upload a file in ZIP format (not RAR) to the elearn platform(`https://elearn5.ut.ac.ir/`).
- It is prohibited to use ~~handwritten material~~ and only material produced through typing in the HW template is permissible.
- To receive a score, it is necessary to submit both the PDF exercise report and R codes together in a zip file. Sending only PDF files or R codes separately will not be evaluated.
- If you submit your exercise in LaTeXformat, please include your complete LaTeXproject as a Zip file.

General Rules

- Please show all the necessary calculations and interpretations for each analysis in Q1. For parts Q1-2, it is required to solve them manually. However, for all other sections of this question, you have the freedom to use R for your analysis.
- You are not allowed to use R for the Q2, Also use only non-parametric statistical tests. Please show all the necessary calculations and interpretations for each analysis.
- You must utilize R for the analysis in Q3,Q4 and refrain from manual calculations. Ensure you show all the necessary steps and interpretations.
- Utilizing a LaTeXto compose the report will grant an additional 5 points.
- You are allowed a total grace period of 3 days to submit late assignments for all of your exercises.
- Once the grace period has ended, you can submit the assignment one week later, but your grade will be reduced by 20%.

Deadline

Thursday 23:59. 1402/03/24.

Contact Information

Please direct your questions regarding Homework 5 only to the teaching assistants, Mohammad Sepehri, and Siavash Razmi, through the course mail (statistical.inference.ut@gmail.com). Use ”HW5” as the subject line.

Good Luck

Attachment (Q4 Dataset Description)

Nodes Dataset:

Feature name	Description	
id	Id of artist in network	
isdone	Whether or not the artist is done with their career.	Boolean
spotifyid	Id of the artist on spotify	
genres	The genres of music the artist is associated with.	String
popularity	The popularity of the artist on Spotify.	Integer
followers	The number of followers the artist has on Spotify.	Integer
historgram	A histogram of the artist's popularity over time.	Array
num release	The number of releases the artist has.	Integer
first release	The date of the artist's first release.	String
last release	The date of the artist's latest release.	String
name	Name of the artist.	String
network rank	Rank of the node in the network	

Edges Dataset:

Feature name	Description
id 1	Id of the first contributing artist
id 2	Id of the second contributing artist
songid	songid on spotify
song	Name of the song