Assignment 1

Data Science Course

1 Amazon Sales Analysis

1.1 Objective

Amazon sales product dataset contains product information, including pricing, discounts, ratings, reviews, and user interactions. The following questions are designed to analyze different aspects of the dataset using statistical methods.

1.2 Statistical Questions

1. Does the discounted price significantly impact product ratings?

Hint: Use a **Spearman Rank Correlation Test** since both *discounted_price* and *rating* are ordinal or continuous variables, and their relationship might not be linear.

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \tag{1}$$

where d_i is the difference between the ranks of corresponding values, and n is the number of observations.

2. Are product categories and high/low ratings independent?

Hint: Use a **Chi-Square Test for Independence** to determine if there is a significant relationship between *category* and *rating* (e.g., defining high rating as ≥ 4 and low rating as < 4).

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \tag{2}$$

where O_{ij} is the observed frequency, and E_{ij} is the expected frequency.

3. Is there a significant difference in ratings between high-discount and low-discount products?

Hint: Use an **Independent Samples T-test**. Split the dataset into two groups: products with a *discount_percentage* above a threshold (e.g., 50%) and below it, then compare the mean ratings.

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{3}$$

where $\bar{X}_i$ is the mean rating of each group, s_i^2 is the variance, and n_i is the sample size.

4. Do more expensive products receive higher ratings on average?

Hint: Use an **ANOVA (Analysis of Variance)** test to check if *actual_price* significantly affects *rating*. Divide products into price groups (e.g., low, medium, high).

$$F = \frac{\text{Between-group variance}}{\text{Within-group variance}} \tag{4}$$

5. Does the distribution of rating counts follow a normal distribution?

Hint: Use a **Kolmogorov-Smirnov Test** (or Shapiro-Wilk Test) to check if the *rating_count* variable follows a normal distribution. This helps determine whether parametric tests can be applied.

For the Kolmogorov-Smirnov test:

$$D_n = \sup_x |F_n(x) - F(x)| \tag{5}$$

where $F_n(x)$ is the empirical cumulative distribution function (ECDF) of the sample, and $F(x)$ is the cumulative distribution function (CDF) of the normal distribution.

2 Customer Personality Analysis

Objective

The goal of this exercise is to analyze a customer personality dataset containing demographic information, spending habits, promotional responses, and purchasing behaviors. You will conduct hypothesis tests to uncover insights about customer behaviors and characteristics.

2.1 Data Exploration and Visualization

Before performing hypothesis tests, it's essential to understand the dataset's structure and the relationships between variables. Begin by exploring the data through appropriate visualizations. This process will help you identify patterns, detect outliers, and formulate meaningful hypotheses.

Recommended Steps

a. **Data Inspection:** Examine the dataset to understand the types of variables (e.g., categorical, numerical) and their distributions.

b. **Visualization:** Utilize various plots to explore relationships between variables. For guidance on selecting suitable visualizations, refer to From Data to Viz, a resource that assists in choosing the appropriate chart types based on your data.

2.2 Hypothesis Testing

After familiarizing yourself with the data, conduct the following hypothesis tests:

1. Do customers from different education levels have different income levels?

To test whether customers with different education backgrounds earn significantly different incomes, use the **Kruskal-Wallis H test**. Since income data may not be normally distributed, this non-parametric test is appropriate for comparing median income levels across multiple education categories.

2. Does the marketing campaign influence spending behavior across customer groups?

Compare the spending behavior of customers who received a marketing campaign versus those who did not. First check the distribution of data and use the **Mann-Whitney U Test** if the data is not normally distributed, or the **Independent Samples T-test** if normality holds.

3. Do customers with children spend differently than those without children?

To determine whether households with children allocate their spending differently compared to those without, use the **Mann-Whitney U Test**. This test compares the spending distributions of two independent groups when the assumption of normality is uncertain.

4. Is there a significant difference in spending on different product categories?

Use the **Friedman Test** to compare customer spending across multiple product categories . This test will help determine whether customers have a preference for spending more on certain product categories over others.

5. Is there a relationship between customer education level and acceptance of promotional campaigns?

Use the **Chi-Square Test for Independence** to determine whether education level is associated with customers accepting promotional campaigns. This test will analyze whether different education groups have significantly different response rates to campaign offers.

2.3 Further Exploration and Data Storytelling

To deepen your analysis:

Additional Tests and Visualizations

- Identify other variables of interest and design new tests to explore potential relationships.
- Create visualizations that effectively communicate your findings.

Data Storytelling

- Develop a narrative that connects your analytical findings to actionable business insights.
- Highlight how your results can inform marketing strategies, product development, or customer engagement initiatives.

3 Submission Guidelines

- **Code Submission:** Provide two .ipynb code files that includes all data exploration, visualization, and analysis steps. These codes must include all your plots and statistical tests results.
- **Report Submission:** Submit two written report (in PDF format) summarizing your methodology, analyses, findings, and recommendations. one for each dataset.
- **Visualization Quality:** Ensure that all plots are clear, properly labeled, and effectively convey the intended information.