Statistical Analysis of Video Game Ratings and Sales
Intermediate Report

Team Members

- Stadnik Oleksandr
- Hutsuliak Oleh
- Zaiats Yulian

1. Project Goal

Our project aims to investigate patterns and factors that influence video game variables using a dataset based on real-world games to test hypotheses about their market performance and popularity trends.

2. Dataset

Our team uses the Video Game Sales dataset (from Kaggle), which contains sales data for over 16,000 video games which have sold over 100,000 copies. Variables for this dataset include game name, platform, release year, genre, publisher, and sales figures across different continents, which allow for a high level of analysis.

3. Preliminary Findings

- Highly Skewed Sales: most games sell under 200,000 units, while a few blockbusters like Wii Sports sell over 80 million.
- Nintendo Dominance: all top 15 best-selling games are Nintendo titles.
- Regional Trends: NA and EU sales are strongly correlated, with NA generally leading in sales.

- Distribution Shape: global sales data distribution is not normal, with high skewness and kurtosis, making non-parametric tests appropriate.

4. Planned Hypotheses

1. Correlation Hypothesis: NA and EU sales are positively correlated – popular games in NA tend to also perform well in EU.

H_0: NA sales do not predict EU sales; no linear relationship.

H_1: higher NA sales predict higher EU sales; positive linear relationship.

Test: linear regression.

2. Platform and Genre are statistically associated: independence between categorical variables.

H_0: platform and Genre are independent.

H_1: platform and Genre are dependent.

Test: chi-squared.

3. The Skewness of Sales Distribution Differs by Platform: some platforms have more “blockbuster-heavy” or “flat” sales profiles.

H_0: all platforms have equal skewness.

H_1: at least one platform differs in skewness.

Test: graphs comparison.

5. Conclusion

Our team will analyze our hypotheses and how different game features can affect game’s performance on the market. We will use linear regression where appropriate. Further analysis may explore genre or publisher effects if time allows.