Statistical Analysis of Player Behaviour in Online Poker

Roman Prokhorov, Maksym Holovin, Nazar Pasichnyk

1. Research Topic

We aim to analyze patterns in online poker player behavior using real transaction and gameplay data. Our study will examine relationships between player demographics, betting patterns, and financial outcomes to understand different player types and their gambling strategies.

2. Data Source

Source: The Transparency Project (www.thetransparencyproject.org)

Provider: Division on Addiction, Cambridge Health Alliance, Harvard Medical School

Dataset: "Second Session at the Virtual Poker Table"

Sample Size: 5,028 online poker players who registered in February 2015

Reference: Tom, M. A., et al. (2022). Second Session at the Virtual Poker Table: A Contemporary Study of Actual Online Poker Activity. *Journal of Gambling Studies*. https://doi.org/10.1007/s10899-022-10147-1

3. Data Description

The dataset consists of five interconnected CSV files:

- **Demographics** (5,028 players): Age, Gender, Country of residence (encoded using ISO 3166-1 country codes)
- **Cash Games** (51,763 daily records): Stakes wagered, winnings, number of sessions
- **Tournaments** (82,831 daily records): Entry fees, prizes, number of tournaments
- **Deposits** (295,119 transactions): Amount, payment method, transaction status
- **Withdrawals** (32,307 transactions): Amount, payment method, transaction status

Key variables include player demographics, daily gambling activity (stakes and winnings), and financial transactions (deposits and withdrawals in Euros). Country information is mapped from numeric ISO 3166-1 codes to country names for analysis.

Sample Data Structure

Table 1: Demographics Sample (with Country Names)

User ID	Age	Gender	Country
11	32	M	Germany
14	39	M	Germany
22	21	M	Belgium
37	25	M	Germany
38	60	F	Germany

Table 2: Deposits Sample

User ID	Amount (€)	Status	Payment Method
11	10.00000	S	PayPal
14	48.74999	S	MAESTRO
167	20.07214	S	VISA
70	61.09888	F	MASTERCARD
208	102.00000	S	VISA

Table 3: Cash Games Sample

User ID	Date	Sessions	Stakes (€)	Winnings (€)
11	2015-02-09	1	0.69	0.73
14	2015-02-01	6	111.78	68.73
14	2015-02-02	3	114.11	71.13
14	2015-02-03	2	15.00	4.38
37	2015-02-01	4	18.15	8.95

Table 4: Top 10 Countries by Player Count

Country	Number of Players
Germany	1416
France	1252
Spain	336
Belgium	300
United Kingdom of Great Britain and Northern Ireland	231
Austria	216
Czechia	190
Switzerland	185
Hungary	184
Netherlands, Kingdom of the	140

4. Preliminary Hypotheses (Revised β_0 Explanation)

Hypothesis 1 (Two-Sample T-Test)

Research Question: Is there a significant difference in the mean total stakes wagered between players who bet frequently (High-Frequency) and players who bet infrequently (Low-Frequency)?

We define two groups based on the median number of bets:

- **Group 1 (Low-Frequency Bettors):** Players with a number of bets below the median (μ_{low})
- **Group 2 (High-Frequency Bettors):** Players with a number of bets at or above the median (μ_{high})

The T-test compares the mean total stakes (Y) between these two population groups.

- H_0 **(Null Hypothesis):** $\mu_{high} = \mu_{low}$. The mean total stakes is the same for both groups.
- H_a **(Alternative Hypothesis):** $\mu_{high} > \mu_{low}$. High-Frequency bettors have a significantly higher mean total stakes than Low-Frequency bettors.

Hypothesis 2 (Multiple Linear Regression)

Research Question: Do demographic characteristics (Age and Gender) predict the total amount a player wagers?

We define Y as the Total Stakes Wagered, X_1 as Age, and X_2 as Gender (encoded as 0 for Female, 1 for Male). The model is:

$$Y = \beta_0 + \beta_1(Age) + \beta_2(Gender) + \epsilon$$

Where:

- β_0 **(Y-Intercept):** Mathematically, this is the predicted total stakes (€) when all predictor variables are zero (Age=0, Gender=0). Since a 0-year-old cannot gamble, this term provides no practical interpretation and should be understood only as the necessary constant to align the fitted model with the observed data.
- β_1 **(Age Coefficient):** The change in total stakes for every one-year increase in age, holding gender constant.
- β_2 **(Gender Coefficient):** The difference in total stakes between male players and the baseline female players (when controlling for age).
- ϵ **(Error Term):** The unexplained variation in total stakes not accounted for by Age and Gender.
- H_0 **(Null Hypothesis):** $\beta_1 = 0$ and $\beta_2 = 0$. Age and Gender have no linear relationship with total stakes wagered.
- H_a **(Alternative Hypothesis):** $\beta_1 < 0$ and $\beta_2 > 0$.
  - We hypothesize that β_1 will be negative (younger players bet more).
  - We hypothesize that β_2 will be positive (male players bet more than female players).

Why We Include β_0 and ϵ in the Multiple Regression Model (Hypothesis 2)

Understanding the roles of the constant (β_0) and the error term (ϵ) is crucial for interpreting the multiple regression model (Hypothesis 2):

1. The Role of β_0 (The Intercept) Even when the literal interpretation of β_0 is meaningless (like predicting the stakes of a 0-year-old), the term is crucial for the mathematical integrity of the model.

- **It Anchors the Line:** β_0 allows the regression line (or plane) to be shifted up or down so that it passes through the mean of the data. Without it, the line would be forced to start at the origin (0, 0), which would almost certainly result in a terrible fit for the actual data points.
- **It Isolates the Effects of the Slopes:** In a multiple regression, the goal is to determine the effect of Age (β_1) and Gender (β_2) independently. β_0 absorbs the overall average stakes and ensures that the coefficients (β_1 and β_2) truly represent the change associated with their respective variables, rather than compensating for a general baseline value.

2. The Role of ϵ (The Error Term) The error term (ϵ) is equally crucial because it acknowledges that our model is an imperfect representation of reality.

- **It Captures Missing Information:** ϵ represents all the variation in total stakes that is not explained by our chosen predictor variables (Age and Gender). This includes factors like individual luck, specific poker strategy, time spent online, or the player's emotional state.
- **It Satisfies Statistical Assumptions:** In linear regression, we must assume this ϵ (the difference between the actual stakes and the predicted stakes) is purely random noise and not systematically related to our predictor variables. This assumption is fundamental for calculating reliable p-values and confidence intervals, which we use to determine if the effects of our β coefficients are statistically significant. ## 5. Initial Insights

Descriptive Statistics

Table 5: Demographics Summary

N Players	Mean Age	SD Age	% Male
5028	28.89	8.84	90.65

Table 6: Financial Activity Overview

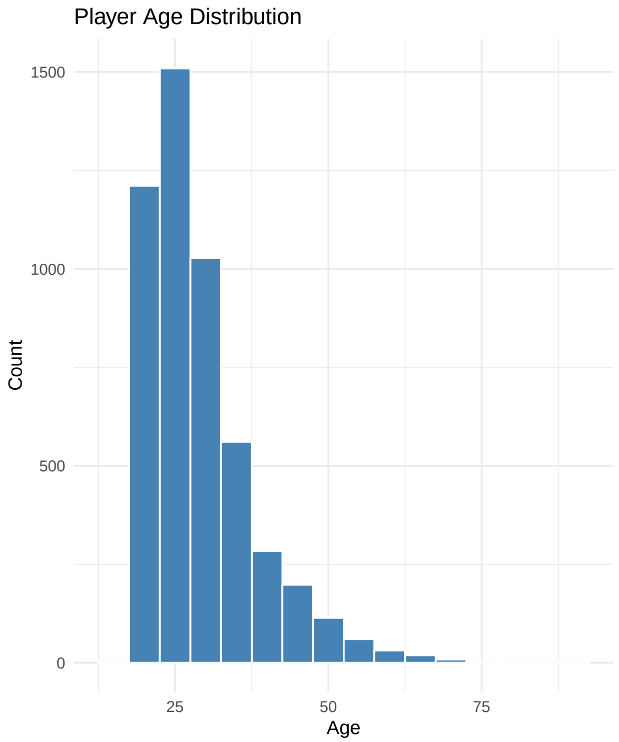
Metric	Value
Total Deposits	295088
Successful Deposits	223601
Total Withdrawals	32307
Successful Withdrawals	17182
Players with Withdrawals	1739

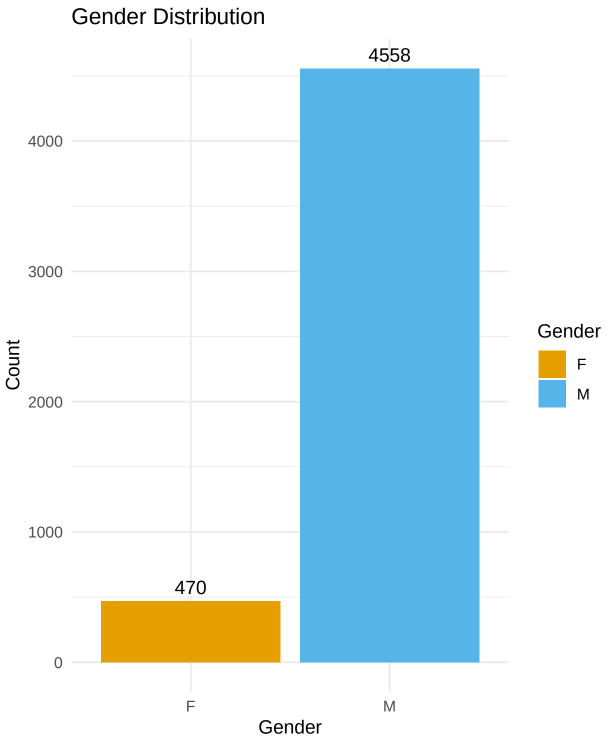
Key Observations

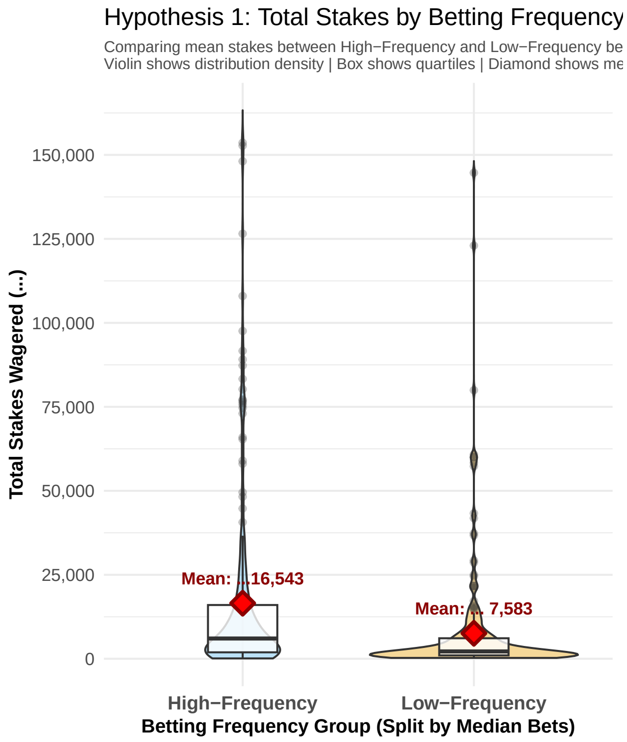
Preliminary exploration reveals several important patterns based on the dataset of 5,028 players:

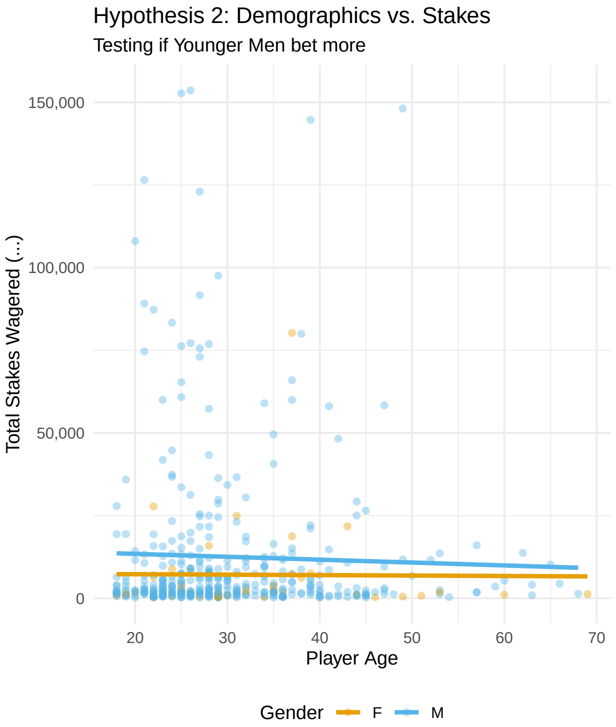
- Only **1739** of **5028** players (**34.6%**) made successful withdrawals, suggesting many players experience net losses.
- **4232** players engaged in cash games, which tracks daily stakes and winnings.
- The dataset shows high variability in betting amounts and player engagement levels.
- Rich temporal structure (daily records from February 2015) allows for analysis of betting patterns over time.

Exploratory Visualizations


Player Age Distribution


Gender Distribution


Hypothesis 1: Total Stakes by Betting Frequency
Comparing mean stakes between High−Frequency and Low−Frequency be
Violin shows distribution density | Box shows quartiles | Diamond shows me


Hypothesis 2: Demographics vs. Stakes
Testing if Younger Men bet more

Interpretation of Plot 3 (Hypothesis 1):

Table 7: Summary Statistics: Total Stakes by Betting Frequency

Group	Mean Stakes (€)	Median Stakes (€)	SD Stakes (€)	N Players
High-Frequency	16,543	6,031	27,795	218
Low-Frequency	7,583	2,190	17,023	220

The visualization reveals several key patterns supporting our hypothesis:

- **Distribution Shape:** The violin plots show the probability density of stakes for each group. High-Frequency bettors have a wider distribution, indicating greater variability in betting amounts.
- **Central Tendency:** The red diamond marks the mean total stakes for each group. Visual inspection suggests High-Frequency bettors wager substantially more on average than Low-Frequency bettors.
- **Spread:** The box plots display quartiles (bottom, middle, and top of the box represent 25th, 50th, and 75th percentiles). The white boxes show that High-Frequency bettors have both higher medians and greater spread in their betting behavior.
- **Outliers:** Individual points outside the "whiskers" represent outliers. Both groups contain some extreme bettors, but High-Frequency bettors show more variability overall.

This visual pattern strongly suggests that betting frequency and total stakes are related, motivating our statistical test to determine if the difference in means is statistically significant.