TennisIndex

title: “Bayesian Tennis Module – Rank vs. Outcome” description: Modeling tennis outcomes with Bayesian statistics categories: - Bayesian Inference

# Module Overview

This SCORE module investigates the relationship between ATP player rankings and match outcomes using Bayesian statistical modeling. Students will explore whether lower-ranked tennis players are likely to upset higher-ranked players, and how rank difference impacts match results. They’ll also practice building and interpreting prior and posterior distributions using real match data from the ATP Tour.

The module is split into two connected parts: - First, students explore a logistic regression model where the outcome is whether the lower-ranked player wins, and the predictor is the rank difference. - Then, they investigate how prior beliefs about serving performance (e.g., Rafael Nadal’s performance against Novak Djokovic) can be formalized and updated using Beta-Binomial Bayesian models.

# Learning Goals

By the end of this module, students will be able to:

* Use dplyr and ggplot2 to manipulate and visualize match-level sports data
* Fit a logistic model with rank difference as a predictor
* Use likelihood surfaces and grid approximation to estimate a posterior
* Create and interpret prior and posterior Beta distributions
* Make plots that compare prior beliefs to observed outcomes
* Build intuition for how prior strength and data interact in Bayesian inference

# Dataset

This module uses ATP match data from 2010–2022, pulled from [Kaggle](https://www.kaggle.com/datasets/edoardominora/atp-tennis-matches-dataset) and filtered to include key variables for Bayesian modeling.

The full dataset includes: - atp\_matches\_till\_2022.csv: Match-level data including winners, losers, rankings, and dates - tennis\_model\_data.csv: Cleaned data used for modeling whether lower-ranked players win - nadal\_djokovic\_2020\_french\_open.csv: Match statistics for Nadal vs. Djokovic in the 2020 French Open final

# Variable Descriptions

### From tennis\_model\_data.csv

| Variable | Description |
| --- | --- |
| winner\_name | Name of the match winner |
| loser\_name | Name of the match loser |
| winner\_rank | ATP ranking of the winner at time of match |
| loser\_rank | ATP ranking of the loser at time of match |
| rank\_diff | Difference between ranks (higher - lower) |
| lower\_rank\_wins | Binary outcome: 1 = lower-ranked player won, 0 = lost |
| lower\_rank\_player | Name of the lower-ranked player in each match |
| surface | Type of court surface (Hard, Clay, Grass) |
| match\_year | Year the match occurred |

### From nadal\_djokovic\_2020\_french\_open.csv

| Variable | Description |
| --- | --- |
| w\_svpt | Number of service points played by Nadal |
| w\_1stWon | Points Nadal won on first serve |
| w\_2ndWon | Points Nadal won on second serve |
| w\_1stIn | First serves in |
| loser\_name | Djokovic |
| winner\_name | Nadal |
| tourney\_name | Tournament name (French Open) |
| tourney\_date | Match date in YYYYMMDD format |

# Materials

* worksheet.qmd: Student worksheet with guided prompts and embedded code chunks
* worksheet\_key.qmd: Instructor key with full code and explanations
* tennis\_model\_data.csv: Cleaned match dataset for Bayesian logistic modeling
* nadal\_djokovic\_2020\_french\_open.csv: Dataset used for prior/posterior Beta updates
* index.qmd: This module summary

# Conclusion

By completing this module, students will have developed practical skills in Bayesian modeling using real-world sports data. The exercise shows how ranking systems translate into win probabilities, and how prior knowledge can be formalized and updated with new data. Students will leave with a deeper appreciation of how data science supports understanding and strategy in professional sports.