TennisWorksheet

# Introduction

This worksheet will guide you through a full Bayesian data analysis using real ATP match data from 2010 to 2022. We’ll start by building a likelihood model for the probability that a lower-ranked tennis player beats a higher-ranked one. Then, we’ll explore how prior beliefs can affect our conclusions — visually comparing **prior** and **posterior** distributions.

Your final goal is to answer this:

**How do rank differences influence tennis match outcomes, and how do different prior beliefs impact the Bayesian analysis of player success?**

# Part 1 – Likelihood: Modeling Match Outcomes with Rank Differences

We’ll begin by creating a cleaned dataset of matches, then modeling the probability that a lower-ranked player wins using a grid-based Bayesian likelihood approach.

library(readr)

Warning: package 'readr' was built under R version 4.4.3

library(dplyr)

Warning: package 'dplyr' was built under R version 4.4.3

Attaching package: 'dplyr'

The following objects are masked from 'package:stats':  
  
 filter, lag

The following objects are masked from 'package:base':  
  
 intersect, setdiff, setequal, union

library(ggplot2)

Warning: package 'ggplot2' was built under R version 4.4.3

library(lubridate)

Warning: package 'lubridate' was built under R version 4.4.3

Attaching package: 'lubridate'

The following objects are masked from 'package:base':  
  
 date, intersect, setdiff, union

library(purrr)

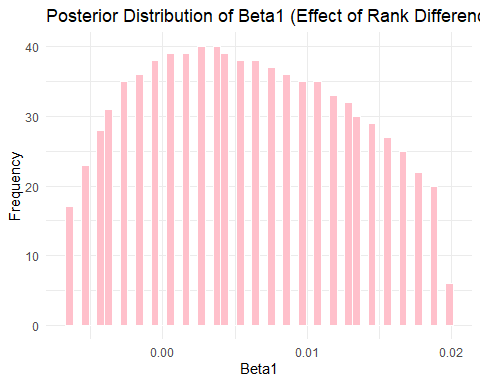
Warning: package 'purrr' was built under R version 4.4.3

# Load and clean matches  
matches <- read\_csv("../Data/atp\_matches\_till\_2022.csv")

Rows: 188161 Columns: 49

── Column specification ────────────────────────────────────────────────────────  
Delimiter: ","  
chr (14): tourney\_id, tourney\_name, surface, tourney\_level, winner\_entry, wi...  
dbl (35): draw\_size, tourney\_date, match\_num, winner\_id, winner\_seed, winner...  
  
ℹ Use `spec()` to retrieve the full column specification for this data.  
ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

cleaned <- matches |>  
 mutate(tourney\_date = ymd(tourney\_date)) |>  
 filter(year(tourney\_date) >= 2010 & year(tourney\_date) <= 2022) |>  
 filter(!is.na(winner\_rank) & !is.na(loser\_rank)) |>  
 mutate(  
 winner\_rank = as.integer(winner\_rank),  
 loser\_rank = as.integer(loser\_rank),  
 lower\_rank\_won = if\_else(winner\_rank > loser\_rank, 1, 0),  
 lower\_rank\_player = if\_else(winner\_rank > loser\_rank, winner\_name, loser\_name),  
 surface = factor(surface),  
 match\_year = year(tourney\_date)  
 ) |>  
 select(tourney\_date, match\_year, surface, winner\_name, loser\_name,  
 winner\_rank, loser\_rank, lower\_rank\_player, lower\_rank\_won)  
  
# Add variables for modeling  
model\_data <- cleaned |>  
 mutate(  
 lower\_rank = pmin(winner\_rank, loser\_rank),  
 higher\_rank = pmax(winner\_rank, loser\_rank),  
 lower\_rank\_wins = ifelse(winner\_rank > loser\_rank, 0, 1),  
 rank\_diff = higher\_rank - lower\_rank  
 ) |>  
 filter(winner\_rank != loser\_rank)  
  
# Sample subset for modeling  
set.seed(123)  
model\_sample <- model\_data |> slice\_sample(n = 5000)  
  
# Create grid of beta values  
beta0\_vals <- seq(-3, 3, length.out = 100)  
beta1\_vals <- seq(-0.05, 0.05, length.out = 100)  
param\_grid <- expand.grid(beta0 = beta0\_vals, beta1 = beta1\_vals)  
  
# Log-likelihood function  
log\_likelihood <- function(beta0, beta1, x, y) {  
 eta <- beta0 + beta1 \* x  
 p <- 1 / (1 + exp(-eta))  
 if (any(p == 0 | p == 1)) return(-Inf)  
 sum(dbinom(y, size = 1, prob = p, log = TRUE))  
}  
  
# Compute likelihoods  
loglik\_vals <- map2\_dbl(  
 param\_grid$beta0,  
 param\_grid$beta1,  
 ~ log\_likelihood(.x, .y, model\_sample$rank\_diff, model\_sample$lower\_rank\_wins)  
)  
  
# Posterior via grid  
loglik\_max <- max(loglik\_vals)  
posterior <- exp(loglik\_vals - loglik\_max)  
posterior <- posterior / sum(posterior)  
param\_grid$posterior <- posterior  
  
# Sample from posterior  
set.seed(321)  
posterior\_samples <- param\_grid |>  
 filter(posterior > 0) |>  
 slice\_sample(n = 5000, weight\_by = posterior)  
  
# Plot posterior of beta1  
ggplot(posterior\_samples, aes(x = beta1)) +  
 geom\_histogram(bins = 50, fill = "pink", color = "white") +  
 labs(title = "Posterior Distribution of Beta1 (Effect of Rank Difference)",  
 x = "Beta1", y = "Frequency") +  
 theme\_minimal()



### Questions:

1. What does a negative beta1 value suggest about rank difference and match outcomes?
2. What is the average value of beta1 in your posterior?
3. How wide is the 90% credible interval?

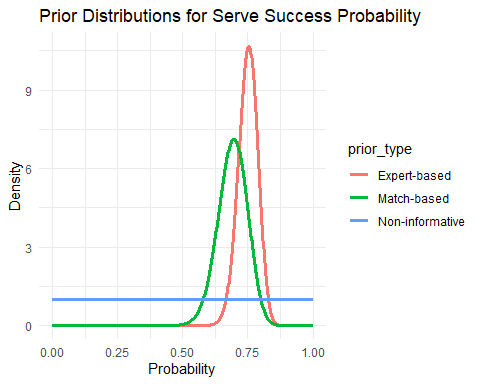
# Part 2 – Introduction to Prior Distributions

Let’s now explore what different **prior beliefs** might look like for a player’s success rate. Below are three types of priors we’ll construct for a serve win probability:

* **Non-informative**: completely flat (Beta(1, 1))
* **Match-Based**: Nadal won 46 of 66 points in a previous match (Beta(47, 21))
* **Expert**: A sports announcer believes Nadal wins 75% of points, with <2% chance he wins less than 70%

x\_vals <- seq(0, 1, length.out = 500)  
prior1 <- dbeta(x\_vals, 1, 1)  
prior2 <- dbeta(x\_vals, 47, 21)  
  
alphas <- seq(0.01, 100, length.out = 2000)  
betas <- alphas \* (1 - 0.75) / 0.75  
prob\_70 <- pbeta(0.70, alphas, betas)  
  
expert\_params <- tibble(alphas, betas, prob\_70) |>  
 mutate(error = abs(prob\_70 - 0.02)) |>  
 slice\_min(error, n = 1)  
  
prior3 <- dbeta(x\_vals, expert\_params$alphas, expert\_params$betas)  
  
priors\_df <- tibble(  
 x = rep(x\_vals, 3),  
 density = c(prior1, prior2, prior3),  
 prior\_type = rep(c("Non-informative", "Match-based", "Expert-based"), each = length(x\_vals))  
)  
  
ggplot(priors\_df, aes(x = x, y = density, color = prior\_type)) +  
 geom\_line(size = 1.2) +  
 labs(title = "Prior Distributions for Serve Success Probability",  
 x = "Probability", y = "Density") +  
 theme\_minimal()

Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.  
ℹ Please use `linewidth` instead.



### Questions:

1. Which prior has the sharpest peak? What does this indicate?
2. How do the match-based and expert-based priors differ?
3. Which prior reflects the most uncertainty?

# Part 3 – Posterior Distributions with Match Data

We’ll now combine our priors with real data to see how beliefs update. From your cleaned match data, we estimated Nadal’s service point performance as follows:

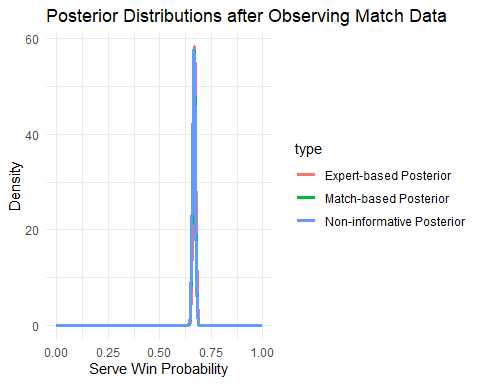
nadal\_data <- read\_csv("../Data/nadal\_djokovic\_2020\_french\_open.csv")

Rows: 59 Columns: 49  
── Column specification ────────────────────────────────────────────────────────  
Delimiter: ","  
chr (12): tourney\_id, tourney\_name, surface, tourney\_level, winner\_name, win...  
dbl (35): draw\_size, tourney\_date, match\_num, winner\_id, winner\_seed, winner...  
lgl (2): winner\_entry, loser\_entry  
  
ℹ Use `spec()` to retrieve the full column specification for this data.  
ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

nadal\_summary <- nadal\_data |> summarise(  
 total\_pts = sum(w\_svpt, na.rm = TRUE),  
 pts\_won = sum(w\_1stWon + w\_2ndWon, na.rm = TRUE)  
)  
  
n\_total <- nadal\_summary$total\_pts  
n\_won <- nadal\_summary$pts\_won

Now let’s update all three priors using a Beta posterior:

posterior\_noninf <- c(1 + n\_won, 1 + n\_total - n\_won)  
posterior\_match <- c(47 + n\_won, 21 + n\_total - n\_won)  
posterior\_expert <- c(expert\_params$alphas + n\_won, expert\_params$betas + n\_total - n\_won)  
  
p\_vals <- seq(0, 1, length.out = 500)  
posterior\_df <- tibble(  
 p = rep(p\_vals, 3),  
 density = c(dbeta(p\_vals, posterior\_noninf[1], posterior\_noninf[2]),  
 dbeta(p\_vals, posterior\_match[1], posterior\_match[2]),  
 dbeta(p\_vals, posterior\_expert[1], posterior\_expert[2])),  
 type = rep(c("Non-informative Posterior", "Match-based Posterior", "Expert-based Posterior"), each = length(p\_vals))  
)  
  
ggplot(posterior\_df, aes(x = p, y = density, color = type)) +  
 geom\_line(size = 1.2) +  
 labs(title = "Posterior Distributions after Observing Match Data",  
 x = "Serve Win Probability", y = "Density") +  
 theme\_minimal()



### Questions:

1. Which posterior is most concentrated (i.e., has least uncertainty)?
2. Which prior had the strongest influence on its posterior?
3. What is the estimated probability Nadal wins a point on serve?

# Wrap-Up

In this worksheet, you:

* Modeled match outcomes using rank differences
* Constructed and visualized different priors
* Updated each prior using Bayesian inference with real match data

**Final Reflection:**  
How did different prior beliefs affect your conclusions about Nadal’s serve performance? Which prior-posterior pair do you trust most, and why?