TennisWorksheetKey

# Introduction

This is the answer key for the module where students explored how Bayesian analysis can be used to evaluate tennis match outcomes, specifically focusing on player ranking and serve success probability. The exercises build from logistic modeling to comparing prior and posterior beliefs about Nadal’s serve performance against Djokovic.

# Part 1: Likelihood Modeling with Rank Difference

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

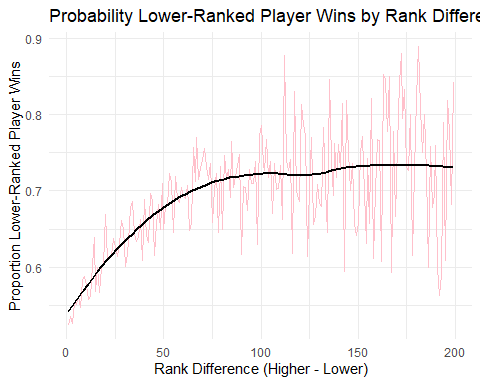
# Read the cleaned dataset  
model\_data <- read\_csv("../Data/tennis\_model\_data.csv")

Rows: 35778 Columns: 13

── Column specification ────────────────────────────────────────────────────────  
Delimiter: ","  
chr (4): surface, winner\_name, loser\_name, lower\_rank\_player  
dbl (8): match\_year, winner\_rank, loser\_rank, lower\_rank\_won, lower\_rank, h...  
date (1): tourney\_date  
  
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.

# Exploratory plot: probability of lower-ranked win vs. rank difference  
model\_data |>   
 group\_by(rank\_diff) |>   
 summarise(prop\_lower\_win = mean(lower\_rank\_wins), n = n()) |>   
 filter(n > 10, rank\_diff < 200) |>   
 ggplot(aes(x = rank\_diff, y = prop\_lower\_win)) +  
 geom\_line(color = "pink") +  
 geom\_smooth(method = "loess", se = FALSE, color = "black") +  
 labs(  
 title = "Probability Lower-Ranked Player Wins by Rank Difference",  
 x = "Rank Difference (Higher - Lower)",  
 y = "Proportion Lower-Ranked Player Wins"  
 ) +  
 theme\_minimal()

`geom\_smooth()` using formula = 'y ~ x'



**Q: What do you notice about the trend as rank difference increases?**  
A: As the rank difference increases, the lower-ranked player’s probability of winning decreases. This makes sense since the skill gap is wider.

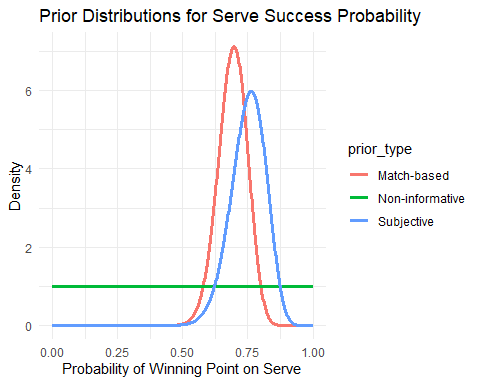
# Part 2: Plotting Prior Beliefs

library(tibble)

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

# Define priors  
x\_vals <- seq(0, 1, length.out = 500)  
  
prior1 <- dbeta(x\_vals, 1, 1) # Non-informative  
prior2 <- dbeta(x\_vals, 47, 21) # Match-based  
alpha3 <- 31  
beta3 <- 10.3  
prior3 <- dbeta(x\_vals, alpha3, beta3) # Expert-based  
  
priors\_df <- tibble(  
 x = rep(x\_vals, 3),  
 density = c(prior1, prior2, prior3),  
 prior\_type = rep(c("Non-informative", "Match-based", "Subjective"), 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 of Winning Point on Serve", y = "Density") +  
 theme\_minimal()

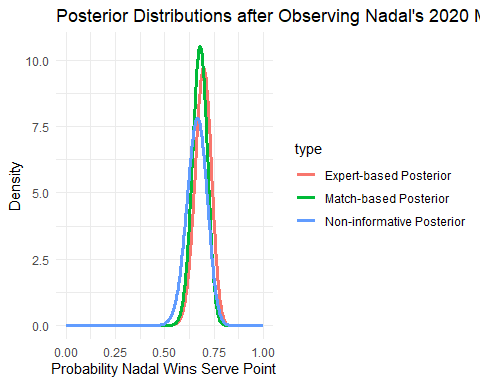
Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.  
ℹ Please use `linewidth` instead.



**Q: Which prior is most skeptical of extreme values?**  
A: The non-informative prior (Uniform) places equal density across all values and doesn’t favor any range.

# Part 3: Posterior Distributions for Nadal vs. Djokovic

# Observed data  
n\_won <- 56  
n\_total <- 84  
  
# Posterior distributions  
posterior\_noninf <- dbeta(x\_vals, 1 + n\_won, 1 + n\_total - n\_won)  
posterior\_match <- dbeta(x\_vals, 47 + n\_won, 21 + n\_total - n\_won)  
posterior\_expert <- dbeta(x\_vals, alpha3 + n\_won, beta3 + n\_total - n\_won)  
  
posteriors\_df <- tibble(  
 x = rep(x\_vals, 3),  
 density = c(posterior\_noninf, posterior\_match, posterior\_expert),  
 type = rep(c("Non-informative Posterior", "Match-based Posterior", "Expert-based Posterior"), each = length(x\_vals))  
)  
  
ggplot(posteriors\_df, aes(x = x, y = density, color = type)) +  
 geom\_line(size = 1.2) +  
 labs(title = "Posterior Distributions after Observing Nadal's 2020 Match",  
 x = "Probability Nadal Wins Serve Point", y = "Density") +  
 theme\_minimal()



**Q: Do the posteriors differ substantially?**  
A: Slightly — all are centered around ~0.665–0.668, but the expert prior results in a more confident (narrower) posterior.

# Posterior Summary Table

summary\_df <- tibble(  
 type = c("Non-informative", "Match-based", "Expert-based"),  
 mean = c(  
 (1 + n\_won) / (1 + n\_won + 1 + n\_total - n\_won),  
 (47 + n\_won) / (47 + n\_won + 21 + n\_total - n\_won),  
 (alpha3 + n\_won) / (alpha3 + n\_won + beta3 + n\_total - n\_won)  
 ),  
 lower\_90 = c(  
 qbeta(0.05, 1 + n\_won, 1 + n\_total - n\_won),  
 qbeta(0.05, 47 + n\_won, 21 + n\_total - n\_won),  
 qbeta(0.05, alpha3 + n\_won, beta3 + n\_total - n\_won)  
 ),  
 upper\_90 = c(  
 qbeta(0.95, 1 + n\_won, 1 + n\_total - n\_won),  
 qbeta(0.95, 47 + n\_won, 21 + n\_total - n\_won),  
 qbeta(0.95, alpha3 + n\_won, beta3 + n\_total - n\_won)  
 )  
)  
  
summary\_df

# A tibble: 3 × 4  
 type mean lower\_90 upper\_90  
 <chr> <dbl> <dbl> <dbl>  
1 Non-informative 0.663 0.577 0.744  
2 Match-based 0.678 0.614 0.738  
3 Expert-based 0.694 0.625 0.760

**Q: Which posterior has the narrowest interval? Why?**  
A: The expert-based posterior, because its prior was more concentrated (higher precision), making the update more confident.

# Final Reflection

* All priors update toward the observed proportion (56/84 ≈ 0.667).
* The choice of prior affects posterior spread but not the central tendency much here.
* The more informative the prior, the more confident the posterior (narrower intervals).
* Bayesian analysis gives us a full distribution of belief, not just a point estimate — that’s the key benefit here!