Bayesian Prediction for the 2019 FIFA Women's World Cup

Dylan Tomasello and Emmanuel Aluo

Mentors: Dr. Valerie Poynor, Jose Toledo, Jessica Romero, John Hong

California State University, Fullerton and Los Angeles City College

Abstract

The purpose of our project is to use a Bayesian modeling approach to predict the outcome of the final game of USA vs Netherlands in the 2019 FIFA Women's World Cup. We constructed our prior beliefs using the outcomes of the matches in the group stage. We considered the goals scored in each game in the knockout stage as our current data. We assumed independent Poisson likelihoods and Gamma priors for each team. Using a simulation technique, Monte Carlo approximation, we are able to obtain posterior inference for the rate of goals scored for each team per game, and further infer on the number of goals to be scored in the final tournament game for each team. Finally, we determined the probability that the USA team will score more goals than the Netherlands.

Intro

Bayes statistics stems from the Bayes' theorem:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

This theorem is based off the idea of updating prior beliefs about an event, A, using new information in the form of another event B.

In terms of data, $\boldsymbol{y} = (y_1, ..., y_n)'$, and model parameters $\boldsymbol{\theta} = (\theta_1, ...\theta_p)'$, Bayes theorem can be written:

$$p(\boldsymbol{\theta}|\boldsymbol{y}) \propto \left(\prod_{i=1}^{n} f(y_i|\boldsymbol{\theta})\right) \pi(\boldsymbol{\theta})$$

where $\pi(\boldsymbol{\theta})$ is the prior distribution, $\left(\prod_{i=1}^{n} f(y_i|\boldsymbol{\theta})\right)$ is the likelihood, and $p(\boldsymbol{\theta}|\boldsymbol{y})$ is the posterior.

Bayes method is often praised for being pertinent even with small sample sizes. However, it is often considered subjective as prior beliefs can vary among people.

Data

- The FIFA World Cup consists of two phases leading up to the final game.
- The first phase is the group round, which consists of 3 games per team. We utilize the number of goals scored in each game as our prior knowledge.
- The second phase is the knockout round, also consisting of three games for each of the two teams (USA and Netherlands). We use the number of goals scored in these games to update our prior in obtaining the posterior.

Team	Group[1] (goals/game)	Knockout[2] (goals/game)
USA (U)	13, 3, 2	2, 2, 2
Netherlands (N)	2, 3, 1	2, 2, 1

[1] https://www.fifa.com/w%omensworldcup/archive/france2019%/matches/groupphase

[2] https://www.fifa.com/womensworldcup/archive/france2019/matches/knockoutphase

Models

- For each team, we assume independent Poisson distributions for modelling the number of goals scored per game. The rate parameter, λ_j, describes the expected number of goals scored per game.
- Gamma distributions are assumed apriori for each rate parameter. The shape and rate parameters of the Gamma distributions are denoted as α_j and β_j, respectively.

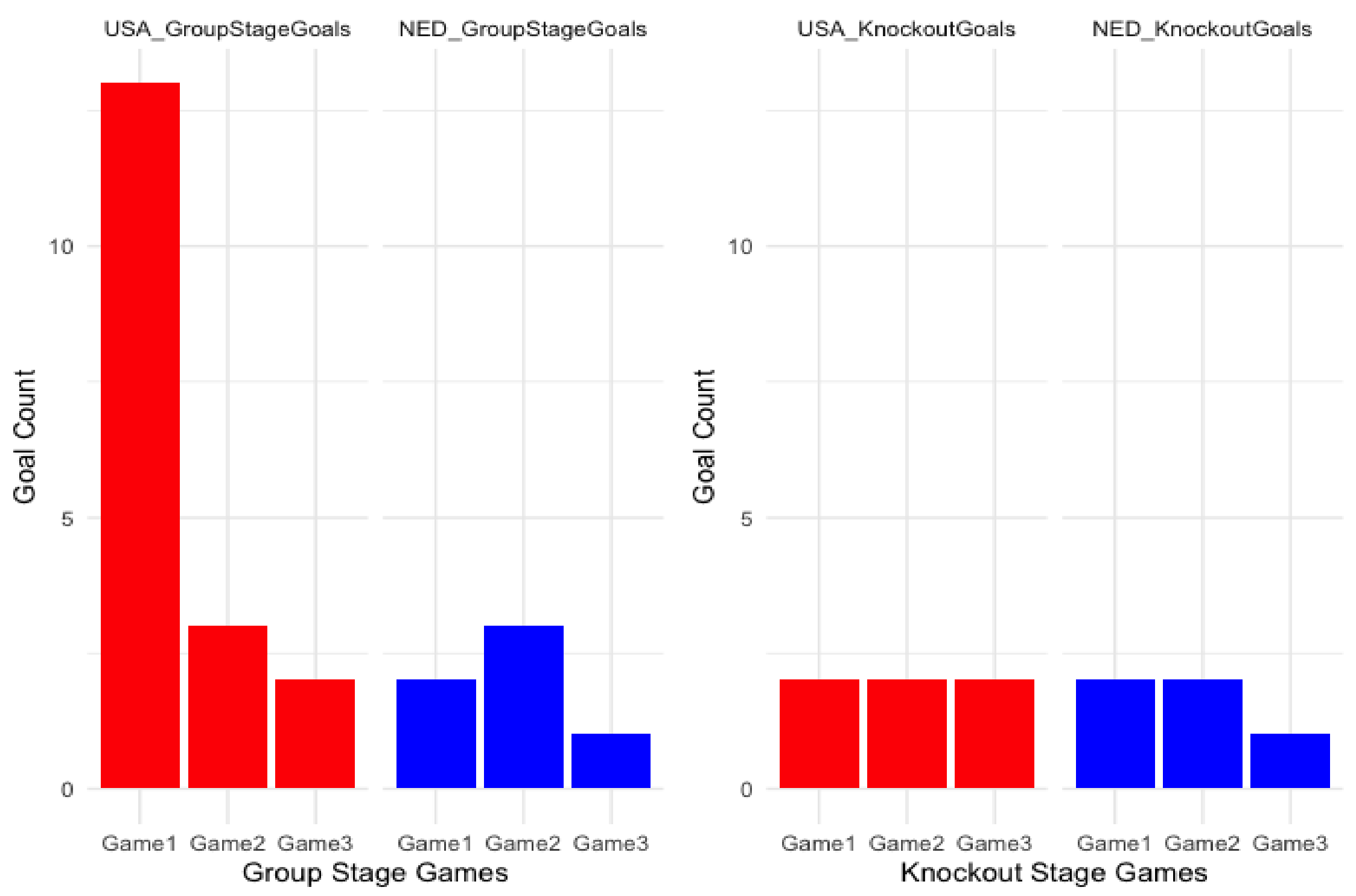
For $i = 1, ..., n_j$ and $j \in \{U, N\}$ we assume the following model:

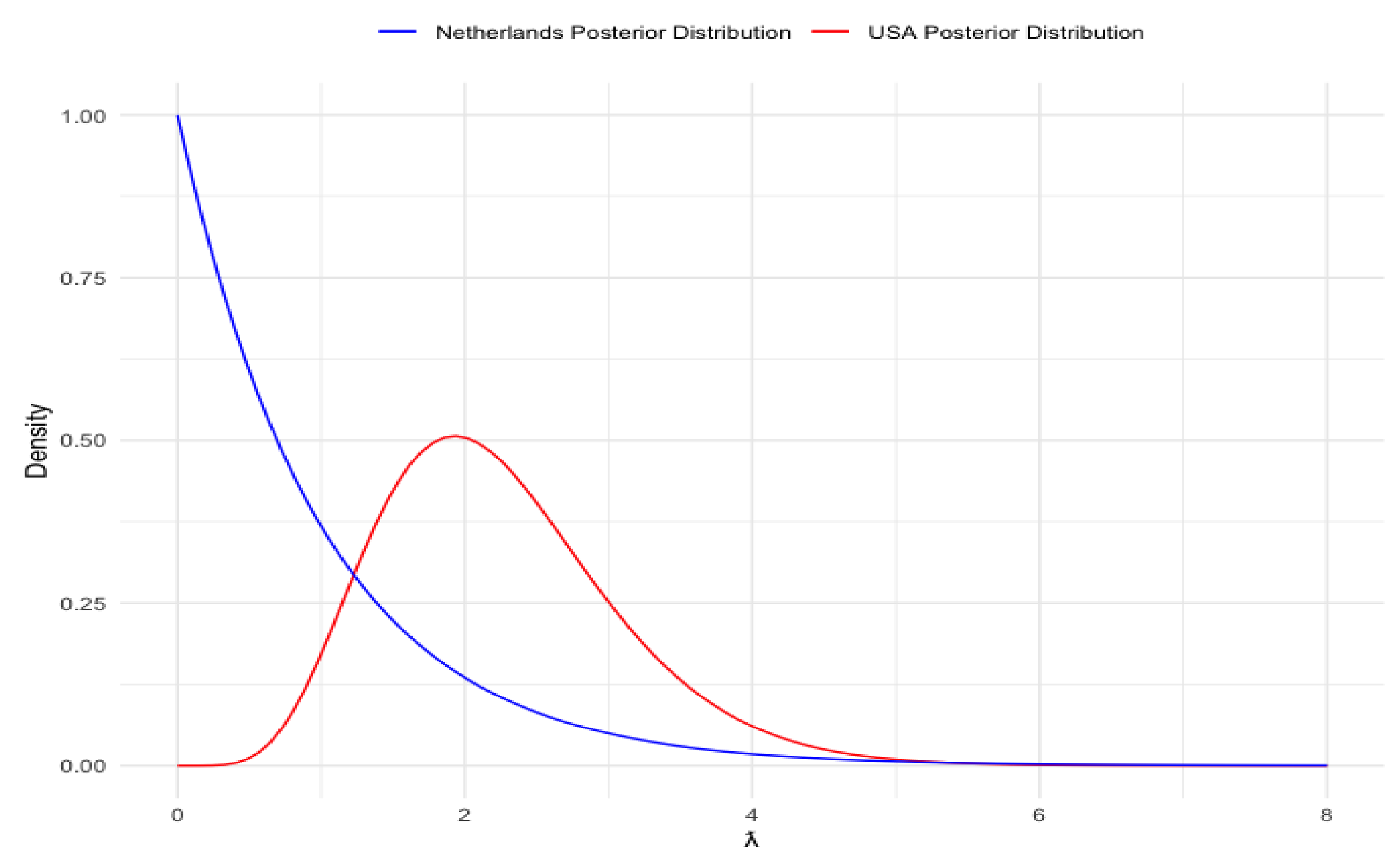
$$y_i|\lambda_j \overset{ind}{\sim} Pois(\lambda_j)$$
$$\lambda_j \overset{ind}{\sim} \Gamma(\alpha_j, \beta_j)$$

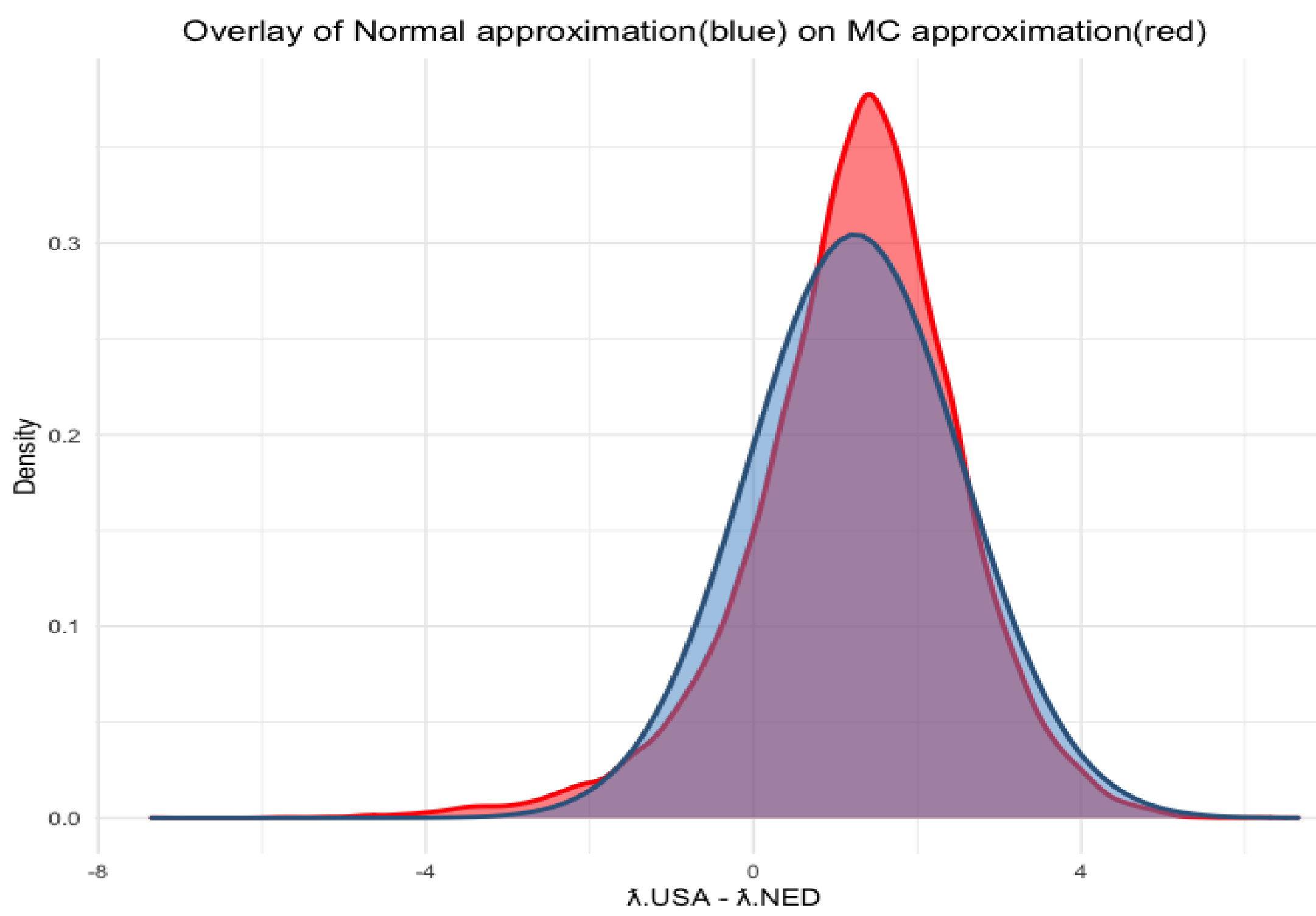
The posterior distributions are thus given by:

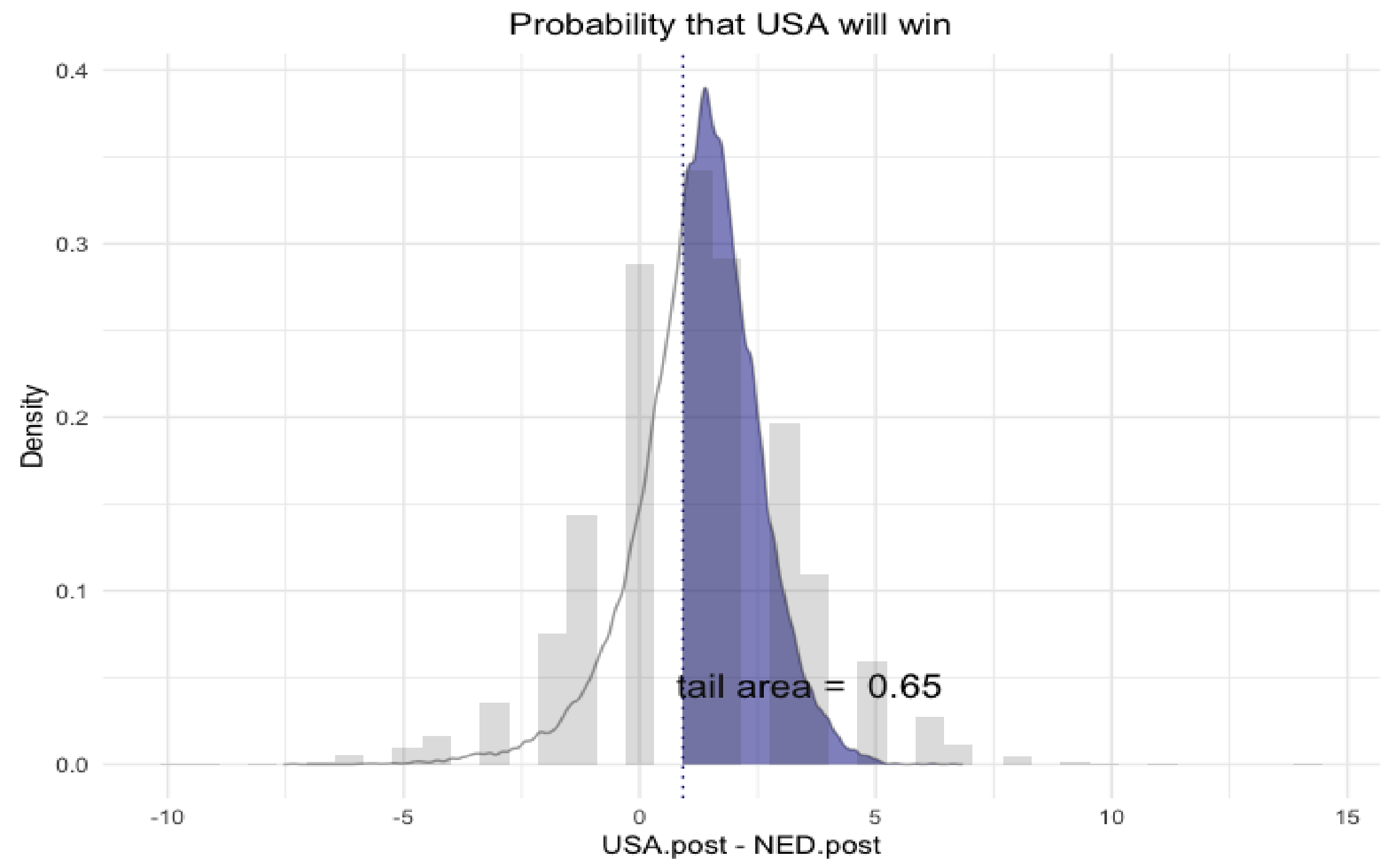
$$\Rightarrow \lambda_j|\boldsymbol{y}_j \overset{ind}{\sim} \Gamma\left(\alpha_j^{post} = \alpha_j + \sum^{n_j} y_i,\ \beta_j^{post} = \beta_j + n_j\right)$$

Figures









Prior Specification/Post Predictive

We utilized the mean and range of the number of goals scored in the group stage to specifying the prior parameters via: $E(Y_j) = \alpha_j/\beta_j$ and $Var(Y_j) = \alpha_j/\beta_j^2 + \alpha_j/\beta_j$.

USA		Netherlands	
Prior	Posterior	Prior	Post
$\alpha_U = 1.16$	$\alpha_U^{post} = 7.16$	$\alpha_N = -4$	$\alpha_N^{post} = 3.17$
$\beta_U = 0.19$	$\beta_U^{post} = 3.19$	$\beta_N = -2$	$\beta_N^{post} = 1$

The posterior predictive distribution of the number of goals scored is defined:

$$p(y_j^{new}|\boldsymbol{y}_j) = \int_0^\infty Pois(y_j^{new}|\lambda_j)\Gamma(\lambda_j|\boldsymbol{y}_j)d\lambda_j$$

we obtained this distribution via Monte Carlo simulation.

Methods

- A hypergeometric series (an analytical approach) and an asymptotic normal approximation (Kawasaki and Miyaoka, 2012) was used to obtain the posterior probability of the USA scoring more goals on average than the Netherlands: $P(\lambda_U >_N |data)$.
- We also utilized Monte Carlo simulation to obtain this same quantity.
- We compared the results under all three approaches (see table in Results).
- Finally, we used Monte Carlo simulation to obtain the posterior predictive probability of the USA winning over the Netherlands in the final game.

Results

Method	$P(\lambda_U > \lambda_N|data)$
Monte Carlo	0.8571
Asymptotic	0.8281
Analytical	0.8578

- From the table above, we can see that the Monte Carlo method provides an approximation closer to the exact probability as opposed to the Normal asymptotic method.
- Through simulation (20,000 iterations) based on our posterior information we obtained a probability of 0.6495 (or 64.95%) that the USA will win the final.

Discussion

- The prior specification for the Netherlands was an improper distribution with negative parameters where α_N = -4 and $\beta_N = -2$. However, the posterior distribution was indeed proper ($\alpha_N^{post}, \beta_N^{post} > 0$).
- Our final result of the USA having a 64.95% probability of winning the final is comparable to that obtained by FiveThirtyEight. Their model, designed by Jay Boice and Nate Silver, predicted that the probability of the USA winning was 66%[3].
- The final game resulted in USA winning over the Netherlands at 2-0 goals.

[3] https://fivethirtyeight.com/features/how-our-2019-womens-world-cup-predictions-work/