Olympic Medals Bayesian Analysis

Apostolos Stamenos

2/12/2022

Background

This analysis is inspired by research questions and data aggregated by NPR and FiveThirtyEight:

- Is There Home-Field Advantage At The Olympics?
- How Home Field Advantage Gives Olympic Host Countries An Edge — And More Gold Medals

This report is my submission for an Applied Bayesian Statistics exam, so all methods and interpretations are strictly Bayesian. As a result, statements like "there is a 95% probability that the parameter is between a and b" or "the probability that the null hypothesis is true" are intentionally different from the frequentist concepts of confidence intervals and p-values.

Aggregate Analysis

Let quantities with a subscript of *1* denote the aggregate quantities for host countries in the host year. Let quantities with a subscript of *0* denote the aggregate quantities for host countries in the previous Olympics. For both the host years and the non-host years, the goal is to study rates of events per unit of effort (i.e., number of medals won per number of participants). Thus, if the events are independent, Poisson likelihoods are reasonable models of the data.

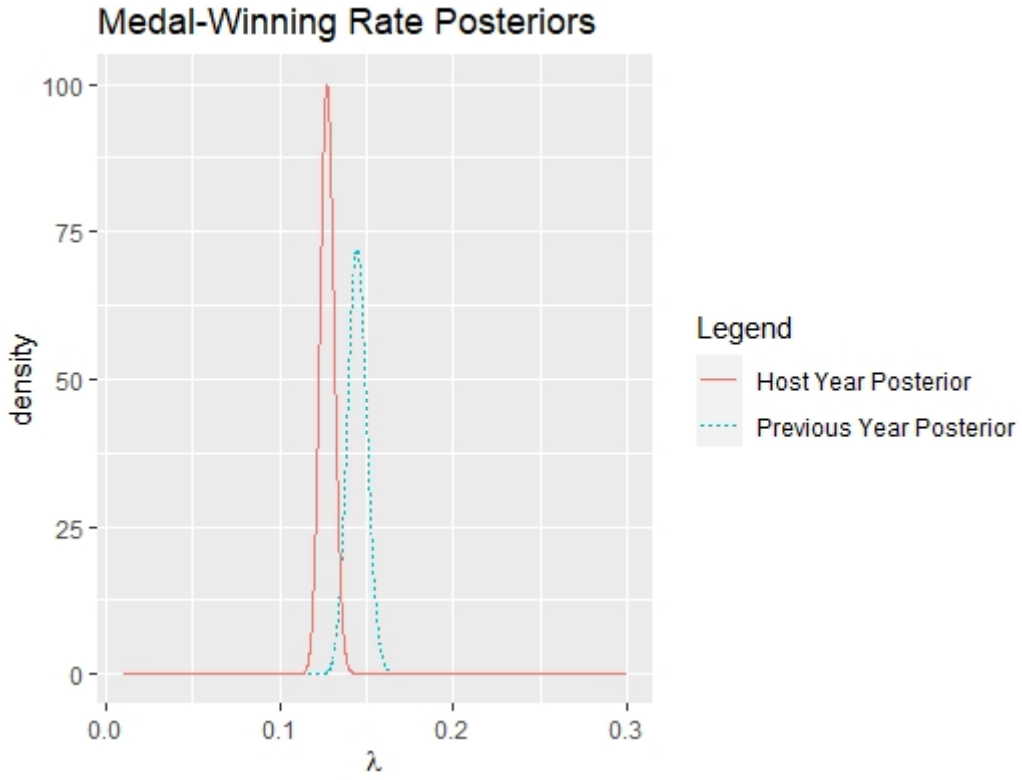
$$Y_0 \mid \lambda_0 \sim Poisson\left(N_0,\ \lambda_0\right)\ and\ Y_1 \mid \lambda_1 \sim Poisson\left(N_1,\ \lambda_1\right)$$

Because the parameter of a Poisson distribution is a positive real number, the prior distributions of the parameters can be modeled as Gamma. Thus, $\lambda_0 \sim Gamma(\epsilon,\ \epsilon)$ and $\lambda_1 \sim Gamma(\epsilon,\ \epsilon)$, where $\epsilon > 0$ is small. For these Poisson-gamma conjugate pairs, $\lambda_0 \mid Y_0 \sim Gamma(Y_0 + \epsilon,\ N_0 + \epsilon)$ and $\lambda_1 \mid Y_1 \sim Gamma(Y_1 + \epsilon,\ N_1 + \epsilon)$. *Figure 1* shows the two posterior distributions. The posterior of the host country in the host year is to the left of the posterior of the host country in the previous Olympics. The main assumptions of my analysis are that:

- The selected likelihoods and prior distributions are reasonable models of the data and of the uncertainty about the values of the parameters
- It makes sense to combine data across all years for the host country in the host year and compare it to the aggregated data for the host country in the previous Olympics

The prior and likelihood were primarily chosen because they have a known posterior (i.e., the choice was convenient). However, given that the parameters are positive and that the data could plausibly have been generated from a Poisson process, the parametric assumptions regarding the prior and likelihood are valid. Aggregating the data and conducting inference using the combined numbers of medals and participants may not be as valid. There may be country-specific or temporal differences, and this kind of aggregation might be discarding useful information.

Figure 1: **The posterior for the host year is further to the left compared to the posterior for the previous year.**



Hypothesis Test

$$\text{Let } H_0 : \lambda_1 > \lambda_0 \quad \text{and} \quad H_1 : \lambda_1 \leq \lambda_0.$$

In order to conduct this hypothesis test, I used Monte Carlo sampling to sample from the posterior distributions $p(\lambda_0 \mid Y_0) \quad \text{and} \quad p(\lambda_1 \mid Y_1)$. I then used the samples to calculate the proportion of $(\lambda_0,\ \lambda_1)$ pairs for which λ_1 exceeds λ_0. This percentage is an approximation of the probability that the null hypothesis is true. Based on the sample and the parameters of the uninformative prior defined in the previous section,

$$P(\lambda_1 > \lambda_0 \mid Y_0,\ Y_1) \approx 0.0053$$

The probability that the null hypothesis is true (i.e., that there is a host country advantage) is only 0.0053, so there is sufficient evidence to conclude that there is no host-country advantage. *Table 1* lists the probability of the null hypothesis for different values $a = b = \epsilon$. The probability that the null hypothesis is true changes slightly, but not enough to change the conclusion that in the aggregate analysis, there is no home-country advantage.

Table 1: **Sensitivity Table**: $P(H_0)$ changes slightly with different ϵ, but not enough to affect the conclusion.

ϵ	$P(H_0)$
0.01	0.0053
0.1	0.0053
1	0.0051
10	0.0040
100	0.0000

Prediction

In order to predict how many medals France will win in 2024, it is important to know how many French athletes will participate in the 2024 Olympics. The Poisson likelihood used throughout this paper assumes that N is fixed or non-random. In order to continue using the same model, I calculated the percent change in the number of participants for all country-year pairs in the dataset:

$$\%\Delta N = \frac{N_1 - N_0}{N_0}$$

I then calculated the approximate median percent change in participants and used it to estimate the number of French participants in the 2024 Olympics:

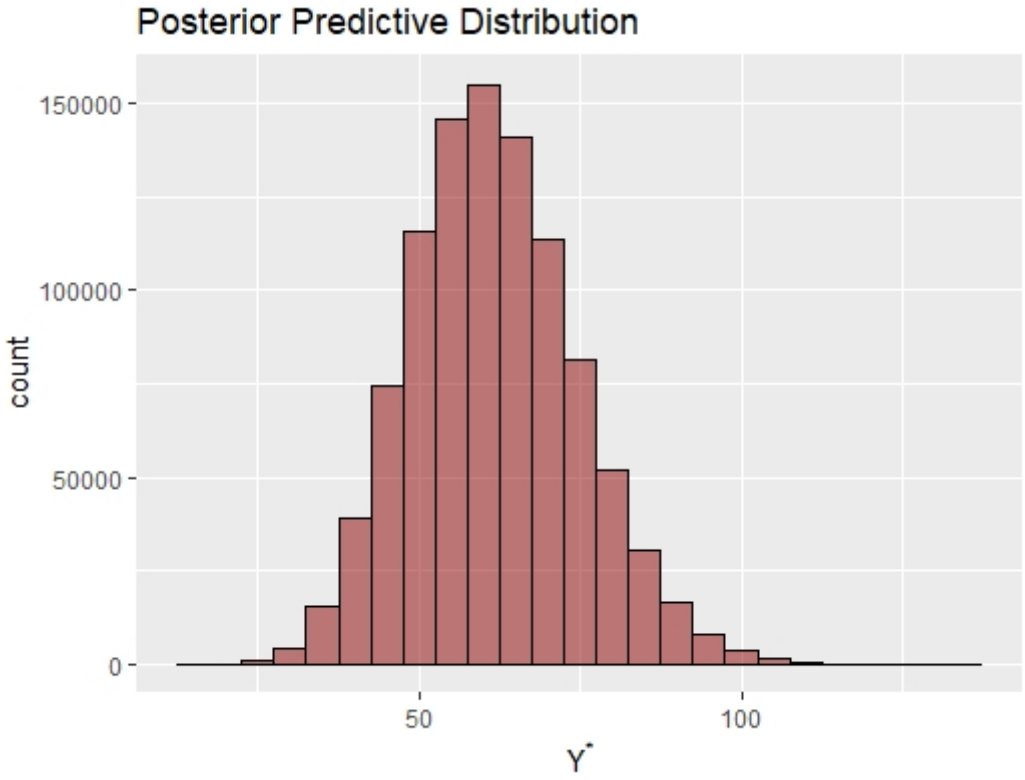
$$N_{1,\,FR} \approx N_{0,\,FR}\,(1 + \%\Delta N) = 398 \cdot (1 + 0.847) = 735.12 \;\; participants$$

I calculated the median number of players and median number of medals won during the host year and used those numbers to generate estimated λ^* from the prior distribution of λ.

With estimates for these numbers, I then used a Poisson-Gamma likelihood-prior model similar to the ones from the previous sections: $Y_{1,\,FR} \mid \lambda_{1,\,FR} \sim Poisson\,(N_{1,\,FR},\ \lambda_{1,\,FR})$ and $\lambda_{1,\,FR} \sim Gamma(\epsilon,\ \epsilon)$, where $\epsilon > 0$ is small. For this Poisson-gamma conjugate pair, $\lambda^*_{1,\,FR} \mid Y_{1,\,FR} \sim Gamma(Y_{1,\,FR} + \epsilon,\ N_{1,\,FR} + \epsilon)$.

In order to make a prediction, I took a sample of 1,000,000 from the posterior distribution. I then plugged each simulated λ^* in the likelihood function and drew a total of 1,000,000 estimated Y^* values. Through this Monte Carlo approximation approach, I essentially sampled values from the posterior predictive distribution (PPD), $f^*(Y^* \mid Y)$. In other words, the 1,000,000 values that are the output of this procedure serve as an approximation of the posterior predictive distribution. *Figure 2* shows a histogram of draws from the PPD. Unlike simply plugging in an estimate and sampling from the posterior, sampling from the PPD properly accounts for parametric uncertainty. I used the PPD to calculate 2.5%, 50%, and 97.5% quantiles. Based on these values, there is a 95% probability that the number of medals that France will win in the 2024 Olympics will be between 38 and 89 medals, with 61 as a reasonable point estimate.

Figure 2: **Approximate Posterior Predictive Distribution (PPD) of the number of medals France will win in 2024.**



Country-specific Analysis

For each country $j \in \{1, \ldots, 15\}$, the goal is to study rates of events per unit of effort (i.e., number of medals won per number of participants). So just like in other parts of this paper, a reasonable model for the data is $Y_{ij} \mid \lambda_{ij} \sim Poisson(N_{ij}\lambda_{ij})$ for $i \in \{0, 1\}$ and $j \in \{1, \ldots, 15\}$. Since the parameters of interest are positive real numbers, $\lambda_{ij} \sim Gamma(\epsilon,\ \epsilon)$ is a reasonable prior. Because Gamma is a conjugate prior for a Poisson likelihood, $\lambda_{ij} \mid Y_{ij} \sim Gamma(Y_{ij} + \epsilon,\ N_{ij} + \epsilon)$. In the country-specific analysis, $r_j = \frac{\lambda_{1j}}{\lambda_{0j}} > 1$ means that there is a home-country advantage for country j. Graphically comparing 15 posteriors is cumbersome, so instead I compared the posterior medians and 95% credible intervals for each country in *Table 2*. Some of the credible intervals (e.g., for Greece and Finland) do not contain *1*, which means that for those countries there is not sufficient evidence of a home-country advantage. Most of the credible intervals contain *1*, but many of these are not symmetric around *1* and have posterior medians that are less than *1*. Based on this analysis, there is evidence that the home-country advantage differs by country.

Table 2: **Country-specific Posterior Summaries**:

Country	95% *CI*	*Posterior Median*
Australia	(0.70, 1.39)	0.98
Brazil	(0.30, 1.12)	0.57
Canada	(0.44, 3.94)	1.22
China	(0.75, 1.40)	1.02
Finland	(0.25, 0.82)	0.46
Great Britain	(0.55, 1.16)	0.79
Greece	(0.19, 0.86)	0.41
Italy	(0.42, 1.17)	0.70
Japan	(0.57, 1.12)	0.80
Mexico	(0.64, 84.40)	3.78
South Korea	(0.44, 1.37)	0.76
Soviet Union	(1.05, 1.64)	1.31
Spain	(1.18, 10.95)	3.13
United States	(0.91, 1.32)	1.10
West Germany	(0.62, 1.67)	1.00

Conclusion

I first conducted an aggregate analysis and discovered that in aggregate, there is no home-country advantage. When taking into consideration the number of participating athletes, there is not enough evidence that the rate of medals per athlete is higher for the host country during the host year. This was determined both by visually inspecting the posterior distributions and by conducting a hypothesis test. The results were not particularly sensitive to the choice of prior. I then used a posterior predictive distribution to predict that in the 2024 Olympics, France has a 95% probability of winning between 38 and 89 medals. Finally, I conducted a country-specific analysis and found evidence that the home-country advantage differs by country.

A main limitation of this study is the choice of a likelihood and prior based on their conjugate relationship. In this analysis, a Poisson-Gamma model is plausible, but perhaps another model without a known posterior distribution would be better suited to studying home-country advantage. Another limitation is that this analysis did not properly account for temporal relationships or potential covariates that may affect the results of the study. A factor other than home-country advantage may be responsible for higher rates of medal-winning by the host country during the host year. Both of these limitations could be addressed in future work by employing more sophisticated computational approaches and/or hierarchical models.

Appendix A.1: R Code

```r
# Import libraries and data
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.1.2
```

```r
library(dplyr)
```

```
##
## 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
```

```r
medals <- read.csv("Medals.csv")
set.seed(21222)

# Aggregate analysis
aggregates <- medals %>%
    summarise(Y0 = sum(MEDALS.WON.DURING.PREVIOUS.OLYMPICS),
        Y1 = sum(MEDALS.WON.DURING.HOST.YEAR), N0 = sum(PARTICIPATING.ATHLETES.DURING.PREVIOUS.OLYMPICS),
        N1 = sum(PARTICIPATING.ATHLETES.DURING.HOST.YEAR))

colnames(medals) <- c("country", "year", "Y0", "Y1", "N0", "N1")
S <- 10^6
N0 <- aggregates$N0
Y0 <- aggregates$Y0
N1 <- aggregates$N1
Y1 <- aggregates$Y1
epsilon <- 0.1

lambda <- seq(0.01, 0.3, 1e-04)
posterior_0 <- dgamma(lambda, Y0 + epsilon, N0 + epsilon)
posterior_1 <- dgamma(lambda, Y1 + epsilon, N1 + epsilon)
df <- data.frame(lambda, posterior_0, posterior_1)

# Visualizations of posterior distributions
p <- ggplot(data = df, aes(x = lambda))
p <- p + ggtitle("Medal-Winning Rate Posteriors") + labs(x = expression(lambda),
    y = "density", color = "Legend", linetype = "Legend")
```

```r
p <- p + geom_line(aes(y = posterior_0, color = "Previous Year Posterior",
    linetype = "Previous Year Posterior"))
p <- p + geom_line(aes(y = posterior_1, color = "Host Year Posterior",
    linetype = "Host Year Posterior"))

# Hypothesis Test
epsilon <- c(0.01, 0.1, 1, 10, 100)
prob_null <- c()

# MC samples from the posteriors
for (e in epsilon) {
    lambda_0 <- rgamma(S, Y0 + e, N0 + e)
    lambda_1 <- rgamma(S, Y1 + e, N1 + e)

    # Approximate probability that theta2 > theta1
    prob_null <- c(prob_null, mean(lambda_1 > lambda_0))
}

# Prediction
medals <- medals %>%
    mutate(pct_change_Y = (Y1 - Y0)/Y0, pct_change_N = (N1 -
        N0)/N0)
median_pct_change_Y <- quantile(medals$pct_change_Y, 0.5)
median_pct_change_N <- quantile(medals$pct_change_N, 0.5)
a <- b <- 0.1
N0_FR <- 398
Y0_FR <- 33
estimated_Y1_FR <- Y0_FR * (1 + median_pct_change_Y)
estimated_N1_FR <- N0_FR * (1 + median_pct_change_N)
lambda_star <- rgamma(S, estimated_Y1_FR + a, estimated_N1_FR +
    b)
ppd <- data.frame(Y1_star = rpois(S, estimated_N1_FR * lambda_star))
plot_ppd <- ggplot(ppd, aes(Y1_star)) + ggtitle("Posterior Predictive Distribution") +
    labs(x = expression(Y^"*"))

plot_ppd <- plot_ppd + geom_histogram(color = "black", fill = "darkred",
    alpha = 0.5, bins = 49)

quantiles <- quantile(ppd$Y1_star, c(0.025, 0.5, 0.975))

# Country-specific Analysis
country_specific <- medals %>%
    group_by(country) %>%
    summarise(Y0 = sum(Y0), Y1 = sum(Y1), N0 = sum(N0), N1 = sum(N1))
```

```r
posterior_summaries <- data.frame(country = c(), CI95_low = c(),
    MAP = c(), CI95_high = c())
e <- 0.1

for (j in 1:15) {
    lambda_0 <- rgamma(S, country_specific$Y0[j] + e, country_specific$N0[j] +
        e)
    lambda_1 <- rgamma(S, country_specific$Y1[j] + e, country_specific$N1[j] +
        e)
    r <- lambda_1/lambda_0
    CI95_low = quantile(r, 0.025)
    med = quantile(r, 0.5)
    CI95_high = quantile(r, 0.975)
    posterior_summaries <- rbind(posterior_summaries, data.frame(country = country_specific$country[j],
        CI95_low = round(CI95_low, 2), med = round(med, 2), CI95_high = round(CI95_high,
            2)))
}
```