Emily Pierce- ST540 Exam 1, 2/17/2022

Introduction

The purpose of this report is to employ a Bayesian approach to analyze the host country advantage in the Summer Olympic games. Data for this analysis includes the number of participants and medals won for the host country during each of the summer Olympic games for games between 1952 and 2000. Data also includes the number of participants and number of medals won for each host country in the prior Olympics which were held somewhere else. Analysis will be split into several components. First, the data will be summed across each year and analysis will be performed in aggregate to test for a host country advantage. Then, the data will be used to make a prediction about the outcome of the coming summer Olympics in France. Finally, the data will be analyzed separately for each year to determine if the host country advantage differs between countries.

Aggregate Analysis

To assess broadly for a host country advantage, data were summed across years to derive the total number of medals won by the host country (Y_1), the total number of host participants (N_1), the total number of medals won by host country during the previous Olympics (Y_0), and the total number of host participants in the previous Olympics (N_0). There are two parameters of interest related to this dataset, λ_1 which represents the expected number of medals per participant in their home country, which is hosting and λ_0, which represents the expected number of medals per participant of the original host country during the previous Olympic games. Because this data is essentially a count of Y events in N units, the likelihood is best modelled with a Poisson distribution:

$$Y_1|\lambda_1 \sim Poisson(N_1\lambda_1) \qquad Y_0|\lambda_0 \sim Poisson(N_0\lambda_0)$$

Assuming an uninformative conjugate distribution for the prior, a beta distribution with a=0.b=0.1 will be utilized for this analysis:

$$\lambda_{1,0} \sim Gamma(a = 0.1, b = 0.1)$$

Given the distribution of the likelihood and the prior, the conjugate distribution of the posterior follows a gamma distribution as shown below:

$$\lambda_1|Y_1 \sim Gamma(Y_1 + a, N_1 + b) \qquad \lambda_0|Y_0 \sim Gamma(Y_0 + a, N_0 + b)$$

The posterior distributions slightly overlap with the center of the λ_0 distribution slightly larger than the center of the λ_1 distribution (Fig. 1).

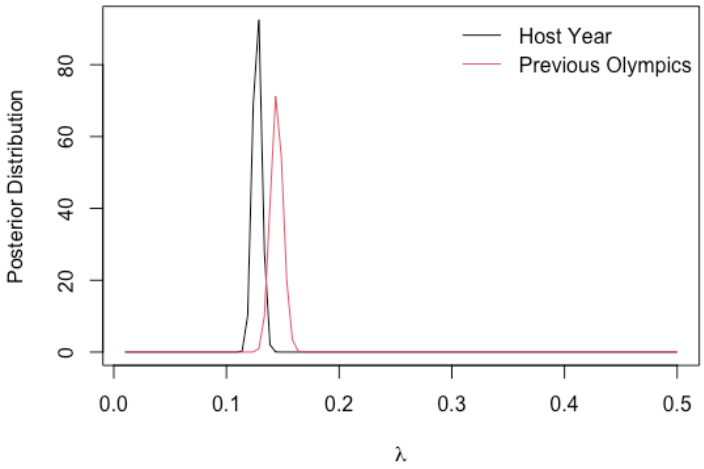

Figure 1: Posterior Distributions for λ_1 (black) and λ_0 (red).

There are several main assumptions made in this analysis. The first assumption is that the likelihood follows a Poisson distribution. The data may at first seem to resemble a number of successes in N trials, which would be best fit with a binomial distribution. However, we assume that each participant may be competing in many events and thus has the potential to win many medals, so there is not a binary outcome for each participant, but rather a medal count which is best modeled with the Poisson distribution. We are also assuming the participants in each of the countries are independent of each other, which is a valid assumption since each participants performance is independent. This assumption may become more complicated if the concept of team sports is involved, but for this analysis, the assumption is that each event is independent.

Hypothesis Test

To test whether there is an advantage to the host country, I tested the hypothesis that $\lambda_1 > \lambda_0$. This was accomplished by first using Monte Carlo sampling to draw 100,000 samples of λ_1 and λ_0 and then calculating the posterior probability that $\lambda_1 > \lambda_0$. The test resulted in a posterior probability equal to 0.00534, indicating that **there is a miniscule probability that there is a host country advantage** under the described prior and likelihood for aggregated data. The posterior probability changed very slightly, within a range of 0.00516-0.00563, when recalculated using different priors, indicating the test results are not very sensitive to the prior (Table 1).

Table 1: Sensitivity Analysis for Changes in Probability as a Result of Changing the Prior

Prior	**$P(\lambda_1 > \lambda_0)$**
Gamma(0.1,0.1)	0.00534
Gamma(0.05,0.05)	0.00548
Gamma(0.15,0.15)	0.00516
Gamma(0.2,0.2)	0.00563

Prediction

The next Summer Olympics are set to be held in France in 2024. In the previous summer Olympics in 2021, hosted in Japan, France sent 398 participants and 33 medals. To predict the number of medals France will win in their home country in 2024, I first predicted the number of participants France would send to the Olympics. The number of participants was calculated using the discussed dataset by taking the difference in participants between host year and previous year for each set of games and averaging that number. On average, a country sent 183 more participants when hosting in their own country. To project the number of participants in France in 2024 (N_1), that average was added to the number of participants they sent in the previous Olympics to get a predicted number of 581 participants. Next, Monte Carlo sampling was used to draw 100,000 samples from λ_0 describing the expected number of medals per participant for France in 2021. The results of sampling were saved as λ_{MC}. Then Monte Carlo sampling was used again to draw 100,000 estimates of French medals in 2024 (Y*) using the following distribution:

$$Y^* \sim Poisson(N_1 * \lambda_{MC})$$

The sample mean of Y* was taken as the posterior predictive mean from posterior predictive distribution (Fig.2) and was equal to 48.26. Because the medal count must be a whole number,

the posterior predictive mean was rounded down to conclude that **France will likely win 48 medals** in 2024 under the assumption that they will send 581 participants. The predicted medal count is more than that won by France during the 2021 Olympics in Japan, but the expected number of medals per participants in 2024 (.0826) is slightly smaller than that from 2021 (.0829), so a host country advantage is not suggested by this prediction. The PPD approach for prediction was chosen over the plug-in approach because it properly accounts for parametric error which is likely to be large. To quantify uncertainty in the prediction, the 95% Credible interval was calculated to be (35,62) which reflects both parametric and random uncertainty.

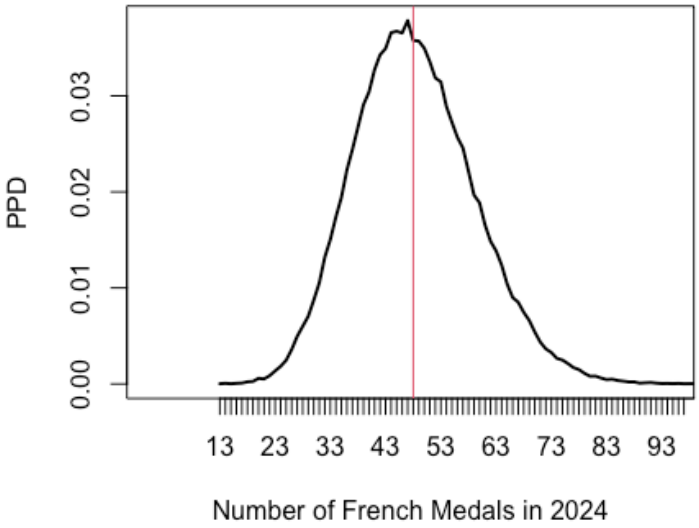

Figure 2: Posterior Predictive Distribution of the Number of French Medals in the future 2024 Olympics

Country Specific Analysis

Thus far, analysis has indicated that when the data are aggregated across all countries, there is no host country advantage. Now, the data will be analyzed without aggregation to determine if the host country advantage varies across countries. The dataset includes three countries that hosted 2 Olympic games: Australia, United States, and Japan). The two data points for each of the countries that hosted twice were averaged to produce a dataset in which each host country is only represented once. As with the aggregated data, the likelihood is best described with a Poisson distribution where *i* represents the index of the 15 countries:

$$Y_{i1}|\lambda_{i1} \sim Poisson(N_{i1}\lambda_{i1}) \qquad Y_{i0}|\lambda_{i0} \sim Poisson(N_{i0}\lambda_{i0}$$

Assuming an uninformative conjugate prior, the prior follows a gamma distribution as does the posterior:

$$\lambda_{i1} \sim Gamma(a = 0.1, b = 0.1) \qquad \lambda_{i0} \sim Gamma(a = 0.1, b = 0.1)$$
$$\lambda_{i1}|Y_{i1} \sim Gamma(Y_{i1} + a, N_{i1} + b) \qquad \lambda_{01}|Y_{01} \sim Gamma(Y_{01} + a, N_{01} + b)$$

The posterior distributions were sampled 100000 times using Mote Carlo sampling for each country. Summaries of each of those posterior distributions indicated there is high variability in the posterior median between countries and between host and non-hosting Olympic games for some countries like Finland and the Soviet Union (Table 2). Additionally, for each country the ratio (R) of the samples of λ_1 and λ_0 and the posterior probability that $\lambda_1 > \lambda_0$ were both computed (Table 2). There is large variability in the ratio and probability between countries indicating that the host country advantage is country dependent based on this analysis. For

example, The Soviet Union has a 99.1% probability of having a higher expected medal count per participant when competing at home compared to .5% for Finland.

Table 2: Comparing Posterior Results Across Countries

	λ_1 (Year of Hosting)	λ_0 (Previous Olympics)		
Host Country	**Posterior Median**	**Posterior Median**	**R**	**Prob ($\lambda i_1 > \lambda i_0$)**
Finland	0.08	0.18	0.36	0.005
Italy	0.13	0.18	0.56	0.088
Mexico	0.03	0.01	0.61	0.928
West Germany	0.09	0.09	1.61	0.509
Canada	0.03	0.02	1.44	0.649
Soviet Union	0.4	0.30	1.46	0.991
South Korea	0.08	0.11	0.69	0.178
Spain	0.05	0.02	2.01	0.991
Greece	0.04	0.09	0.56	0.010
China	0.17	0.16	0.95	0.550
Great Britain	0.12	0.15	0.84	0.118
Brazil	0.04	0.07	1.10	0.050
Japan	0.08	0.10	0.43	0.178
Australia	0.10	0.10	0.72	0.472
United States	0.23	0.21	1.15	0.760

Conclusions

Assuming a Poisson-gamma conjugate distribution pair and independent observation, results indicate that on aggregate, when data are summed across countries, there is no host-country advantage; however, country by country analysis that there is evidence of a country dependent host advantage in the Summer Olympics. Additionally, results indicate that there will likely not be a host-country advantage for France who is set to host the 2024 games and is predicted to win 48 medals. This prediction was made using a PPD approach and an assumption on the number of participants France will send.

One limitation of this analysis is that ignores the variety of different Olympic events. Because this dataset only included a count of medals per number of participants, the assumption was made based on the Olympic format that each participant could've won multiple medals which informed the choice of a Poisson likelihood distribution. However, if more data were collected to reflect the binary medal outcome of each participant in each event, a binomial likelihood distribution would better suit the data and may be more informative. Additionally, another limitation of this analysis was that it had to be based on an uninformative prior, placing significant weight on the data. It may improve analysis to consult with an expert in the field of sports statistic or someone with the IOC to determine if there is a more informative prior that may expand analysis.

Appendix Code

```
#Load in data
medals<-read.csv('Medals.csv', header = TRUE, sep = ',')
##Aggregate provided data
Y1=sum(medals$MEDALS.WON.DURING.HOST.YEAR)
Y0=sum(medals$MEDALS.WON.DURING.PREVIOUS.OLYMPICS)
N1=sum(medals$PARTICIPATING.ATHLETES.DURING.HOST.YEAR)
N0=sum(medals$PARTICIPATING.ATHLETES.DURING.PREVIOUS.OLYMPICS)

##Determine and Plot the Posterior
lambda<-seq(0.01,1,length=100)
a<-b<-.01
plot(lambda, dgamma(lambda,Y1+a,N1+b), type='l',
     xlab=expression(lambda),
     ylab='Posterior Distribution')
lines(lambda, dgamma(lambda,Y0+a,N0+b), col=2)
legend('topright',c('Host Year','Previous Olympics'), lty=1,col=1:2,bty='n')

lambda1<-rgamma(100000,Y1+a,N1+b)
lambda0<-rgamma(100000,Y0+a,N0+b)

##Hypothesis Test & Sensitivity Analysis
mean(lambda1>lambda0)

## [1] 0.00597

a.05<-b.05<-.05
a.15<-b.15<-.15
a.2<-b.2<-.2
lambda1.05<-rgamma(100000,Y1+a.05,N1+b.05)
lambda0.05<-rgamma(100000,Y0+a.05,N0+b.05)
mean(lambda1.05>lambda0.05)

## [1] 0.00554

lambda1.15<-rgamma(100000,Y1+a.15,N1+b.15)
lambda0.15<-rgamma(100000,Y0+a.15,N0+b.15)
mean(lambda1.15>lambda0.15)

## [1] 0.00483

lambda1.2<-rgamma(100000,Y1+a.2,N1+b.2)
lambda0.2<-rgamma(100000,Y0+a.2,N0+b.2)
mean(lambda1.2>lambda0.2)

## [1] 0.00537

##French Prediction
Y<-seq(1,70,1)
Y0_Fr<-33
N0_Fr<-398
N1_Fr<-581
```

```r
plot(lambda, dgamma(lambda,Y0_Fr+a,N0_Fr+b), type='l',
     xlab=expression(lambda),
     ylab='Posterior Distribution')

lam_mc<-rgamma(100000,Y0_Fr+a,N0_Fr+b)
Ymc<-rpois(100000,(N1_Fr*lam_mc))
mean(Ymc)
```

```
## [1] 48.20565
```

```r
plot(table(Ymc)/100000,type='l',ylab='PPD',xlab='Number of French Medals in 2
024',xlim=c(0,95))+
  abline(v=48,col=2)
```

```
## numeric(0)
```

```r
mean(qpois(.025,(N1_Fr*lam_mc)))
```

```
## [1] 35.14594
```

```r
mean(qpois(.975,(N1_Fr*lam_mc)))
```

```
## [1] 62.20942
```

```r
##Country by country analysis
mc<-read.csv('Medals_Combined.csv', header = TRUE, sep = ',')

Lambda1_Countries<-{}
Lambda0_Countries<-{}
for(i in 1:15){
  Lambda1_Countries<-rgamma(100000, mc$MEDALS.WON.DURING.HOST.YEAR[i]+a,mc$PA
RTICIPATING.ATHLETES.DURING.HOST[i]+b)
  Lambda0_Countries<-rgamma(100000, mc$MEDALS.WON.DURING.PREVIOUS.OLYMPICS[i]
+a,mc$PARTICIPATING.ATHLETES.DURING.PREVIOUS.OLYMPICS[i]+b)
}
q50.1    <- qgamma(0.500,mc$MEDALS.WON.DURING.HOST.YEAR+a,mc$PARTICIPATING.AT
HLETES.DURING.HOST+b)
q_low.1  <- qgamma(0.025,mc$MEDALS.WON.DURING.HOST.YEAR+a,mc$PARTICIPATING.AT
HLETES.DURING.HOST+b)
q_high.1 <- qgamma(0.975,mc$MEDALS.WON.DURING.HOST.YEAR+a,mc$PARTICIPATING.AT
HLETES.DURING.HOST+b)
q50.0    <- qgamma(0.500,mc$MEDALS.WON.DURING.PREVIOUS.OLYMPICS+a,mc$PARTICIP
ATING.ATHLETES.DURING.PREVIOUS.OLYMPICS+b)
q_low.0  <- qgamma(0.025,mc$MEDALS.WON.DURING.PREVIOUS.OLYMPICS+a,mc$PARTICIP
ATING.ATHLETES.DURING.PREVIOUS.OLYMPICS+b)
q_high.0 <- qgamma(0.975,mc$MEDALS.WON.DURING.PREVIOUS.OLYMPICS+a,mc$PARTICIP
ATING.ATHLETES.DURING.PREVIOUS.OLYMPICS+b)
out    <- round(cbind(q50.1,q_low.1,q_high.1),2)
rownames(out) <- mc$HOST.COUNTRY
out
```

```
##                   q50.1 q_low.1 q_high.1
## Finland            0.08    0.05     0.12
## Italy              0.13    0.09     0.17
## Mexico             0.03    0.01     0.06
## West Germany       0.09    0.07     0.13
## Canada             0.03    0.01     0.05
## Soviet Union       0.40    0.34     0.46
## South Korea        0.08    0.06     0.11
## Spain              0.05    0.03     0.08
## Greece             0.04    0.02     0.06
## China              0.17    0.14     0.20
## Great Britain      0.12    0.09     0.15
## Brazil             0.04    0.02     0.06
## Japan_Avg          0.08    0.06     0.11
## Australia_Avg      0.10    0.07     0.13
## United States_Avg  0.23    0.20     0.28
```

```
r={}
for(i in 1:15){
  r[i]=(rgamma(100000,mc$MEDALS.WON.DURING.HOST.YEAR[i]+a,mc$PARTICIPATING.AT
HLETES.DURING.HOST[i]+b)/(Lambda0_Countries<-rgamma(100000, mc$MEDALS.WON.DUR
ING.PREVIOUS.OLYMPICS[i]+a,mc$PARTICIPATING.ATHLETES.DURING.PREVIOUS.OLYMPICS
[i]+b)))
}
```

```
r
```

```
##  [1]  0.3563162  0.3611546 32.3549982  0.9871239  0.8402364  1.0576365
##  [7]  0.4375176  1.8046812  0.3227774  0.7884704  1.0914827  0.6092098
## [13]  0.7492811  1.3242950  1.1541089
```

```
prob={}
for(i in 1:15){
  prob[i]=mean(rgamma(100000,mc$MEDALS.WON.DURING.HOST.YEAR[i]+a,mc$PARTICIPA
TING.ATHLETES.DURING.HOST[i]+b)>(Lambda0_Countries<-rgamma(100000, mc$MEDALS.
WON.DURING.PREVIOUS.OLYMPICS[i]+a,mc$PARTICIPATING.ATHLETES.DURING.PREVIOUS.O
LYMPICS[i]+b)))
}
prob
```

```
##  [1] 0.00455 0.08670 0.92666 0.50724 0.65006 0.99097 0.18107 0.99082 0.010
84
## [10] 0.54969 0.11756 0.05065 0.18266 0.47111 0.76165
```

Medals_Combined.csv data:

HOST COUNTRY	YEAR	MEDALS WON DURING	MEDALS WON DURING	PARTICIPATING ATHLETES DURING	PARTICIPATING ATHLETES DURING HOST YEAR

		PREVIOUS OLYMPICS	HOST YEAR	PREVIOUS OLYMPICS	
Finland	1952	24	22	129	258
Italy	1960	25	36	135	280
Mexico	1968	1	9	94	275
West Germany	1972	26	40	275	423
Canada	1976	5	11	208	385
Soviet Union	1980	125	195	410	489
South Korea	1988	19	33	175	401
Spain	1992	4	22	229	422
Greece	2004	13	16	140	426
China	2008	63	100	384	599
Great Britain	2012	47	65	304	530
Brazil	2016	17	19	236	462
Japan_Avg	2021	29.5	40	278.5	474.5
Australia_Avg	1996	26	46.5	249	455.5
United States_Avg	2000	101	137.5	470.5	584.5