ST440/540 – Applied Bayesian Analysis: Midterm exam 1
Spring 2022
Tilekbek Zhoroev

Introduction

We have given the data of the summer Olympic games since 1952. The data consists the name of host country and results also the results of the game four years earlier. Using the given data we created the columns of the ratio medals per participants and the aggregated results, it's presented at Table 1. In this work we would like to analyze the advantage of the home-country in the summer Olympics. Using table we observe the average number of the medals are increased in home-country. Even though the number of the medals of host country tend to increase, the overall ratio of the medals per participants are decreased due to the increase of the number of participants. Here we would like to analyse this using Bayesian Analysis.

Let Y_{i1} be the number of medals won by the host country during the Olympics i and Y_{i0} be the number of medals won by the county in the previous Olympics. Similarly, let N_{i0} and N_{i1} be the number of participates from the country in the corresponding Olympics. For example, Finland hosted the Olympics in 1952. They had $N_{11} = 258$ participants in 1952 and $N_{10} = 129$ participants in the 1948; they won $Y_1 1 = 22$ medals in 1952 and $Y_{10} = 24$ medals in the 1948. Then we can compute the total number of the results, i.e

$$Y_1 = \sum_{i=1}^{18} Y_{i1} = 1016, N_1 = \sum_{i=1}^{18} N_{i1} = 7979$$

and

$$Y_0 = \sum_{i=1}^{18} Y_{i0} = 682, N_1 = \sum_{i=1}^{18} N_{i0} = 4715$$

		Previous			Host		
Host country	Year	Medals	Participants	Ratio	Medals	Participants	Ratio
Finland	1952	24	129	0.19	22	258	0.09
Australia	1956	11	81	0.14	35	294	0.12
Italy	1960	25	135	0.19	36	280	0.13
Japan	1964	18	162	0.11	29	328	0.09
Mexico	1968	1	94	0.01	9	275	0.03
West Germany	1972	26	275	0.09	40	423	0.09
Canada	1976	5	208	0.02	11	385	0.03
Soviet Union	1980	125	410	0.30	195	489	0.40
United States	1984	94	396	0.24	174	522	0.33
South Korea	1988	19	175	0.11	33	401	0.08
Spain	1992	4	229	0.02	22	422	0.05
United States	1996	108	545	0.20	101	647	0.16
Australia	200	41	417	0.10	58	517	0.09
Greece	2004	13	140	0.09	16	426	0.04
China	2008	63	384	0.16	100	599	0.17
Great Britain	2012	47	304	0.15	65	530	0.12
Brazil	2016	17	236	0.07	19	462	0.04
Japan	2021	41	395	0.10	51	621	0.08
Total		682	4715	0.145	1016	7979	0.127

Table 1: The summer Olympic results since 1952

Aggregate analysis

Let λ_0 and λ_1 be the number of medal per participant in previous and home-country games respectively. Our first step will be looking to find both a reasonable likelihood for the data, and a conjugate prior distribution for λ. Since one participant can obtain more than one medals we would use the Poisson likelihood

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

We assume that the observations are independent. Subsequently, we assume uninformative conjugate prior, $Gamma(0.1, 0.1)$, for both λ_0 and λ_1. Then using Bayes' rule the posterior distributions are become Gamma distribution,

$$\lambda_1|Y_1 \propto Gamma(Y_1 + 0.1, N_1 + 0.1) \qquad \lambda_0|Y_0 \propto Gamma(Y_0 + 0.1, N_0 + 0.1)$$

The graph of the posterior distributions are given in Figure 1 and the code is given in the Appendix.

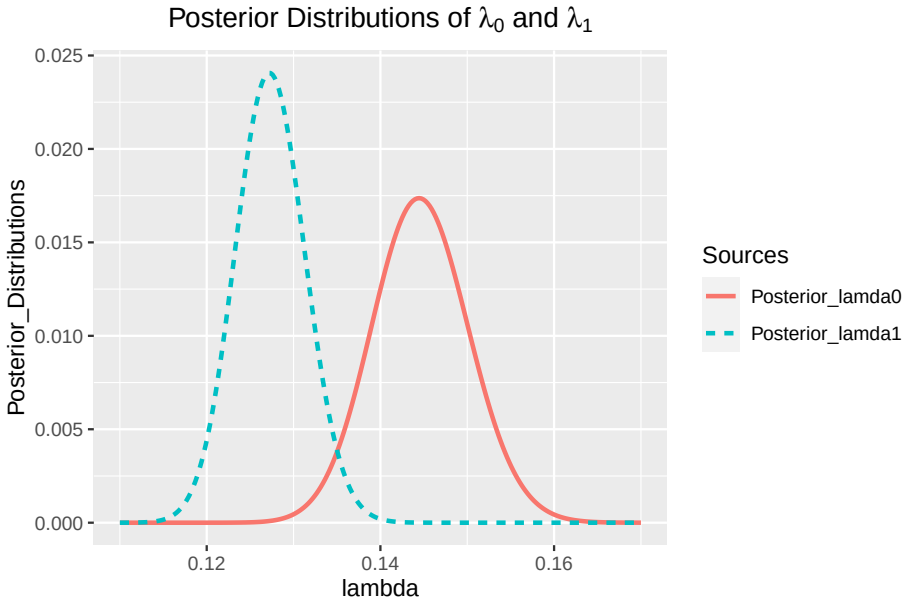


Figure 1

Hypothesis test

We would like to conduct the Bayesian hypothesis test that there is advantage of home country i.e

$$H_0 : \lambda_1 > \lambda_0, \qquad H_a : \lambda_1 \leq \lambda_0.$$

To find the $P(H_0|data)$ we use MC sampling method with $N = 10^6$ sample and to check sensitivity of the our results we used different priors. The results are presented in Table 2. We observe that the results are

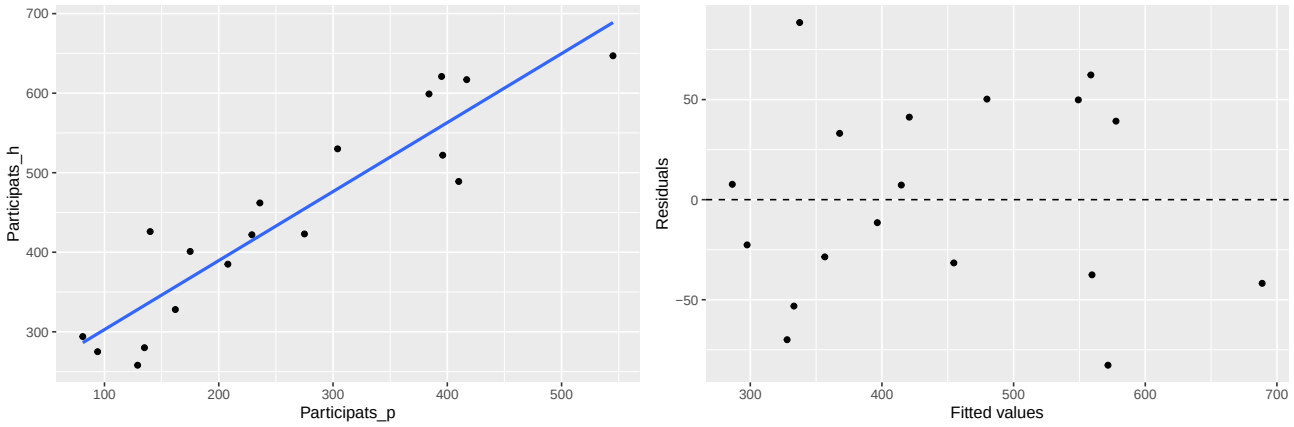
Prior	$Gamma(0.1, 0.1)$	$Gamma(0.5, 0.5)$	$Gamma(1, 1)$	$Gamma(2, 2)$	$Gamma(4, 4)$
$P(H_0\|data)$	0.005293	0.005264	0.005133	0.005101	0.004773

Table 2: Sensitivity table of hypothesis test to the choice of prior

changed slightly as we change the priors, so results are not sensitive to the choice of prior. Moreover the small probability means that we reject null hypothesis in sake of alternative. Therefore, we conclude that there no evidence to support the advantage of the home-country in summer Olympics.

Prediction

To predict the number of participants in France in 2024, we would like to find linear regression with number of participants in previous game as an explanatory variable and number of participant in host game as a response variable. First we obtained the scatter plot shown in Figure 2 (a), and observe that there is linear relationship between explanatory and response variable. Next, we checked the residuals and they are identically and independent distributed. Hence, we can use linear regression. Using linear regression we obtain,



(a) (b)

Figure 2: (a) The linear regression model; (b) Errors of linear regression model

$$\hat{N}_{i1} = 0.86740 * N_{i0} + 216.06833$$

where $\hat{N}_{1i}$ is predicted number of participant from the country in the i' the Olympic. Using this linear regression we predicted that the number of participants from France in 2024 will be 561. Even if we knew parameters, we could not predict Y exactly because there is inherent randomness in which participant get medal. This is quantified using the likelihood distribution $Y|\lambda$ and we can never know λ exactly. This uncertainty is quantified by its prior and posterior distributions. Hence to predict the number of the medals France will win in 2024, we used the Posterior Predictive Distribution (PPD),

$$f(Y^*|Y) = \int f(Y^*|\lambda)p(\lambda|Y)$$

$$p(\lambda|Y) \propto Gamma(33+1, 398+1), \quad f(Y^*|\lambda) \propto Poisson(561\lambda).$$

To approximate PPD we used the Monte Carlo sampling with $N = 10^5$ and the distribution plot is given in Figure 3.

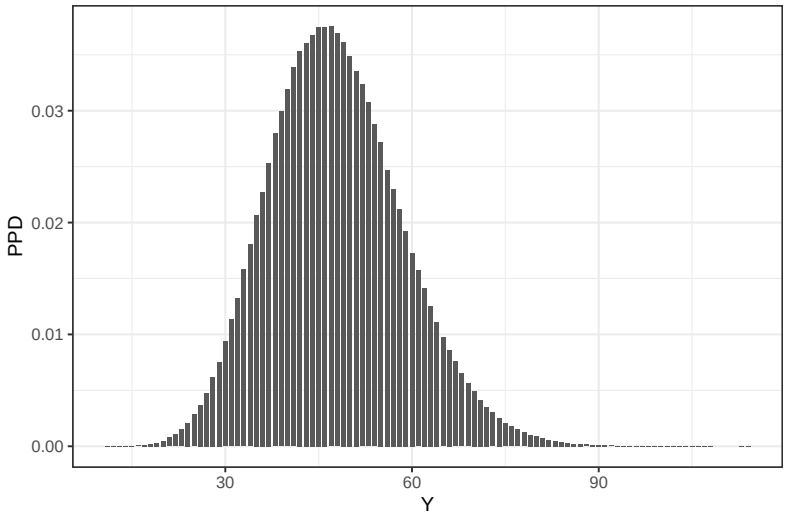


Figure 3: Posterior Predictive Distribution

We observe that the distribution is nearly normal hence the predicted number of the medals France will win is,

$$\mathbb{E}(Y^*|Y) \approx 48.$$

Country-specific analysis

To get the posterior distribution of the of ratio $r = \frac{\lambda_1}{\lambda_0}$ for each country, we use posterior distributions of λ_1 and λ_0. We used the prior for each parameter $Gamma(0.1, 0.1)$ distribution, and Poisson likelihood distribution. Then the the posterior distribution of the of ratio $r = \frac{\lambda_1}{\lambda_0}$ is ratio of posterior distributions of λ_1 and λ_0. The posterior distributions of each country is given in Figure 4. The plots shows strong evidence that the advantage of home-country differs by country.

Conclusions

Using given data and Bayesian analyse we found that there is no string evidence to support overall home-country advantage. However, observe that home-country advantage is different for each country. Home-advantage in summer Olympics is interesting to research and to do better analysis we need much more detailed data. For example:

1. We need data about how many participants have from host country to one position? It may happen that number of participants are increased but the number of different sport positions are stay constant and many participants are played to win one medal.

2. Also we need to discuss about political status of the host country. Due to countries political situation some participants are rejects to participate and it may increase the results of the host country.

3. We also should consider the geographic location of the host country.

After collecting much more detailed data we can do reliable analysis for this question.

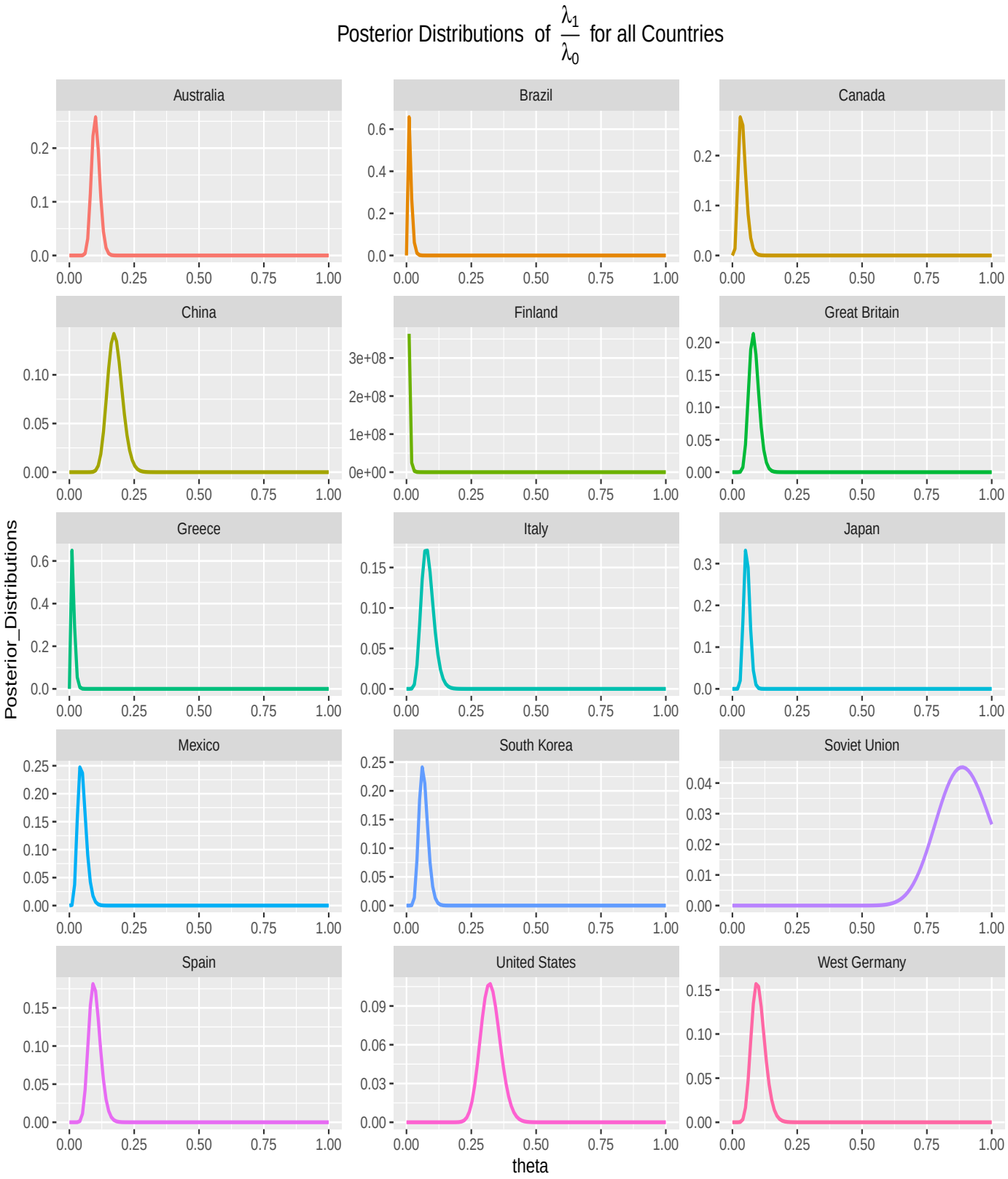


Figure 4: Posterior Distribution of $\frac{\lambda_1}{\lambda_0}$ for each country

Appendix

```
#load given data
Medals <- read_csv("Downloads/R_programming/ST-540/Medals.csv")
```

```r
colnames(Medals)<-c("Country", "year","Medal_p", "Medal_h", "Participats_p","Participats_h")
#extract the totals of each columns
Y0<-sum(Medals$Medal_p)
Y1<-sum(Medals$Medal_h)
N0<-sum(Medals$Participats_p)
N1<-sum(Medals$Participats_h)
# uninformative prior
a<-b<-.1
# posterior parameters
alpha0<-Y0+a
beta0<-N0+b
alpha1<-Y1+a
beta1<-N1+b
# posterior means and variance
alpha0/beta0
alpha1/beta1

alpha0/beta0^2
alpha1/beta1^2
#posterior distributions
lambda = seq(from = 0.11, to = .17, length.out = 250)
dt<-data.frame(lambda)
dt["Posterior_lamda1"] = sapply(lambda, dgamma, alpha1, beta1)/
   + sum(sapply(lambda, dgamma, alpha1, beta1))
dt["Posterior_lamda0"] = sapply(lambda, dgamma, alpha0, beta0)/
   + sum(sapply(lambda, dgamma, alpha0, beta0))

#plot of the posterior distributions
dt%>%
  pivot_longer(cols = "Posterior_lamda1":"Posterior_lamda0", names_to = "Sources",
               values_to = "Posterior_Distributions")%>%
  ggplot(aes(x=lambda, y = Posterior_Distributions, color = Sources, linetype = Sources))+
  geom_line(size=1)+theme(plot.title = element_text(hjust = 0.5))+
  ggtitle(TeX("Posterior Distributions of $\\lambda_0$ and $\\lambda_1$"))

# Prob(lamda1s>lamda0s|data)
S = 1e+6
lamda1s = rgamma(S, Y1+4, N1+4)
lamda0s = rgamma(S, Y0+4, N0+4)
mean(lamda1s>lamda0s)

# scatter plot
Medals%>%
  ggplot(aes(x=Participats_p, y = Participats_h ))+
  geom_point()+
  stat_smooth(method = "lm", se = FALSE)
# linear regression
m1<-lm( Participats_h ~ Participats_p , data = Medals)
 summary(m1)
```

```r
 # Residual plot
 ggplot(data = m1, aes(x = .fitted, y = .resid)) +
   geom_point() +
   geom_hline(yintercept = 0, linetype = "dashed") +
   xlab("Fitted values") +
   ylab("Residuals")

 # Q_Q plot
 ggplot(data = m1, aes(sample = .resid)) +
   stat_qq()
 # predicion
 216.06833 + 0.86740 * 398

 #PPD parameters
 P_fr_host = 561
 medal_fr=33
 P_fr_prev = 398
 A = medal_fr +1
 B = P_fr_prev+1

 # MC sampling to approximate PPD
 lambda<- rgamma(S,A,B)       # Samples from posterior
 Y<-rpois(S,P_fr_host*lambda)
 data.frame(Y)%>%
   ggplot(aes(x=Y))+
   geom_bar(aes(y = ..prop..), stat = "count",width = .65)+
   ylab("PPD")+theme_bw()

# agrregate results of the countries
Medals_uni<-aggregate(.~ Country, data = Medals, sum)%>%
  subset (select = -year)

# posterior distribution of each
theta = seq(from = 0, to = 1, length.out = 100)
df<-data.frame(theta)
for(i in 1:15){
  alpha0 = Medals_uni$Medal_p[i]+.1
  alpha1 = Medals_uni$Medal_h[i]+.1
  beta0 = Medals_uni$Participats_p[i]+.1
  beta1 = Medals_uni$Participats_h[i]+.1
  if(alpha1-alpha0+1>0 && beta1-beta0>0){
    df[Medals_uni$Country[i]] = sapply(theta, dgamma, alpha1-alpha0+1, beta1-beta0)/
    + sum(sapply(theta, dgamma, alpha1-alpha0+1, beta1-beta0))
  }else{

  df[Medals_uni$Country[i]] = (sapply(theta, dgamma, alpha1, beta1)/
  + sapply(theta, dgamma, alpha0, beta0))
  }
}
df%>%
```

```
 pivot_longer(cols = "Australia":"West Germany", names_to = "Countries",
              values_to = "Posterior_Distributions")%>%
 ggplot(aes(x=theta, y = Posterior_Distributions, colour = Countries, ))+
 geom_line(size=1,aes(colour = Countries) )+theme(plot.title = element_text(hjust = 0.5))+
 facet_wrap(Countries~., ncol = 3,scales = "free")+
 ggtitle(TeX("Posterior Distributions  of  $\\frac{\\lambda_1}{\\lambda_0}$ for all Countries"))
)
```