Introduction

Every four years the world's greatest athletes gather together at a host country to compete in a variety of sports to decide who are the three best performers to be awarded a medal for their representing country. A common question that has been asked about the Olympics is if there is a home-country advantage, where it is believed that the rate of medals per participant increases during the host year. The provided data is the number of participants and the number of medals they won for various host countries the Olympics they hosted and the previous Olympics they participated in. The data is for the Summer Olympics from 1952 to 2021. This data can be described with Y_{i1} being the number of medals won by the host country during Olympics i and Y_{i0} being the number of medals won by the same country in the previous Olympics. Similarly, N_{i1} and N_{i0} are the number of participants by the host country during Olympics i and the same country in the previous Olympics respectively. Using this data, this summary will detail the Bayesian analysis to determine if there is statistical backing for the claim that hosting the Olympics leads to a greater medal rate, predicting the number of medals France will win in the 2024 Olympics when they host, and if there is evidence that specific countries have more of a home-country advantage over others.

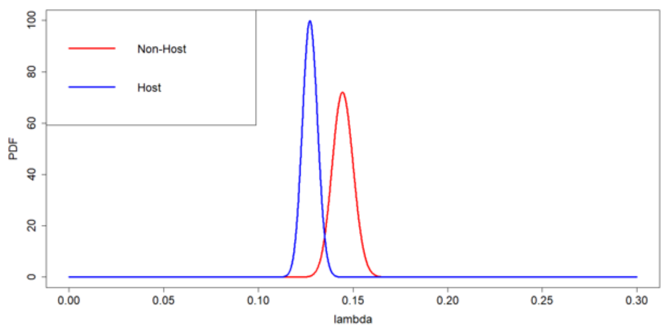
Aggregate Analysis

Beginning with the claim that hosting the Olympics leads to a greater medal rate, the data across all years is aggregated, i.e. $Y_1 = \sum_{i=1}^{18} Y_{i1} = 1016$, $Y_0 = \sum_{i=1}^{18} Y_{i0} = 682$, $N_1 = \sum_{i=1}^{18} N_{i1} = 7979$, and $N_0 = \sum_{i=1}^{18} N_{i0} = 4715$ being the total number of medals won by the host country, the total number of medals won by the same country during the previous Olympics, the total number of participants by the host country, and the total number of participants by the same country during the previous Olympics. To obtain posterior distributions for λ_1 and λ_0, i.e. the expected number of medals per participate by the home country during the host year and during the previous Olympics respectively, the first step is to define what our likelihood and prior distributions are. The data can be reasonably modelled with a Poisson distribution, i.e. $P(Y|\lambda) \sim Poisson(N * \lambda)$, since the number of medals that can be won is a discrete positive real value, multiple medals can be won by each individual so that λ can be greater than 1, and each medal won is independent of each other. A conjugate prior for a Poisson rate is the gamma distribution, $P(\lambda) \sim Gamma(a, b)$, with parameters a and b both equaling 0.10 so that it is uninformative. As stated previously, λ could theoretically be any positive real number and thus the gamma distribution fulfills this support. With this assumed likelihood and prior, the posterior distribution will follow a gamma distribution with parameters $A = a + Y$ and $B = b + N$. Thus, the posterior distribution is $P(\lambda|Y) \sim Gamma(A, B)$.

The main assumptions of this analysis are (1) that there is no difference between the participants in the different countries so that the participants can be aggregated together, (2) the increased participation of the host country is adding participants of equal caliber as those that would be normally competing during a non-host year so that they each have equal probability of winning a medal, and (3) each medal won is independent of each other. Assumption (1) is believed to be valid since all athletes receive similar nutrition and training so that they all have equal probability of winning a medal, although this is debatable when comparing the level of funding athletes receive from country to country. Assumption (2) is also believed to be valid since all athletes need to qualify in order to participate in the Olympics and thus should all have equal probability of winning a medal, but it has been reported in the 538 reference that the qualifications are lowered in the host country so that more athletes can participate. Just because the qualifications are lowered, does not necessarily change the equal probability that everyone has to winning a medal. Assumption (3) is also believed to be valid, taking swimming as a working example, since winning the gold medal in free style does not give you any advantage in your competition in the butterfly stroke.

With the posterior defined to be $P(\lambda|Y) \sim Gamma(a+Y, b+N)$ and the assumptions of the model discussed above, the posterior distributions and their respective summaries for both the host and non-host medal rates for the aggregated data are shown in Figure 1. Comparing the two posterior distributions, one can see that the majority of the non-host posterior distribution, i.e. red curve, is above the host posterior distribution, i.e. blue curve. The spread of the host posterior is less than the spread of the non-host posterior due to more data being avaible to construct the host distribution. Looking at the 2.5% quantile for the non-host distribution and the 97.5% host distribution, there is overlap of the distributions.

Figure 1: Posterior distributions and summary of host and non-host medal rate for the aggregated data.



	λ_0	λ_1
Mean	0.14466	0.12735
Std	0.00003	0.00002
Quantile (2.5%)	0.13401	0.11963
Quantile (97.5%)	0.15572	0.13529

Hypothesis Test

With the posterior distributions from the previous section, a hypothesis test can be conducted to determine the probability that the host medal rate is greater than the non-host medal rate given the data, i.e. $P(\lambda_1 > \lambda_0|Y_1, Y_2)$, to see if there is a home-country advantage. The null hypothesis, H_0, is if $P(\lambda_1 > \lambda_0|Y_1, Y_2) \geq 0.95$ then we can conclude that the home-country advantage does exist. The alternative hypothesis, H_1, is if $P(\lambda_1 > \lambda_0|Y_1, Y_2) < 0.95$ then we can conclude that the home-country advantage does not exist. To obtain an estimate of $P(\lambda_1 > \lambda_0|Y_1, Y_2)$, Monte Carlo (MC) sampling is employed with 100,000 samples from both posterior distributions to determine how much area in the λ_1 distribution, i.e. blue distribution, is greater than the λ_0 distribution, i.e. red distribution. From these simulations it was found that 0.506% of the λ_1 distribution is greater than the λ_0 distribution, so that one can say that there is a 0.506% chance that the true value of λ_1 is greater than the true value for λ_0. Thus, one can be confident from these results that the home-country advantage does not contribute to a significantly greater medal rate per participant.

To determine if the above conclusion is sensitive to the selected prior of Gamma(a=0.1,b=0.1), $P(\lambda_1 > \lambda_0|Y_1, Y_2)$ and summarizing statistics for $P(\lambda_0|Y_0)$ and $P(\lambda_1|Y_1)$ using different priors is shown in Table 1. Note that the summarizing statistics for $P(\lambda_0|Y_0)$ and $P(\lambda_1|Y_1)$ are separated with "/". As can be seen from these results, they are found to not change significantly with the chosen prior. So that we can conclude that the results are somewhat sensitive to the prior but not to a degree that would significantly impact our conclusions since the amount of data we have overpowers the prior.

Table 1: Probability that the host medal rate is greater than the non-host medal rate, i.e. $P(\lambda_1 > \lambda_0|Y_1, Y_2)$, and summary of non-host / host posterior distributions, i.e. $P(\lambda_0|Y_0)$ / $P(\lambda_1|Y_1)$, for different priors

a	b	$P(\lambda_1 > \lambda_0|Y_1, Y_2)$	Mean	Quantile (2.5%)	Quantile (97.5%)
0.10	0.10	0.00506	0.145 / 0.127	0.134 / 0.120	0.156 / 0.135
1	1	0.00491	0.145 / 0.127	0.134 / 0.120	0.156 / 0.135
2	2	0.00475	0.145 / 0.127	0.134 / 0.120	0.156 / 0.135
5	5	0.00465	0.146 / 0.128	0.135 / 0.120	0.157 / 0.136

Prediction

Knowing that France had $N_{F0} = 398$ participants and won $Y_{F0} = 33$ medals in the 2021 summer Olympics, the posterior predictive distribution (PPD) can be derived to estimate the number of medals France will win in the 2024 Olympics when they host and quantify the associated uncertainty. In order to derive the PPD, the number of participants N_{F1} must first be predicted. This is accomplished by utilizing linear regression with the provided data to predict N_1 as a function of N_0. This fit is shown in blue in the left graph of Figure 2. The 95% prediction intervals are also shown in green and are used to include the uncertainty of N_{F1} in the PPD. With this information, it is determined that France should have about 561 ± 114 participants in 2024. Also, in Figure 2 is the linear regression fit of the provided data to predict Y_1 as a function of Y_0 and shows that France should win about 50 medals in 2024.

With an estimate of N_{F1}, the next step is to define the posterior $P(\lambda|Y)$ and the likelihood $f(Y^*|\lambda)$ to be used in a MC sampling procedure to estimate the PPD $f^*(Y^*|Y)$. The posterior is a gamma distribution, $P(\lambda|Y) \sim Gamma(0.10 + Y_{F0}, 0.10 + N_{F0})$, for the reasons explained in the aggregate study. In this analysis, the French participants and medals won in the previous Olympics are being used instead of the aggregate values calculated previously, i.e. N_1 and Y_1, since (1) the previous French performance is assumed to be a better indicator of future performance and (2) the aggregate analysis and the country specific analysis conducted in the next section showed that the home-country advantage does not contribute to a significantly greater medal rate per participant. The likelihood is given with a Poisson distribution, $P(Y^*|\lambda) \sim Poisson(N_{F1} * \lambda)$, for the same reasons as discussed during the aggregate analysis, where N_{F1} follows a normal distribution, $N_{F1} \sim Normal(561{,}52)$, from the linear regression analysis. With 100,000 MC samples, the PPD in red assuming N_{F1} to have uncertainty and the PPD in blue assuming N_{F1} to be equal to 561 is shown in Figure 3. From these results we see that the PPD is not very sensitive to the N_{F1} estimate. Including N_{F1} uncertainty, there is a 95% probability that France will win between 26-72 medals in 2024 with a mean medal count of 46.64, which is close to the predicted 50 medals from linear regression.

Figure 2: Linear regression fit of number of participants (left) and the number of medals won (right)

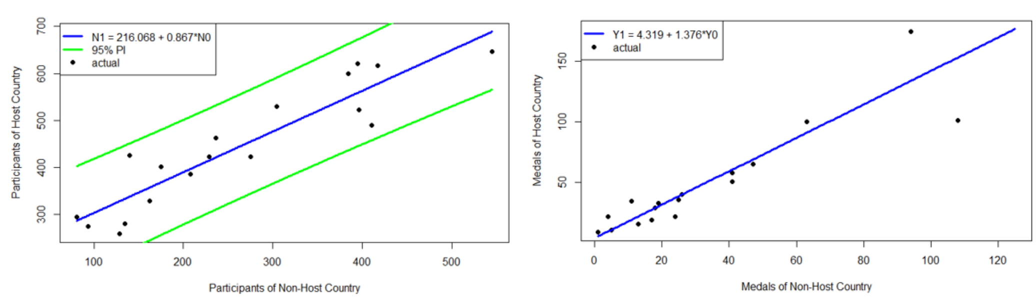
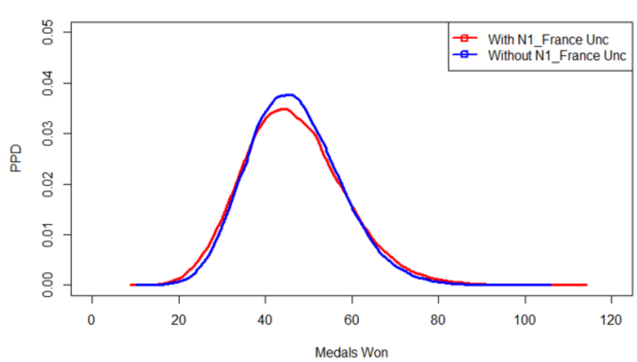


Figure 3: PPD and summary of Y_{F1} when including and not including N_{F1} uncertainty shown in red and blue



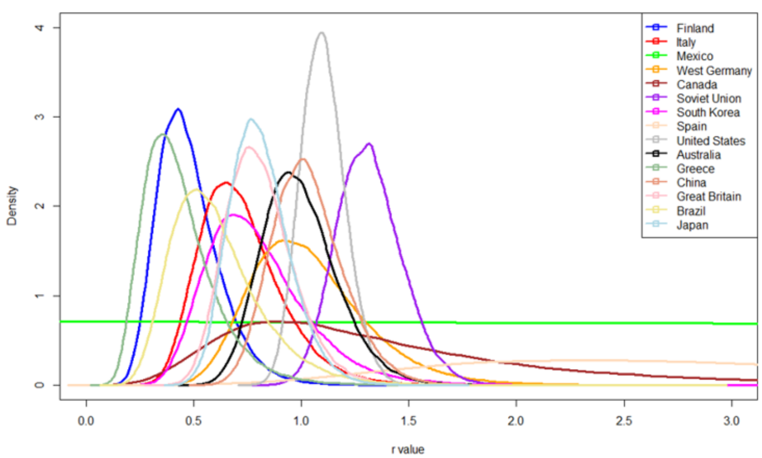
	With N_{F1} Uncertainty	Without N_{F1} Uncertainty
Mean	46.64	46.67
Std	11.61	10.58
Quantile (2.5%)	26	28
Quantile (97.5%)	72	69

Country-Specific Analysis

For the country specific analysis, the medal rate ratio $r_i = \lambda_{i1}/\lambda_{i0}$ is compared to determine if the home-country advantage differs by country. Similarly, to the aggregate analysis, the posterior distributions of λ_{i1} and λ_{i0} are defined to be $P(\lambda_{i1}|Y_{i1}) \sim Gamma(a + Y_{i1}, b + N_{i1})$ and $P(\lambda_{i0}|Y_{i0}) \sim Gamma(a + Y_{i0}, b + N_{i0})$. To obtain these posterior distributions, the likelihood is a Poisson distribution, i.e. $P(Y_i|\lambda_i) \sim Poisson(N_i * \lambda_i)$, while the conjugate prior is the gamma distribution, $P(\lambda_i) \sim Gamma(a, b)$, with parameters $a = b = 0.10$ so that it is uninformative. The likelihood and prior were chosen for the same reasons outlined previously in the aggregate analysis. With 100,000 MC samples, the posterior distributions of the medal rate ratio for each country are shown in Figure 4 and are constructed by sampling from both $P(\lambda_{i1}|Y_{i1})$ and $P(\lambda_{i0}|Y_{i0})$ to obtain a sample from the r_i posterior. From the curves, a medal rate ratio of 1 signifies there is not a home-country advantage, less than 1 signifies the athletes performed worse in their home country, and greater than 1 signifies the athletes performed better in their home country.

To determine if the medal rate ratio is different between countries, a new transformation parameter for the difference between the medal rate ratio of country i and country j, i.e. $\Delta r_{ij} = r_i - r_j$, is defined. Utilizing the 100,000 MC samples utilized to construct the posterior distributions, equal-tailed 95% credible intervals are constructed for Δr_{ij}. If the medal rate ratio is different between country i and country j, then this interval should not contain 0. The countries that were found to differ from one another are those listed in Figure 4.

Figure 4: Posterior distributions of $r_i = \lambda_{i1}/\lambda_{i0}$ and countries that show a difference in r_i (right)



Country 1	Country 2	Country 1	Country 2	Country 1	Country 2
Finland	Mexico	Mexico	Greece	Spain	Australia
Finland	West Germany	Italy	Spain	Spain	Greece
Finland	Soviet Union	West Germany	Greece	Spain	China
Finland	Spain	Soviet Union	Greece	Spain	Great Britain
Finland	USA	Soviet Union	Great Britain	Spain	Brazil
Finland	Australia	Soviet Union	Brazil	Spain	Japan
Finland	China	Soviet Union	Japan	USA	Greece
Italy	Soviet Union	South Korea	Spain	Australia	Greece
West Germany	Spain	Spain	USA	Greece	China

Conclusions

From the analysis performed in this report, the question of if there is a home-country advantage was investigated. From this study it was found that from the aggregated data, there is only a 0.506% chance that the true value of λ_1 is greater than the true value for λ_0. When a similar analysis was done on a country-by-country basis, there is an observed difference with countries performing significantly better than other countries when they host the Olympics. However, the majority of the countries analyzed in this study did not show significant differences from each other or from the year they hosted and the previous year. When attempting to predict the number of medals with uncertainty France will win in 2024 when they host, a PPD was constructed and used to show that there is a 95% probability that France will win between 26-72 medals in 2024 with a mean medal count of 46.64. Two limitations of this analysis are (1) the number of predictors and (2) not having a detailed breakdown of the events being held at each Olympics. For future work, more predictors can be used, such as country population or country GDP and the specific sports at each Olympics can be further explored since host countries tend to add sports that are popular in their country.

```r
##########################################################
#                                                        #
#                      Midterm 1                         #
#                                                        #
#                     Andy Rivas                         #
#                                                        #
#                                                        #
#                                                        #
##########################################################

setwd('C:/Users/aknri/OneDrive/Desktop/Applied Bayesian Stats/Midterm1/')

#Load the data into a Dataframe
medals <- read.csv("Medals.csv")
attach(medals)
set.seed(1)

#####
# a #
#####
# Read in the data for aggregate analysis
N0 = sum(medals$PARTICIPATING.ATHLETES.DURING.PREVIOUS.OLYMPICS)
N1 = sum(medals$PARTICIPATING.ATHLETES.DURING.HOST.YEAR)
Y0 = sum(medals$MEDALS.WON.DURING.PREVIOUS.OLYMPICS)
Y1 = sum(medals$MEDALS.WON.DURING.HOST.YEAR)

lambda0 = Y0/N0
lambda1 = Y1/N1

# define uninformative priors
a = 0.1
b = 0.1

# Determine mean, variance, median, and 95% lower/upper bounds for non-host year
A0 = Y0 + a
B0 = N0 + b
mean_lambda_0 = A0/B0
variance_lambda_0 = A0/B0^2
median_lambda_0 = qgamma(0.50, Y0+a,N0+b)
upper_bound_lambda_0 = qgamma(0.975, Y0+a,N0+b)
lower_bound_lambda_0 = qgamma(0.025, Y0+a,N0+b)

# Determine mean, variance, median, and 95% lower/upper bounds for host year
A1 = Y1 + a
B1 = N1 + b
mean_lambda_1 = A1/B1
variance_lambda_1 = A1/B1^2
median_lambda_1 = qgamma(0.50, Y1+a,N1+b)
upper_bound_lambda_1 = qgamma(0.975, Y1+a,N1+b)
lower_bound_lambda_1 = qgamma(0.025, Y1+a,N1+b)
```

```r

# Show results in a table
result_table_a = round(matrix(c(mean_lambda_0,variance_lambda_0,median_lambda_0,lower_bound_lambda_0,upper_bound_lambda_0,mean_lambda_1
,variance_lambda_1,median_lambda_1,lower_bound_lambda_1,upper_bound_lambda_1), nrow=2, ncol=5, byrow=T),5)
rownames(result_table_a) <- c("lambda_0","lambda_1")
colnames(result_table_a) <- c("Mean","Variance","Median","95% Lower Bound","95% Uppper Bound")
result_table_a<- as.table(result_table_a)
result_table_a

# Graph the posteriors
lambda = seq(0,0.3,0.0001)
posterior_non_host = dgamma(lambda,Y0+a,N0+b)
posterior_host = dgamma(lambda,Y1+a,N1+b)
plot(lambda,posterior_non_host,col="red", type="l",lwd=3, xlab="lambda", ylab="PDF", cex.lab = 1.5, cex.axis = 1.5, ylim = c(0,100))
lines(lambda,posterior_host,col="blue", lwd=3, xlab="lambda", ylab="PDF", cex.lab = 1.5, cex.axis = 1.5)
legend("topleft",legend=c("Non-Host","Host"),col=c("red","blue"),lwd=c(3,3), cex=1.5, pt.cex = 1.5)

#####
# b #
#####
S = 100000

# Determine if there is a home country advantage
# Prior 1
lambda_0_MC <- rgamma(S,Y0+a,N0+b)
lambda_1_MC <- rgamma(S,Y1+a,N1+b)

mean_lambda_0 = (Y0+a)/(N0+b)
upper_bound_lambda_0 = qgamma(0.975, Y0+a,N0+b)
lower_bound_lambda_0 = qgamma(0.025, Y0+a,N0+b)

mean_lambda_1 = (Y1+a)/(N1+b)
upper_bound_lambda_1 = qgamma(0.975, Y1+a,N1+b)
lower_bound_lambda_1 = qgamma(0.025, Y1+a,N1+b)

# Determine if there is a home country advantage using different priors
# Prior 2
a1 = 1
b1 = 1
lambda_0_MC_1 <- rgamma(S,Y0+a1,N0+b1)
lambda_1_MC_1 <- rgamma(S,Y1+a1,N1+b1)

mean_lambda_0_1 = (Y0+a1)/(N0+b1)
upper_bound_lambda_0_1 = qgamma(0.975, Y0+a1,N0+b1)
lower_bound_lambda_0_1 = qgamma(0.025, Y0+a1,N0+b1)

mean_lambda_1_1 = (Y1+a1)/(N1+b1)
upper_bound_lambda_1_1 = qgamma(0.975, Y1+a1,N1+b1)
lower_bound_lambda_1_1 = qgamma(0.025, Y1+a1,N1+b1)
```

```r

# Prior 3
a2 = 2
b2 = 2
lambda_0_MC_2 <- rgamma(S,Y0+a2,N0+b2)
lambda_1_MC_2 <- rgamma(S,Y1+a2,N1+b2)

mean_lambda_0_2 = (Y0+a2)/(N0+b2)
upper_bound_lambda_0_2 = qgamma(0.975, Y0+a2,N0+b2)
lower_bound_lambda_0_2 = qgamma(0.025, Y0+a2,N0+b2)

mean_lambda_1_2 = (Y1+a1)/(N1+b1)
upper_bound_lambda_1_2 = qgamma(0.975, Y1+a2,N1+b2)
lower_bound_lambda_1_2 = qgamma(0.025, Y1+a2,N1+b2)

# Prior 4
a3 = 5
b3 = 5
lambda_0_MC_3 <- rgamma(S,Y0+a3,N0+b3)
lambda_1_MC_3 <- rgamma(S,Y1+a3,N1+b3)

mean_lambda_0_3 = (Y0+a3)/(N0+b3)
upper_bound_lambda_0_3 = qgamma(0.975, Y0+a3,N0+b3)
lower_bound_lambda_0_3 = qgamma(0.025, Y0+a3,N0+b3)

mean_lambda_1_3 = (Y1+a3)/(N1+b3)
upper_bound_lambda_1_3 = qgamma(0.975, Y1+a3,N1+b3)
lower_bound_lambda_1_3 = qgamma(0.025, Y1+a3,N1+b3)

# Show results in a tables
result_table_b = round(matrix(c(a,b,100*mean(lambda_1_MC>lambda_0_MC),a1,b1,100*mean(lambda_1_MC_1>lambda_0_MC_1),a2,b2,100*mean
(lambda_1_MC_2>lambda_0_MC_2),a3,b3,100*mean(lambda_1_MC_3>lambda_0_MC_3)), nrow=4, ncol=3, byrow=T),3)
rownames(result_table_b) <- c("Prior_1","Prior_2","Prior_3","Prior_4")
colnames(result_table_b) <- c("a","b","Prob")
result_table_b <- as.table(result_table_b)
result_table_b

result_table_b_0 = round(matrix(c(a,b,mean_lambda_0,upper_bound_lambda_0,lower_bound_lambda_0,a1,b1,mean_lambda_0_1,upper_bound_lambda_0_1
,lower_bound_lambda_0_1,a2,b2,mean_lambda_0_2,upper_bound_lambda_0_2,lower_bound_lambda_0_2,a3,b3,mean_lambda_0_3,upper_bound_lambda_0_3
,lower_bound_lambda_0_3), nrow=4, ncol=5, byrow=T),3)
rownames(result_table_b_0) <- c("Prior_1","Prior_2","Prior_3","Prior_4")
colnames(result_table_b_0) <- c("a","b","Mean","Quantile (2.5%)","Quantile (97.5%)")
result_table_b_0 <- as.table(result_table_b_0)
result_table_b_0

result_table_b_1 = round(matrix(c(a,b,mean_lambda_1,upper_bound_lambda_1,lower_bound_lambda_1,a1,b1,mean_lambda_1_1,upper_bound_lambda_1_1
,lower_bound_lambda_1_1,a2,b2,mean_lambda_1_2,upper_bound_lambda_1_2,lower_bound_lambda_1_2,a3,b3,mean_lambda_1_3,upper_bound_lambda_1_3
,lower_bound_lambda_1_3), nrow=4, ncol=5, byrow=T),3)
rownames(result_table_b_1) <- c("Prior_1","Prior_2","Prior_3","Prior_4")
colnames(result_table_b_1) <- c("a","b","Mean","Quantile (2.5%)","Quantile (97.5%)")
result_table_b_1 <- as.table(result_table_b_1)
result_table_b_1

```

```r
#####
# c #
#####
# Load known data
N0_France = 398
Y0_France = 33
a = 0.10
b = 0.10

# First predict N1_France, i.e. the number of participants in 2024 from France, using LR
# Plot the fit with data points of other countries and 95% prediction intervals
x = medals$PARTICIPATING.ATHLETES.DURING.PREVIOUS.OLYMPICS
y = medals$PARTICIPATING.ATHLETES.DURING.HOST.YEAR
fit <- lm(y~x)
summary(fit)   # fit summary
samples <- seq(min(x),max(x),1)
Host_Participants = fit$coefficients[1] + fit$coefficients[2]*samples
pred_interval_N0_France = predict(fit,data.frame(x=N0_France),level=0.95,interval='prediction')
pred_interval = predict(fit,data.frame(x=c(samples)),level=0.95,interval='prediction')
plot(x,y, pch=19,xlab="Participants of Non-Host Country", ylab="Participants of Host Country", xlim=c(min(x),max(x)),ylim=c(min(y),max
(Host_Participants)))
lines(samples,Host_Participants,col='blue',lwd=3)
lines(samples,pred_interval[,2],col='green',lwd=3)
lines(samples,pred_interval[,3],col='green',lwd=3)
legend("topleft",legend=c("N1 = 216.068 + 0.867*N0","95% PI","actual"),lwd=c(3,3,NA),pch=c(NA,NA,19),lty=c(1,1,NA),col=c("blue","green"
,"black"))

# Predict Y1_France, i.e. the number of medals won in 2024 from France, using LR to compare with PPD
x_medals = medals$MEDALS.WON.DURING.PREVIOUS.OLYMPICS
y_medals = medals$MEDALS.WON.DURING.HOST.YEAR
fit_medals <- lm(y_medals~x_medals)
summary(fit_medals)   # fit summary
samples_medals <- seq(min(x_medals),max(x_medals),1)
Host_medals = fit_medals$coefficients[1] + fit_medals$coefficients[2]*samples_medals
plot(x_medals,y_medals, pch=19,xlab="Medals of Non-Host Country", ylab="Medals of Host Country", xlim=c(min(x_medals),max(x_medals)),ylim=c
(min(y_medals),max(Host_medals)))
lines(samples_medals,Host_medals,col='blue',lwd=3)
legend("topleft",legend=c("Y1 = 4.319 + 1.376*Y0","actual"),lwd=c(3,NA),pch=c(NA,19),lty=c(1,NA),col=c("blue","black"))

# Including N1_France uncertainty in the PPD, MC sampling is performed and summary statistics are computed
S = 100000
set.seed(1)
Y_star_w_unc = rep(0,S)
index = 1
for (i in 1:S){
  lambda_star = rgamma(1,Y0_France+a,N0_France+b)
  N1_France_unc = rnorm(1,pred_interval_N0_France[,1],(pred_interval_N0_France[,3]-pred_interval_N0_France[,1])/2)
  Y_star_w_unc[index] = rpois(1,N1_France_unc*lambda_star)
  index = index + 1
}
hist(Y_star_w_unc)
Y_star_w_unc_mean = mean(Y_star_w_unc)
Y_star_w_unc_sd = sd(Y_star_w_unc)
Y_star_w_unc_2_5_quantile = quantile(Y_star_w_unc,0.025)
Y_star_w_unc_97_5_quantile = quantile(Y_star_w_unc,0.975)
```

```r

# Not including N1_France uncertainty in the PPD, MC sampling is performed and summary statistics are computed
lambda_star = rgamma(S,Y0_France+a,N0_France+b)
N1_France = fit$coefficients[1] + fit$coefficients[2]*N0_France
Y_star = rpois(S,N1_France*lambda_star)
hist(Y_star)
Y_star_mean = mean(Y_star)
Y_star_sd = sd(Y_star)
Y_star_2_5_quantile = quantile(Y_star,0.025)
Y_star_97_5_quantile = quantile(Y_star,0.975)

# Plot with and without N1_France uncertainty in PPD
plot(NULL, xlim=c(0,120), ylim=c(0,0.05), ylab="PPD", xlab="Medals Won")
dens_w_N1_unc = density(Y_star_w_unc)
dens_wo_N1_unc = density(Y_star)
lines(dens_w_N1_unc$x,length(data)*dens_w_N1_unc$y,type="l",col='red',lwd=3)
lines(dens_wo_N1_unc$x,length(data)*dens_wo_N1_unc$y,type="l",col='blue',lwd=3)
legend("topright",legend=c('With N1_France Unc','Without N1_France Unc'), pch = rep(0,2),lwd=rep(2,2),col=c('red','blue'), cex=1, pt.cex = 1
)

# Show results in a table
result_table_PPD = matrix(c(Y_star_w_unc_mean,Y_star_w_unc_sd,Y_star_w_unc_2_5_quantile,Y_star_w_unc_97_5_quantile,Y_star_mean,Y_star_sd
,Y_star_2_5_quantile,Y_star_97_5_quantile), nrow=2, ncol=4, byrow=T)
rownames(result_table_PPD) <- c("With N1_France Uncertainty","Without N1_France Uncertainty")
colnames(result_table_PPD) <- c("Mean","Std","Quantile (2.5%)","Quantile (97.5%)")
result_table_PPD <- as.table(result_table_PPD)
result_table_PPD

#####
# d #
#####
# Load in Data
N0_per_country = c(129, 135, 94, 275, 208, 410, 175, 229, 941, 498, 140, 384, 304, 236, 557)
N1_per_country = c(258, 280,  275,  423,  385, 489, 401, 422, 1169, 911, 426, 599, 530, 462, 949)
Y0_per_country = c(24, 25, 1, 26, 5, 125, 19, 4, 202, 52, 13, 63, 47, 17, 59)
Y1_per_country = c(22, 36, 9, 40, 11, 195, 33, 22, 275, 93, 16, 100, 65, 19, 80)
Country_names = c('Finland',  'Italy',  'Mexico', 'West Germany', 'Canada', 'Soviet Union', 'South Korea',  'Spain',  'United States',
'Australia',  'Greece', 'China',  'Great Britain',  'Brazil', 'Japan')

lambda_0_per_country = Y0_per_country/N0_per_country
lambda_1_per_country = Y1_per_country/N1_per_country
r_per_country = lambda_1_per_country/lambda_0_per_country

# Determine if there is a home country advantage
S = 100000
set.seed(1)
a = 0.10
b = 0.10

```

```r
# Define vectors
r_mean_per_country = rep(0,length(Country_names))
r_std_per_country = rep(0,length(Country_names))
r_2_5_quantile_per_country = rep(0,length(Country_names))
r_97_5_quantile_per_country = rep(0,length(Country_names))

r_MC_list = list()

# Plot posteriors of r for all countries and save summary statistics in vectors
plot(NULL, xlim=c(0,3), ylim=c(0,4), ylab="Density", xlab="r value")
colors = c('blue','red','green','orange','brown','purple','magenta','peachpuff','grey','black','darkseagreen','darksalmon','pink','khaki','lightblue')
for (i in 1:length(Country_names)){
  lambda_0_MC_per_country <- rgamma(S,Y0_per_country[i]+a,N0_per_country[i]+b)
  lambda_1_MC_per_country <- rgamma(S,Y1_per_country[i]+a,N1_per_country[i]+b)
  r_MC_per_country = lambda_1_MC_per_country/lambda_0_MC_per_country

  r_MC_list[i] = list(r_MC_per_country)

  r_mean_per_country[i] = mean(r_MC_per_country)
  r_std_per_country[i] = sd(r_MC_per_country)
  r_2_5_quantile_per_country[i] = quantile(r_MC_per_country,0.025)
  r_97_5_quantile_per_country[i] = quantile(r_MC_per_country,0.975)
  dens = density(r_MC_per_country)
  lines(dens$x,length(data)*dens$y,type="l",col=colors[i],lwd=3)
}
legend("topright",legend=Country_names[1:15], pch = rep(0,15),lwd=rep(2,15),col=colors, cex=1, pt.cex = 1)

# Show results in a table
result_table_r = matrix(c(r_mean_per_country,r_std_per_country,r_2_5_quantile_per_country,r_97_5_quantile_per_country), nrow=length(Country_names), ncol=4, byrow=F)
rownames(result_table_r) <- Country_names
colnames(result_table_r) <- c("Mean","Std","Quantile (2.5%)","Quantile (97.5%)")
result_table_r <- as.table(result_table_r)
result_table_r

```

```r
# Determine which countries significantly differ from one another in their medal rate ratio r
# Only output result of pairings once, e.g. Finland vs USA is the same conclusion as USA vs Finland
Country_name_i = rep(0,(length(Country_names)*length(Country_names)-length(Country_names))/2)
Country_name_j = rep(0,(length(Country_names)*length(Country_names)-length(Country_names))/2)
r_diff_mean_per_country = rep(0,(length(Country_names)*length(Country_names)-length(Country_names))/2)
r_diff_std_per_country = rep(0,(length(Country_names)*length(Country_names)-length(Country_names))/2)
r_diff_2_5_quantile_per_country = rep(0,(length(Country_names)*length(Country_names)-length(Country_names))/2)
r_diff_97_5_quantile_per_country = rep(0,(length(Country_names)*length(Country_names)-length(Country_names))/2)
r_diff_decision_per_country = rep(0,(length(Country_names)*length(Country_names)-length(Country_names))/2)
index = 1
number_of_true_differences = 0
for (i in 1:length(Country_names)){
  for (j in 1:length(Country_names)){
    if (i==j) {
      next
    }
    if (i>j) {
      next
    }
    Country_name_i[index] = Country_names[i]
    Country_name_j[index] = Country_names[j]
    r_diff_mean_per_country[index] = mean(r_MC_list[[i]]-r_MC_list[[j]])
    r_diff_std_per_country[index] = sd(r_MC_list[[i]]-r_MC_list[[j]])
    r_diff_2_5_quantile_per_country[index] = quantile(r_MC_list[[i]]-r_MC_list[[j]],0.025)
    r_diff_97_5_quantile_per_country[index] = quantile(r_MC_list[[i]]-r_MC_list[[j]],0.975)
    r_diff_decision_per_country[index] = ifelse(quantile(r_MC_list[[i]]-r_MC_list[[j]],0.025)*quantile(r_MC_list[[i]]-r_MC_list[[j]],0.975) > 0,"True","False")
    if (quantile(r_MC_list[[i]]-r_MC_list[[j]],0.025)*quantile(r_MC_list[[i]]-r_MC_list[[j]],0.975) > 0) {
      number_of_true_differences = number_of_true_differences + 1
    }
    index = index + 1
  }
}

# Show results in a table. If final column has True then there is a difference in r for the two countries
result_table_r_diff = matrix(c(Country_name_i,Country_name_j,r_diff_mean_per_country,r_diff_std_per_country,r_diff_2_5_quantile_per_country,r_diff_97_5_quantile_per_country,r_diff_decision_per_country), nrow=length( Country_name_i), ncol=7, byrow=F)
colnames(result_table_r_diff) <- c("Country 1","Country 2","Mean","Std","Quantile (2.5%)","Quantile (97.5%)","Difference?")
result_table_r_diff <- as.table(result_table_r_diff)
result_table_r_diff
number_of_true_differences


detach(medals)
```