ST540 Midterm1

Shi Cen

02/17 2022

Introduction

There is heated discussion of Olympic Games that Olympic atheletes from the host country would take more gold medals in summer Olympics than they did when they were competing in other countries. Based on the previous data, it does seem that countries that are hosting the Summer Olympics will tend to win more medals. This could come from many results. However, the expected number of medals per participant for each country in Olympics has not been investigated yet. In this paper, we compared the expected number of medals per participant between countries that were hosting and were not using medal and participant data from 1952 to 2021.

1 Aggregate Analysis

To start with, according to the definition, we would have $Y_1 = \sum_{i=1}^{18} Y_{i1} = 1016$ and $N_1 = \sum_{i=1}^{18} N_{i1} = 7979$. Let λ_1 be the the expected number of medals per participant in their home country. The data could be modelled with a Poisson distribution with $Y_1 = 1016$ successes in $N_1 = 7979$ events. Therefore, we have the likelihood function for Y_1 given λ_1

$$Y_1|\lambda_1 \sim Poisson(N_1\lambda_1).$$

Here, we choose a Gamma prior

$$\lambda_1 \sim Gamma(a = 0.1, b = 0.1)$$

as a uninformative conjugate prior because it has a high variance $var(\lambda_1) = \frac{a}{b^2} = 10$ compared to the range of expected value. Therefore, we have the posterior distribution

$$p(\lambda_1|Y_1) \propto f(Y_1|\lambda_1)\pi(N_1\lambda_1) \propto exp(-N_1\lambda_1)\lambda_1^{N_1Y_1}\lambda_1^{a-1}exp(b\lambda)$$

$$\Rightarrow p(\lambda_1|Y_1) \propto exp(-(N_1+b)\lambda_1)\lambda_1^{Y_1+a-1}$$

$$\Rightarrow \lambda_1|Y_1 \sim Gamma(Y_1+a, N_1+b)$$

Similarly, we could have

$$\lambda_0|Y_0 \sim Gamma(Y_0+a, N_0+b)$$

The distributions of these two λ were shown in Figure 1.

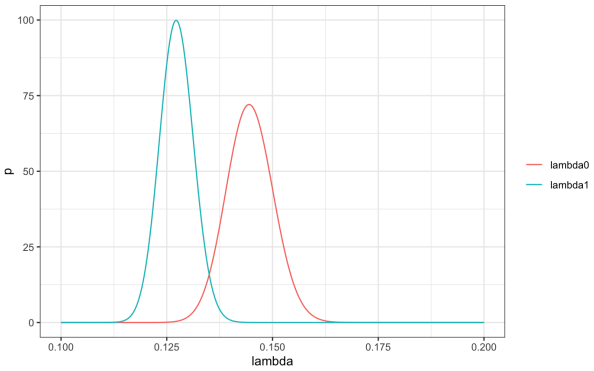


Figure 1: The distribution of λ_1 and λ_0. Here, red lines stands for λ_0 and blue lines stands for λ_1

Here, the main assumption is that every participant has the same rate to get a medal and that the probability for every participant to get medals is independent. Under the purpose of analyzing aggregate data, we could believe that these assumption were valid.

2 Hypothesis Test

$H_0 : \lambda_1 - \lambda_0 > 0$, that is, the the expected number of medals per participant won by home country in Olympics is larger than that of the same country won in the previous Olympics.
$H_\alpha : \lambda_1 - \lambda_0 \leq 0$, that is, the the expected number of medals per participant won by home country in Olympics is no larger than that of the same country won in the previous Olympics.

We can use the Monte Carlo sampling to perform the hypothesis test and calculate the difference and its distribution from posterior distributions. The distribution of the difference between λ_1 and λ_0 from 100,000 MC sampling was shown in Figure 2. The probability of $\lambda_1 > \lambda_0$ in MC sampling data is 0.0052. This probability is < 0.95. Therefore, we do not reject the null hypothesis. In other words, we are only 0.52% certain that $\lambda_1 > \lambda_0$.

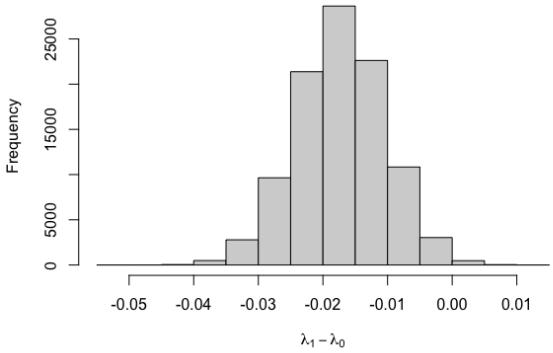


Figure 2: Distribution of $\lambda_1 - \lambda_0$ from MC sampling

We could set $a_2 = b_2 = 1$, $a_3 = b_3 = 10$ and $a_4 = b_4 = 0.01$. Based on these parameter values probability for each the sampling set was shown in table 1. The result showed that when we changed a and b, the probability changed

	$a = b = 0.1$	$a = b = 1$	$a = b = 10$	$a = b = 0.01$
probability of $\lambda_1 > \lambda_0$	0.0052	0.00521	0.0039	0.0049

Table 1: probability of $\lambda_1 > \lambda_0$ under different a and b value

a bit, especially when a and b increase, which would causing the variance of the distribution close to the range of expected value of λ. Therefore, this result was somewhat sensitive to the prior.

3 Prediction

By looking at the scatterplot between N_{i0} and N_{i1} (Figure 3), we could see a correlation between the number of participant for each country in Olympics when they were hosting and the number in the previous one. Using this data distribution pattern, we could make a regression model and use the model to predict the participants in 2024 in France based on their 2021's participant number, which is 561.

When predicting, we still assume that the success rate for each athlete is the same including athletes in France when they're hosting. Therefore, they share a same λ_1 distribution $\lambda_1 \sim Gamma(Y_1 + a, N_1 + b)$. To predict the medal number in 2024 for France, we could use the Posterior Predictive Distribution (PPD) method. Using MC sampling,

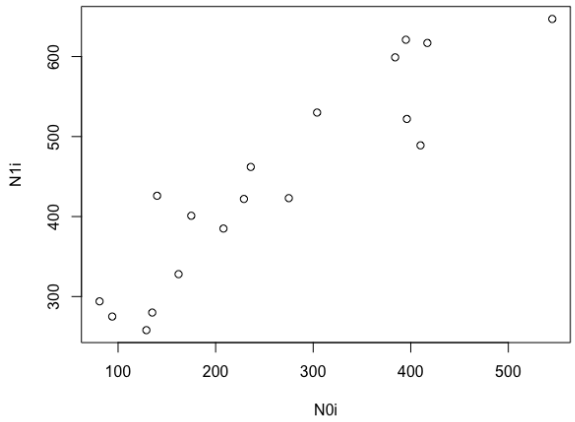


Figure 3: Scatterplot between N_{i0} and N_{i1}

we could generate 100,000 samples $\lambda_1^{(1)}, \lambda_1^{(2)} ... \lambda_1^{(100,000)}$ from the posterior. Since we have $Y^{*(s)} \sim f(Y|\lambda_1^{(s)})$ for each $\lambda_1^{(s)}$, the posterior predictive mean is approximated by the sample mean of $Y^{*(s)}$. Here,

$$f(Y|\lambda_1^{(s)}) \sim Poisson(N_1\lambda_1^{(s)}).$$

Through MC sampling, we could have $\bar{Y^*} = 71$, which is the posterior predictive value for the medal number of France in 2024 Olympics. The 95% credible interval is $(55, 89)$. The distribution of Y^* was shown in Figure 4.

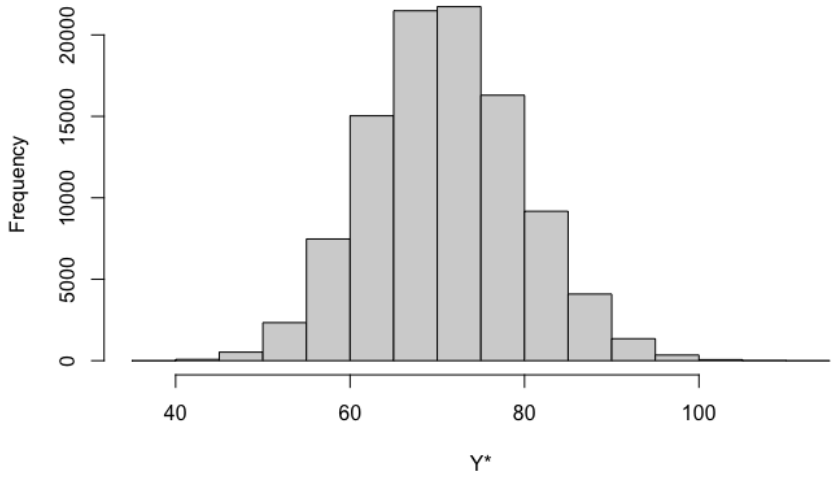


Figure 4: Distribution of Y^*

4 Country-specific analysis

For each country i, we have the likelihood

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

which is the medal numbers given the number of participants for hosting and the previous Olympics. The conjugate priors

$$\lambda_{i1} \sim Gamma(a = 0.1, b = 0.1), \lambda_{i0} \sim Gamma(a = 0.1, b = 0.1).$$

indicate the uninformative distribution of λ_{i1} and λ_{i0} Therefore, we would have the posterior of the rate of winning medals for each participant in each country.

$$\lambda_{i1}|Y_{i1} \sim Gamma(Y_{i1} + a, N_{i1} + b), \lambda_{i1}|Y_{i0} \sim Gamma(Y_{i0} + a, N_{i0} + b).$$

Therefore, we would have $r_i = \frac{\lambda_{i1}}{\lambda_{i0}}, (i = 1, 2, ..., 15)$ for 15 countries. Using MC sampling, we could generate an approximate distribution of r_i (Figure 5).

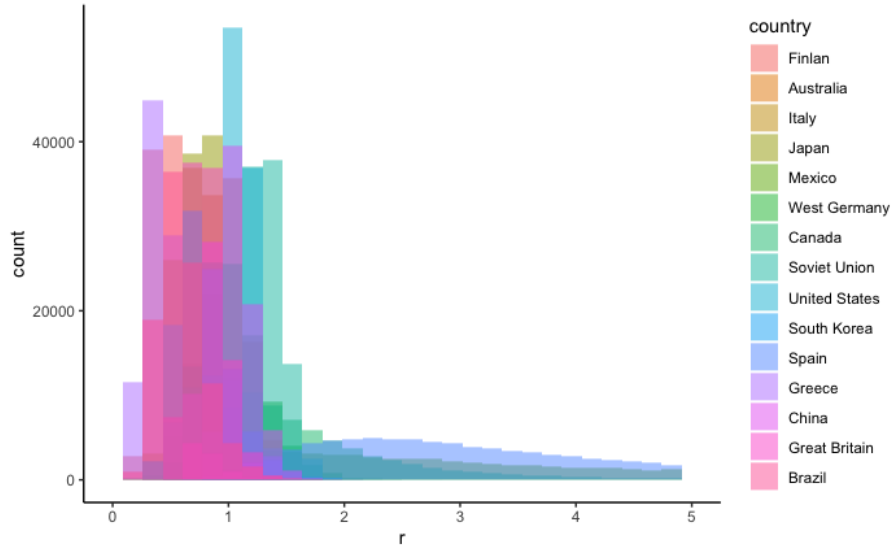


Figure 5: Distribution of r_i

However, it would be difficult to estimate the home-country advantage difference by only looking at their distribution. Hence, we could consider their 95% CI to compare their difference, which were shown in table 2.

	FIN	AUS	ITA	JPN	MEX	FRG	CAN	URS	USA	KOR	ESP	GRE	CHN	GBR	BRA
2.5%	0.26	0.70	0.42	0.57	0.64	0.62	0.44	1.05	0.91	0.44	1.18	0.19	0.75	0.55	0.30
97.5%	0.82	1.39	1.17	1.12	83.71	1.67	3.93	1.64	1.32	1.38	11.03	0.86	1.40	1.16	1.12

Table 2: 95% CI for each r_i

We can also do pair-wise analysis, calculating the 95% CI for every $r_i - r_j$ $(i = 1, ..., 15, j = 1, ..., 15, i \neq j)$. Twenty-seven CIs out of $\binom{15}{2} = 105$ did not contain 0 (data not shown, see the code in appendix). All these evidence supported that the ratio r between the winning rate will differ from country to country.

5 Conclusions

In this paper, we discussed whether there is a home-country advantage in Olympic Games. If we only look at the number of medal that each country gets in Olympics, we might draw the conclusion that there was a home-country advantage. However, if we took the number of participants into consideration and compared the rate of winning a medal for each participants, our results showed that one country will perform no better when it is hosting the Olympics. That is because that when a country is hosting the Olympics, the number of participants would also increase dramatically. Furthermore, if we explore the country-specific data, for the 15 countries in the data set, the ratios of the rates of winning medals in the Olympics they were hosting and in the previous one were also different from each other. Some countries, like USA, would have higher ratio, indicating a stronger home-country advantage.

However, there were also some limitation in this paper. Firstly, we assumed that every participant had the same rate of winning the medal in the aggregate analysis, which is not realistic. In other words, in the future work when we discuss the home-country advantage, we should do it case by case or classify those countries into different categories based on their competitiveness for aggregate analysis and also separate athletes when we do country-specific analysis to see the home-country advantage. Secondly, in this study, we only considered the previous Olympics performance for a country when comparing to the Olympics that was hosted by that country. If we want to perform a more comprehensive analysis, we should also consider a country's performance throughout all Olympics.

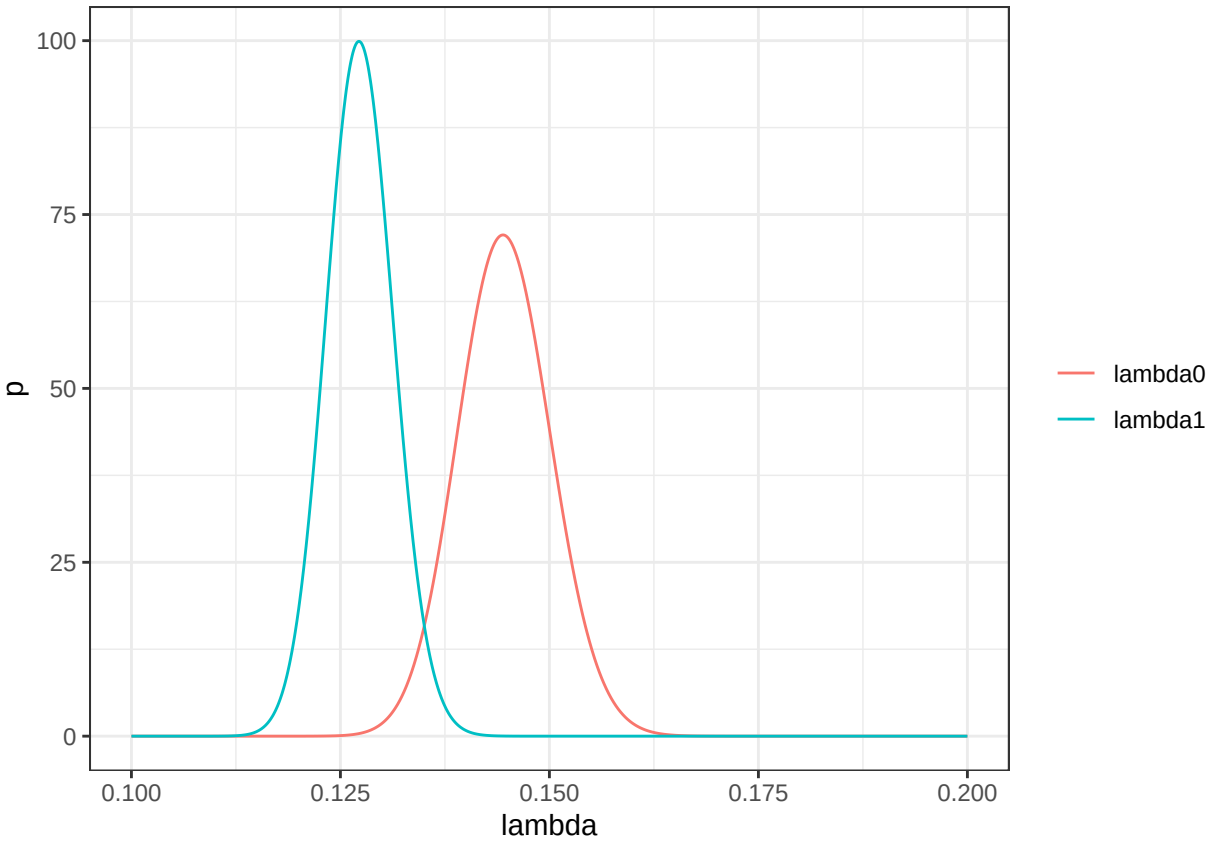
Midterm 1

Shi Cen

2/15/2022

1. Aggregate analisys

```
Y1i <- c(22,35,36,29,9,40,11,195,174,33,22,101,58,16,100,65,19,51)
Y0i <- c(24,11,25,18,1,26,5,125,94,19,4,108,41,13,63,47,17,41)
N1i <- c(258,294,280,328,275,423,385,489,522,401,422,647,617,426,599,530,462,621)
N0i <- c(129,81,135,162,94,275,208,410,396,175,229,545,417,140,384,304,236,395)

Y1 <- sum(Y1i)
N1 <- sum(N1i)

a <- 0.1
b <- 0.1

A1 <- Y1 + a
B1 <- N1 + b

# lambda1_hat <- A1 / B1

Y0 <- sum(Y0i)
N0 <- sum(N0i)

A0 <- Y0 + a
B0 <- N0 + b

# lambda0_hat <- A0 / B0
x <- seq(0.1,0.2,0.0001)
lambda <- rep(x,2)
p <- c(dgamma(x,A1,B1),dgamma(x,A0,B0))
group <- c(rep('lambda1',length(x)),rep('lambda0',length(x)))
df <- data.frame(lambda=lambda,p=p,group=group)

library(ggplot2)
g <- ggplot(df, aes(x=lambda,y=p,group=group)) + geom_line(aes(color=group)) +
  theme_bw() + theme(legend.title = element_blank())
g
```

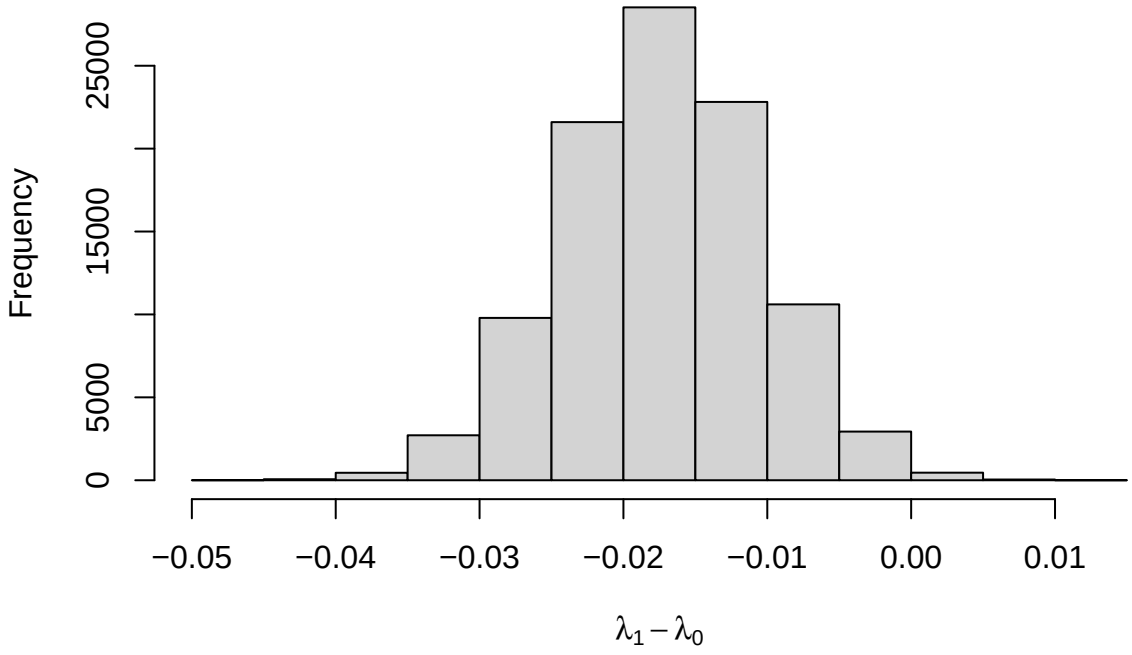


2. Hypothesis test

```
S <- 100000
lambda_0 <- rgamma(S, A0, B0)
lambda_1 <- rgamma(S, A1, B1)
mean(lambda_1 > lambda_0)
```

```
## [1] 0.00501
```

```
hist(lambda_1-lambda_0, xlab = expression(lambda[1]-lambda[0]), main = NULL)
```




```
a2 <- 1
b2 <- 1

lambda_0_2 <- rgamma(S, Y0 + a2, N0 + b2)
lambda_1_2 <- rgamma(S, Y1 + a2, N1 + b2)
mean(lambda_1_2 > lambda_0_2)
```

```
## [1] 0.00538
```

```
a4 <- 10
b4 <- 10

lambda_0_4 <- rgamma(S, Y0 + a4, N0 + b4)
lambda_1_4 <- rgamma(S, Y1 + a4, N1 + b4)
mean(lambda_1_4 > lambda_0_4)
```

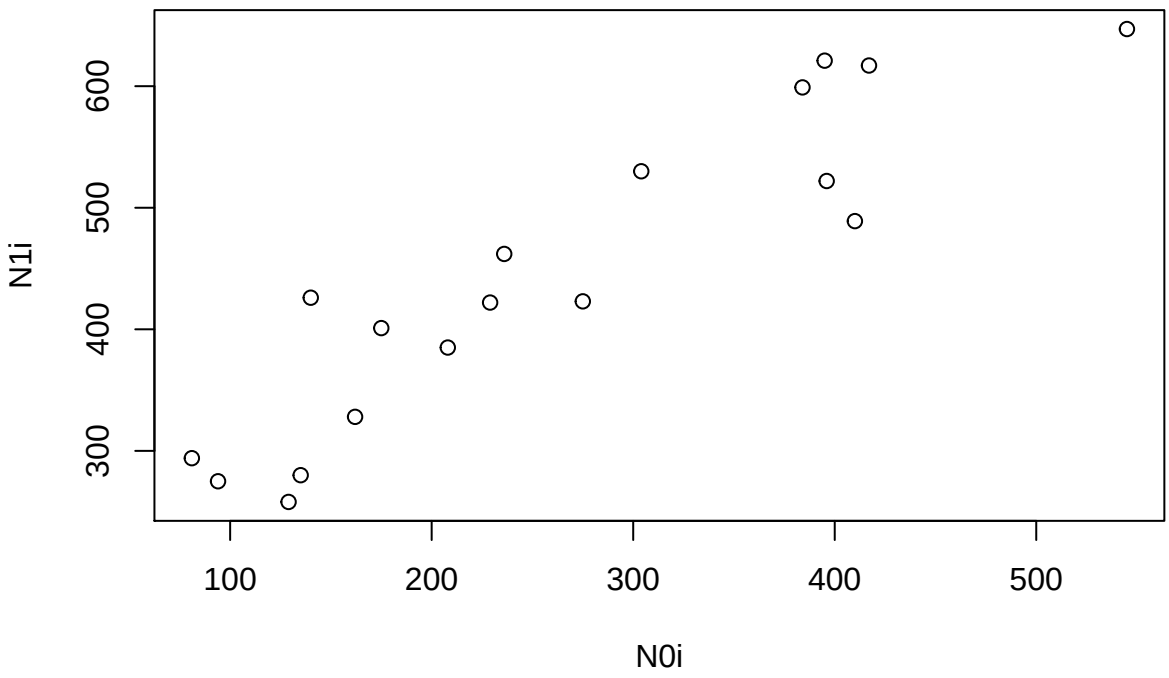
```
## [1] 0.00416
```

```
a3 <- 0.01
b3 <- 0.01

lambda_0_3 <- rgamma(S, Y0 + a3, N0 + b3)
lambda_1_3 <- rgamma(S, Y1 + a3, N1 + b3)
mean(lambda_1_3 > lambda_0_3)
```

```
## [1] 0.0053
```

3. Prediction

```
plot(N0i,N1i)
```




```
model.medal <- lm(N1i ~ N0i)
round(predict.lm(model.medal,newdata = data.frame(N0i=398)))
```

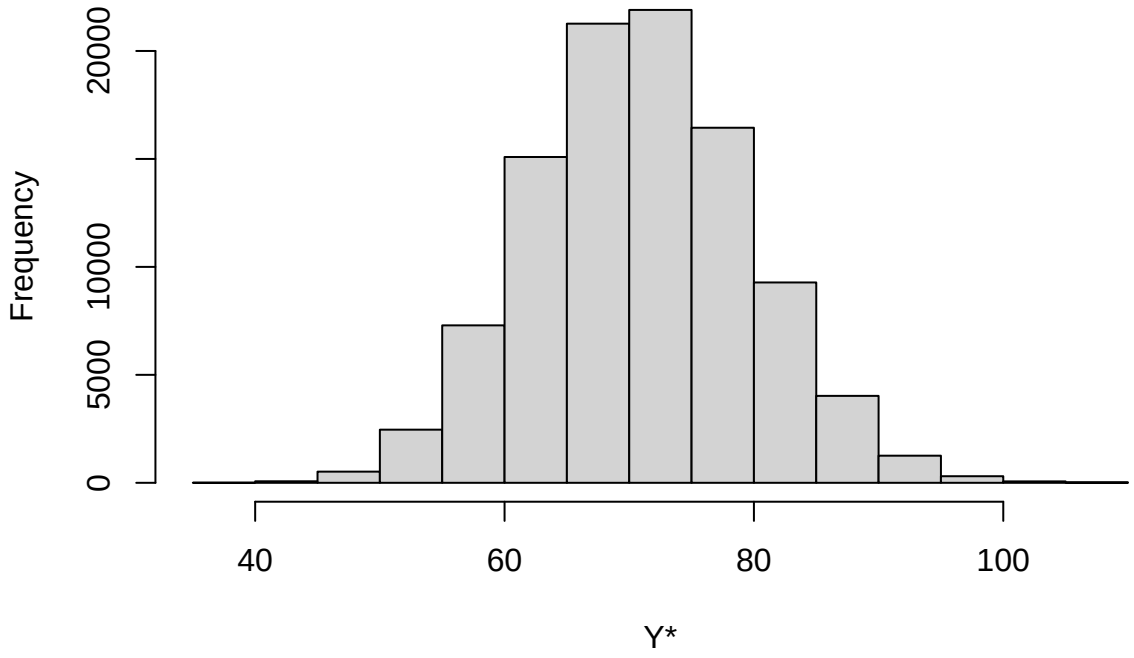
```
##   1
## 561
```

assume there are 561 participants in 2024 in France.

```
N1_f <- 561
# lambda_1 <- rgamma(S, 33, 398)
# Ystar <- rpois(S, lambda_1)*N1_f
# round(mean(Ystar))

lambda_1 <- rgamma(S, Y1, N1)
Ystar <- rpois(S, lambda_1*N1_f)
round(mean(Ystar))
```

```
## [1] 71
```

```r
hist(Ystar,xlab = 'Y*',main = NULL)
```




```r
quantile(Ystar,c(0.025,0.975))
```

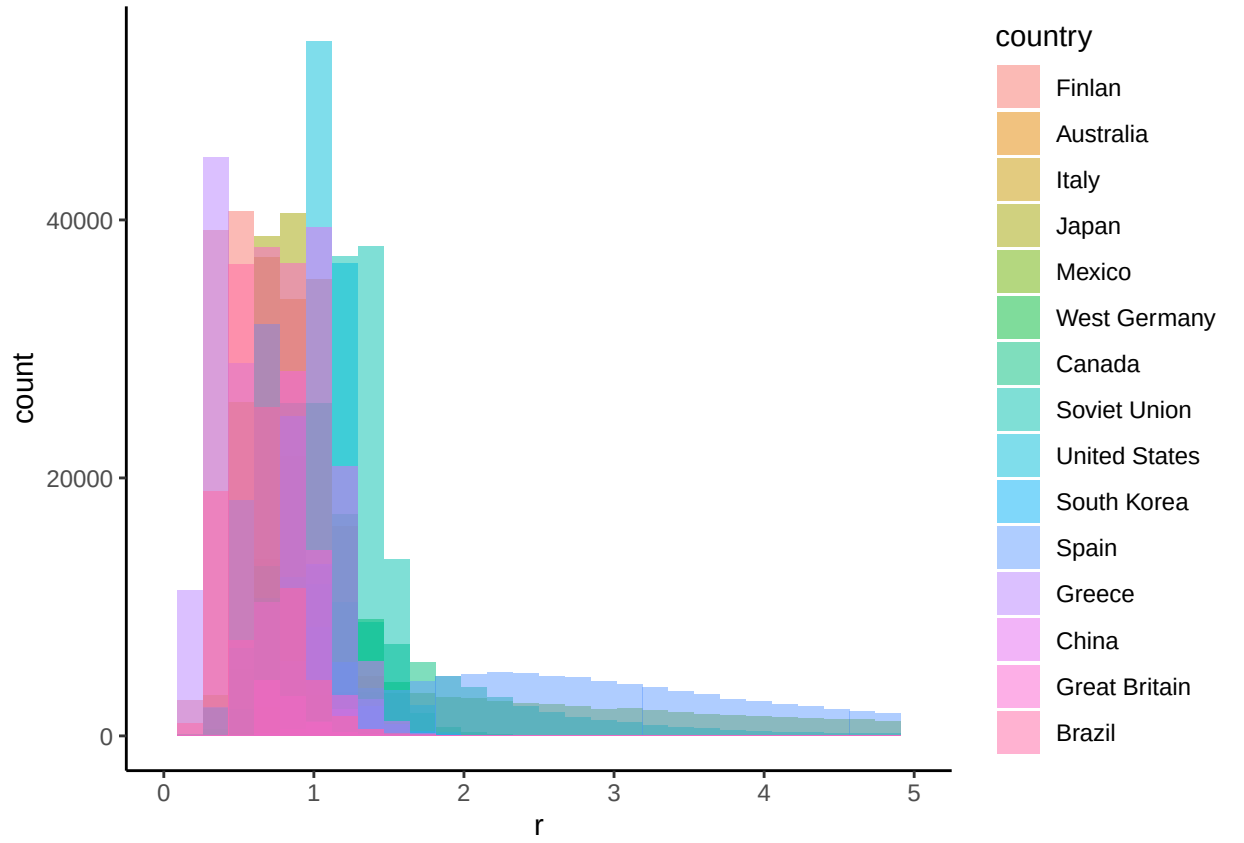
```
##  2.5% 97.5%
##    55    89
```

4. Country-specific analisys

```r
Y1_country <- c(22,93,36,80,9,40,11,195,275,33,22,16,100,65,19)
Y0_country <- c(24,52,25,59,1,26,5,125,202,19,4,13,63,47,17)
N1_country <- c(258,911,280,949,275,423,385,489,1169,401,422,426,599,530,462)
N0_country <- c(129,498,135,557,94,275,208,410,941,175,229,140,384,304,236)
country <- c('Finlan','Australia','Italy','Japan','Mexico','West Germany','Canada','Soviet Union',
             'United States','South Korea','Spain','Greece','China','Great Britain','Brazil')


df <- c()
for (i in 1:15) {
  a <- b <- 0.1
  lambda1 <- rgamma(S, Y1_country[i]+a, N1_country[i]+b)
  lambda0 <- rgamma(S, Y0_country[i]+a, N0_country[i]+b)
```

```
  r <- lambda1/lambda0

  df <- cbind(df,r)
}

r_country_CI <- round(apply(df, 2, function(x){
  return(quantile(x,c(0.025,0.975)))
}),2)

df <- as.data.frame(df)
colnames(df) <- country
```

```
library(reshape2)
df2 <- melt(df, variable.name = 'country', value.name = 'r')

library(ggplot2)
g <- ggplot(df2, aes(x=r,fill=country)) + geom_histogram(alpha=0.5,position = 'identity')+
      # stat_bin(binwidth = 0.5) +
      xlim(c(0,5)) + theme_classic()
g
```




```
c1 <- c()
c2 <- c()
q_2.5 <- c()
q_97.5 <- c()
```

```r
for (i in 1:14) {
  for (j in (i+1):15) {
    l <- quantile(df[[i]] - df[[j]],c(0.025,0.975))[1]
    u <- quantile(df[[i]] - df[[j]],c(0.025,0.975))[2]
    c1 <- c(c1, country[i])
    c2 <- c(c2, country[j])
    q_2.5 <- c(q_2.5, round(l,2))
    q_97.5 <- c(q_97.5, round(u,2))
  }
}

diff <- as.data.frame(cbind(c1,c2,q_2.5,q_97.5))
diff$q_2.5 <- as.numeric(diff$q_2.5)
diff$q_97.5 <- as.numeric(diff$q_97.5)

diff_sig <- diff[(diff$q_2.5)*(diff$q_97.5)>0,]
nrow(diff_sig)
```

```
## [1] 27
```