STAT 251 - Project

Joshua Carpenter, Cecilia Fu

3/24/2021

Introduction

Between the years of 2000 and 2019, the World Health Organization (WHO) has collected data on causes of death in 183 countries. We are interested particularly in Cardiovascular Disease. The purpose of this analysis will be to determine if cardiovascular disease is a more or less prevalent cause of death in the United States than elsewhere, that is if the proportion of deaths due to cardiovascular disease is greater or smaller in the US than in other countries. We are also interested in comparing death due to cardiovascular disease in the US vs China, the countries of origin of the authors.

To do this, we will model the proportion of deaths due to cardiovascular disease in the United States as the probability of success of a binomial random variable, where the population consists of all deaths in the United States and a trial consists of sampling one death and determining whether or not the cause was cardiovascular disease. We will consider a success to be that the death was caused by cardiovascular disease and a failure that it was not. We will compare that to a similar model of deaths due to cardiovascular disease in China. We will also model the proportion of deaths outside the United States caused by cardiovascular disease as the probability of success of a binomial random variable where the population is all deaths that occurred outside of the United States.

We will determine an appropriate beta prior distribution, which we will use for all three data distributions; we will run a Bayesian update based on data from WHO; and we will compare the posterior distribution for the USA to the posterior distributions for China and for all countries except the USA combined using Monte-Carlo methods. Our null hypothesis in the first case is that $\mu_{\text{USA}} - \mu_{\text{CHN}} = 0$ and the alternative hypothesis is that $\mu_{\text{USA}} - \mu_{\text{CHN}} \neq 0$. In the second case, our null hypothesis is that $\mu_{\text{USA}} - \mu_{\text{OTH}} = 0$ and the alternative hypothesis is that $\mu_{\text{USA}} - \mu_{\text{OTH}} \neq 0$. Based on the Monte-Carlo estimated posterior distribution for the difference in proportions, we will determine a 95% confidence interval for each test and conclude whether the proportions are significantly different.

Data

Below is a summary of the data to be used. The variable `Total_Deaths` is the total number of deaths, measured in thousands of deaths, in that country during the time period of data collection. The variable `Cardio_Disease` is the number of those deaths that were caused by cardiovascular disease, also measured in thousands of deaths.

ID	Country	Cardio_Disease	Total_Deaths
1	AFG	71.26378	254.8099
2	ALB	19.4825	31.1542
3	DZA	91.51461	203.3004
...	...	...	...
35	CHN	4306.53601	10105.5956
...	...	...	...
175	USA	873.20014	2949.2139
...	...	...	...

ID	Country	Cardio_Disease	Total_Deaths
182	ZMB	16.6686	121.1049
183	ZWE	17.3354	117.7098

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Cardio_Disease	0.1548	5.9390	17.5439	97.6165	59.7985	4306.536
Total_Deaths	0.6186	20.5604	72.8564	302.8184	187.4312	10105.596

Prior Distribution

In the absence of a professional opinion, we will rely on our own limited knowledge to construct a prior distribution for p. Given the many possible ways to die, we think that it is very unlikely that the proportion of deaths that are caused by cardiovascular disease is greater than 0.5. In fact, we believe that the proportion will be much less that 0.5, so we will therefore choose a distribution that assigns most of the weight close to zero and almost no weight greater than 0.5. Since we are not experts on the subject, we will choose low values for the a and b and expect most of our information to come from the data. For our prior, we will use a beta distribution with $a = 1.5$ and $b = 10$.

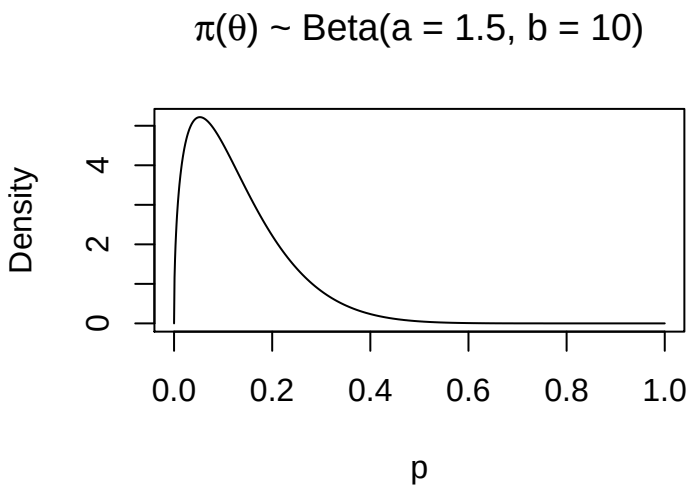

Figure 1: Prior distribution of the proportion of deaths that are caused by cardiovascular disease

Analysis: USA vs China

Comparison for people who passed away from Cardiovascular Disease between the USA and China

Country	Cardio_Disease	Total_Deaths
CHN	4306536	10105596
USA	873200	2949214

Posterior Distribution

Now, based on that data, we will update the prior distributions and call the new parameters a^* and b^*.

$$a^*_{\text{USA}} = a + y = 1.5 + 873,200 = 873,201.5$$

$$b^*_{\text{USA}} = b + n - y = 10 + 2,949,214 - 873,200 = 2,076,024$$

$$a^*_{\text{CHN}} = a + y = 1.5 + 4,306,536 = 4,306,537.5$$

$$b^*_{\text{CHN}} = b + n - y = 10 + 10,105,596 - 4,306,536 = 5,799,070$$

Posterior distribution for USA and CHINA

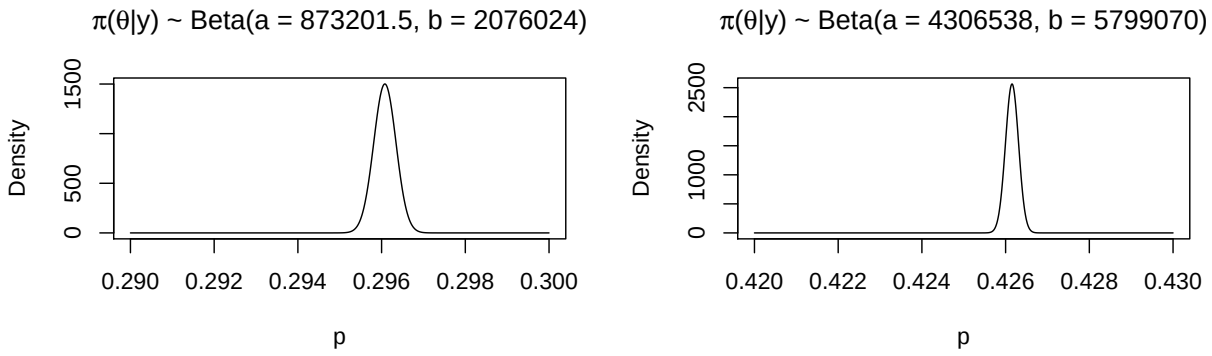

Figure 2: Posterior distribution of Death by Cardiovascular Disease in the United States (left) and China (right)

Posterior Difference between USA and CHN

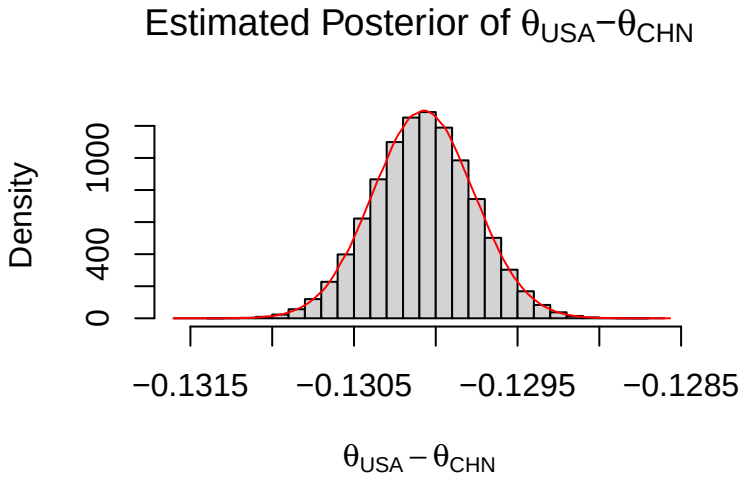

Figure 3: Monte Carlo approximation of the posterior distribution of the difference in proportions of deaths due to cardiovascular disease in the USA compared to other countries

Conclusion of the posterior difference from USA and China: We compared the the proportion of deaths caused by cardiovascular disease in the USA and China, and conclude, based on our prior belief and data collection from WHO, that the proportion of deaths in the United States due to cardiovascular disease is between 0.1295 and 0.1307 lower than the proportion of deaths due to cardiovascular disease in China.

Analysis: USA vs All Other Countries

In order to make a comparison between the United States and all other countries, we will summarize the data in just two rows.

Category	Cardio_Disease	Total_Deaths
USA	873200	2949214
OTH	16990627	52466558

Posterior Distributions for USA and OTH

Now, based on that data, we will update the prior distributions and call the new parameters a^* and b^*.

$$a^*_{\text{USA}} = a + y = 1.5 + 873,200 = 873,201.5$$

$$b^*_{\text{USA}} = b + n - y = 10 + 2,949,214 - 873,200 = 20,76,024$$

$$a^*_{\text{OTH}} = a + y = 1.5 + 16,990,627 = 16,990,628.5$$

$$b^*_{\text{OTH}} = b + n - y = 10 + 52,466,558 - 16,990,627 = 35,475,941$$

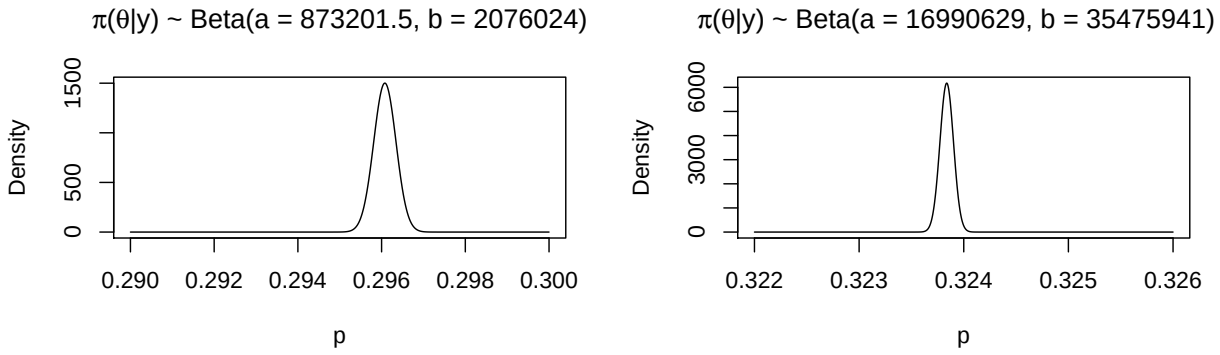

Figure 4: Posterior distribution of Death by Cardiovascular Disease in the United States (left) and the rest of the world (right)

Posterior Difference between USA and OTH

We will now use Monte Carlo sampling to take samples from the posterior distribution of $p_{\text{USA}} - p_{\text{OTH}}$ and compute a 95% confidence interval.

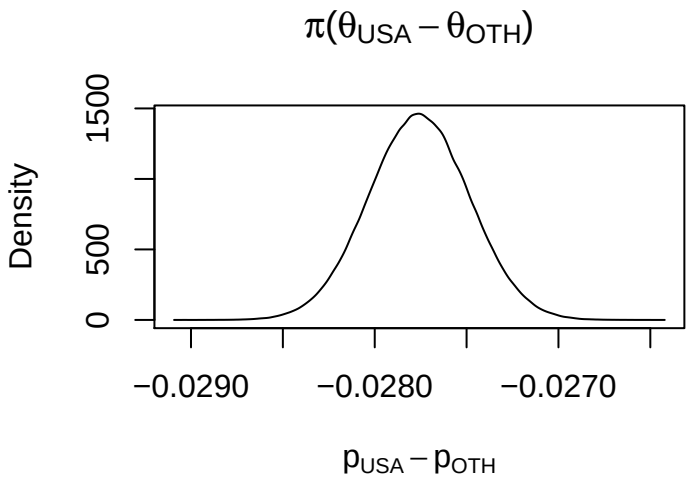

Figure 5: Monte Carlo approximation of the posterior distribution of the difference in proportions of deaths due to cardiovascular disease in the USA compared to other countries

According to our credible interval, there is a 95% probability that the proportion of deaths in the United States due to cardiovascular disease is between 0.0272 and 0.0283 lower than the proportion of deaths due to cardiovascular disease in other countries. We therefore conclude that the proportion of deaths caused by cardiovascular disease outside the United States is greater than in the United States.

Conclusion/Discussion

According to our data character, we chose a Bayesian approach to analyze our parameters of interest, which are whether cardiovascular disease is a more or less prevalent cause of death in the United States than in China or elsewhere.

The model we chose is binomial-beta model, and our prior belief for the proportion of deaths caused by cardiovascular disease follows a beta distribution with shape = 1.5 and rate = 10. We have found the conjugate posterior distribution for the proportion of deaths caused by cardiovascular disease in the USA, China, and all other countries.

We did posterior distribution comparison between the USA vs China and the USA vs all other countries. According to the large number of draws from the posterior difference, we reject the null hypothesis that the proportion of deaths due to cardiovascular disease is the same in the USA as in China as in the world. In the USA vs China comparison, there is a 95% probability that the difference in the proportion of deaths in the United States vs China due to cardiovascular disease is between -0.1307 and -0.1295. Since zero is not in the interval and the mean difference in proportions is less than zero, we conclude that the proportion of deaths due to cardiovascular disease in the USA is smaller than in China. Similarly, in the comparison between the USA and all other countries, there is a 95% probability that the proportion of deaths in the United States due to cardiovascular disease is between -0.0283 and -0.0272 and we conclude that the proportion of deaths due to cardiovascular disease is lower in the United States than the world average.

Although we conclude that the proportion of deaths in the United States due to cardiovascular disease is lower than same cause in China and all other countries as a whole, we do not know the reason behind it. It might be hard to draw a conclusion for why the USA has a lower proportion of death due to cardiovascular disease, but we could do further research to attempt to determine a cause. Our hypothesis is that maybe the health care system is better in the United States than in other countries. In follow up, perhaps we could measure the quality of health care in different countries and make a comparison.

Apendix A: Data Source

Global health estimates: Leading causes of death

Cause-specific mortality, 2000–2019

Downloaded from:

https://www.who.int/data/gho/data/themes/mortality-and-global-health-estimates/ghe-leading-causes-of-death

See "Global summary estimates" under "Global and by Region".

Apendix B: Code

```
## Setup
library(knitr)
opts_chunk$set(echo = FALSE, comment=NA)

library(readxl)
library(tidyverse)
set.seed(3812)


## Read in the data
country_codes <- read_xlsx("deaths2019.xlsx",
                           range = "'Deaths All ages'!H8:GH8",
                           col_names = FALSE) %>%
  pivot_longer(everything(), names_to = "Names", values_to = "Country") %>%
  select(-Names)
cardio_disease_vals <- read_xlsx("deaths2019.xlsx",
          range = "'Deaths All ages!H148:GH148",
          col_names = FALSE) %>%
  pivot_longer(everything(), names_to = "Names",
               values_to = "Cardio_Disease") %>%
  select(-Names)
total_deaths_vals <- read_xlsx("deaths2019.xlsx",
                               range = "'Deaths All ages!H11:GH11",
                               col_names = FALSE) %>%
  pivot_longer(everything(), names_to = "Names", values_to = "Total_Deaths") %>%
  select(-Names)
cardio <- bind_cols(country_codes, cardio_disease_vals, total_deaths_vals)


## Summarise the data
cardio_tail <- cardio %>%
  rownames_to_column("ID") %>%
  tail(2) %>%
  mutate(Cardio_Disease = as.character(round(Cardio_Disease, 4)),
         Total_Deaths = as.character(round(Total_Deaths, 4)))
row_USA <- cardio %>%
  rownames_to_column("ID") %>%
  filter(Country == "USA") %>%
  mutate(Cardio_Disease = as.character(round(Cardio_Disease, 5)),
         Total_Deaths = as.character(round(Total_Deaths, 4)))
row_CHN <- cardio %>%
```

```
  rownames_to_column("ID") %>%
  filter(Country == "CHN") %>%
  mutate(Cardio_Disease = as.character(round(Cardio_Disease, 5)),
         Total_Deaths = as.character(round(Total_Deaths, 4)))
cardio_head <- cardio %>%
  rownames_to_column("ID") %>%
  head(3) %>%
  mutate(Cardio_Disease = as.character(round(Cardio_Disease, 5)),
         Total_Deaths = as.character(round(Total_Deaths, 4))) %>%
  add_row(ID = "...", Country = "...",
          Cardio_Disease = "...", Total_Deaths = "...") %>%
  add_row(row_CHN) %>%
  add_row(ID = "...", Country = "...",
          Cardio_Disease = "...", Total_Deaths = "...") %>%
  add_row(row_USA) %>%
  add_row(ID = "...", Country = "...",
          Cardio_Disease = "...", Total_Deaths = "...") %>%
  bind_rows(cardio_tail)
kable(cardio_head, align = "c")

cardio_summ <- cardio$Cardio_Disease %>%
  summary() %>%
  as.matrix() %>%
  t()
death_summ <- cardio$Total_Deaths %>%
  summary() %>%
  as.matrix() %>%
  t()
overall_summ <- rbind(cardio_summ, death_summ) %>%
  round(4)
row.names(overall_summ) <- c("Cardio_Disease", "Total_Deaths")
kable(overall_summ, align = "c")


## Plot prior distribution
x <- seq(0, 1, length = 1001)
a <- 1.5
b <- 10
plot(x, dbeta(x, a, b), type = "l",
     main = substitute(paste(pi, "(", theta, ")", " ~ Beta(a = ",
                             a, ", b = ", b, ")",  sep = ""),
                       list(a = a, b = b)),
     ylab = "Density",
     xlab = "p")


# create a table that have information for both USA and China
cardio_USA_CHN <- cardio %>%

  filter(Country == "USA"| Country == "CHN") %>%
  mutate(Cardio_Disease = round(Cardio_Disease * 1000),
         Total_Deaths = Total_Deaths * 1000)
```

```r
kable(cardio_USA_CHN, align = "c")


# Number of people passed away from Cardio Disease in the USA from 2000 to 2019
NumCardio.USA <- as.numeric(cardio_USA_CHN[2,"Cardio_Disease"])

# Number of people passed away from other causes in the USA from 2000 to 2019
NumNonCardio.USA <- as.numeric(cardio_USA_CHN[2, "Total_Deaths"]) - NumCardio.USA

# Posterior parameter
astar.USA <- a + NumCardio.USA
bstar.USA <- b + NumNonCardio.USA

# Posterior distribution for people passed away due to Cardio Disease
J <- 10^6
post.cardio.USA <- rbeta(J, astar.USA, bstar.USA)

# Number of people passed away from Cardio Disease in the China from 2000 to 2019
NumCardio.CHN <- as.numeric(cardio_USA_CHN[1,"Cardio_Disease"])

# Number of people passed away from other causes in the China from 2000 to 2019
NumNonCardio.CHN <- as.numeric(cardio_USA_CHN[1, "Total_Deaths"]) - NumCardio.CHN

# Posterior parameter for China
astar.CHN <- a + NumCardio.CHN
bstar.CHN <- b + NumNonCardio.CHN

# Posterior distribution for people passed away due to Cardio Disease
post.cardio.CHN <- rbeta(J, astar.CHN, bstar.CHN)


par(mfrow = c(1, 2))
x <- seq(0.29, 0.3, length = 1001)
plot(x, dbeta(x, astar.USA, bstar.USA), type = "l",
     main = substitute(paste(pi, "(", theta, "|y)", " ~ Beta(a = ",
                             a, ", b = ", b, ")",  sep = ""),
                       list(a = astar.USA, b = bstar.USA)),
     ylab = "Density",
     xlab = "p")

y <- seq(0.42, 0.43, length = 1001)
plot(y, dbeta(y, astar.CHN, bstar.CHN), type = "l",
     main = substitute(paste(pi, "(", theta, "|y)", " ~ Beta(a = ",
                             a, ", b = ", b, ")",  sep = ""),
                       list(a = astar.CHN, b = bstar.CHN)),
     ylab = "Density",
     xlab = "p")


# Posterior Difference
post.diff <- post.cardio.USA - post.cardio.CHN

# Histogram of the posterior Difference
```

```r
hist(post.diff, freq = FALSE, main = expression(paste("Estimated Posterior of ", theta[USA], "-", theta
     xlab = expression(theta[USA] - theta[CHN]), ylab = "Density")
lines(density(post.diff), col = "red")


# mean of post.diff
mean(post.diff)

# 95% credible interval for Posterior difference
CI_USA_CHN <- round(quantile(post.diff, c(0.025, 0.975)),4)



## Summarize data for USA vs all other countries
cardio_comp <- cardio %>%
  mutate(Category = fct_collapse(cardio$Country,
                                 USA = "USA",
                                 other_level = "OTH")) %>%
  group_by(Category) %>%
  summarise(Cardio_Disease = round(sum(Cardio_Disease) * 1000),
            Total_Deaths = sum(Total_Deaths) * 1000)
kable(cardio_comp, align = "c")

USvAll <- cardio_comp %>%
  select(-Category) %>%
  as.matrix()
row.names(USvAll) <- cardio_comp$Category


par(mfrow = c(1, 2))

## Update the prior for USA and plot the posterior
y <- USvAll["USA", "Cardio_Disease"]
n <- USvAll["USA", "Total_Deaths"]
astar_USA <- a + y
bstar_USA <- b + n - y

x <- seq(0.29, 0.3, length = 1001)
plot(x, dbeta(x, astar_USA, bstar_USA), type = "l",
     main = substitute(paste(pi, "(", theta, "|y)", " ~ Beta(a = ",
                             a, ", b = ", b, ")",  sep = ""),
                       list(a = astar_USA, b = bstar_USA)),
     ylab = "Density",
     xlab = "p")


## Update the prior for the rest of the world and plot the posterior
y <- USvAll["OTH", "Cardio_Disease"]
n <- USvAll["OTH", "Total_Deaths"]
astar_OTH <- a + y
bstar_OTH <- b + n - y

x <- seq(0.322, 0.326, length = 1001)
```

```r
# Note the two extra spaces in the main label are intentional to fix formatting
plot(x, dbeta(x, astar_OTH, bstar_OTH), type = "l",
     main = substitute(paste(pi, "(", theta, "|y)", " ~ Beta(a = ",
                             a, ", b = ", b, ")  ",  sep = ""),
                       list(a = astar_OTH, b = bstar_OTH)),
     ylab = "Density",
     xlab = "p")


## Sample from the posterior of USA - OTH and plot the estimated posterior
J <- 10^6
sample_USA <- rbeta(J, astar_USA, bstar_USA)
sample_OTH <- rbeta(J, astar_OTH, bstar_OTH)
sample_USAvAll <- sample_USA - sample_OTH

plot(density(sample_USAvAll), zero.line = FALSE,
     main = expression(paste(pi, "(", theta[USA] - theta[OTH], ")", sep = "")),
     xlab = expression(p[USA] - p[OTH]))


## Compute a 95% credible interval for the difference in proportions
CI <- round(quantile(sample_USAvAll, c(0.025, 0.975)), 4)
```