POLI502-Midterm_JackJeffrey

Jack Jeffrey

POLI502 Midterm

```
library(tidyverse) # load tidyverse package for midterm
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.1     v tibble    3.2.1
v lubridate 1.9.3     v tidyr     1.3.1
v purrr     1.0.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

Task 1

```
world.data <- read.csv("/Users/jackjeffrey/Documents/Poli502_Jeffrey/Data/world.csv")
# Loaded and stored world.data as an object
```

Task 2

```
ft.oecd <- as.data.frame(table(world.data$oecd)) # Created a frequency table for the OECD variable and stored it as an object
names(ft.oecd)[1] <- "OECD Member?"
# Chaged column name to OECD Member?
ft.oecd$percentage <- (ft.oecd$Freq/sum(ft.oecd$Freq)) * 100
# Created a column name called percentage and calculated the percentage of countries in the OECD
print(ft.oecd)
```

```
       OECD Member? Freq percentage
1        Not member  161   84.29319
2 OECD Member state   30   15.70681
```

```
# View Changes
```

Task 3

(A) 30 members

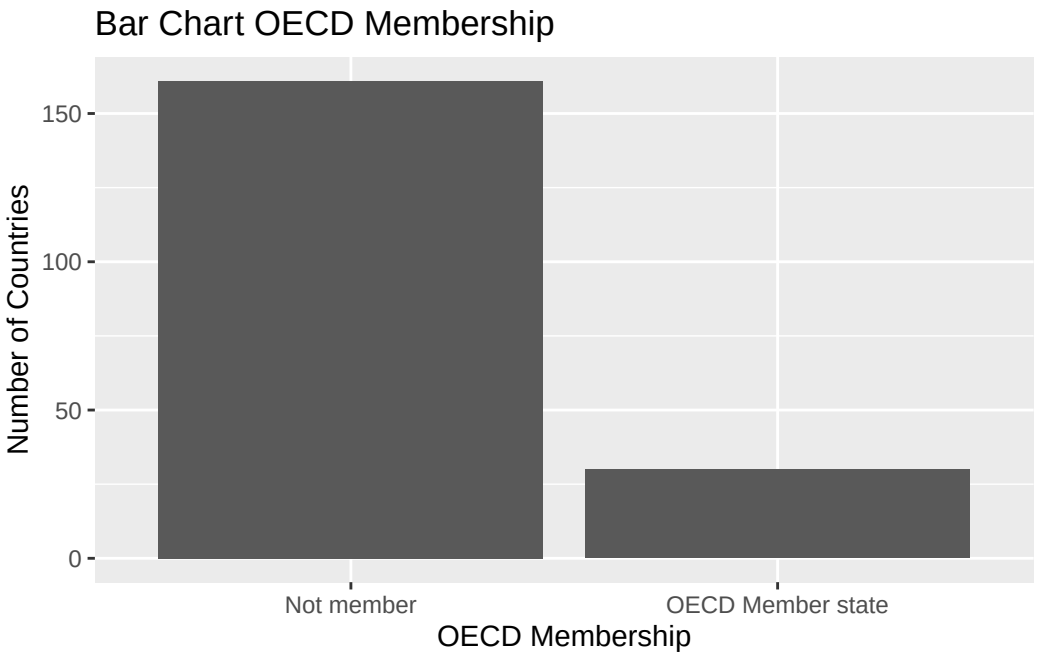
(B) 161 non members

(C) 15.7% are members

(D) 84.2% are non members

Task 4

```
ggplot(data = world.data, aes(x = oecd)) +
  geom_bar() + xlab("OECD Membership") +
  ylab("Number of Countries") +
  ggtitle("Bar Chart OECD Membership")
```




```
# created a bar chart for the OECD variable, changed the x and y labels, and changed the title
```

Task 5

```
view(world.data)
# investigate data for OECD members and non democratic countries
# 3 OECD Members coded in the data: Australia, Austria, Belgium
# 3 Non-democratic countries coded in the data: Afghanistan, Algeria, Angola
```

Task 6

```
summary(world.data$gdp_10_thou)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
 0.0090  0.0503  0.1897  0.6018  0.6320  4.7354      14
```

```
sd(world.data$gdp_10_thou, na.rm = TRUE)
```

```
[1] 0.9433982
```

numerical summary of GDP variable, Maximum GDP per capita is 4.7354 indicating over 40,000 US dollars per capita for the wealthiest country The mean indicates an average of 6,000 US dollars per capita for all countries. The max indicates that the wealthiest country is over 40,000 US dollars per capita. There is also 14 countries without data accoridng the NA value.

produce standard deviation with removed NA's equating to 0.943

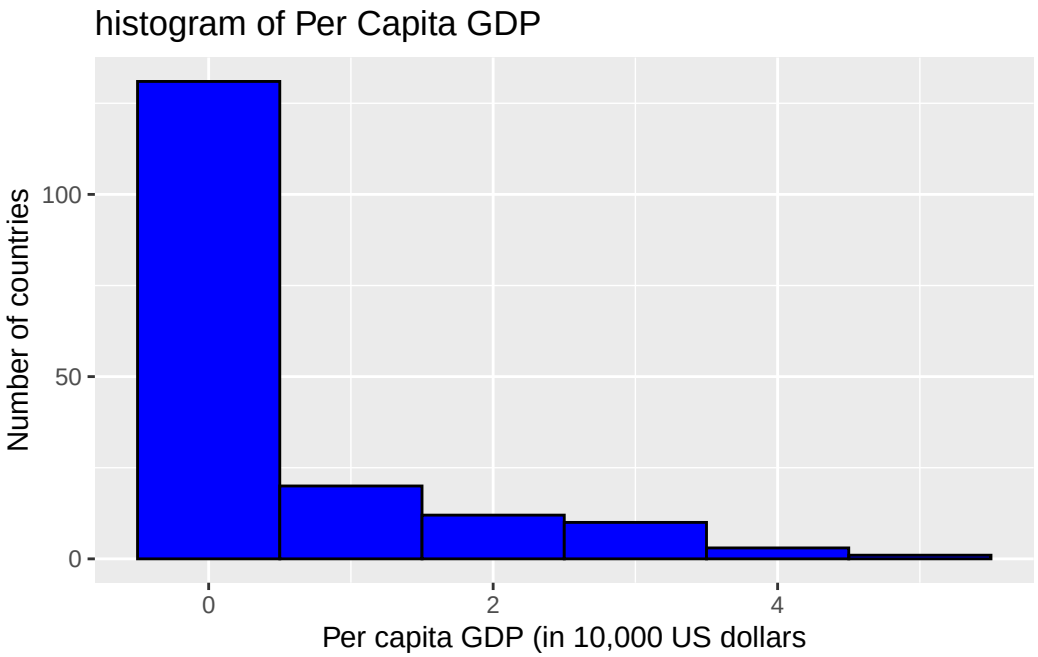
Task 7

the distribution is positively skewed, meaning most of the data points are concentrated in the lower left, having lower GDP than the average GDP. The higher mean than median indicates that there are a smaller number of extremely wealthy countries which are causing the mean to be so much higher than the median.

Task 8

```
ggplot(world.data, aes(x = gdp_10_thou)) +
  geom_histogram(binwidth = 1, fill = "blue", color = "black") +
  xlab("Per capita GDP (in 10,000 US dollars")+
  ylab("Number of countries") +
  ggtitle("histogram of Per Capita GDP")
```

```
Warning: Removed 14 rows containing non-finite outside the scale range
(`stat_bin()`).
```




```
# created a histogram displaying the distribution of
# countries and their gdp per capita
# both x and y labels were updated as well as the title
# colored the bars blue and their outlines black
```

Task 9

```
high_gdp_countries <- world.data %>%
  filter(gdp_10_thou > 4) %>% # filtered countries in the variable greater than 4
  select(country, gdp_10_thou) #selected those countries within the gdp variable
print(high_gdp_countries) # Luxembourg and Norway are the 2 countries with GDP per capita higher than 4 or 40,000 US dollars.
```

```
     country gdp_10_thou
1 Luxembourg      4.7354
2     Norway      4.1974
```

Task 10

```
std_dev_gdp <- sd(world.data$gdp_10_thou, na.rm = TRUE) # calculate the standard dev and store it as an object
n <- 177 # define the sample size of countries
standard_error <- std_dev_gdp/sqrt(n) # calculate the standard error of the mean
print(standard_error) # view results
```

```
[1] 0.07091015
```

Task 11

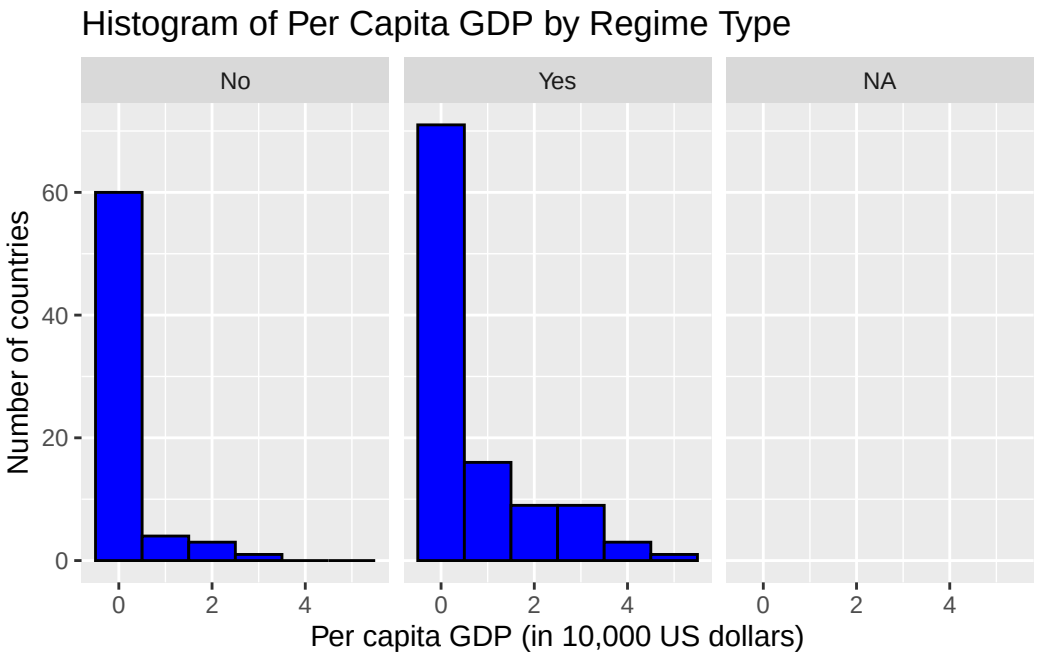
```
mean_gdp <- mean(world.data$gdp_10_thou, na.rm = TRUE) # store mean for the variable as an object
z_value <- 1.96 # define 95% confidence interval
lower_bound <- mean_gdp - z_value * standard_error
upper_bound <- mean_gdp + z_value * standard_error # calculate upper and lower bounds of confidence interval
cat("95% Confidence Interval: [",lower_bound, ", ", upper_bound, "]\n") #display confidence intervals: lower bound = 0.462, upper bound = 0.740
```

```
95% Confidence Interval: [ 0.4628347 ,  0.7408025 ]
```

Task 12

```
ggplot(world.data, aes(x = gdp_10_thou)) +
  geom_histogram(binwidth = 1, fill = "blue", color = "black") + facet_wrap(~ democ_regime) +
  xlab("Per capita GDP (in 10,000 US dollars)") +
  ylab("Number of countries") + ggtitle("Histogram of Per Capita GDP by Regime Type") # created histograms of countries per capita GDP based on democr
```

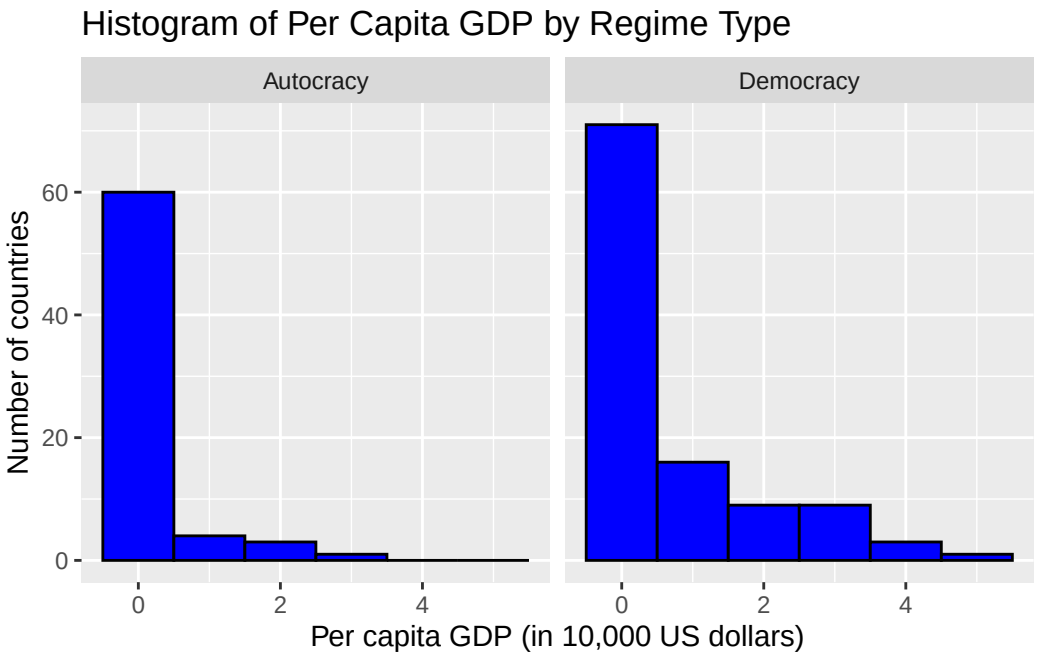
```
Warning: Removed 14 rows containing non-finite outside the scale range
(`stat_bin()`).
```



Task 13

```
dem.gdp <- world.data[!is.na(world.data$democ_regime),] #exclude mssing values from data frame
dem.gdp$dem.dum <- ifelse(dem.gdp$democ_regime == "Yes", "Democracy", "Autocracy") # created new variable with intuitive labels
ggplot(dem.gdp, aes(x = gdp_10_thou)) +
  geom_histogram(binwidth = 1, fill = "blue", color = "black",) + facet_wrap(~dem.dum) + xlab("Per capita GDP (in 10,000 US dollars)") + ylab("Number
```

```
Warning: Removed 12 rows containing non-finite outside the scale range
(`stat_bin()`).
```



Task 14

```
democracy_data <- dem.gdp %>%
  filter(dem.dum == "Democracy") # Filter to keep only democracies
mean_gdp <- mean(democracy_data$gdp_10_thou, na.rm = TRUE) #remove Na
ci_gdp <- t.test(democracy_data$gdp_10_thou, conf.level = 0.95)$conf.int # Use t test to calculate the mean and 95% confidence intervals
mean_gdp
```

```
[1] 0.8013927
```

```
ci_gdp
```

```
[1] 0.5961327 1.0066527
attr(,"conf.level")
[1] 0.95
```

```
# display results: Mean = 0.8, lower bound = 0.596, upper bound = 1.00
```

Task 15

```
dem.gdp <- world.data[!is.na(world.data$democ_regime),]
dem.gdp$dem.dum <- ifelse(dem.gdp$democ_regime == "Yes", "Democracy", "Autocracy")
autocracy_data <- dem.gdp %>% filter(dem.dum == "Autocracy") # Filter for autocracies
mean_gdp_autocracy <- mean(autocracy_data$gdp_10_thou, na.rm = TRUE)
ci_gdp_autocracy <- t.test(autocracy_data$gdp_10_thou, conf.level = 0.95)$conf.int
mean_gdp_autocracy
```

```
[1] 0.2819132
```

```
ci_gdp_autocracy
```

```
[1] 0.1526567 0.4111697
attr(,"conf.level")
[1] 0.95
```

```
#display results: mean = 0.282, lower bound = 0.152, and upper bound = 0.411
```

Task 16

given values

P(R) = 0.3

P(C|R) = 0.95

P(C|~R) = 0.25

P(~R) = 0.7 # probability its not raining because 1-0.3 = 0.7

P(C) = P(C|R)*P(R)+P(C~R)*(P~R)

probability of clouds, equals probability of clouds if it is raining, times probability of rain, plus probability of clouds if its not raining, times probability of no rain

```
(0.95*0.30) + (0.25*0.70)
```

```
[1] 0.46
```

```
# 0.46 = P(C)
# Plug values into Bayes Theorem to find P(R|C)
(0.95)*(0.3)/(0.46)
```

```
[1] 0.6195652
```

```
# Result of 0.619 or about a 62% chance of rain after seeing clouds in October
```

Task 17

Finding prior mean and prior standard deviation

```
# To find prior mean
(1.5)/(1.5+1.5) # prior mean = 0.5
```

```
[1] 0.5
```

```
# To find prior variance
(1.5*1.5)/((1.5 + 1.5 +1)*(1.5 +1.5)^2) # prior variance equals 0.0625
```

```
[1] 0.0625
```

```
# To find prior SD
sqrt(0.0625) # Prior SD = 0.25
```

```
[1] 0.25
```

```
# part(a) prior mean equals 0.5 & prior SD equals 0.25
```

part(b)

```
a <- 1.5
b <- 1.5
c <- 0.6
probability <- pbeta(c, a, b)
probability
```

```
[1] 0.62647
```

```
# part(b) prior probability that theta is less than 0.6 equals 0.626
```

part(c)

```
# define yes and no
n_yes <- 37
n_no <- 13
# define likelihood function
likelihood <- function(theta) {
  return(theta^n_yes * (1 - theta)^n_no)
}
# calculate likelihood of specific value of theta
theta_value <- 0.5
# calculate likelihood for different thetas
likelihood_value <- likelihood(theta_value)
likelihood_value
```

```
[1] 8.881784e-16
```

```
# part(c) likelihood equals 8.881784e-16
```

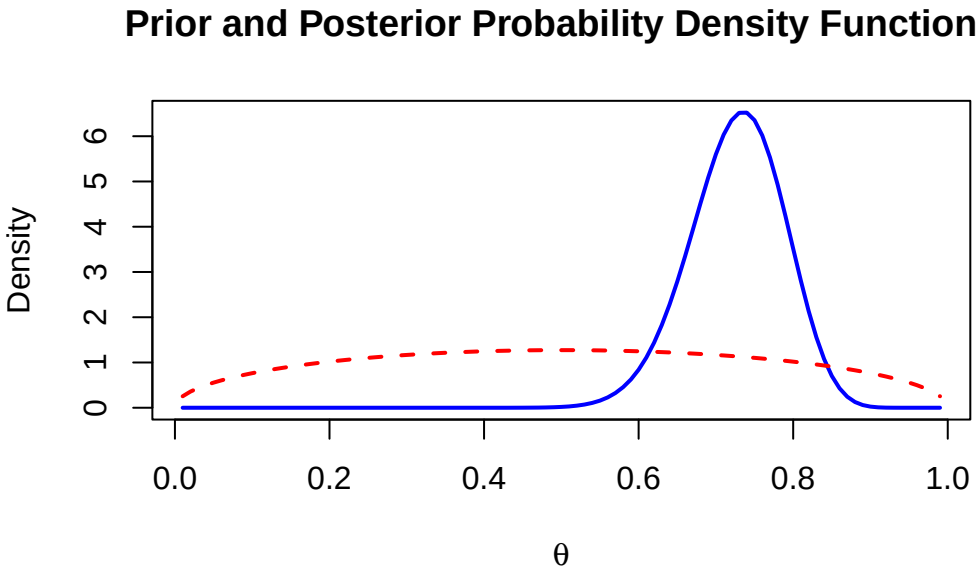
part(d)

combine the prior distribution (1.5,1.5) to the observed sample (37,13)

part (d) the posterior distribution equals (38.5, 14.5) because our prior distribution has been incorporated into our observed sample and now we have an updated distribution of Yes and No answers

part(e)

```
#Plot a graph showing the prior and posterior probability density functions
a <- 1.5
b <- 1.5
# defined prior parameters
c <- 38.5
d <- 14.5
# defined posterior parameters
theta <- seq(0.01, 0.99, 0.01)
# created a sequence of theta values
prior <- dbeta(theta, a, b)
posterior <- dbeta(theta, c, d)
# calculated densities
plot(theta, posterior, xlab=expression(theta),
     ylab="Density", type="l", col="blue", lwd=2, main="Prior and Posterior Probability Density Functions")
lines(theta, prior, lty=2, col="red", lwd=2)
```



```
# Ploted the prior and posterior densities
```

part(f) Final Interpretations

By generating this graph I have plotted the prior and posterior distributions. The prior distributins are depicted by the red dotted graph. This red dotted line is fairly stagnant and does not vary much in denisty indicating that the professor was very unsure about the chances of a student enjoying the class. After incorporating the observed values and plotting the posterior distribution on the graph the professors beliefs can be updated. The blue line represents the posterior distributin. This line has a much higher density around 0.6 and 0.8. This indicates an updated belief and probability of students enjoying the class. The professor can be more confident that the probability of a student enjoying the class is around 60% to 80% rather than the professors prior belief of about 50/50 after incorporating the observations of students given responses. This shows the power of Bayesian statistics, as more data is incorporated, the stronger one can feel about the probability being accurate to what will be observed.