DA1

### Problem 16

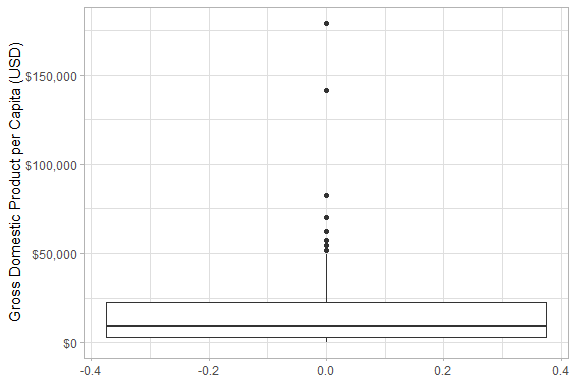
***Gross Domestic Product (GDP) per Capita.*** *The data file ex0116 contains the gross domestic product per capita for 228 countries - the data used to construct the box plot in Display 1.11.*

#### 16a)

*Make a box plot for the per capita GDPs with a statistical computer program. Include a y-axis label (for example, "Gross Domestic Product per Capita in $U.S.).*

R Code:

dat16 %>%   
 ggplot() +  
 geom\_boxplot(aes(y = PerCapitaGDP)) +  
 scale\_y\_continuous(labels = scales::dollar\_format()) +  
 labs(y = "Gross Domestic Product per Capita (USD)") +  
 theme\_light()



##### Explanation/Interpretation

The boxplot shows the distribution of GDP per capita in 2010, with each observation being a country (total of 228 countries). The median (50th percentile) is approximately $10,000 USD with the largest value being approximately $180,000 USD and the smallest value being $300 USD. Based on the box plot, it is clear that the data is right-skewed.

#### 16b)

*In what ways, if any, is the display from your software different from Display 1.11?*

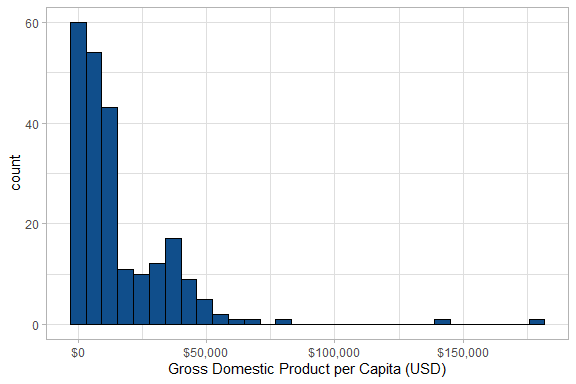
My plot shows 8 outliers, while the plot in the book shows only 7. Also, in my plot the upper whisker stops just short of $50,000, whereas the upper whisker goes just past $50,000. It seems that whatever software the authors used to create their plot used a different method than R (ggplot2) does for determining outliers for box plots. Lastly, there are a few formatting and stylistic differences.

#### 16c)

*Use a statistical computer program to draw a histogram of the per capita GDPs. Include an x-axis label. Use the program’s default bin width. Report the bin width.*

R Code:

dat16 %>%   
 ggplot() +  
 geom\_histogram(aes(x = PerCapitaGDP), color = "black", fill = "dodgerblue4") +  
 scale\_x\_continuous(labels = scales::dollar\_format()) +  
 labs(x = "Gross Domestic Product per Capita (USD)") +  
 theme\_light()



Bin width = $5956.67

##### Explanation/Interpretation

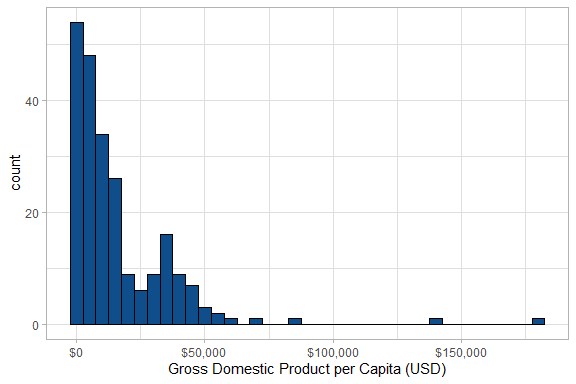
Like the box plot, this histogram shows that the data is right skewed with the majority of countries having less than $25,000 GDP per capita. Unlike the box plot, however, this plot shows that the data is slightly bimodal with a small second peak around $30,000 or so.

#### 16d)

*The program’s default bin width for histograms is not necessarily the best choice. If it’s possible with your computer program, redraw the histogram of part (c) using bin widths of $5,000.*

R Code:

dat16 %>%   
 ggplot() +  
 geom\_histogram(aes(x = PerCapitaGDP), binwidth = 5000,   
 color = "black", fill = "dodgerblue4") +  
 scale\_x\_continuous(labels = scales::dollar\_format()) +  
 labs(x = "Gross Domestic Product per Capita (USD)") +  
 theme\_light()



##### Explanation/Interpretation

There isn’t a lot of difference between this plot and the one that used the default bin width value, mostly because there wasn’t a big difference between the default value much different than $5,000. I would say that the biggest difference the change causes is that the slight bimodal effect is a bit more pronounced in the $5,000 bin width plot.

### Problem 17

*Seven students volunteered for a comparison of study guides for an advanced course in mathematics. They were randomly assigned, four to study guide A and three to study guide B. All were instructed to study independently. Following a two-day study period, all students were given an examination about the material covered by the study guides, with the following results:*

Study Guide A scores: 68, 77, 82, 85  
Study Guide B scores: 53, 64, 71

*Perform a randomization test by listing all possible ways that these students could have been randomized into two groups. There are 35 ways. For each outcome, calculate the difference between sample averages. Finally, calculate the two-sided p-value for the observed outcome.*

R Code:

# Find all combinations of groups  
dat17 <- c(68, 77, 82, 85, 53, 64, 71) # all scores  
a <- combn(dat17, 4) %>% data.frame()  
b <- data.frame(matrix(NA, nrow = 3, ncol = 35))  
  
# If scores from dat17 aren't in group A, assign them to group B  
j <- 1  
for (i in a[1:35]){  
 b[j] <- dat17[!dat17 %in% i]  
 j <- j + 1  
}  
  
# Join A and B into one data.frame  
x <- bind\_rows(a, b) %>% t() %>% data.frame() %>%   
 rename("A1" = X1,  
 "A2" = X2,  
 "A3" = X3,  
 "A4" = X4,  
 "B1" = X5,  
 "B2" = X6,  
 "B3" = X7) %>%   
 add\_column(combination = 1:35) %>%   
 relocate(combination)  
  
# Pivot longer to prepare for t.test()  
x\_long <- x %>% pivot\_longer(cols = A1:B3, names\_to = c("group", "number"),  
 names\_pattern = "(.)(.)", values\_to = "score")  
  
# Function to run t-test and return p-values  
get\_pvalue <- function(dat) {  
   
 p <- rep(NA, 35)  
   
 for (k in c(1:35)) {  
 test\_dat <- dat %>% filter(combination == k)  
   
 temp <- t.test(score ~ group, data = test\_dat)  
 p[k] <- temp$p.value %>% round(4)  
 }  
   
 return (p)  
}  
  
# get p\_values  
p\_values <- get\_pvalue(x\_long)  
  
# Join p-values to dataframe, and compute sample means and the difference  
results <- x %>% add\_column(p\_values) %>%   
 rowwise() %>%   
 mutate(mean\_a = round(mean(c(A1, A2, A3, A4)), 2),  
 mean\_b = round(mean(c(B1, B2, B3)), 2),  
 mean\_diff = mean\_a - mean\_b) %>%   
 relocate(p\_values, .after = last\_col()) %>%   
 select(-c(mean\_a, mean\_b))  
  
knitr::kable(results)

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| combination | A1 | A2 | A3 | A4 | B1 | B2 | B3 | mean\_diff | p\_values |
| 1 | 68 | 77 | 82 | 85 | 53 | 64 | 71 | 15.33 | 0.0776 |
| 2 | 68 | 77 | 82 | 53 | 85 | 64 | 71 | -3.33 | 0.7228 |
| 3 | 68 | 77 | 82 | 64 | 85 | 53 | 71 | 3.08 | 0.7822 |
| 4 | 68 | 77 | 82 | 71 | 85 | 53 | 64 | 7.17 | 0.5318 |
| 5 | 68 | 77 | 85 | 53 | 82 | 64 | 71 | -1.58 | 0.8617 |
| 6 | 68 | 77 | 85 | 64 | 82 | 53 | 71 | 4.83 | 0.6494 |
| 7 | 68 | 77 | 85 | 71 | 82 | 53 | 64 | 8.92 | 0.4108 |
| 8 | 68 | 77 | 53 | 64 | 82 | 85 | 71 | -13.83 | 0.0884 |
| 9 | 68 | 77 | 53 | 71 | 82 | 85 | 64 | -9.75 | 0.3037 |
| 10 | 68 | 77 | 64 | 71 | 82 | 85 | 53 | -3.33 | 0.7788 |
| 11 | 68 | 82 | 85 | 53 | 77 | 64 | 71 | 1.33 | 0.8788 |
| 12 | 68 | 82 | 85 | 64 | 77 | 53 | 71 | 7.75 | 0.4326 |
| 13 | 68 | 82 | 85 | 71 | 77 | 53 | 64 | 11.83 | 0.2289 |
| 14 | 68 | 82 | 53 | 64 | 77 | 85 | 71 | -10.92 | 0.1934 |
| 15 | 68 | 82 | 53 | 71 | 77 | 85 | 64 | -6.83 | 0.4625 |
| 16 | 68 | 82 | 64 | 71 | 77 | 85 | 53 | -0.42 | 0.9708 |
| 17 | 68 | 85 | 53 | 64 | 77 | 82 | 71 | -9.17 | 0.2780 |
| 18 | 68 | 85 | 53 | 71 | 77 | 82 | 64 | -5.08 | 0.5748 |
| 19 | 68 | 85 | 64 | 71 | 77 | 82 | 53 | 1.33 | 0.9027 |
| 20 | 68 | 53 | 64 | 71 | 77 | 82 | 85 | -17.33 | 0.0148 |
| 21 | 77 | 82 | 85 | 53 | 68 | 64 | 71 | 6.58 | 0.4397 |
| 22 | 77 | 82 | 85 | 64 | 68 | 53 | 71 | 13.00 | 0.1415 |
| 23 | 77 | 82 | 85 | 71 | 68 | 53 | 64 | 17.08 | 0.0378 |
| 24 | 77 | 82 | 53 | 64 | 68 | 85 | 71 | -5.67 | 0.5291 |
| 25 | 77 | 82 | 53 | 71 | 68 | 85 | 64 | -1.58 | 0.8680 |
| 26 | 77 | 82 | 64 | 71 | 68 | 85 | 53 | 4.83 | 0.6660 |
| 27 | 77 | 85 | 53 | 64 | 68 | 82 | 71 | -3.92 | 0.6563 |
| 28 | 77 | 85 | 53 | 71 | 68 | 82 | 64 | 0.17 | 0.9855 |
| 29 | 77 | 85 | 64 | 71 | 68 | 82 | 53 | 6.58 | 0.5358 |
| 30 | 77 | 53 | 64 | 71 | 68 | 82 | 85 | -12.08 | 0.1639 |
| 31 | 82 | 85 | 53 | 64 | 68 | 77 | 71 | -1.00 | 0.9074 |
| 32 | 82 | 85 | 53 | 71 | 68 | 77 | 64 | 3.08 | 0.7242 |
| 33 | 82 | 85 | 64 | 71 | 68 | 77 | 53 | 9.50 | 0.3306 |
| 34 | 82 | 53 | 64 | 71 | 68 | 77 | 85 | -9.17 | 0.2941 |
| 35 | 85 | 53 | 64 | 71 | 68 | 77 | 82 | -7.42 | 0.3911 |

dat25 %>%   
 ggplot() +  
 geom\_boxplot(aes(x = Group, y = Zinc, fill = Group)) +  
 geom\_label\_repel(data = five\_num, aes(x = Group, y = five, label = five, color = Group),  
 nudge\_x = 0.5, nudge\_y = c(-0.1, -0.05, 0.025, 0.05, 0.1), segment.color = "black") +  
 scale\_fill\_manual(values = c("firebrick1", "deepskyblue1")) +  
 labs(title = "Effect of calcium supplement on zinc levels in rats",  
 subtitle = "Group A received the dietary supplement, group B did not",  
 y = "Zinc concentration (mg/ml)") +  
 theme\_light() +  
 theme(legend.position = "none")