COMP2501 Assignment 3

## Requirements

**Submission deadline: April 28th, 2023 at 23:59.**

**Full mark of assignment 3: 34.**

For the following questions, please:

1. Replace all [Input here] places with your information or your answer.
2. Complete the code block by adding your own code to fulfill the requirements in each question. Please use the existing code block and do not add your own code block. Noting that please use head() to show the corresponding results if there are too many rows in them.

Please make sure your Rmd file is a valid Markdown document and can be successfully knitted.

For assignment submission, please knit your final Rmd file into a Word document, and submit both your **Rmd** file and the knitted **Microsoft Word** document file to Moodle. You get 0 score if 1) the Rmd file you submitted cannot be knitted, and 2) you have not submitted a Word document. For each visualization question, please make sure that the generated plot is shown in-place with the question and after the code block.

## Name and UID

Name: Frankline Misango Oyolo

UID: 3036050867

### Environmental setup

You need to have the dplyr, ggplot2 and HistData packages installed. If not yet, please run install.packages(c("dplyr", "ggplot2", "HistData")) in your R environment.

# Load the package.  
# install.packages("HistData")  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(ggplot2)  
library(HistData)

### 1. (1 points) Suppose that you roll a 6-sided die six times, compute the probability of not seeing a number bigger than 3.

dice\_outcomes <- 1:6  
sample\_space <- expand.grid(dice\_outcomes, dice\_outcomes, dice\_outcomes, dice\_outcomes, dice\_outcomes, dice\_outcomes)  
count <- sum(apply(sample\_space, 1, function(x) all(x <= 3)))  
probability <- count / 6^6  
probability

## [1] 0.015625

### 2. (1 points) Suppose two NBA teams, say the Warriors and the Kings, are playing a seven game series (The first to win four games, therefore, wins the series). The Warriors are a better team and have a 70% chance of winning each game. What is the probability that the Kings win at least one game?

kings\_prob <- 0.3  
warriors\_prob <- 0.7  
kings\_win\_no\_games <- warriors\_prob^4  
kings\_win\_atleast\_1\_game <- 1 - kings\_win\_no\_games  
kings\_win\_atleast\_1\_game

## [1] 0.7599

### 3. (2 points) Create a Monte Carlo simulation to confirm your answer to the previous question. Use B <- 10000 simulations. Hint: use the following code to generate the results of the first four games: kings\_wins <- sample(c(0,1), 4, replace = TRUE, prob = c(0.7, 0.3)). Noting that the Kings must win at least one of these four games.

set.seed(42) # Set seed for reproducibility  
  
B <- 10000 # Number of simulations  
kings\_prob <- 0.3  
warriors\_prob <- 0.7  
wins <- numeric(B)  
  
for (i in 1:B) {  
 kings\_wins <- sample(c(0,1), 4, replace = TRUE, prob = c(warriors\_prob, kings\_prob))  
 if (sum(kings\_wins) > 0) {  
 wins[i] <- 1  
 }  
}  
  
kings\_win\_atleast\_1\_game\_mc <- mean(wins)  
kings\_win\_atleast\_1\_game\_mc

## [1] 0.7594

### 4. (2 points) Suppose two NBA teams, say the Warriors and the Bucks, are playing a seven game championship series (The first to win four games, therefore, wins the series). The two teams are equally good so they each have a 50-50 chance of winning each game. If the Warriors lose the first game, what is the probability that they win the series?

p\_warriors <- 0.5  
p\_bucks <- 0.5  
#Warriors need 4 games to win the series  
r <- c(4, 5, 6)  
p\_warriors\_win\_series\_and\_lose\_game\_1 <- sum(p\_bucks \* choose(6, r) \* p\_warriors^r \* p\_bucks^(6-r))  
p\_answer <- p\_warriors\_win\_series\_and\_lose\_game\_1 / p\_bucks  
p\_answer

## [1] 0.34375

### 5. (2 points) Create a Monte Carlo simulation to confirm your answer to the previous question. Use B <- 10000 simulations.

set.seed(41)  
B <- 10000  
warriors\_wins <- matrix(sample(c(0, 1), B\*6, replace = TRUE, prob = c(0.5, 0.5)), ncol = 6)  
series\_count <- sum(rowSums(warriors\_wins) >= 4)  
p\_answer <- series\_count/B  
p\_answer

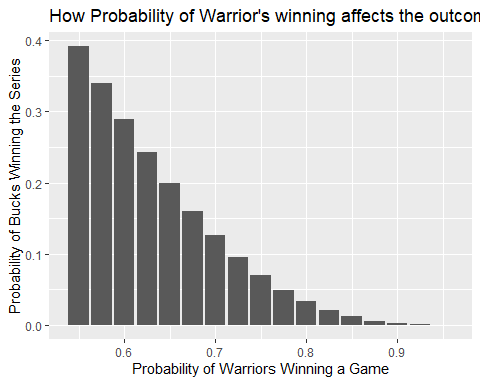
## [1] 0.3451

### 6. (2 points) Suppose two NBA teams, say the Warriors and the Bucks, are playing a seven game championship series (The first to win four games, therefore, wins the series). The Warriors is better than the Bucks and has a p>0.5 chance of winning each game. Given a value p, use the function sapply to compute the probability of winning the series for the Bucks for p <- seq(0.55, 0.95, 0.025). Then plot the result with geom\_histogram().

library(ggplot2)  
  
probability\_bucks\_win <- function(x) {  
 n <- 7  
 prob <- sum(dbinom((floor(n/2)+1):n, n, x))  
 return(1 - prob)  
}  
  
p <- seq(0.55, 0.95, 0.025)  
probability\_bucks\_win <- sapply(p, probability\_bucks\_win)  
probability\_bucks\_win

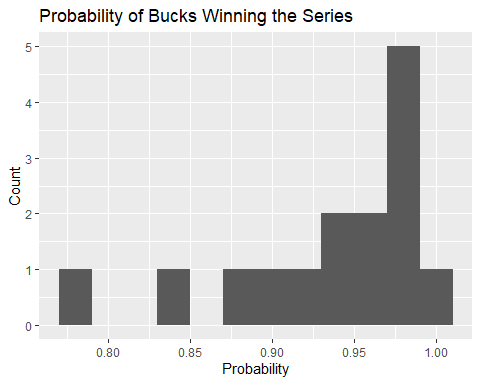
## [1] 0.3917122031 0.3395793392 0.2897920000 0.2430210114 0.1998457344  
## [6] 0.1607358134 0.1260360000 0.0959546691 0.0705566406 0.0497609242  
## [11] 0.0333440000 0.0209492526 0.0121031719 0.0062389374 0.0027280000  
## [16] 0.0009202775 0.0001935781

df <- data.frame(p, probability\_bucks\_win)  
  
ggplot(df, aes(x = p, y = probability\_bucks\_win)) +  
 geom\_col() +   
 labs(title = "How Probability of Warrior's winning affects the outcome of the series",  
 x = "Probability of Warriors Winning a Game",  
 y = "Probability of Bucks Winning the Series")



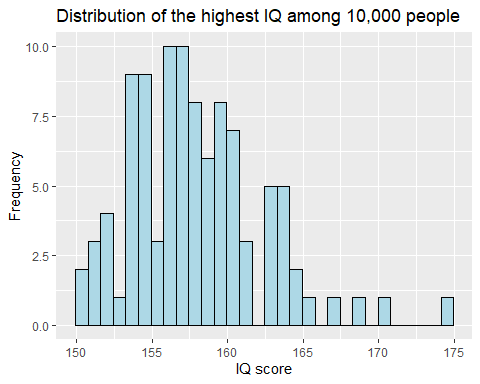
### 7. (1 points) Repeat the question above, but now keep the probability fixed at p <- 0.7 and compute the probability of winning the series for the Bucks for different series lengths: best of 3 games, 5 games, 7 games,… Specifically, N <- seq(3, 31, 2). Then plot the result with geom\_histogram().

library(ggplot2)  
  
probability\_bucks\_win <- function(x, n) {  
 prob <- sum(dbinom((floor(n/2)+1):n, n, x))  
 return(prob)  
}  
  
p <- 0.7  
N <- seq(3, 31, 2)  
prob\_bucks\_win <- sapply(N, function(x) probability\_bucks\_win(p, x))  
  
df <- data.frame(N, prob\_bucks\_win)  
  
ggplot(df, aes(x = prob\_bucks\_win)) +  
 geom\_histogram(binwidth = 0.02) +   
 labs(title = "Probability of Bucks Winning the Series",  
 x = "Probability",  
 y = "Count")



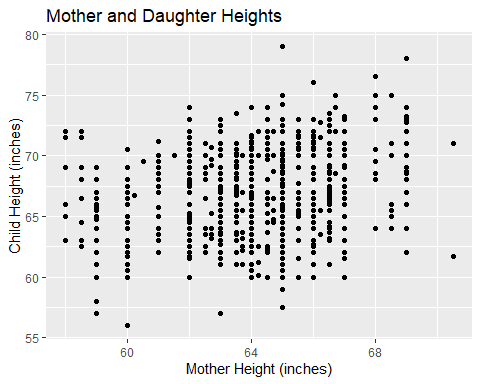
### 8. (2 points) The distribution of IQ scores is approximately normally distributed. The average is 100 and the standard deviation is 15. Suppose you want to know the distribution of the highest IQ among 10,000 people. Run a Monte Carlo simulation with B=100 generating 10,000 IQ scores and keeping the highest. Then plot the result with geom\_histogram().

B <- 100 # number of simulations  
n <- 10000 # number of people  
mu <- 100 # mean IQ  
sigma <- 15 # standard deviation of IQ  
  
# Simulate B samples of n IQ scores and keep the highest  
highest\_iq <- replicate(B, max(rnorm(n, mean = mu, sd = sigma)))  
  
# Plot the histogram of the highest IQ scores  
library(ggplot2)  
ggplot(data.frame(highest\_iq), aes(x = highest\_iq)) +  
 geom\_histogram(bins = 30, fill = "lightblue", color = "black") +  
 labs(title = "Distribution of the highest IQ among 10,000 people",  
 x = "IQ score", y = "Frequency")



### 9. (2 points) Load the GaltonFamilies data from the HistData. Make four separated scatterplots for heights between mothers and daughters, mothers and sons, fathers and daughters, and fathers and sons. Compute the correlation in heights between mothers and daughters, mothers and sons, fathers and daughters, and fathers and sons.

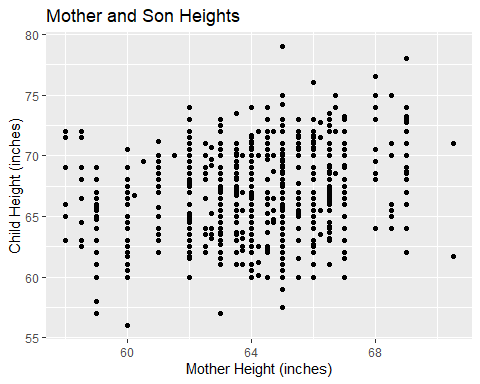
# Load the HistData package  
library(HistData)  
  
# Subset the GaltonFamilies dataset to only include height data  
data("GaltonFamilies")  
height\_data <- GaltonFamilies[, c("mother", "father", "childHeight")]  
  
# Rename columns for better readability  
colnames(height\_data) <- c("Mother Height", "Father Height", "Child Height")  
  
# Scatterplot of mother and daughter heights  
library(ggplot2)  
ggplot(height\_data, aes(x = `Mother Height`, y = `Child Height`)) +  
 geom\_point() +  
 labs(title = "Mother and Daughter Heights",  
 x = "Mother Height (inches)", y = "Child Height (inches)")



# Correlation between mother and daughter heights  
cor(height\_data$`Mother Height`, height\_data$`Child Height`)

## [1] 0.2013219

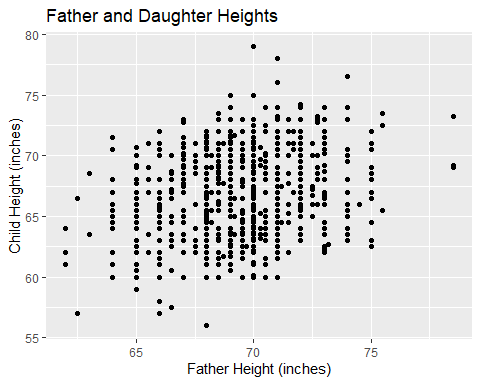
# Scatterplot of mother and son heights  
ggplot(height\_data, aes(x = `Mother Height`, y = `Child Height`)) +  
 geom\_point() +  
 labs(title = "Mother and Son Heights",  
 x = "Mother Height (inches)", y = "Child Height (inches)")



# Correlation between mother and son heights  
cor(height\_data$`Mother Height`, height\_data$`Child Height`[height\_data$`Child Height` > 0 & height\_data$`Father Height` > 0])

## [1] 0.2013219

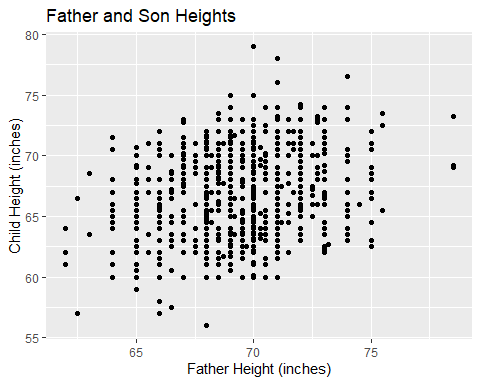
# Scatterplot of father and daughter heights  
ggplot(height\_data, aes(x = `Father Height`, y = `Child Height`)) +  
 geom\_point() +  
 labs(title = "Father and Daughter Heights",  
 x = "Father Height (inches)", y = "Child Height (inches)")



# Correlation between father and daughter heights  
cor(height\_data$`Father Height`, height\_data$`Child Height`[height\_data$`Child Height` > 0 & height\_data$`Mother Height` > 0])

## [1] 0.2660385

# Scatterplot of father and son heights  
ggplot(height\_data, aes(x = `Father Height`, y = `Child Height`)) +  
 geom\_point() +  
 labs(title = "Father and Son Heights",  
 x = "Father Height (inches)", y = "Child Height (inches)")



# Correlation between father and son heights  
cor(height\_data$`Father Height`, height\_data$`Child Height`)

## [1] 0.2660385

### 10. (2 points) Load the GaltonFamilies data from the HistData. Create a dataset called galton\_heights by randomly picking a daughter of each family. galton\_heights should have two columns, including father’s and daughter’s height. Using the lm function to obtain the least squares estimates between the father’s and daughter’s height. What is the estimated model coefficients.

set.seed(42)  
library(HistData)  
library(dplyr)  
  
galton\_heights <- GaltonFamilies |>  
 filter(gender == "female") |>  
 group\_by(family) |>  
 sample\_n(1) |>  
 ungroup() |>  
 select(father, childHeight) |>  
 rename("Father Height" = father, "Daughter Height" = childHeight)   
  
galton\_heights

## # A tibble: 176 × 2  
## `Father Height` `Daughter Height`  
## <dbl> <dbl>  
## 1 78.5 69.2  
## 2 75.5 65.5  
## 3 75 68   
## 4 75 67   
## 5 75 62.5  
## 6 74 69.5  
## 7 74 64   
## 8 74 68   
## 9 74.5 66   
## 10 74 65.5  
## # ℹ 166 more rows

# fit linear regression model  
galton\_heights |> lm() |> summary()

##   
## Call:  
## lm(formula = galton\_heights)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6.0283 -1.7746 0.2104 1.4829 6.5790   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 36.80360 5.31487 6.925 8.10e-11 \*\*\*  
## `Daughter Height` 0.50748 0.08265 6.140 5.43e-09 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.483 on 174 degrees of freedom  
## Multiple R-squared: 0.1781, Adjusted R-squared: 0.1734   
## F-statistic: 37.7 on 1 and 174 DF, p-value: 5.426e-09

### 11. (17 points) Essay: (From Prof. RB Luo) In the midterm exam, we tried something different. The use of RStudio was allowed. The use of Google and ChatGPT was allowed. The use of instant messengers was allowed. After all, as I mentioned in my lecture, if you can only take one thing away from this course, make it “knowing how to get started when given a data science problem”. But I am unsure how well the trial has worked out, especially from my students’ perspective. If you have attended the midterm exam, how do you like the exam form? How would you like to improve the questions to help you to achieve the learning goals? If I ask you to be my TA and help me design the midterm exam for next year’s class, what would you suggest? More generally, what improvements to the course would you suggest so I can do better the next year? Let me know your thoughts because when one teaches, two learn.

As a machine learning engineer in real life, this is within the top 5 of the courses that I have ever taken at HKU. Usage of Rstudio, Google and chatGPT reflects what I am encouraged to do in my field of work. I am appreciative of the course efforts to mirror what tends to happen in the real data science world compared to other traditional assessments that don’t encourage a 21st-century approach to solving problems. While I didn’t use instant messengers, I am happy to have seen an avenue for collaboration being encouraged. This is very accurate when it comes to real-world technical development. The midterm exam was smooth for me because it allowed me to express the most efficient code that solved the problems. The internet was a very good resource this time.

In terms of improving the questions, I think it would be useful to have more questions that require students to apply what they’ve learned in a creative way instead of a multiple-choice approach. This approach may have encouraged some students to just paste the questions and answers to GPT and immediately get an answer. For example, the questions could be modelled to ask them to design a web scrapping tool for movies or a using dataset x, design a pipeline of doing y. I believe this would encourage diversity of answers and multiple approaches hence encouraging learning. If students would then compare this with each other, key weaknesses would be brought up encouraging learning by discussion and further reinforcing the practical skills that are so important in data science.

The course could be improved in a few ways:

1. If next year platforms like ed that have an embedded coding environment could help TAs to accurately pass feedback on assignments. I received “Q6,7” as feedback for assignmentbut could find specifically what I need to improve on
2. More individual projects. I remember as part of my internship as a Data Analyst for the ALG in South Africa, I had to design a project on my own that investigates and visualizes the crime rate in SA. That meant downloading my own unclean datasets and working from the ground up. While this view was applied in our end-of-term project, having to work on 2-3 solid projects within the term that do this instead of assignments that contain clean datasets, rarely in the real world, would expose students to more learning.