# Midterm

Michael Ghattas

## Instructions

Please work these problems on your own. You may use web searches and refer to your notes, lectures, and past problem sets, but not interactive methods such as asking others online or in person.

## Question 1

### Context for Question 1

A medical researcher investigates whether systolic blood pressure readings depend on the arm, left or right, from which an individual's blood pressure is measured. The researcher takes readings from both arms of 100 individuals and records $L$ if the measurement from the left arm is higher and $R$ if the measurement from the right arm is higher. Consider the model that the number of $L$s follows a binomial distribution with the size equal to 100 and the probability of success equal to 0.5. That is, the probability that $k$ individuals have measurements that result in $L$ is $f(k) = \left( \begin{array}{c} 100 \\ k \end{array} \right) (0.5)^k (1-0.5)^{100-k} = \left( \begin{array}{c} 100 \\ k \end{array} \right) (0.5)^{100}$. Denote this probability space by $binomial(100, 0.5)$.

### Question 1.1

(3 points)

What is the probability of the event that the number of $L$s is equal to 50 under the $binomial(100, 0.5)$ model?

**Your code and answer here:** For this specific problem, we have $n = 100$, $k = 50$, and $p = 0.5$: $P(X = 50) = \binom{100}{50}(0.5)^{50}(0.5)^{50} = \binom{100}{50}(0.5)^{100}$. The binomial coefficient $\binom{100}{50}$ can be calculated as: $\binom{100}{50} = \frac{100!}{50! \cdot 50!}$ Thus, the probability is: $P(X = 50) = \frac{100!}{50! \cdot 50!} \cdot (0.5)^{100}$. Using R, we find:

```
# Calculate the probability of getting exactly 50 Ls out of 100
prob_50_Ls <- dbinom(50, size = 100, prob = 0.5)
prob_50_Ls
```

```
## [1] 0.07958924
```

### Question 1.2

(3 points)

What is the probability of the event that the number of $L$s is less than or equal to 40 under the $binomial(100, 0.5)$ model?

**Your code and answer here:** To find the probability that the number of $L$s is less than or equal to 40, we need to sum the probabilities for $k = 0$ to $k = 40$: $P(X \leq 40) = \sum_{k=0}^{40} \binom{100}{k}(0.5)^k(0.5)^{100-k} = \sum_{k=0}^{40} \binom{100}{k}(0.5)^{100}$. Using R, we can calculate this cumulative probability:

```
# Calculate the cumulative probability of getting 40 or fewer Ls out of 100
prob_40_or_less_Ls <- pbinom(40, size = 100, prob = 0.5)
prob_40_or_less_Ls
```

```
## [1] 0.02844397
```

## Question 1.3

(3 points)

If the researcher treats $binomial(100, 0.5)$ as the null distribution for the null hypothesis that neither arm consistently produces larger measurements, is an observed count of $L$s equal to 60 strong evidence against the null hypothesis? Please explain.

**Your code and answer here:** We can use the cumulative distribution function (CDF) of the binomial distribution to find the probability of observing 60 or more $L$s: $P(X \geq 60) = 1 - P(X \leq 59)$. Where $P(X \leq 59)$ is the cumulative probability up to 59. Using R, we can calculate this cumulative probability and the resulting $p$-value. A small $p$-value (typically less than 0.05) indicates strong evidence against the null hypothesis.

```
# Calculate the cumulative probability of getting 59 or fewer Ls out of 100
prob_59_or_less_Ls <- pbinom(59, size = 100, prob = 0.5)
# Calculate the probability of getting 60 or more Ls out of 100
p_value <- 1 - prob_59_or_less_Ls
p_value
```

```
## [1] 0.02844397
```

Since the calculated $p - value$ is less than 0.05, it indicates that an observed count of $L$'s equal to 60 is strong evidence against the null hypothesis that neither arm consistently produces larger measurements.

## Question 1.4

(3 points)

If the researcher treats $binomial(100, 0.5)$ as the null distribution for the null hypothesis that neither arm consistently produces larger measurements, what would be an unusually large or small observation under the null hypothesis? Please explain your answer and how you quantify unusual?

**Your code and answer here:** We need to find the critical values $k$ such that: $P(X \leq k_{\text{low}}) \leq 0.025$ and $P(X \geq k_{\text{high}}) \leq 0.025$. Using the cumulative distribution function (CDF) for the binomial distribution, we can find these critical values. In R, this can be achieved using the `qbinom` function.

```
# Define the significance level
alpha <- 0.05
# Find the critical value for the lower tail (2.5%)
k_low <- qbinom(alpha / 2, size = 100, prob = 0.5)
# Find the critical value for the upper tail (97.5%)
k_high <- qbinom(1 - alpha / 2, size = 100, prob = 0.5)
list(k_low = k_low, k_high = k_high)
```

```
## $k_low
## [1] 40
##
## $k_high
## [1] 60
```

Therefore, an observation of 40 or fewer $L$'s would be unusually small, and an observation of 60 or more $L$'s would be unusually large under the null hypothesis. This quantification is based on the typical 5 significance level, meaning that such observations would occur by chance less than 5 of the time under the null hypothesis.

# Question 2

Consider the experiment of tossing two fair 6-sided dice.

## Question 2.1

(3 points)

Please define a reasonable sample space and event space to model this experiment.

**Your answer here:** The sample space $S$ for tossing two fair 6-sided dice consists of all possible ordered pairs $(d_1, d_2)$ where $d_1$ and $d_2$ represent the outcomes on the first and second die, respectively. Since each die has 6 faces, there are a total of $6 \times 6 = 36$ possible outcomes: $S = \{(d_1, d_2) \mid d_1, d_2 \in \{1, 2, 3, 4, 5, 6\}\}$. An event $E$ is a subset of the sample space $S$. Here are all possible outcomes: $S = \{(1,1), (1,2), (1,3), (1,4), (1,5), (1,6), (2,1), (2,2), (2,3), (2,4), (2,5), (2,6),$
$(3,1), (3,2), (3,3), (3,4), (3,5), (3,6), (4,1), (4,2), (4,3), (4,4), (4,5), (4,6),$
$(5,1), (5,2), (5,3), (5,4), (5,5), (5,6), (6,1), (6,2), (6,3), (6,4), (6,5), (6,6)\}$

## Question 2.2

(3 points)

What is the probability of an outcome $(a, b)$ where a is the outcome of the first toss of the die, and b is the outcome of the second toss of the die?

**Your answer here:** Since each die is fair and has 6 faces, each face has an equal probability of appearing. The probability of any specific outcome on one die is $\frac{1}{6}$. When tossing two fair 6-sided dice, the outcome of each die is independent of the other. Therefore, the probability of any specific pair $(a, b)$ occurring, where $a$ is the outcome of the first die and $b$ is the outcome of the second die, is the product of the probabilities of each individual outcome. $P((a, b)) = P(a) \times P(b)$. Given that $P(a) = \frac{1}{6}$ and $P(b) = \frac{1}{6}$, we have: $P((a, b)) = \left(\frac{1}{6}\right) \times \left(\frac{1}{6}\right) = \frac{1}{36}$. Therefore, the probability of any specific outcome $(a, b)$ where $a$ is the outcome of the first toss and $b$ is the outcome of the second toss is: $P((a, b)) = \frac{1}{36}$.

## Question 2.3

(3 points)

What is the probability of the event $E = \{(a, b) \,|a \geq 3b\}$?

**Your answer here:** We need to determine the pairs $(a, b)$ where $a$ is at least three times $b$. Since both $a$ and $b$ can take values from 1 to 6 (because the dice are 6-sided), we identify the pairs that satisfy $a \geq 3b$: 1. $b = 1 : \{(3,1), (4,1), (5,1), (6,1)\}$ 2. $b = 2 : \{(6,2)\}$ For $b \geq 3$, there are no possible values of $a$ that satisfy $a \geq 3b$ because $a$ can only go up to 6. Thus, the event $E$ consists of the following pairs:

$E = \{(3,1),(4,1),(5,1),(6,1),(6,2)\}$. There are 5 such pairs. The total number of possible outcomes when tossing two fair 6-sided dice is 36. The probability of the event $E$ is the number of favorable outcomes divided by the total number of possible outcomes: $P(E) = \frac{|E|}{|S|} = \frac{5}{36}$. Therefore, the probability of the event $E$ is: $P(E) = \frac{5}{36}$.

### Question 2.4

(3 points)

What is the probability of the event $B = \{(a,1)\,|b = 1\}$ that the second toss $b$ is has a value of 1?

**Your answer here:** To determine the probability of the event $B$ where the second die shows a 1, we need to identify all the pairs $(a,1)$ where $a$ can be any value from 1 to 6, and $b$ is fixed at 1. The possible pairs are: $B = \{(1,1),(2,1),(3,1),(4,1),(5,1),(6,1)\}$. There are 6 such pairs. The total number of possible outcomes when tossing two fair 6-sided dice is 36. The probability of the event $B$ is the number of favorable outcomes divided by the total number of possible outcomes: $P(B) = \frac{|B|}{|S|} = \frac{6}{36} = \frac{1}{6}$. Therefore, the probability of the event $B$ is: $P(B) = \frac{1}{6}$.

### Question 2.5

(3 points)

Are the events $B = \{(a,1)\,|b = 1\}$ and $E = \{(a,b)\,|a \geq 3b\}$ independent?

**Your answer here:** To determine if the events $B$ and $E$ are independent, we need to check if: $P(B \cap E) = P(B) \cdot P(E)$. First, recall the definitions and probabilities of the events: 1. Event $B$: The second die shows a 1, $B = \{(1,1),(2,1),(3,1),(4,1),(5,1),(6,1)\}$, $P(B) = \frac{6}{36} = \frac{1}{6}$. 2. Event $E$: The first die is at least three times the second die, $E = \{(3,1),(4,1),(5,1),(6,1),(6,2)\}$, $P(E) = \frac{5}{36}$. Now, consider the intersection $B \cap E = \{(3,1),(4,1),(5,1),(6,1)\}$. There are 4 outcomes in $B \cap E$. The probability of the intersection is: $P(B \cap E) = \frac{|B \cap E|}{|S|} = \frac{4}{36} = \frac{1}{9}$. Now, we compare $P(B \cap E)$ with $P(B) \cdot P(E) = \frac{1}{6} \cdot \frac{5}{36} = \frac{5}{216}$. We need to see if: $P(B \cap E) = P(B) \cdot P(E)$, $\frac{1}{9} = \frac{5}{216}$. Clearly, $\frac{1}{9} \neq \frac{5}{216}$. Therefore, the events $B$ and $E$ are not independent because $P(B \cap E) \neq P(B) \cdot P(E)$.

## Question 3

Consider a continuous random variable $X$ with the probability density function defined by $f(x) = 4x^3$ for $x \in [0,1]$ and $f(x) = 0$ otherwise.

### Question 3.1

(3 points)

What is cumulative distribution of $X$?

**Your answer here:** Given the PDF $f(x) = 4x^3$ for $x \in [0,1]$ and $f(x) = 0$ otherwise, we can find the CDF $F(x)$ for different ranges of $x$. 1. For $(x < 0) : F(x) = 0$ 2. For $(0 \leq x \leq 1) : F(x) = \int_0^x 4t^3\,dt$ 3. For $x > 1$: $F(x) = 1$ (We calculate the integral: $F(x) = 4\int_0^x t^3\,dt = 4\left[\frac{t^4}{4}\right]_0^x = 4\left(\frac{x^4}{4} - 0\right) = x^4$). Thus, the cumulative distribution function $F(x)$ of $X$ is:

$$F(x) = \begin{cases} 0 & \text{if } x < 0 \\ x^4 & \text{if } 0 \leq x \leq 1 \\ 1 & \text{if } x > 1 \end{cases}$$

### Question 3.2

(4 points)

What is the expected value of $X$, $E[X]$?

**Your answer here:** Given the PDF $f(x) = 4x^3$ for $x \in [0, 1]$ and $f(x) = 0$ otherwise, we can find the expected value $E[X]$ by integrating over the range where $f(x)$ is non-zero: $E[X] = \int_0^1 x \cdot 4x^3 \, dx$. Simplify the integrand: $E[X] = \int_0^1 4x^4 \, dx$. Now, calculate the integral: $E[X] = 4 \int_0^1 x^4 \, dx = 4 \left[ \frac{x^5}{5} \right]_0^1 = 4 \left( \frac{1^5}{5} - \frac{0^5}{5} \right) = 4 \cdot \frac{1}{5} = \frac{4}{5}$. Therefore, the expected value of $X$ is: $E[X] = \frac{4}{5}$.

### Question 3.3

(4 points)

What is the variance of $X$, $Var[X]$?

**Your answer here:** Given the PDF $f(x) = 4x^3$ for $x \in [0, 1]$ and $f(x) = 0$ otherwise, we can find $E[X^2]$ by integrating over the range where $f(x)$ is non-zero: $E[X^2] = \int_0^1 x^2 \cdot 4x^3 \, dx$. Simplify the integrand: $E[X^2] = \int_0^1 4x^5 \, dx$. Now, calculate the integral: $E[X^2] = 4 \int_0^1 x^5 \, dx = 4 \left[ \frac{x^6}{6} \right]_0^1 = 4 \left( \frac{1^6}{6} - \frac{0^6}{6} \right) = 4 \cdot \frac{1}{6} = \frac{4}{6} = \frac{2}{3}$. We already found the expected value $E[X] = \frac{4}{5}$. Now, we can find the variance $\text{Var}[X] = E[X^2] - (E[X])^2 = \frac{2}{3} - \left( \frac{4}{5} \right)^2 = \frac{2}{3} - \frac{16}{25} = \frac{2}{75}$. Therefore, the variance of $X$ is: $\text{Var}[X] = \frac{2}{75}$.

## Question 4

Consider the hypothetical scenario in which a pharmaceutical company is determining effect of an experimental drug on a rare disease. Individuals in this study (800 individuals) are randomly assigned to the treatment (in which the individual receives the drug) and (1000 are assigned) to the placebo. The matrix m below represents the data in this hypothetical scenario. Consider the null hypothesis that individuals who receive the placebo and individuals who receive the treatment have the same susceptibility to the disease. One possible probability model for this is that each case in the pooled `Placebo` and `Treatment` cases can be viewed as being assigned to the `Treatment` group with probability equal to the proportion of the `Treatment` population in the pooled `Placebo` and `Treatment` population. In the questions below we will assess the claim that the drug has no effect.

```
d<-rbind(c(800,3), c(1000,13))
m<-as.matrix(d)
colnames(m)= c("Population","Disease")
rownames(m) = c("Treatment", "Placebo")
m
```

```
##           Population Disease
## Treatment        800       3
## Placebo         1000      13
```

### Question 4.1

(3 points)

Using the same null model described above, please calculate the probability that the disease cases in the treatment group under the null model is less than or equal to `m$Disease[1]` directly. Recall that the function pbinom(x,size,prob) returns the probability of the event that the number of successes is in the set $\{0, 1, ...x\}$.

**Your code and answer here:**

```
# Define the parameters
size = 16
prob = (800 / 1800)
x = 3

# Calculate the probability using the binomial distribution function
probability <- pbinom(x, size = size, prob = prob)
probability
```

```
## [1] 0.03107061
```

## Question 4.2

(3 points)

Use the function `rbinom(n,size,prop)` to simulate the number of individuals with the disease in the treatment group under the null model and null hypothesis. Use the function to generate 10,000 random numbers representing the instances of the disease in the treatment group under the null model and null hypothesis, and assign these 10,000 random numbers to `sim_counts` in the code chunk below. Discuss each parameter used in the rbinomial function x, size, and prob.

**Your code and answer here:**

```
set.seed(1)
n = 10000
size = 16
prob = (800 / 1800)
sim_counts <- rbinom(n, size = size, prob = prob)
```

## Question 4.3

(3 points)

Use ggplot to display a histogram of `sim_counts` with the appropriate number of bins. On the same graph with the histogram display the number of actual observed counts of the disease in the treatment group.

**Your code and answer here:**

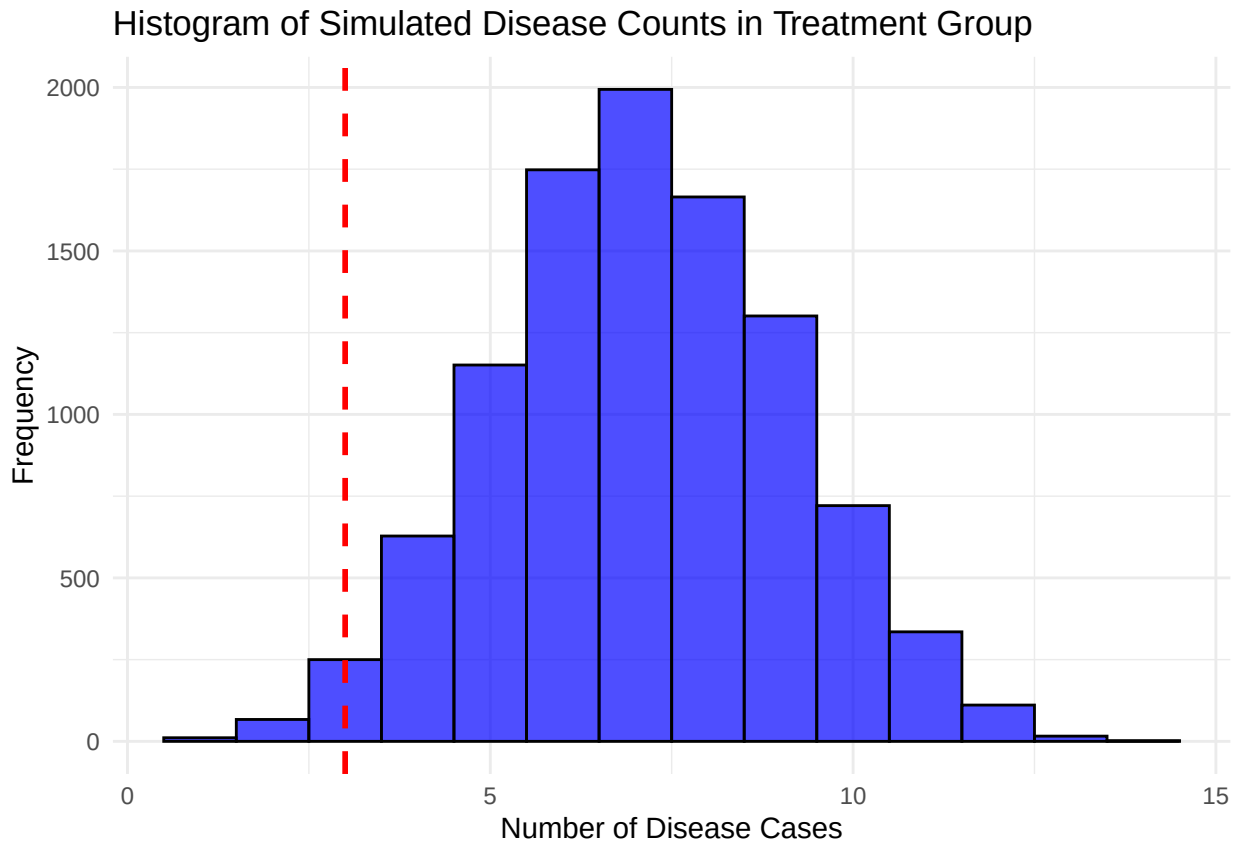
```
# Load necessary library
library(ggplot2)

# Create a data frame for ggplot
sim_data <- data.frame(sim_counts = sim_counts)

# Create the histogram
ggplot(sim_data, aes(x = sim_counts)) +
  geom_histogram(binwidth = 1, fill = "blue", color = "black", alpha = 0.7) +
  geom_vline(aes(xintercept = 3), color = "red", linetype = "dashed", size = 1) +
```

```
  labs(title = "Histogram of Simulated Disease Counts in Treatment Group",
       x = "Number of Disease Cases",
       y = "Frequency") +
  theme_minimal()
```

```
## Warning: Using ‘size‘ aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use ‘linewidth‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```




## Question 4.4

(3 points)

What would be an usually small number of observed counts of the disease in the treatment group under the null model/hypothesis? Quantify what you mean by unusual.

**Your code and answer here:**

```
# Calculate the 5th percentile of the simulated counts
lower_bound <- quantile(sim_counts, 0.05)
lower_bound
```

```
## 5%
##  4
```

An unusually small number of observed counts of the disease in the treatment group under the null model/hypothesis would be any number less than or equal to 4. This is because the 5th percentile of the distribution of simulated counts is 4, meaning that only 5% of the simulations resulted in 4 or fewer cases. Therefore, we can consider observing 4 or fewer cases as unusual under the null hypothesis.

### Question 4.5

(3 points)

Assess the null model and hypothesis. Is the drug effective, and how certain can you be of the results? Can the drug company conclude that the drug causes a reduction in how likely an individual is to contract this rare disease?

**Your answer here:** Given the low probability of observing 3 or fewer disease cases under the null hypothesis, and the fact that this observation is considered unusually low, we have strong evidence to reject the null hypothesis. This indicates that the drug is effective in reducing the likelihood of contracting the disease. Certainty of Results: The results are statistically significant at the 5% level. We can be reasonably certain (with a confidence level of at least 95%) that the drug has an effect. Conclusion: Based on this analysis, the drug company can conclude that the drug likely causes a reduction in the likelihood of contracting this rare disease. However, it is important to consider other factors such as sample size, potential biases, and the need for further studies to confirm the findings.

## Question 5

(6 points)

Consider a continuous random variable $X$ with the probability density function defined by $f(x) = \frac{10}{x^2}$ for $x \in [10, \infty)$ and $f(x) = 0$ otherwise. Please give the expected value $E[X]$ of $X$ if possible, or explain why it is not possible.

**Your answer here:** Given the PDF $f(x) = \frac{10}{x^2}$ for $x \in [10, \infty)$ and $f(x) = 0$ otherwise, we can find the expected value $E[X]$ by integrating over the range where $f(x)$ is non-zero: $E[X] = \int_{10}^{\infty} x \cdot \frac{10}{x^2}\, dx$. Simplify the integrand: $E[X] = \int_{10}^{\infty} \frac{10}{x}\, dx$. Now, calculate the integral: $E[X] = 10 \int_{10}^{\infty} \frac{1}{x}\, dx = 10\,[\ln|x|]_{10}^{\infty}$. Evaluate the integral: $E[X] = 10\,(\lim_{x\to\infty} \ln|x| - \ln|10|)$. Since $\ln|x|$ approaches infinity as $x$ approaches infinity: $E[X] = 10\,(\infty - \ln|10|) = \infty$. Therefore, the expected value $E[X]$ is not finite; it is infinite. This means that the expected value does not exist in the conventional sense because the integral diverges.

## Question 6

Consider the scenarios below on estimation, and answer the questions in each scenario.

### Question 6.1

(5 points)

If you model the data below (represented as b) as the result of 8 independent Bernoulli trials with probability of success equal to $p$, what is the maximum likelihood estimate of $p$? Please show your work mathematically.

```
b <- c("success", "success", "failure", "failure", "failure", "success","failure", "failure")
```

**Your answer here:** Given the data $b$ as the result of 8 independent Bernoulli trials, we want to find the maximum likelihood estimate (MLE) of the probability of success $p$. Let $X_i$ be the outcome of the $i^{th}$ trial, where $X_i = 1$ if the trial is a success and $X_i = 0$ if the trial is a failure. The probability mass function for a Bernoulli random variable is: $P(X_i = x_i) = p^{x_i}(1-p)^{1-x_i}$. The likelihood function for the 8 trials is: $L(p) = \prod_{i=1}^{8} P(X_i = x_i) = \prod_{i=1}^{8} p^{x_i}(1-p)^{1-x_i}$. Given the data we have: $\sum_{i=1}^{8} X_i = 3$ (number of successes). Thus, the likelihood function becomes: $L(p) = p^3(1-p)^5$. To find the maximum likelihood estimate of $p$, we take the natural logarithm of the likelihood function to obtain the log-likelihood function: $\ell(p) = \ln L(p) = \ln(p^3(1-p)^5) = 3\ln p + 5\ln(1-p)$. Next, we differentiate the log-likelihood function with respect to $p$, set the derivative to zero to find the critical points, and solve for $p$:

$$\frac{d\ell(p)}{dp} = \frac{3}{p} - \frac{5}{1-p} = 0$$

$$\frac{3}{p} = \frac{5}{1-p}$$

$$3(1-p) = 5p$$

$$3 - 3p = 5p$$

$$3 = 8p$$

$$p = \frac{3}{8}$$

Thus, the maximum likelihood estimate of $p$ is: $\hat{p} = \frac{3}{8}$.

## Question 6.2

(5 points)

If you model the data below as a sample from a Normal distribution with mean $\mu = 1.5$, what is the maximum likelihood estimate for the $\sigma^2$ in the density $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$ for the data given in the sample? Find the MLE mathematically. Then in code, use the estimate of $\sigma^2$ to determine the value of the cumulative distribution $\phi(x)$ such that $\phi(x) = .85$.

```
sample_data = c(-1.1,  3.7,  1.6,  9.2,  1.9,  4.1)
```

**Your answer and code here:**

```
# Given sample data
sample_data = c(-1.1, 3.7, 1.6, 9.2, 1.9, 4.1)
mu = 1.5

# Calculate the MLE for sigma^2
n <- length(sample_data)
sigma2_hat <- (sum((sample_data - mu)^2) / n)
sigma2_hat
```

```
## [1] 12.97
```

```r
# Determine the value of the cumulative distribution phi(x) such that phi(x) = 0.85
phi_x <- qnorm(0.85, mean = mu, sd = sqrt(sigma2_hat))
phi_x
```

```
## [1] 5.232599
```

# Question 7

The following questions use the data "usa_00022.csv" provided with this assignment. These data are from The Census Bureau's American Community Survey (ACS) Public Use Microdata Sample (PUMS) for 2021. They were downloaded from IPUMS: Steven Ruggles, Sarah Flood, Ronald Goeken, Megan Schouweiler and Matthew Sobek. IPUMS USA: Version 12.0 [dataset]. Minneapolis, MN: IPUMS, 2022. https://doi.org/10.18128/D010.V12.0

The PUMA to county relationships were generated on https://mcdc.missouri.edu/applications/geocorr2022.html

```r
dat <- read.csv("usa_00022.csv")
```

## Question 7.1

(3 points)

Please generate a data frame from the full data "usa_00022.csv" which is restricted to respondents in the Public Use Microsample Areas (PUMAs) 812 through 816 inclusive, that is, cases for which the value of the variable "PUMA" is greater than or equal to 812 and less than or equal to 816. These are predominantly in **Denver**, CO. How many cases are in the restricted data set?

**Your answer and code here:**

```r
# Load the necessary library
library(dplyr)

# Filter the data for PUMAs 812 through 816
restricted_data <- dat %>% filter(PUMA >= 812 & PUMA <= 816)

# Get the number of cases in the restricted data set
num_cases <- nrow(restricted_data)
num_cases
```

```
## [1] 6333
```

## Question 7.2

(3 points)

For the data set below of employed individuals not living in group quarters, please report the slope and intercept of the least squares best fit line for the model $TRANTIME = m(INCWAGE) + b$, a linear model of "TRANTIME", top-coded total daily commute time in minutes, on "INCWAGE", top-coded wage and salary income. The values 999999 and 999998 code missing data for INCWAGE.

**Your code and answer here**

```r
dat.trans <- filter(dat, EMPSTAT == 1, GQ == 1 | GQ == 2, !(INCWAGE %in% c(999999, 999998)))
# Fit the linear model
model <- lm(TRANTIME ~ INCWAGE, data = dat.trans)
# Get the slope and intercept of the model
slope <- coef(model)[2]
intercept <- coef(model)[1]
# Print the results
print(c(intercept, slope))
```

```
##   (Intercept)       INCWAGE
##  1.905176e+01 -1.181462e-05
```

## Question 7.3

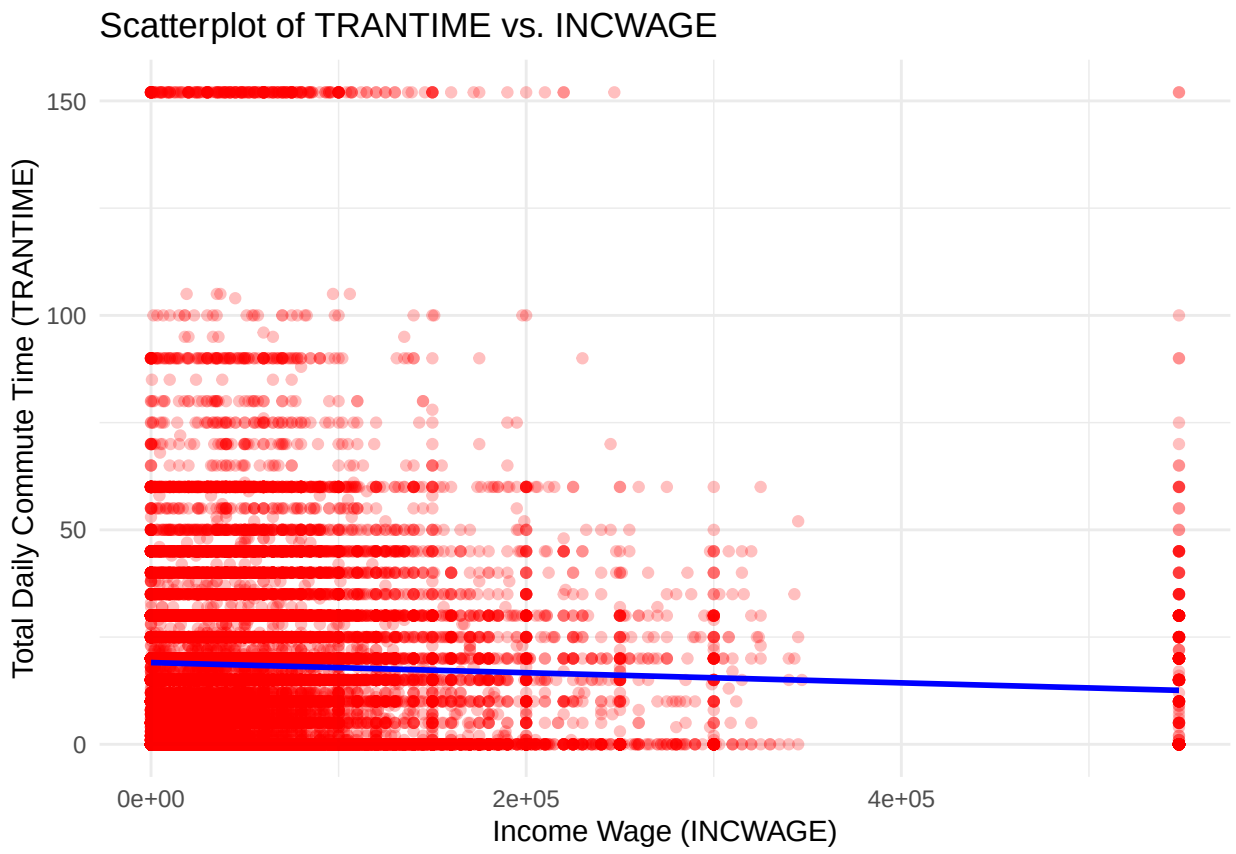
(3 points)

Make a scatterplot with the "INCWAGE" variable on the horizontal axis and the "TRANTIME" variable on the vertical axis. Add the line computed in above. (If you used the built-in function, please extract the values of the slope and intercept from the fitted model object, rather than using copy-paste.) For full credit, please use a strategy to aid visualization of regions of the plot with many points.

**Your code here:**

```r
# Create the scatterplot
ggplot(dat.trans, aes(x = INCWAGE, y = TRANTIME)) +
  geom_point(alpha = 0.25, color = "red") +  # Use transparency to aid visualization
  geom_smooth(method = "lm", se = FALSE, color = "blue") +  # Add the regression line
  labs(title = "Scatterplot of TRANTIME vs. INCWAGE",
       x = "Income Wage (INCWAGE)",
       y = "Total Daily Commute Time (TRANTIME)") +
  theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




# Question 8

Data on a quantity can come in the form of the maximum of measurement of the quantity over multiple observations. To explore this, using the random seed below, please sample 4 values from the uniform distribution on [0,1] using "runif". Repeat this 8 times and create a matrix of the values with the first sample of 4 in the first row, the second in the second row, and so on. The "byrow" argument to the "matrix" function may be useful.

## Question 8.1

(3 points)

Please create and print/display the matrix.

**Your code here:**

```
set.seed(12345)
# Sample 4 values from uniform distribution on [0,1], repeated 8 times
samples <- replicate(8, runif(4))
# Create a matrix with the samples, filling by row
sample_matrix <- matrix(samples, nrow = 8, byrow = TRUE)
# Print/display the matrix
print(sample_matrix)
```

```
##           [,1]        [,2]       [,3]       [,4]
## [1,] 0.7209039 0.875773193 0.76098233 0.88612457
## [2,] 0.4564810 0.166371785 0.32509539 0.50922434
## [3,] 0.7277053 0.989736938 0.03453544 0.15237349
## [4,] 0.7356850 0.001136587 0.39120334 0.46249465
## [5,] 0.3881440 0.402485142 0.17896358 0.95165875
## [6,] 0.4537281 0.326752409 0.96541532 0.70748188
## [7,] 0.6445426 0.389828485 0.69854364 0.54405786
## [8,] 0.2264672 0.484557755 0.79300717 0.00598763
```

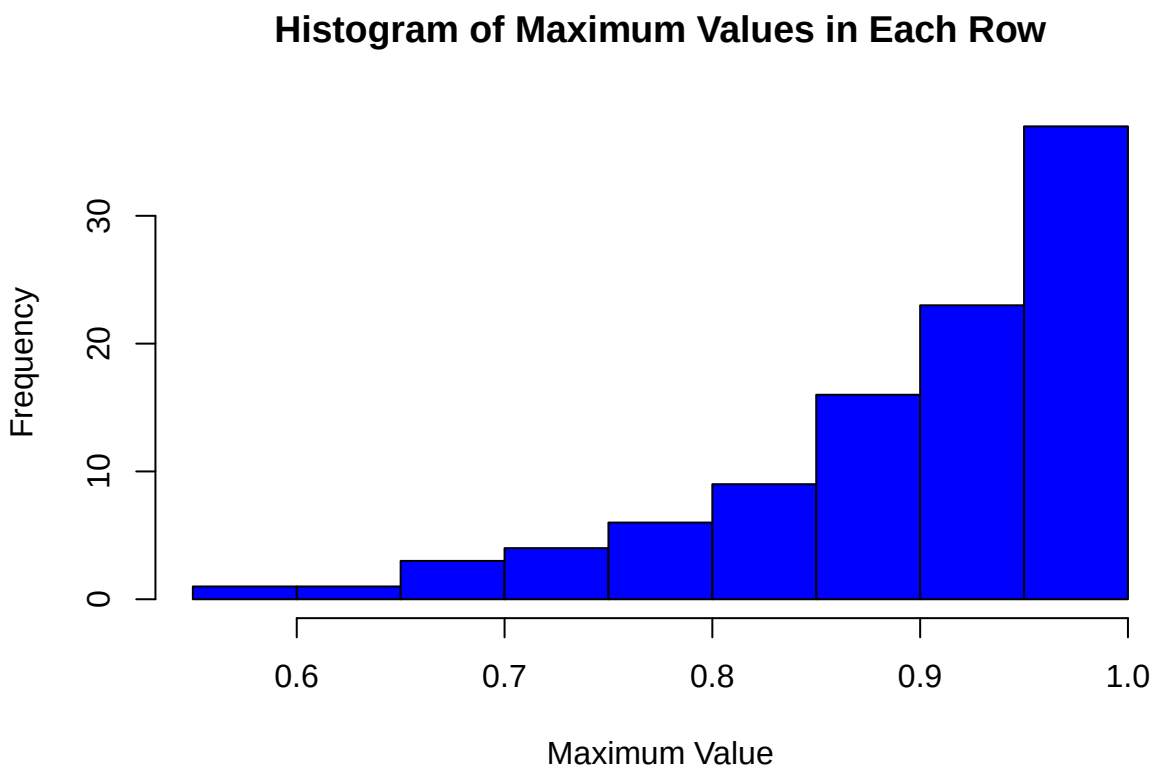
## Question 8.2

(4 points)

For the matrix created below, please display a histogram of the maximum value in each row using the default binwidth.

**Your code here:**

```
set.seed(12345)
mat2 <- matrix(runif(800), ncol = 8)
# Calculate the maximum value in each row
row_max <- apply(mat2, 1, max)
# Display a histogram of the maximum values
hist(row_max, main = "Histogram of Maximum Values in Each Row",
     xlab = "Maximum Value", ylab = "Frequency", col = "blue")
```



# Question 9

Two discrete random variables $X$ and $Y$ have a joint distribution of

$$f(x,y) = \frac{x+y+1}{27}$$

where the values of $x$ can be $0, 1, 2$ and $y$ can be $0, 1, 2$.

*Hint for this question:* For a continuous random variable defined on sample space $\mathcal{S}$, $p_X(x) = \int_{y\in\mathcal{S}} p_{X|Y}(x,y)p_Y(y)dy = E_Y[p_{X|Y}(x|y)]$

## Question 9.1

(3 points)

What is $P(X \leq 1, Y = 1)$?

**Your answer here:** Given the joint distribution $f(x,y) = \frac{x+y+1}{27}$, we can find the required probabilities: 1. $P(X = 0, Y = 1) : f(0,1) = \frac{0+1+1}{27} = \frac{2}{27}$ 2. $P(X = 1, Y = 1) : f(1,1) = \frac{1+1+1}{27} = \frac{3}{27}$ Now, we sum these probabilities: $P(X \leq 1, Y = 1) = P(X = 0, Y = 1) + P(X = 1, Y = 1) = \frac{2}{27} + \frac{3}{27} = \frac{5}{27}$. Therefore, the probability $P(X \leq 1, Y = 1) = \frac{5}{27}$.

## Question 9.2

(3 points)

What is the marginal distribution corresponding to $Y$? I.e. $f(y)$?

**Your answer here:** Given the joint distribution $f(x,y) = \frac{x+y+1}{27}$ for $x, y \in \{0, 1, 2\}$, the marginal distribution $f_Y(y)$ is: $f_Y(y) = \sum_{x=0}^{2} f(x,y)$. Calculate $f_Y(y)$ for each value of $y$:

1. For $y = 0 : f_Y(0) = f(0,0) + f(1,0) + f(2,0) = \frac{0+0+1}{27} + \frac{1+0+1}{27} + \frac{2+0+1}{27} = \frac{1}{27} + \frac{2}{27} + \frac{3}{27} = \frac{6}{27} = \frac{2}{9}$
2. For $y = 1 : f_Y(1) = f(0,1) + f(1,1) + f(2,1) = \frac{0+1+1}{27} + \frac{1+1+1}{27} + \frac{2+1+1}{27} = \frac{2}{27} + \frac{3}{27} + \frac{4}{27} = \frac{9}{27} = \frac{1}{3}$
3. For $y = 2 : f_Y(2) = f(0,2) + f(1,2) + f(2,2) = \frac{0+2+1}{27} + \frac{1+2+1}{27} + \frac{2+2+1}{27} = \frac{3}{27} + \frac{4}{27} + \frac{5}{27} = \frac{12}{27} = \frac{4}{9}$

Therefore, the marginal distribution of $Y$ is:

$$f_Y(y) = \begin{cases} \frac{2}{9} & \text{if } y = 0 \\ \frac{1}{3} & \text{if } y = 1 \\ \frac{4}{9} & \text{if } y = 2 \end{cases}$$

## Question 9.3

(3 points)

What is $P(Y = 1)$?

**Your answer here:** From the previous question, we determined the marginal distribution of $Y$. Therefore, $P(Y = 1)$ is given directly by the marginal distribution of $P(Y = 1) = f_Y(1) = \frac{1}{3}$.

## Question 9.4

(3 points)

What is $E(Y)$?

**Your answer here:** From the previous question, we determined the marginal distribution of $Y$. We can calculate the expected value $E(Y)$ as follows: $E(Y) = 0 \cdot \frac{2}{9} + 1 \cdot \frac{1}{3} + 2 \cdot \frac{4}{9}$, then simplify the calculation $E(Y) = 0 + \frac{1}{3} + \frac{8}{9} = \frac{1}{3} + \frac{8}{9}$ and convert $\frac{1}{3}$ to a common denominator $\frac{1}{3} = \frac{3}{9}$. Now, sum the fractions $E(Y) = \frac{3}{9} + \frac{8}{9} = \frac{11}{9}$. Therefore, the expected value $E(Y) = \frac{11}{9}$.

## Question 9.5

(3 points)

Are the random variables $X$ and $Y$ independent? Answer by calculating the marginal distributions of $X$ and $Y$ for full credit.

**Your answer here:** From the previous question, we determined the marginal distribution of $Y$. To check independence, we need to see if $f(x, y) = f_X(x) \cdot f_Y(y)$ for all values of $x$ and $y$. For $(x = 0, y = 0) : f(0, 0) = \frac{0+0+1}{27} = \frac{1}{27}$ and $f_X(0) \cdot f_Y(0) = \frac{2}{9} \cdot \frac{2}{9} = \frac{4}{81}$. Clearly, $\frac{4}{81} \neq \frac{1}{27}$. Since $f_X(0) \cdot f_Y(0) \neq f(0, 0)$, we can conclude that $X$ and $Y$ are not independent.

# End.