Homework 5

Han Nguyen - TXN200004

2024

Problem 1

Light Bulb Defects - Geometric Distribution

Max owns a light bulb manufacturing company where 3 out of every 75 bulbs are defective.

```
# Define the probability of defect
p <- 3/75  # probability of finding a defective bulb
q <- 1 - p  # probability of finding a non-defective bulb

# Display probabilities
cat("Probability of defect (p):", p, "\n")
```

```
## Probability of defect (p): 0.04
```

```
cat("Probability of non-defect (q):", q, "\n")
```

```
## Probability of non-defect (q): 0.96
```

a)

Question: What is the probability that Max will find the first faulty light bulb on the 6th one that he tested?

Derivation:

This follows a geometric distribution. The probability of finding the first defective bulb on the k-th trial is:

$$P(X = k) = (1 - p)^{k-1} \cdot p$$

For $k = 6$:

$$P(X = 6) = (1 - \frac{3}{75})^{6-1} \cdot \frac{3}{75} = (\frac{72}{75})^5 \cdot \frac{3}{75} = (0.96)^5 \cdot 0.04$$

```
# Probability of first defect on 6th trial
k <- 6
prob_1a <- (q^(k-1)) * p

cat("P(X = 6) =", round(prob_1a, 3), "\n")
```

```
## P(X = 6) = 0.033
```

Answer: The probability is **0.033**

b)

Question: What is the probability of taking at least four trials to find the first defective light bulb?

Derivation:

We need to find $P(X \geq 4)$, which equals $1 - P(X < 4) = 1 - P(X \leq 3)$.

Alternatively, using the complement rule:

$$P(X \geq 4) = P(\text{first 3 trials are non-defective}) = (1-p)^3 = (0.96)^3$$

```r
# Using complement of CDF
prob_1b <- 1 - pgeom(2, prob = p)  # pgeom(k-1) gives P(X <= k)
cat("P(X >= 4):", round(prob_1b, 3), "\n")
```

```
## P(X >= 4): 0.885
```

Answer: The probability is **0.885**

c)

Question: What is the probability of taking at most 10 trials to find the first defective light bulb?

Derivation:

We need to find $P(X \leq 10)$, which is the cumulative probability:

$$P(X \leq 10) = \sum_{k=1}^{10} P(X = k) = 1 - P(X > 10) = 1 - (1-p)^{10}$$

```r
# Using complement
prob_1c<- 1 - q^10
cat("P(X <= 10):", round(prob_1c, 3), "\n")
```

```
## P(X <= 10): 0.335
```

Answer: The probability is **0.335**

Problem 2

Yahtzee Simulation - Binomial Distribution

In this simplified Yahtzee game, we roll 5 fair six-sided dice and count the number of ones. We repeat this process 10,000 times.

```r
# Load required library
library(ggplot2)

# Set seed for reproducibility
set.seed(20220707)

# Simulate rolling 5 dice 10,000 times
n_simulations <- 10000
n_dice <- 5

# For each simulation, roll 5 dice and count the number of ones
X <- replicate(n_simulations, {
```

```r
  dice_rolls <- sample(1:6, size = n_dice, replace = TRUE)
  sum(dice_rolls == 1)  # Count how many ones
})

# Display first few values
cat("First 20 values of X:", head(X, 20), "\n")
```

```
## First 20 values of X: 2 0 3 2 2 1 3 1 1 0 0 1 1 0 0 0 1 0 0 1
```

```r
cat("Summary statistics:\n")
```

```
## Summary statistics:
```

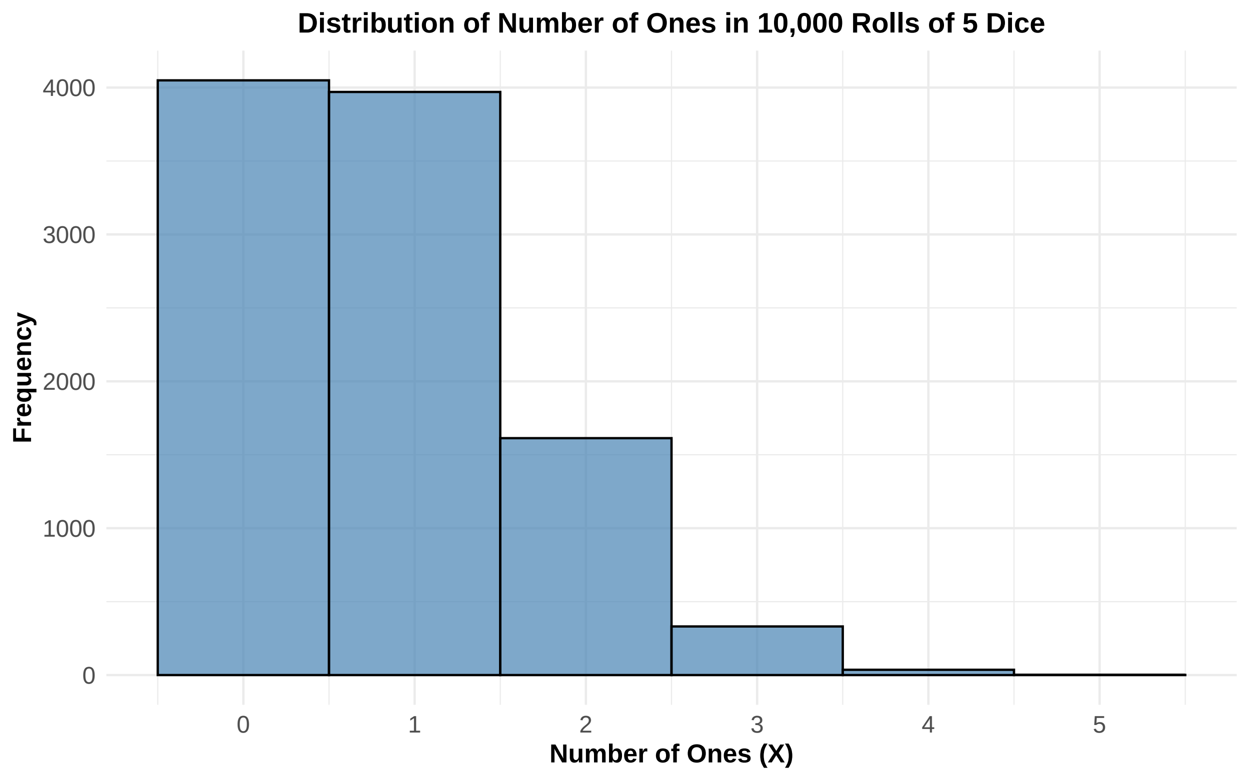
```r
summary(X)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  1.0000  0.8338  1.0000  5.0000
```

Histogram of X

```r
# Create a data frame for ggplot
data_X <- data.frame(X = X)

# Create histogram using ggplot
ggplot(data_X, aes(x = X)) +
  geom_histogram(binwidth = 1, fill = "steelblue", color = "black", alpha = 0.7) +
  scale_x_continuous(breaks = 0:5) +
  labs(
    title = "Distribution of Number of Ones in 10,000 Rolls of 5 Dice",
    x = "Number of Ones (X)",
    y = "Frequency"
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold"),
    axis.text = element_text(size = 11),
    axis.title = element_text(size = 12, face = "bold")
  )
```



Sample Mean and Sample Variance

Formulas:

The **sample mean** is calculated as:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \cdots + x_{10000}}{10000}$$

The **sample variance** is calculated as:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{\sum_{i=1}^{10000}(x_i - \bar{x})^2}{10000 - 1}$$

```r
# Calculate sample mean
sample_mean <- sum(X) / length(X)

# Calculate sample variance
sample_variance <- sum((X - sample_mean)^2) / (length(X) - 1)

# Display results
cat("Sample Mean (x-bar) =", round(sample_mean, 3), "\n\n")
```

```
## Sample Mean (x-bar) = 0.834
```

```r
cat("Sample Variance (s^2) =", round(sample_variance, 3), "\n\n")
```

```
## Sample Variance (s^2) = 0.705
```

Results:

- **Sample Mean:** $\bar{x} = 0.834$
- **Sample Variance:** $s^2 = 0.705$

Note: This follows a binomial distribution with $n = 5$ trials and $p = 1/6$ probability of success (rolling a one). The theoretical mean is $np = 5 \times \frac{1}{6} \approx 0.833$ and theoretical variance is $np(1-p) = 5 \times \frac{1}{6} \times \frac{5}{6} \approx 0.694$. We can see that the sample mean and sample variance are very close to the theoretical values.

Problem 3

Traffic Congestion - Poisson Distribution

On average, 180 cars per hour pass a specified point on a road during morning rush hour. Congestion occurs if more than 5 cars pass in any one minute.

```
# Calculate lambda for one minute
# Average cars per hour = 180
# Average cars per minute (lambda)
lambda <- 180 / 60
cat("Average cars per minute (lambda):", lambda, "\n")
```

```
## Average cars per minute (lambda): 3
```

Probability of Congestion

Question: What is the probability that congestion will occur in any minute (i.e., more than 5 cars pass)?

Solution:

The number of cars passing in one minute follows a Poisson distribution with $\lambda = \frac{180}{60} = 3$.

We need to find $P(X > 5)$ where $X \sim \text{Poisson}(\lambda = 3)$.

Using the complement rule:

$$P(X > 5) = 1 - P(X \leq 5) = 1 - \sum_{k=0}^{5} \frac{e^{-\lambda}\lambda^k}{k!}$$

```
# Probability of congestion (more than 5 cars)
# P(X > 5) = 1 - P(X <= 5)
prob_congestion <- 1 - ppois(5, lambda = lambda)

cat("P(X > 5) = P(congestion) =", round(prob_congestion, 3), "\n")
```

```
## P(X > 5) = P(congestion) = 0.084
```

Answer: The probability of congestion occurring in any one minute is **0.084**

Bar Chart of Poisson Probabilities

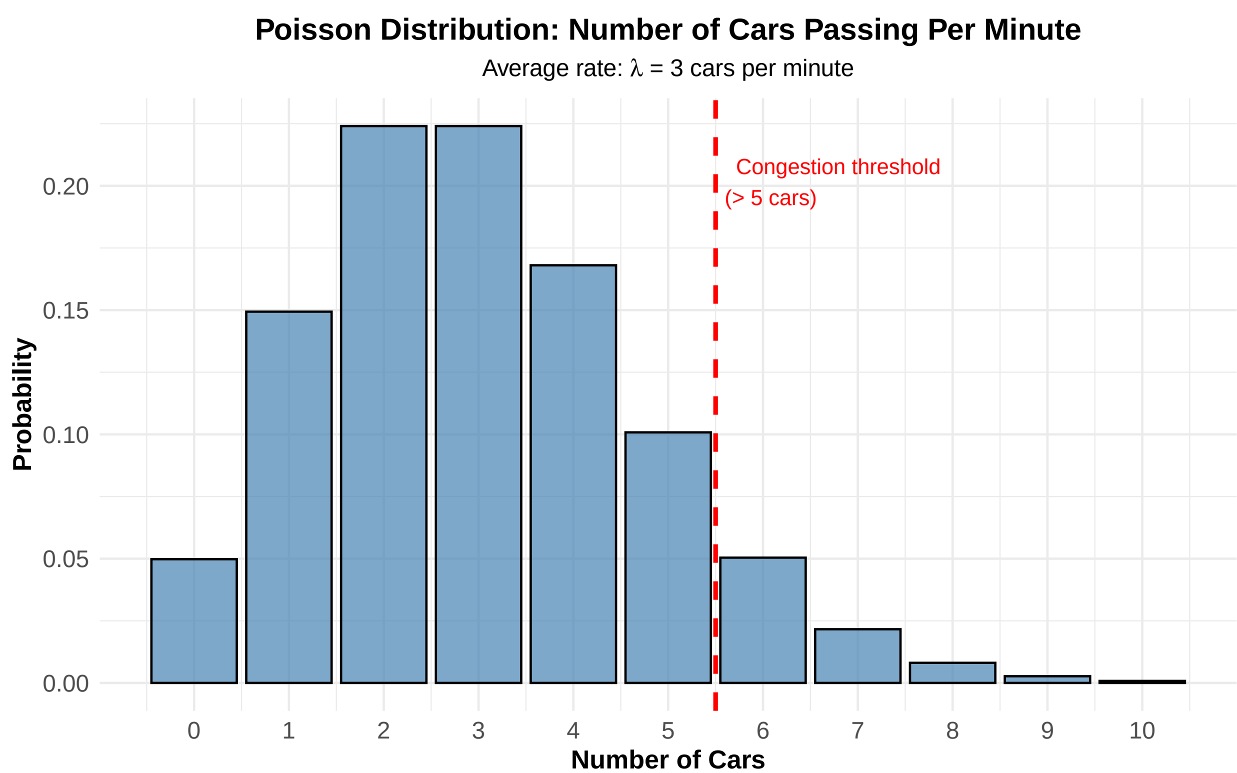
Question: Create a bar chart showing the probability distribution for 0 to 10 cars passing in one minute.

```
# Create data for the bar chart
cars <- 0:10
probabilities <- dpois(cars, lambda = lambda)

# Create data frame
poisson_data <- data.frame(
  Cars = cars,
```

```
  Probability = probabilities
)

# Create bar chart using ggplot
ggplot(poisson_data, aes(x = Cars, y = Probability)) +
  geom_bar(stat = "identity", fill = "steelblue", color = "black", alpha = 0.7) +
  geom_vline(xintercept = 5.5, linetype = "dashed", color = "red", linewidth = 1) +
  annotate("text", x = 5.5, y = max(probabilities) * 0.9,
           label = "Congestion threshold\n(> 5 cars)",
           color = "red", hjust = -0.1, size = 3.5) +
  scale_x_continuous(breaks = 0:10) +
  labs(
    title = "Poisson Distribution: Number of Cars Passing Per Minute",
    subtitle = expression(paste("Average rate: ", lambda, " = 3 cars per minute")),
    x = "Number of Cars",
    y = "Probability"
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold", size = 14),
    plot.subtitle = element_text(hjust = 0.5, size = 11),
    axis.text = element_text(size = 11),
    axis.title = element_text(size = 12, face = "bold")
  )
```




```
# Display the probability table
cat("\nProbability Distribution Table:\n")
```

##

```
## Probability Distribution Table:
```

```r
print(poisson_data, digits = 3)
```

```
##    Cars Probability
## 1     0     0.04979
## 2     1     0.14936
## 3     2     0.22404
## 4     3     0.22404
## 5     4     0.16803
## 6     5     0.10082
## 7     6     0.05041
## 8     7     0.02160
## 9     8     0.00810
## 10    9     0.00270
## 11   10     0.00081
```

Interpretation:

- The red dashed line shows the congestion threshold (5 cars)
- Cars to the right of this line (6, 7, 8, 9, 10, ...) represent congestion scenarios
- The distribution is centered around $\lambda = 3$ cars per minute
- The probability of exactly 3 cars is highest at 0.224

Problem 4

University Entrance Test - Normal Distribution

Entry to a certain University is determined by a national test. The scores on this test are normally distributed with a mean of 500 and a standard deviation of 100.

```r
# Define the parameters of the normal distribution
mu <- 500      # mean
sigma <- 100   # standard deviation

cat("Distribution: X ~ N(mu =", mu, ", sigma =", sigma, ")\n")
```

```
## Distribution: X ~ N(mu = 500 , sigma = 100 )
```

a) Probability of Scoring 585 or Less

Question: What is the probability that someone will score 585 or less on this national test?

Solution:

We need to find $P(X \leq 585)$ where $X \sim N(500, 100)$.

This can be calculated using the cumulative distribution function (CDF):

$$P(X \leq 585) = \Phi\left(\frac{585 - 500}{100}\right) = \Phi(0.85)$$

where Φ is the standard normal CDF.

```r
# Score threshold
score <- 585

# Calculate probability using pnorm
```

```r
prob_585_or_less <- pnorm(score, mean = mu, sd = sigma)
cat("P(X <= 585) =", round(prob_585_or_less, 3), "\n")
```

```
## P(X <= 585) = 0.802
```

Answer: The probability of scoring 585 or less is **0.802**

Interpretation: This means approximately 80.2% of test-takers score 585 or below.

b) Quartiles of the Distribution

Question: Find the lower quartile (Q1), median (Q2), and upper quartile (Q3) of the normal distribution.

Solution:

For a normal distribution: - **Lower Quartile (Q1):** 25th percentile, $P(X \leq Q_1) = 0.25$ - **Median (Q2):** 50th percentile, $P(X \leq Q_2) = 0.50$ - **Upper Quartile (Q3):** 75th percentile, $P(X \leq Q_3) = 0.75$

We use the quantile function (inverse CDF):

$$Q_p = \mu + \sigma \cdot \Phi^{-1}(p)$$

```r
# Calculate quartiles using qnorm
Q1 <- qnorm(0.25, mean = mu, sd = sigma)  # Lower quartile (25th percentile)
Q2 <- qnorm(0.50, mean = mu, sd = sigma)  # Median (50th percentile)
Q3 <- qnorm(0.75, mean = mu, sd = sigma)  # Upper quartile (75th percentile)

# Display results
cat("Lower Quartile (Q1, 25th percentile):", round(Q1, 3), "\n")
```

```
## Lower Quartile (Q1, 25th percentile): 432.551
```

```r
cat("Median (Q2, 50th percentile):", round(Q2, 3), "\n")
```

```
## Median (Q2, 50th percentile): 500
```

```r
cat("Upper Quartile (Q3, 75th percentile):", round(Q3, 3), "\n")
```

```
## Upper Quartile (Q3, 75th percentile): 567.449
```

Answers:

- **Lower Quartile (Q1):** 432.551
- **Median (Q2):** 500
- **Upper Quartile (Q3):** 567.449

Interpretation:

- 25% of students score below 432.551
- 50% of students score below 500 (the median)
- 75% of students score below 567.449

Problem 5

```r
# This is where my code for this question goes
```

Problem 6

```
# This is where my code for this question goes
```