Assignment 2 (10%) - CIND 123

Jivko Uzoff

November 10, 2024

Data Analytics: Basic Methods

Instructions

This assignment can be submitted using either Python or R, whichever you prefer.

- **If using R**, you must submit an RMD file with its knitted file (PDF or HTML). To learn more about knitting and R markdown, visit R Markdown (http://rmarkdown.rstudio.com).
- **If using Python**, you must submit an IPYNB file and its exported PDF/HTML with clearly printed/shown answers.

Failing to submit both files ({RMD + knitted PDF/HTML} OR {IPYNB + PDF/HTML}) will be subject to a 30% mark deduction.

NOTE: IF YOU USE `R STUDIO` , YOU SHOULD NEVER HAVE `install.packages` IN YOUR CODE; OTHERWISE, THE `Knit` OPTION WILL RAISE AN ERROR. COMMENT OUT ALL PACKAGE INSTALLATIONS BUT KEEP `library()` CALLS.

NOTE: If you answer the questions in R, all your answers should be in R (ignore Python questions). If you answer the questions in Python, all your answers should be in Python (ignore R questions). You are not allowed to switch languages in this assignment.

Question 1 (45 points)

The Titanic Passenger Survival Data Set provides information on the fate of passengers on the fatal maiden voyage of the ocean liner “Titanic.” The dataset is available from the kaggle website (Titanic Dataset (https://www.kaggle.com/c/titanic/data)) in several formats. Download the Titanic Data Set `titanic.csv` which is given on assignment 2’s page. Then, read it using the appropriate commands in `R` or `Python` .

Column Name	Description	Values
survival	Survival	0 = No, 1 = Yes
pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
sex	Sex	
age	Age in years	
sibsp	# of siblings / spouses aboard the Titanic	
parch	# of parents / children aboard the Titanic	
ticket	Ticket number	

Column Name	Description	Values
fare	Passenger fare	
cabin	Cabin number	
embarked	Port of Embarkation	C = Cherbourg, Q = Queenstown, S = Southampton

```
file_path <- "C:\\Users\\jivko\\Documents\\Data Analytics, Big Data, and Predictive Analytics
\\CIND 123\\Assignment 2\\titanic.csv"
```

```
titanic_data <- read.csv(file_path, header = TRUE)
```

```
head(titanic_data)
```

```
##   pclass survived                                             name    sex   age
## 1      1        1                    Allen, Miss. Elisabeth Walton female 29.00
## 2      1        1                   Allison, Master. Hudson Trevor   male  0.92
## 3      1        0                     Allison, Miss. Helen Loraine female  2.00
## 4      1        0             Allison, Mr. Hudson Joshua Creighton   male 30.00
## 5      1        0 Allison, Mrs. Hudson J C (Bessie Waldo Daniels) female 25.00
## 6      1        1                               Anderson, Mr. Harry   male 48.00
##   sibsp parch ticket     fare   cabin embarked boat body
## 1     0     0  24160 211.3375      B5        S    2   NA
## 2     1     2 113781 151.5500 C22 C26        S   11   NA
## 3     1     2 113781 151.5500 C22 C26        S        NA
## 4     1     2 113781 151.5500 C22 C26        S       135
## 5     1     2 113781 151.5500 C22 C26        S        NA
## 6     0     0  19952  26.5500     E12        S    3   NA
##                         home.dest
## 1                     St Louis, MO
## 2 Montreal, PQ / Chesterville, ON
## 3 Montreal, PQ / Chesterville, ON
## 4 Montreal, PQ / Chesterville, ON
## 5 Montreal, PQ / Chesterville, ON
## 6                     New York, NY
```

Q1a (5 points)

Extract and show the columns `name`, `fare`, `sibsp`, and `parch` into a new data frame named `titanicSubset`.

Show the `head` of the dataframe.

```
titanicSubset <- data.frame(
  name = titanic_data$name,
  fare = titanic_data$fare,
  sibsp = titanic_data$sibsp,
  parch = titanic_data$parch
)
```

```
head(titanicSubset)
```

```
##                                             name     fare sibsp parch
## 1                    Allen, Miss. Elisabeth Walton 211.3375     0     0
## 2                   Allison, Master. Hudson Trevor 151.5500     1     2
## 3                     Allison, Miss. Helen Loraine 151.5500     1     2
## 4             Allison, Mr. Hudson Joshua Creighton 151.5500     1     2
## 5 Allison, Mrs. Hudson J C (Bessie Waldo Daniels) 151.5500     1     2
## 6                               Anderson, Mr. Harry  26.5500     0     0
```

Q1b (5 points)

Numerical data: Calculate the total number of passengers who were children (age less than 18) and survived.

Print the value.

```
titanic_data_1b <- data.frame(
  age = titanic_data$age,
  survived = titanic_data$survived)
```

```
head(titanic_data_1b)
```

```
##     age survived
## 1 29.00        1
## 2  0.92        1
## 3  2.00        0
## 4 30.00        0
## 5 25.00        0
## 6 48.00        1
```

```
children_survived <- subset(titanic_data_1b, age < 18, survived == 1)
```

```
print(children_survived)
```

```
## data frame with 0 columns and 154 rows
```

```
total_children_survived <- nrow(children_survived)
```

```
print(total_children_survived)
```

```
## [1] 154
```

Q1c (5 points)

Categorical data: Calculate the number of passengers by sex.

Print the value.

```
titanic_data_1c_ <- data.frame(
  sex = titanic_data$sex)
```

```
titanic_data_1c_female <- subset(titanic_data_1c_, sex == "female")
```

```
total_titanic_data_1c_female <- nrow(titanic_data_1c_female)
```

```
print(total_titanic_data_1c_female)
```

```
## [1] 466
```

```
titanic_data_1c_male <- subset(titanic_data_1c_, sex == "male")
```

```
total_titanic_data_1c_male <- nrow(titanic_data_1c_male)
```

```
print(total_titanic_data_1c_male)
```

```
## [1] 843
```

Q1d (5 points)

Find the passengers in the data frame whose age information is missing, and fill them with the median age of passengers.

Show the head of the dataframe.

```
missing_values <- is.na(titanic_data$age)
```

```
median_age <- median(titanic_data$age, na.rm = TRUE)
```

```
print(median_age)
```

```
## [1] 28
```

```
titanic_data$age[missing_values] <- median_age
```

```
head(titanic_data)
```

```
##   pclass survived                                             name    sex   age
## 1      1        1                    Allen, Miss. Elisabeth Walton female 29.00
## 2      1        1                   Allison, Master. Hudson Trevor   male  0.92
## 3      1        0                     Allison, Miss. Helen Loraine female  2.00
## 4      1        0             Allison, Mr. Hudson Joshua Creighton   male 30.00
## 5      1        0 Allison, Mrs. Hudson J C (Bessie Waldo Daniels) female 25.00
## 6      1        1                               Anderson, Mr. Harry   male 48.00
##   sibsp parch ticket     fare   cabin embarked boat body
## 1     0     0  24160 211.3375      B5        S    2   NA
## 2     1     2 113781 151.5500 C22 C26        S   11   NA
## 3     1     2 113781 151.5500 C22 C26        S        NA
## 4     1     2 113781 151.5500 C22 C26        S       135
## 5     1     2 113781 151.5500 C22 C26        S        NA
## 6     0     0  19952  26.5500     E12        S    3   NA
##                         home.dest
## 1                     St Louis, MO
## 2 Montreal, PQ / Chesterville, ON
## 3 Montreal, PQ / Chesterville, ON
## 4 Montreal, PQ / Chesterville, ON
## 5 Montreal, PQ / Chesterville, ON
## 6                     New York, NY
```

Q1e (5 points)

Use the `aggregate()` function to calculate the survival count of each passenger class (`pclass`) and calculate the survival rate of passengers in each class. Draw a conclusion on which passenger class has the highest survival rate.

Print the value and type your response as a comment.

```
print(survival_count_by_class <- aggregate(survived ~ pclass, data = titanic_data, sum))
```

```
##   pclass survived
## 1      1      200
## 2      2      119
## 3      3      181
```

```
print(sum_of_allclass_survived <- sum(survival_count_by_class$survived))
```

```
## [1] 500
```

```
print(pclass_1_survivor_rate <- survival_count_by_class[survival_count_by_class$pclass == 1, "survived"] / sum_of_allclass_survived)
```

```
## [1] 0.4
```

```
print(pclass_2_survivor_rate <- survival_count_by_class[survival_count_by_class$pclass == 2, "survived"] / sum_of_allclass_survived)
```

```
## [1] 0.238
```

```
print(pclass_3_survivor_rate <- survival_count_by_class[survival_count_by_class$pclass == 3, "survived"] / sum_of_allclass_survived)
```

```
## [1] 0.362
```

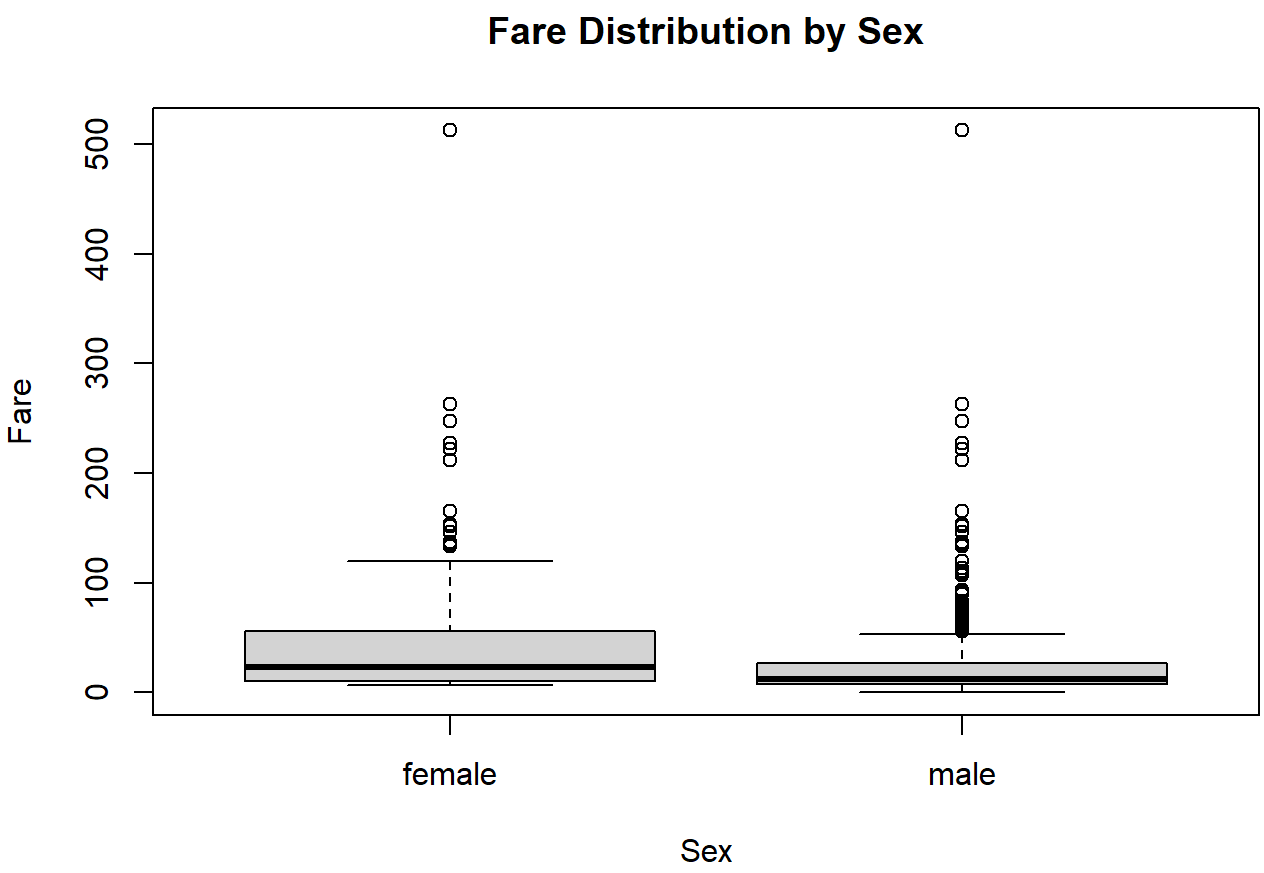
```
#Conclusion: Based on the above calculations, the 1st class passenger ticket has the highest survival rate at 0.4, or 40%.
```

Q1f (5 points)

Use a boxplot to display the distribution of fare for each sex. Infer which gender tends to pay higher fares.

Have the plot and then your comment.

```
boxplot(fare ~ sex, data=titanic_data, main = "Fare Distribution by Sex", xlab = "Sex", ylab = "Fare")
```




```
#Conclusion: The boxplot suggests that female passengers tend to pay higher fares on average compared to male passengers, with a slightly higher median fare and more outliers in the uppe r range.In addition, the upper quartile (Q3) for females is slightly higher, meaning that the top 25% of female passengers paid more than males.
```

Q1g (5 points)

Calculate the mean fare for each sex. Describe if the calculation aligns with the boxplot.

Print the value and comment on it.

```
print(mean_fare_by_sex <- aggregate(fare ~ sex, data = titanic_data, mean))
```

```
##      sex    fare
## 1 female 46.1981
## 2   male 26.1546
```

```
#Conclusion: The output aligns with the boxplot, indicating that female passengers have a hig her mean fare than male passengers. This is supported by the slightly elevated median fare an d the higher upper quartile (Q3) for females shown in the boxplot, suggesting that a larger p roportion of female passengers paid significantly more for their tickets.
```

Q1h (10 points)

Use a `for` loop and `if` control statements to list the names of women, aged 50 or older, who embarked from Southampton (S) on the Titanic. Ensure these women have non-empty home destinations.

Print first 5 people only.

```
women_over_50_from_S <- c()
```

```
for (i in 1:nrow(titanic_data)) {
  if (titanic_data$sex[i] == "female" &&
      titanic_data$age[i] >= 50 &&
      titanic_data$embarked[i] == "S" &&
      titanic_data$home.dest[i] != "") {

    women_over_50_from_S <- c(women_over_50_from_S, titanic_data$name[i])
  }
}
```

```
print(head(women_over_50_from_S, 5))
```

```
## [1] "Andrews, Miss. Kornelia Theodosia"
## [2] "Appleton, Mrs. Edward Dale (Charlotte Lamson)"
## [3] "Bonnell, Miss. Elizabeth"
## [4] "Brown, Mrs. John Murray (Caroline Lane Lamson)"
## [5] "Cavendish, Mrs. Tyrell William (Julia Florence Siegel)"
```

Question 2 (15 points)

100 computers work together in a network. Based on historical data, each computer has a probability of 0.03 of encountering a software issue. If a computer encounters an issue, it affects the network's performance.

Q2a (5 points)

Determine the probability that the network operates without any computer encountering a software issue.

```
total_computers <- 100

probability_issue <- 0.03

probability_no_issue <- 1 - probability_issue

probability_all_no_issues <- probability_no_issue ^ total_computers

print(round(probability_all_no_issues, 3))
```

```
## [1] 0.048
```

Q2b (5 points)

Utilize the Binomial approximation to estimate the probability that at least 5 computers out of 100 encounter software issues. Print the value.

```
prob_at_least_5 <- 1 - pbinom(4, size = 100, prob = 0.03)
print(prob_at_least_5)
```

```
## [1] 0.1821452
```

Q2c (5 points)

Assume the first and second computers are independent. Calculate the conditional probability that the second computer (Computer B) encounters a software issue given that the first computer (Computer A) does not encounter any issue.

```
first_computer <- 0.97
```

```
second_computer <- 0.03
```

```
conditional_probability <- (second_computer)
print(conditional_probability)
```

```
## [1] 0.03
```

Question 3 (25 points)

On average, John receives 3 emails a day.

Q3a (5 points)

Calculate the probabilities that John receives 2, 3, …, up to 9 emails in a day.

```
John <- dpois(2:9, 3)
print(John)
```

```
## [1] 0.224041808 0.224041808 0.168031356 0.100818813 0.050409407 0.021604031
## [7] 0.008101512 0.002700504
```

Q3b (5 points)

Determine the probability that John receives 4 emails or more in a day.

```
Johnb <- ppois(3, 3)
print(John4ormore <- (1-(Johnb)))
```

```
## [1] 0.3527681
```

Q3c1 (5 points)

Generate 50,000 samples for a Binomial random variable using parameters described in Question 2.

No need to print anything. Just the code.

```
binomsamples <- rbinom(50000, size = 100, prob = 0.03)
```

Q3c2 (5 points)

Generate 50,000 samples for a Poisson random variable using parameters described in Question 3.

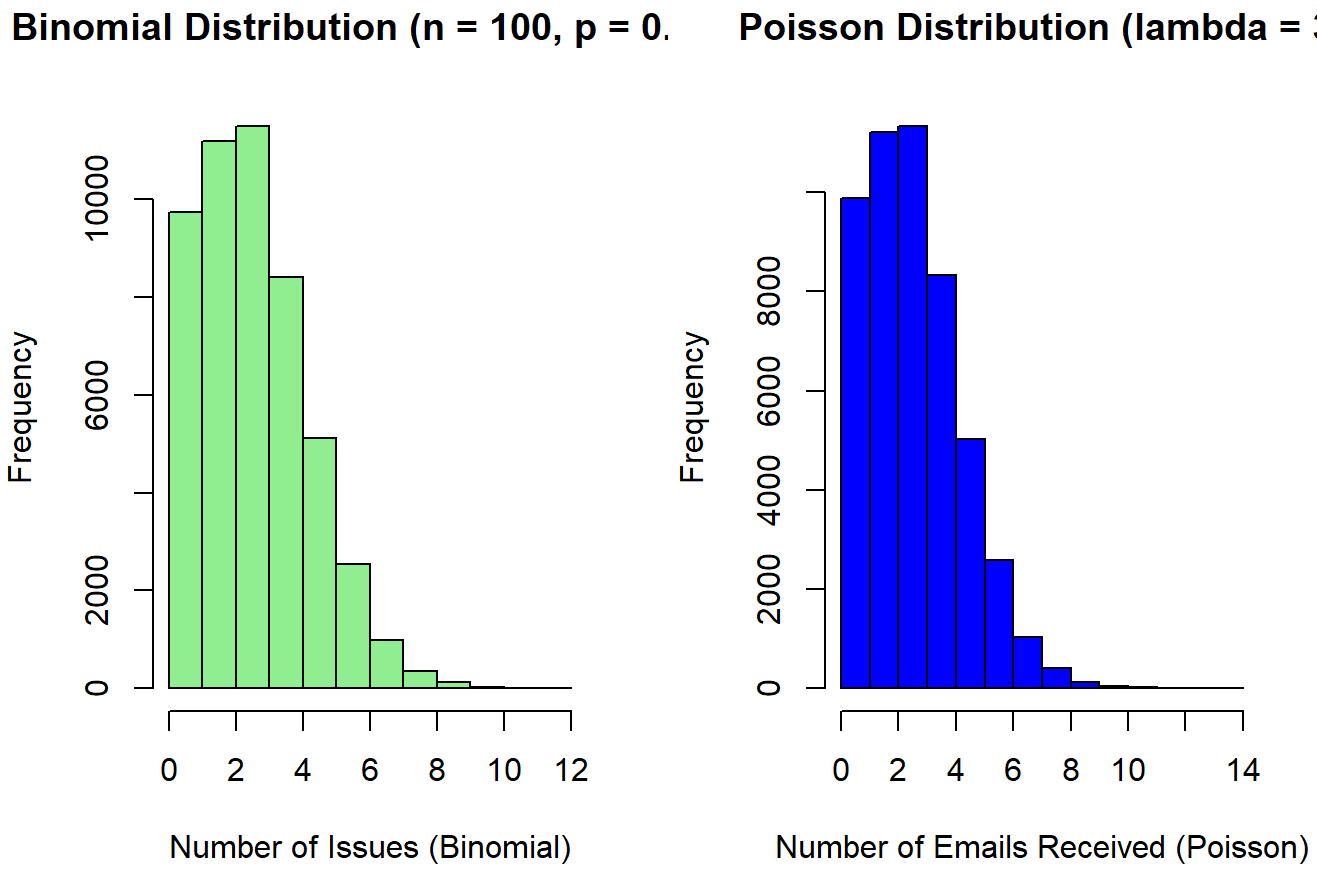
```
poissonsamples <- rpois(50000, 3)
```

Q3c3 (5 points)

Illustrate how well the Poisson probability distribution approximates the Binomial probability distribution in the previous questions.

Hint: Use histograms or other visualization tools to show them side by side.

```
par(mfrow=c(1,2))

hist(binomsamples, col = "lightgreen", xlab = "Number of Issues (Binomial)", ylab = "Frequenc
y", main = "Binomial Distribution (n = 100, p = 0.03)")

hist(poissonsamples, col = "blue", xlab = "Number of Emails Received (Poisson)", ylab = "Freq
uency", main = "Poisson Distribution (lambda = 3)")
```



Question 4 (15 points)

Q4a (5 points)

Generate 2000 random values from two different normal distributions:

- The first distribution has a mean of 8 and a standard deviation of 1.5.
- The second distribution has a mean of 5 and a standard deviation of 2.

Then, combine the two datasets into one and calculate the sample mean and standard deviation for the combined dataset.

```
firstnormaldist <- rnorm(2000,mean=8,sd=1.5)
secondnormaldist <- rnorm(2000, mean=5, sd=2)
```

```
combinednormaldist <- c(firstnormaldist, secondnormaldist)
```

```
print(mean(combinednormaldist))
```

```
## [1] 6.474987
```

```
print(sd(combinednormaldist))
```

```
## [1] 2.356155
```

Q4b (10 points)

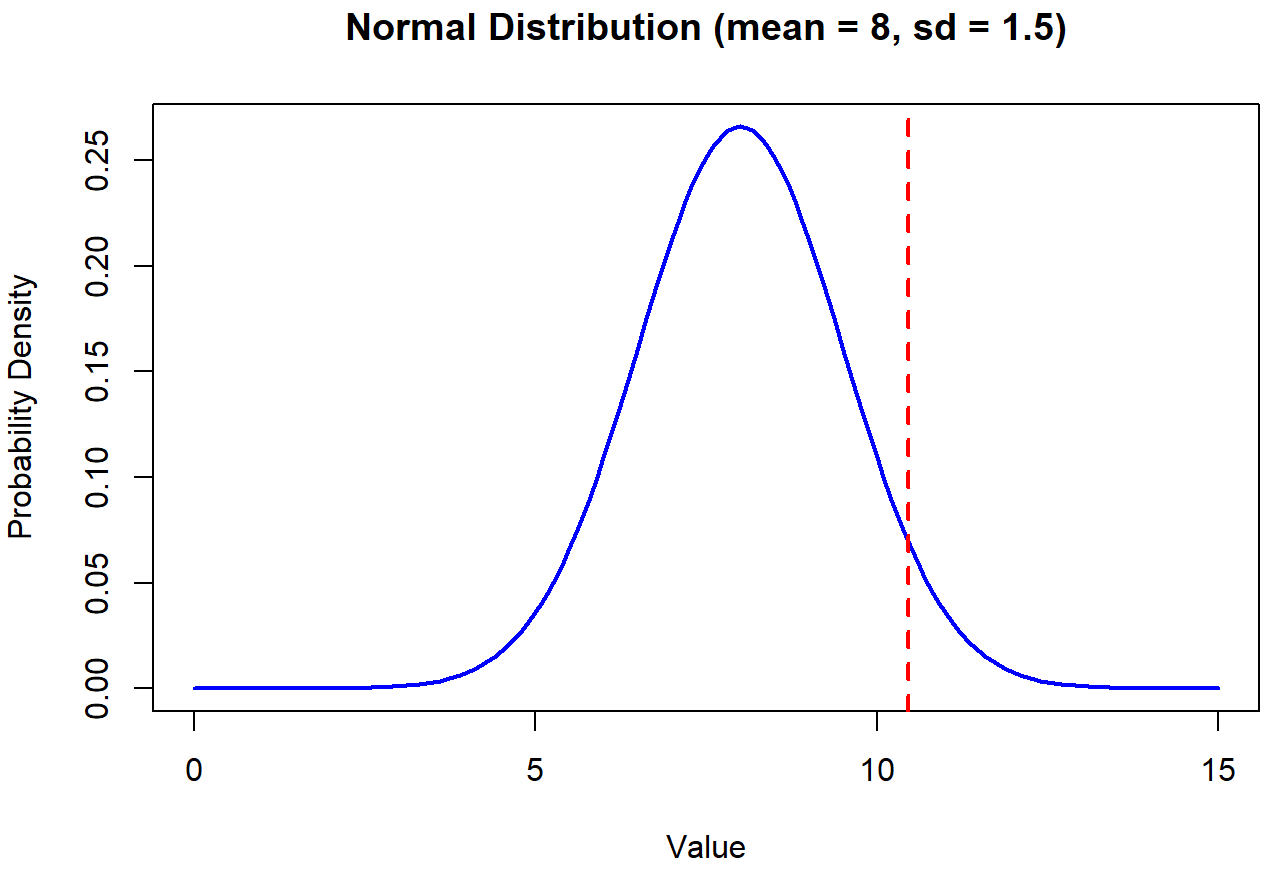
For a normal distribution with a mean of 8 and a standard deviation of 1.5, compute and plot the probability density function (PDF) for values ranging from 0 to 15.

- Identify the value at which the CDF equals 0.95.
- Explain the significance of this value in the context of this distribution.
- Draw a vertical line at the value where the CDF is 0.95.

```
mean_value <- 8
sd_value <- 1.5

x <- seq(0, 15, by = 0.1)

pdf_values <- dnorm(x, mean = mean_value, sd = sd_value)

plot(x, pdf_values, type = "l", col = "blue", lwd = 2,
     xlab = "Value", ylab = "Probability Density",
     main = "Normal Distribution (mean = 8, sd = 1.5)")

print(quantile_95 <- qnorm(0.95, mean = mean_value, sd = sd_value))
```

```
## [1] 10.46728
```

```
abline(v = quantile_95, col = "red", lwd = 2, lty = 2)
```




```
#Explanation of significance:
#The value at which the CDF equals 0.95 represents the 95th percentile of the normal distribu
tion. This means that 95% of the values in this distribution are less than or equal to this v
alue. In the context of this distribution (mean = 8, SD = 1.5), this value helps identify the
upper limit of the lower 95% of the data.
```

```
print(quantile_95 <- qnorm(0.95, mean = mean_value, sd = sd_value))
```

```
## [1] 10.46728
```