NAME: Shreeya Chatterji
CLASS: CO16
ROLL NO: 102103447

Probability and Statistics (UCS410)

Experiment 2: Descriptive statistics, Sample space, definition of Probability

(1) (a) Suppose there is a chest of coins with 20 gold, 30 silver and 50 bronze coins. You randomly draw 10 coins from this chest. Write an R code which will give us the sample space for this experiment. (use of sample(): an in-built function in R)

CODE:

```
#(1)
coins <- c(rep("gold", 20), rep("silver", 30), rep("bronze", 50))

sampleSpace <- sample(coins, size = 10, replace = FALSE)

print(sampleSpace)

```

OUTPUT:

```
> #(1)
> coins <- c(rep("gold", 20), rep("silver", 30), rep("bronze", 50))
> sampleSpace <- sample(coins, size = 10, replace = FALSE)
> print(sampleSpace)
 [1] "bronze" "bronze" "gold"   "bronze"
 [5] "silver" "silver" "silver" "bronze"
 [9] "bronze" "silver"
>
```

(b) In a surgical procedure, the chances of success and failure are 90% and 10% respectively. Generate a sample space for the next 10 surgical procedures performed. (use of prob(): an in-built function in R)

CODE:

```
#(2)
outcomes <- c("Success", "Failure")
probab <- c(0.9, 0.1)

# Generate a sample space for 10 surgical procedures
sample_space <- sample(outcomes, size = 10, replace = TRUE, prob = probab)

# Display
cat("Sample space for next 10 Procedures:\n", sample_space)

```

OUTPUT:

NAME: Shreeya Chatterji
CLASS: CO16
ROLL NO: 102103447

```
> #(2)
> outcomes <- c("Success", "Failure")
> probab <- c(0.9, 0.1)
> # Generate a sample space for 10 surgical procedures
> sample_space <- sample(outcomes, size = 10, replace = TRUE, prob = probab)
> # Display
> cat("Sample space for next 10 Procedures:\n", sample_space)
Sample space for next 10 Procedures:
 Success Success Success Success Success Success Success Success Success Failure
```

(2) A room has n people, and each has an equal chance of being born on any of the 365 days of the year. (For simplicity, we'll ignore leap years). What is the probability that two people in the room have the same birthday?

CODE:

```
##(3)
n <- as.integer(readline("Number of people in the room: "))

# Calculate probability
prob_no_shared <- 1
for (i in 1:n) {
  prob_no_shared <- prob_no_shared * (365 - i + 1) / 365
}
prob_shared <- 1 - prob_no_shared

cat("Probability that at least two people share a birthday in a room with", n, "people:", prob_shared, "\n")
```

OUTPUT:

```
> n <- as.integer(readline("Number of people in the room: "))
Number of people in the room: 10
> # Calculate probability
> prob_no_shared <- 1
> for (i in 1:n) {
+   prob_no_shared <- prob_no_shared * (365 - i + 1) / 365
+ }
> prob_shared <- 1 - prob_no_shared
> cat("Probability that at least two people share a birthday in a room with", n, "people:", prob_shared, "\n")
Probability that at least two people share a birthday in a room with 10 people: 0.1169482
```

(a) Use an R simulation to estimate this for various n.

CODE:

```
##USING AN ARRAY OF VALUES FOR N
num_simulations <- 10000

for (n in c(5, 10, 15, 20, 25)) {
  shared_birthday_count <- 0

  for (sim in 1:num_simulations) {
    birthdays <- sample(1:365, size = n, replace = TRUE)
    if (length(birthdays) != length(unique(birthdays))) {
      shared_birthday_count <- shared_birthday_count + 1
    }
  }

  prob_shared <- shared_birthday_count / num_simulations
  cat("Estimated probability of shared birthday with", n, "people:", prob_shared, "\n")
}
```

OUTPUT:

NAME: Shreeya Chatterji
CLASS: CO16
ROLL NO: 102103447

```
+   cat("Estimated probability of shared birthday with", n, "peopl
+ }
Estimated probability of shared birthday with 5 people: 0.0257
Estimated probability of shared birthday with 10 people: 0.1115
Estimated probability of shared birthday with 15 people: 0.2542
Estimated probability of shared birthday with 20 people: 0.4149
Estimated probability of shared birthday with 25 people: 0.5709
```

(b) Find the smallest value of n for which the probability of a match is greater than .5.

CODE:

```r
##SMALLEST VALUE OF n FOR WHICH THE PROBABILITY IS GREATER THAN 0.5
num_simulations <- 10000
n <- 1
while (TRUE) {
  shared_birthday_count <- 0

  for (sim in 1:num_simulations) {
    birthdays <- sample(1:365, size = n, replace = TRUE)
    if (length(birthdays) != length(unique(birthdays))) {
      shared_birthday_count <- shared_birthday_count + 1
    }
  }

  prob_shared <- shared_birthday_count / num_simulations
  if (prob_shared > 0.5) {
    break
  }

  n <- n + 1
}

cat("Smallest value of n for which the probability is greater than 0.5:", n, "\n")
```

OUTPUT:

```
+     break
+   }
+
+   n <- n + 1
+ }
>
> cat("Smallest value of n for which the probability is greater than 0.5:", n, "\n")
Smallest value of n for which the probability is greater than 0.5: 23
> 
```

(3) Write an R function for computing conditional probability. Call this function to do the following problem: Suppose the probability of the weather being cloudy is 40%. Also suppose the probability of rain on a given day is 20% and that the probability of clouds on a rainy day is 85%. If it's cloudy outside on a given day, what is the probability that it will rain that day?

CODE:

NAME: Shreeya Chatterji
CLASS: CO16
ROLL NO: 102103447

```r
###(3)
conditional_probability <- function(prob_a, prob_b_given_a, prob_b) {
  prob_a_given_b <- (prob_b_given_a * prob_a) / prob_b
  return(prob_a_given_b)
}

# Given probabilities
prob_cloudy <- 0.4
prob_rain <- 0.2
prob_clouds_given_rain <- 0.85

# Compute the probability of rain given that it's cloudy
prob_rain_given_cloudy <- conditional_probability(prob_rain, prob_clouds_given_rain, prob_cloudy)

cat("Probability of rain given that it's cloudy:", prob_rain_given_cloudy, "\n")
```

OUTPUT:

```
> cat("Probability of rain given that it's cloudy:", p
Probability of rain given that it's cloudy: 0.425
>
```

(4) The iris dataset is a built-in dataset in R that contains measurements on 4 different attributes (in centimeters) for 150 flowers from 3 different species. Load this dataset and do the following:

```r
###(4)
#Loading the dataset
data(iris)
```

(a) Print first few rows of this dataset.

CODE:

```r
# (a) Print first few rows of the dataset
head(iris)
```

OUTPUT:

```
> head(iris)
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2  setosa
2          4.9         3.0          1.4         0.2  setosa
3          4.7         3.2          1.3         0.2  setosa
4          4.6         3.1          1.5         0.2  setosa
5          5.0         3.6          1.4         0.2  setosa
6          5.4         3.9          1.7         0.4  setosa
>
```

(b) Find the structure of this dataset.

CODE:

```r
# (b) Find the structure of the dataset
str(iris)
```

OUTPUT:

NAME: Shreeya Chatterji
CLASS: CO16
ROLL NO: 102103447

```
> str(iris)
'data.frame':	150 obs. of  5 variables:
 $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
 $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
 $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
 $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
 $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 1
 ...
```

(c) Find the range of the data regarding the sepal length of flowers.

CODE:

```r
# (c) Find the range of sepal length
range_sepal_length <- range(iris$Sepal.Length)
cat("Range of sepal length:", range_sepal_length, "\n")
```

OUTPUT:

```
> cat("Range of sepal length:", range_sepal_length, "\n")
Range of sepal length: 4.3 7.9
```

(d) Find the mean of the sepal length.

CODE:

```r
# (d) Find the mean of sepal length
mean_sepal_length <- mean(iris$Sepal.Length)
cat("Mean of sepal length:", mean_sepal_length, "\n")
```

OUTPUT:

```
> cat("Mean of sepal length:", mean_sepal_length, "\n")
Mean of sepal length: 5.843333
```

(e) Find the median of the sepal length.

CODE:

```r
# (e) Find the median of sepal length
median_sepal_length <- median(iris$Sepal.Length)
cat("Median of sepal length:", median_sepal_length, "\n")
```

OUTPUT:

```
> cat("Median of sepal length:", median_sepal_length, "\n")
Median of sepal length: 5.8
```

(f) Find the first and the third quartiles and hence the interquartile range.

CODE:

NAME: Shreeya Chatterji
CLASS: CO16
ROLL NO: 102103447

```
# (f) Find the first and third quartiles and the interquartile range
quartiles_sepal_length <- quantile(iris$Sepal.Length, probs = c(0.25, 0.75))
iqr_sepal_length <- quartiles_sepal_length[2] - quartiles_sepal_length[1]
cat("First Quartile:", quartiles_sepal_length[1], "\n")
cat("Third Quartile:", quartiles_sepal_length[2], "\n")
cat("Interquartile Range:", iqr_sepal_length, "\n")
```

OUTPUT:

```
> iqr_sepal_length <- quartiles_sepal_length[2] - quartiles_se
> cat("First Quartile:", quartiles_sepal_length[1], "\n")
First Quartile: 5.1
> cat("Third Quartile:", quartiles_sepal_length[2], "\n")
Third Quartile: 6.4
> cat("Interquartile Range:", iqr_sepal_length, "\n")
Interquartile Range: 1.3
>
```

(g) Find the standard deviation and variance.

CODE:

```
# (g) Find the standard deviation and variance of sepal length
sd_sepal_length <- sd(iris$Sepal.Length)
var_sepal_length <- var(iris$Sepal.Length)
cat("Standard Deviation of sepal length:", sd_sepal_length, "\n")
cat("Variance of sepal length:", var_sepal_length, "\n")
```

OUTPUT:

```
> cat("Standard Deviation of sepal length:", sd_sepal_length, "\n")
Standard Deviation of sepal length: 0.8280661
> cat("Variance of sepal length:", var_sepal_length, "\n")
Variance of sepal length: 0.6856935
```

(h) Try doing the above exercises for sepal.width, petal.length and petal.width.

SEPAL.WIDTH

CODE:

NAME: Shreeya Chatterji
CLASS: CO16
ROLL NO: 102103447

```
# (h) Perform the above exercises for other attributes (sepal.width, petal.l
##SEPAL.WIDTH
#Range
range_sepal_width <- range(iris$Sepal.Width)
cat("Range of sepal width:", range_sepal_width, "\n")
#Mean
mean_sepal_width <- mean(iris$Sepal.Width)
cat("Mean of sepal width:", mean_sepal_width, "\n")
#Median
median_sepal_width <- median(iris$Sepal.Width)
cat("Median of sepal width:", median_sepal_width, "\n")
#Quartiles
quartiles_sepal_width <- quantile(iris$Sepal.Width, probs = c(0.25, 0.75))
iqr_sepal_width <- quartiles_sepal_width[2] - quartiles_sepal_width[1]
cat("First Quartile of sepal width:", quartiles_sepal_width[1], "\n")
cat("Third Quartile of sepal width:", quartiles_sepal_width[2], "\n")
cat("Interquartile Range of sepal width:", iqr_sepal_width, "\n")
##Standard Deviation and Variance
sd_sepal_width <- sd(iris$Sepal.Width)
var_sepal_width <- var(iris$Sepal.Width)
cat("Standard Deviation of sepal width:", sd_sepal_width, "\n")
cat("Variance of sepal width:", var_sepal_width, "\n")
```

OUTPUT:

```
> range_sepal_width <- range(iris$Sepal.Width)
> cat("Range of sepal width:", range_sepal_width, "\n")
Range of sepal width: 2 4.4
> #Mean
> mean_sepal_width <- mean(iris$Sepal.Width)
> cat("Mean of sepal width:", mean_sepal_width, "\n")
Mean of sepal width: 3.057333
> #Median
> median_sepal_width <- median(iris$Sepal.Width)
> cat("Median of sepal width:", median_sepal_width, "\n")
Median of sepal width: 3
> #Quartiles
> quartiles_sepal_width <- quantile(iris$Sepal.Width, probs = c(0.25, 0.75))
> iqr_sepal_width <- quartiles_sepal_width[2] - quartiles_sepal_width[1]
> cat("First Quartile of sepal width:", quartiles_sepal_width[1], "\n")
First Quartile of sepal width: 2.8
> cat("Third Quartile of sepal width:", quartiles_sepal_width[2], "\n")
Third Quartile of sepal width: 3.3
> cat("Interquartile Range of sepal width:", iqr_sepal_width, "\n")
Interquartile Range of sepal width: 0.5
> ##Standard Deviation and Variance
> sd_sepal_width <- sd(iris$Sepal.Width)
> var_sepal_width <- var(iris$Sepal.Width)
> cat("Standard Deviation of sepal width:", sd_sepal_width, "\n")
Standard Deviation of sepal width: 0.4358663
> cat("Variance of sepal width:", var_sepal_width, "\n")
Variance of sepal width: 0.1899794
```

PETAL.LENGTH

INPUT:

NAME: Shreeya Chatterji
CLASS: CO16
ROLL NO: 102103447

```r
##PETAL.LENGTH
#Range
range_petal_length <- range(iris$Petal.Length)
cat("Range of petal length:", range_petal_length, "\n")
#Mean
mean_petal_length <- mean(iris$Petal.Length)
cat("Mean of petal length:", mean_petal_length, "\n")
#Median
median_petal_length <- median(iris$Petal.Length)
cat("Median of petal length:", median_petal_length, "\n")
#Quartiles
quartiles_petal_length <- quantile(iris$Petal.Length, probs = c(0.25, 0.75))
iqr_petal_length <- quartiles_petal_length[2] - quartiles_petal_length[1]
cat("First Quartile of petal length:", quartiles_petal_length[1], "\n")
cat("Third Quartile of petal length:", quartiles_petal_length[2], "\n")
cat("Interquartile Range of petal length:", iqr_petal_length, "\n")
##Standard Deviation and Variance
sd_petal_length <- sd(iris$Petal.Length)
var_petal_length <- var(iris$Petal.Length)
cat("Standard Deviation of petal length:", sd_petal_length, "\n")
cat("Variance of petal length:", var_petal_length, "\n")
```

OUTPUT:

```
Range of petal length: 1 6.9
> #Mean
> mean_petal_length <- mean(iris$Petal.Length)
> cat("Mean of petal length:", mean_petal_length, "\n")
Mean of petal length: 3.758
> #Median
> median_petal_length <- median(iris$Petal.Length)
> cat("Median of petal length:", median_petal_length, "\n")
Median of petal length: 4.35
> #Quartiles
> quartiles_petal_length <- quantile(iris$Petal.Length, probs = c(0.25, 0.75))
> iqr_petal_length <- quartiles_petal_length[2] - quartiles_petal_length[1]
> cat("First Quartile of petal length:", quartiles_petal_length[1], "\n")
First Quartile of petal length: 1.6
> cat("Third Quartile of petal length:", quartiles_petal_length[2], "\n")
Third Quartile of petal length: 5.1
> cat("Interquartile Range of petal length:", iqr_petal_length, "\n")
Interquartile Range of petal length: 3.5
> ##Standard Deviation and Variance
> sd_petal_length <- sd(iris$Petal.Length)
> var_petal_length <- var(iris$Petal.Length)
> cat("Standard Deviation of petal length:", sd_petal_length, "\n")
Standard Deviation of petal length: 1.765298
> cat("Variance of petal length:", var_petal_length, "\n")
Variance of petal length: 3.116278
```

PETAL.WIDTH

INPUT:

NAME: Shreeya Chatterji
CLASS: CO16
ROLL NO: 102103447

```
##PETAL.WIDTH
#Range
range_petal_width <- range(iris$Petal.Width)
cat("Range of petal width:", range_petal_width, "\n")
#Mean
mean_petal_width <- mean(iris$Petal.Width)
cat("Mean of petal width:", mean_petal_width, "\n")
#Median
median_petal_width <- median(iris$Petal.Width)
cat("Median of petal width:", median_petal_width, "\n")
#Quartiles
quartiles_petal_width <- quantile(iris$Petal.Width, probs = c(0.25, 0.75))
iqr_petal_width <- quartiles_petal_width[2] - quartiles_petal_width[1]
cat("First Quartile of petal width:", quartiles_petal_width[1], "\n")
cat("Third Quartile of petal width:", quartiles_petal_width[2], "\n")
cat("Interquartile Range of petal width:", iqr_petal_width, "\n")
##Standard Deviation and Variance
sd_petal_width <- sd(iris$Petal.Width)
var_petal_width <- var(iris$Petal.Width)
cat("Standard Deviation of sepal width:", sd_petal_width, "\n")
cat("Variance of sepal width:", var_petal_width, "\n")
```

OUTPUT:

```
> cat("Range of petal width:", range_petal_width, "\n")
Range of petal width: 0.1 2.5
> #Mean
> mean_petal_width <- mean(iris$Petal.Width)
> cat("Mean of petal width:", mean_petal_width, "\n")
Mean of petal width: 1.199333
> #Median
> median_petal_width <- median(iris$Petal.Width)
> cat("Median of petal width:", median_petal_width, "\n")
Median of petal width: 1.3
> #Quartiles
> quartiles_petal_width <- quantile(iris$Petal.Width, probs = c(0.25, 0
> iqr_petal_width <- quartiles_petal_width[2] - quartiles_petal_width[1
> cat("First Quartile of petal width:", quartiles_petal_width[1], "\n")
First Quartile of petal width: 0.3
> cat("Third Quartile of petal width:", quartiles_petal_width[2], "\n")
Third Quartile of petal width: 1.8
> cat("Interquartile Range of petal width:", iqr_petal_width, "\n")
Interquartile Range of petal width: 1.5
> ##Standard Deviation and Variance
> sd_petal_width <- sd(iris$Petal.Width)
> var_petal_width <- var(iris$Petal.Width)
> cat("Standard Deviation of sepal width:", sd_petal_width, "\n")
Standard Deviation of sepal width: 0.7622377
> cat("Variance of sepal width:", var_petal_width, "\n")
Variance of sepal width: 0.5810063
> 
```

(i) Use the built-in function summary on the dataset Iris.

CODE:

NAME: Shreeya Chatterji
CLASS: CO16
ROLL NO: 102103447

```r
# (i) Use the built-in function summary on the dataset Iris
summary(iris)
```

OUTPUT:

```
> summary(iris)
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width   
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100  
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300  
 Median :5.800   Median :3.000   Median :4.350   Median :1.300  
 Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199  
 3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800  
 Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500  
       Species  
 setosa    :50  
 versicolor:50  
 virginica :50  
```

(5) R does not have a standard in-built function to calculate mode. So, we create a user function to calculate mode of a data set in R. This function n takes the vector as input and gives the mode value as output.

CODE:

```r
#This function n takes the vector as input and gives the mode value as o
calculate_mode <- function(data) {
  table_data <- table(data)
  mode <- as.numeric(names(table_data[table_data == max(table_data)]))
  return(mode)
}

# Test the function
dataset <- c(1,0,2,1,0,3,4,4,7)
mode_value <- calculate_mode(dataset)
cat("Mode of the dataset:", mode_value, "\n")
```

OUTPUT:

```
> cat("Mode of the dataset:", mode_value, '
Mode of the dataset: 0 1 4
```