Midterm2

Elham

2023-03-22

1. You are given a dataset containing the number of hours spent by a group of students studying different subjects. The dataset is as follows:

Subject,Hours Math,20 English,15 Science,10 History,5 Geography,5

(a)Create a data frame in R named “data” using the provided (b) Calculate the percentage of total study hours for each subject, rounded to one decimal place. (c) Create a pie chart titled “Study Hours by Subject” and assign colors to each slice: red, orange,yellow, green, and blue. Set t he x axis and y axis scales to be between 1 and 1. The label of each slice should be the name of subjects along with the percentage of total study for each subject.

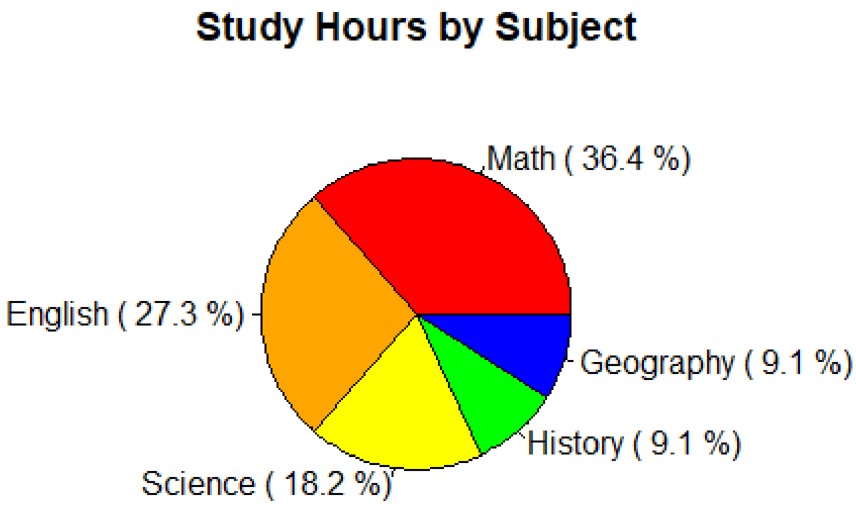
```
# Generate the dataset
data <- data.frame(Subject=c("Math", "English", "Science", "History",
"Geography"),
                   Hours= c(20, 15, 10, 5, 5))
data

##     Subject Hours
## 1      Math    20
## 2   English    15
## 3   Science    10
## 4   History     5
## 5 Geography     5

# Calculate the percentage of total study hours for each subject and round
with one decimal
data$Percentage <- round((data$Hours / sum(data$Hours)*100), 1)

# Create a pie chart with the title of Study Hours by Subject and assign
colors to each slice: red, orange,yellow, green, and blue and  limits on x-
axis and y-axis between -1 and 1.
# Label each slice with the name of subjects and the percentage of total
study for each subject.

pie(data$Hours, labels = paste(data$Subject, "(", data$Percentage, "%)"),
    col = c("red", "orange", "yellow", "green", "blue"),
    xlim = c(-1, 1), ylim = c(-1, 1),
    main = "Study Hours by Subject")
```



2.Use the built-in Titanic dataset. Please refer to the Titanic dataset description. (a) If we want to know the mean of children passengers on the Titanic, what is the parameter of interest? (b) Identify the variables in the dataset and describe their types. (c) Create a variable in R called “totalChildren” that contains the total number of children on the Titanic (across all genders, classes, and survival outcomes). (d) Create a variable in R called “totalSurvival” that contains the total number of survivors in our sample (children who survived the Titanic). (e) What is the observed value of the statistic that we should use to estimate the population mean of interest (survived children on the Titanic)? (f) What is the estimated standard error of the interest? (g) What is critical value for a 90% confidence interval for the population mean? (h) What is the margin of error for our estimate? (i) Determine a 90% confidence interval for the true value of the population mean.

```
#(a)If we want to know the mean of children passengers on the Titanic, what
is the parameter of
#interest?

head(Titanic)

## , , Age = Child, Survived = No
##
##       Sex
## Class  Male Female
##   1st     0      0
##   2nd     0      0
##   3rd    35     17
##   Crew    0      0
```

```
##
## , , Age = Adult, Survived = No
##
##       Sex
## Class  Male Female
##   1st   118      4
##   2nd   154     13
##   3rd   387     89
##   Crew  670      3
##
## , , Age = Child, Survived = Yes
##
##       Sex
## Class  Male Female
##   1st     5      1
##   2nd    11     13
##   3rd    13     14
##   Crew    0      0
##
## , , Age = Adult, Survived = Yes
##
##       Sex
## Class  Male Female
##   1st    57    140
##   2nd    14     80
##   3rd    75     76
##   Crew  192     20
```

```
?Titanic
```

```
## starting httpd help server ... done
```

```
#(b) Identify the variables in the dataset and describe their types.

# Categorical
#   Class   1st, 2nd, 3rd, Crew
#   Sex Male, Female
#   Age Child, Adult
#   Survived    No, Yes

#(c)Create a variable in R called "totalChildren" that contains the total
number of children on the
#Titanic (across all genders, classes, and survival outcomes).


totalchildren= Titanic[ , ,1 , ]
totalchildren
```

```
## , , Survived = No
##
```

```
##       Sex
## Class  Male Female
##   1st     0      0
##   2nd     0      0
##   3rd    35     17
##   Crew    0      0
##
## , , Survived = Yes
##
##       Sex
## Class  Male Female
##   1st     5      1
##   2nd    11     13
##   3rd    13     14
##   Crew    0      0
```

```
#(d)Create a variable in R called "totalSurvival" that contains the total
number of survivors in our sample (children who survived on the Titanic).

totalSurvival= Titanic[ , ,1,2]
totalSurvival
```

```
##       Sex
## Class  Male Female
##   1st     5      1
##   2nd    11     13
##   3rd    13     14
##   Crew    0      0
```

```
#(e)What is the observed value of the statistic that we should use to
estimate the population mean of interest (survived children on the Titanic)?

m=mean(totalSurvival)
m
```

```
## [1] 7.125
```

```
#(f)What is the estimated standard error of the interest?
s=sd(totalSurvival)
n=length(totalSurvival)

ese= s/sqrt(n)
ese
```

```
## [1] 2.215509
```

```
#(g)What is critical value for a 90% confidence interval for the population
mean?
cv= qnorm(0.95)
cv
```

```
## [1] 1.644854

#(h)What is the margin of error for our estimate?
moe= cv*ese
moe

## [1] 3.644189

#(i)Determine a 90% confidence interval for the true value of the population
mean.
upper_bound=m+moe
upper_bound

## [1] 10.76919

lower_bound=m-moe
lower_bound

## [1] 3.480811
```

3. You have been given a dataset named houses.csv that contains information about the prices of houses in a particular neighborhood. The dataset includes the following variables:

Price: The price of the house in dollars SquareFeet: The size of the house in square feet Bedrooms: The number of bedrooms in the house Bathrooms: The number of bathrooms in the house YearBuilt: The year the house was built

Your task is to build a linear regression model that predicts the price of a house based on its size and the number of bedrooms and bathrooms it has. However, before building the model, you need to clean the dataset by removing any missing values.

Using R and performs the following tasks: 1. Build a linear regression model that predicts the price of a house based on its size, number of bedrooms, and number of bathrooms. 2. Create a new variable called TotalRooms that is the sum of the Bedrooms and Bathrooms variables using the for() statement. 3. Build a linear regression model using the lm() function that predicts the Price of a house based on its SquareFeet and TotalRooms. 4. Identify which independent variable should be removed from the model and explain the reason.

```
# Load the dataset
houses <- read.csv(file.choose())
head(houses)

##    Price SquareFeet Bedrooms Bathrooms YearBuilt
## 1 120000        800        1         1      1950
## 2 180000       1200        2         1      1960
## 3 220000       1500        2         2      1970
## 4 280000       2000        3         2      1980
## 5 350000       2500        3         2      1990
## 6 410000       3000        4         3      2000
```

```
# Create a new variable called TotalRooms using the for() statement
houses$TotalRooms <- numeric(nrow(houses))
for (i in 1:nrow(houses)) {
  houses$TotalRooms[i] <- houses$Bedrooms[i] + houses$Bathrooms[i]
}

# Build a linear regression model using the lm() function that predicts the
Price of a house based on its SquareFeet and TotalRooms

model <- lm(houses$Price ~ houses$SquareFeet + houses$TotalRooms)
model

##
## Call:
## lm(formula = houses$Price ~ houses$SquareFeet + houses$TotalRooms)
##
## Coefficients:
##       (Intercept)  houses$SquareFeet  houses$TotalRooms
##          -13150.0              185.8           -15162.4

# Print out the summary of the model to identify which independent variable
should be removed from the model and explain the reason.

summary(model) # TotalRooms should be removed because its p-value is greater
than 0.05.

##
## Call:
## lm(formula = houses$Price ~ houses$SquareFeet + houses$TotalRooms)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -30869 -21223   5571  15088  35980
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -13149.98   29001.13  -0.453  0.66396
## houses$SquareFeet    185.76      42.78   4.342  0.00339 **
## houses$TotalRooms -15162.44   23484.43  -0.646  0.53909
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 25760 on 7 degrees of freedom
## Multiple R-squared:  0.9902, Adjusted R-squared:  0.9874
## F-statistic: 353.5 on 2 and 7 DF,  p-value: 9.333e-08
```

4. Consider the built-in “esoph” dataset in R. Please refer to the “esoph” dataset description. The dataset contains data from a case-control study of esophageal cancer. Use the column named “ncontrols” in the “esoph” dataset and answer the following questions:

(a) Bootstrap 10,000 samples to find the 95th percentile and save the bootstrapped 95th percentiles to a vector called “Bootstrap.”

(b) Plot the sampling distribution of the 95th percentile. Your plot should include a title and label for the x-axis.

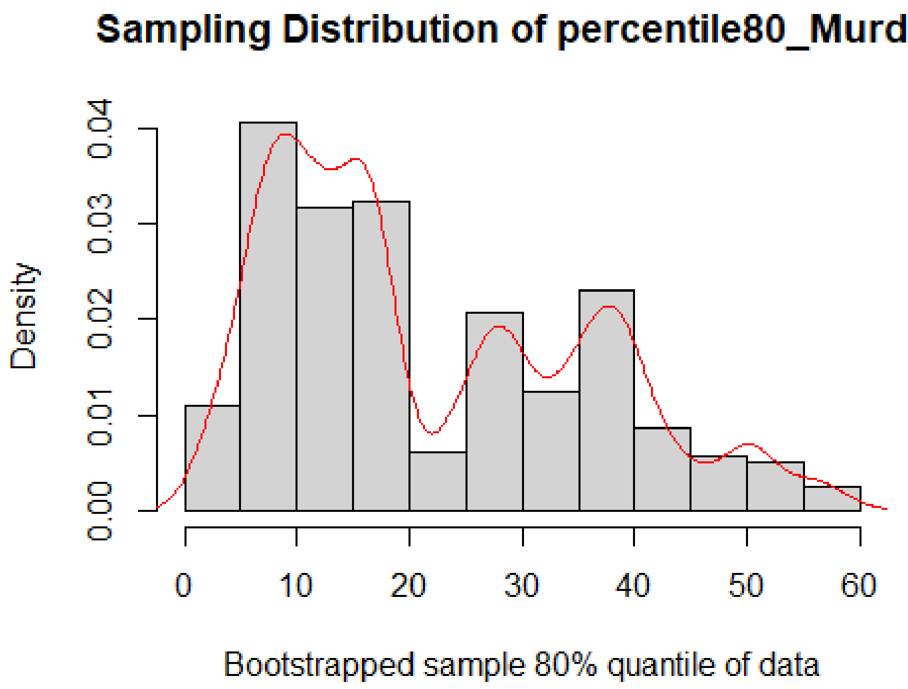
(c) Compute a 92% confidence interval for the 95th percentile.

```
?esoph

n = length(esoph)
n

## [1] 5

Bootstrap = numeric()
for (i in 1:10000) {
  Bootstrap[i] = quantile(sample(esoph$ncontrols, n, replace = TRUE), .95)
}
# plot the sampling distribution of the 95th percentile. Your plot should
include a title and label for the x-axis.
hist(Bootstrap, prob = TRUE, main = "Sampling Distribution of
percentile80_Murder",
     xlab = "Bootstrapped sample 80% quantile of data")
lines(density(Bootstrap), col = "red") # optional
```



```
#Compute a 92% confidence interval for the 95th percentile.
quantile(Bootstrap, c(0.04, 0.96))
```

```
##   4%  96%
##  4.4 50.0
```