Effects of Advertising on Sales: Television, Radio and Newspaper Advertisements

Authors:

Atena Jafari Parsa 2101183
Ava Arabi 2104906
Peyman Khodabandehlouei 2104987

Course Name: Statistics in Engineering

Course Code: INE2002

Theoretical Course Section: 2

Lab Course Section: 903

Table of Contents

- 1.0 Introduction

- 1.1 Define the Problem
- 1.2 How the Related Data Was Collected
- 1.3 Data Analysis

- 2.0 Data Collection Description

  - 2.1 Sample Selection Method

- 3 Diagrams, Charts (Sample Statistics)

- 4.0 Normality Tests

  - 4.1 Shapiro-Wilk Test
  - 4.2 Examples

- 5. Point Estimations and Confidence Intervals

- 6. 2-3 Hypothesis Tests

- 7. Goodness of Fits Tests and other Checks for Detecting the Distribution

- 8. Linear Regression Model Design
- 9. Analysis of Variance (ANOVA)
- 10. Applications of Nonparametric Tests
- 11. References and Image Citations

1.0 Introduction

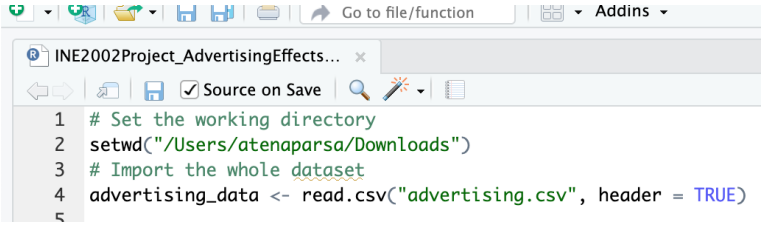
1.1 Define the Problem

Advertising builds companies' reputation among the consumers and obviously, good commercials and advertising would promote the sales and increase the general profitability. If an advertising is effective it would result in promoting the sales whereas harm could be caused to a company's reputation and sales in the case of a poor advertising. In this problem, the effects of three of the most popular means of advertising on sales (and possibly their interaction) would be discussed in detail.

1.2 How the Related Data Was Collected

The dataset chosen is Advertising Dataset of Sales with respect to TV ads, radio ads and Newspaper ads retrieved from Advertising Dataset | Kaggle . The sales are in thousands of units and the budget is in thousands of dollars.

After collecting the data, the dataset is imported to the project in Rstudio and saved as the variable 'advertising_data' through the code below:



1.3 Data Analysis

To check the overall structure of the data, we use the following code and get a several data points of each variable in the output:

```
> str(advertising_data)
'data.frame':   200 obs. of  4 variables:
 $ TV       : num  230.1 44.5 17.2 151.5 180.8 ...
 $ Radio    : num  37.8 39.3 45.9 41.3 10.8 48.9 32.8 19.6 2.1 2.6 ...
 $ Newspaper: num  69.2 45.1 69.3 58.5 58.4 75 23.5 11.6 1 21.2 ...
 $ Sales    : num  22.1 10.4 12 16.5 17.9 7.2 11.8 13.2 4.8 15.6 ...
```

As shown in the output above, there 200 observations of 4 variables in total.
We can also check the output of the summary function to get the minimum and maximum values, 1st quartile and the 3rd quartile, the median and the mean o all the existing variables to have an overview of the data. The code and the output of the code are provided on the next page.

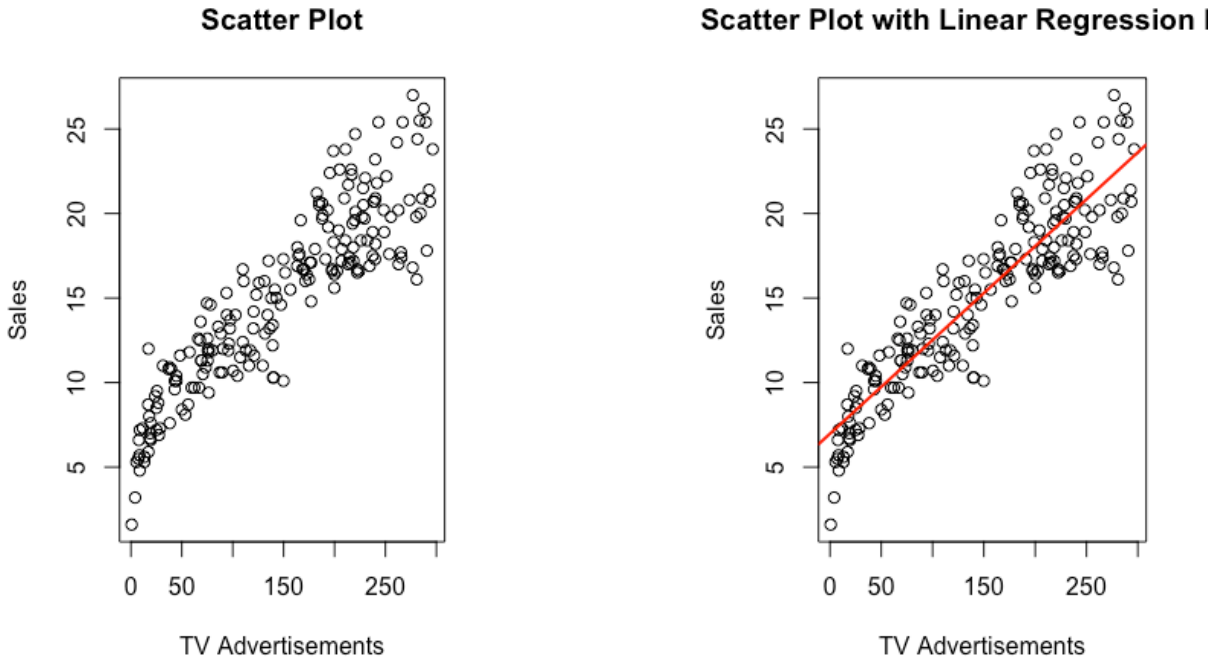
```
> summary(advertising_data)
       TV              Radio          Newspaper          Sales
 Min.   :  0.70   Min.   : 0.000   Min.   :  0.30   Min.   : 1.60
 1st Qu.: 74.38   1st Qu.: 9.975   1st Qu.: 12.75   1st Qu.:11.00
 Median :149.75   Median :22.900   Median : 25.75   Median :16.00
 Mean   :147.04   Mean   :23.264   Mean   : 30.55   Mean   :15.13
 3rd Qu.:218.82   3rd Qu.:36.525   3rd Qu.: 45.10   3rd Qu.:19.05
 Max.   :296.40   Max.   :49.600   Max.   :114.00   Max.   :27.00
```

To check the relationship of each variable on the dependent variable which is "Sales", the columns other than the one whose relation with "Sales" we are observing would be removed and saved to a new file. We would use the new file to create charts and diagrams and eyeball the relation. For instance,), to check the relationship between TV ads and sales and see if it is linear or it needs transformations we need to remove the Radio and Newspaper columns and save the new dataset that only has 2 columns instead of four. As shown in the code below, first we specify the columns that need to be removed in a variable called "columns_to_remove". Then we

remove that variable which indicates the columns “Radio” and “Newspaper” and save the modified dataset as new file called “TV_vs_Sales. With this approach we did not overwrite the original data as we will need it.

```r
#To check the relationship between TV ads and sales and see if it is linear or it needs transformations
#we need to remove the Radio and Newspaper columns as we do not need them for now.

# Identify and remove the columns
columns_to_remove <- c("Radio", "Newspaper")
TV_vs_Sales <- advertising_data[, -which(colnames(advertising_data) %in% columns_to_remove)]



# Save the modified dataset as a new file
write.csv(TV_vs_Sales, "TV_vs_Sales.csv", row.names = FALSE)
```

Then, we can create a scatter plot and try to apply a regression line to see whether the relation between TV ads and sales is linear.



The reason why this dataset is chosen is that linear regression is applicable to it and no complex transformations would be necessary for that purpose.

The scatter plot is obtained through the code:

```r

# Create a scatter plot to visualize the relationship between years of experience and salary
plot(TV_vs_Sales$TV, TV_vs_Sales$Sales, xlab = "TV Advertisements", ylab = "Sales", main = "Scatter Plot")
```

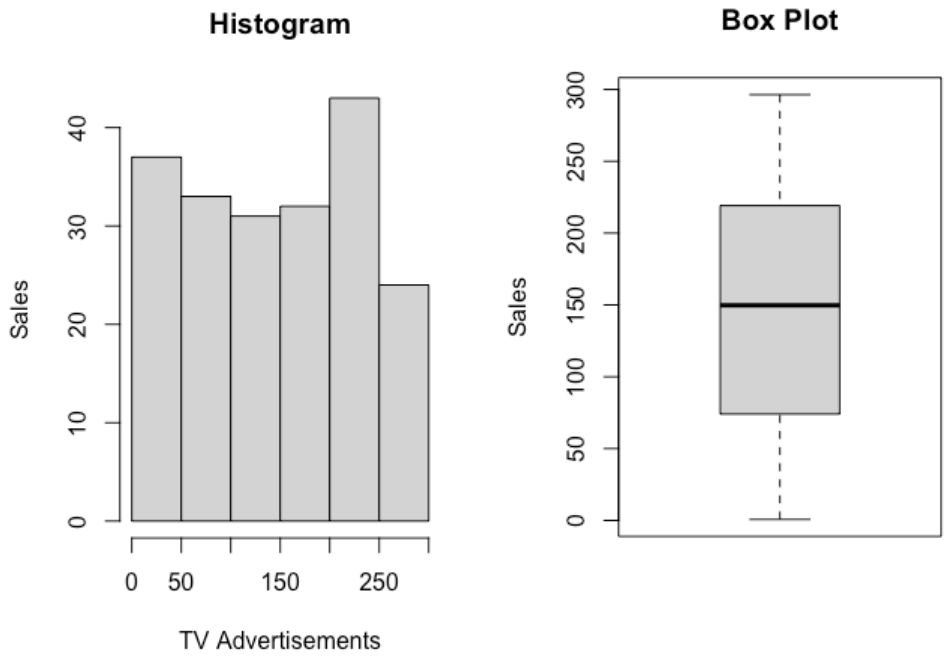
To apply a regression line to the plot, first we should fit the model:

```
24  # Fit a linear regression model
25  lm_model <- lm(Sales ~ TV, data = TV_vs_Sales)
```

Then, we create the plot and add the regression line:

```
27  # Create a scatter plot
28  plot(TV_vs_Sales$TV, TV_vs_Sales$Sales,
29       xlab = "TV Advertisements", ylab = "Sales",
30       main = "Scatter Plot with Linear Regression Line")
31  # Add the regression line to the scatter plot
32  abline(lm_model, col = "red", lwd = 2)
```

We can also use other methods of visualizations to have a general idea of the population distribution. For instance, below is a histogram that shows that the population (TV_vs_Sales is our population) which is obtained through the code on the next page:

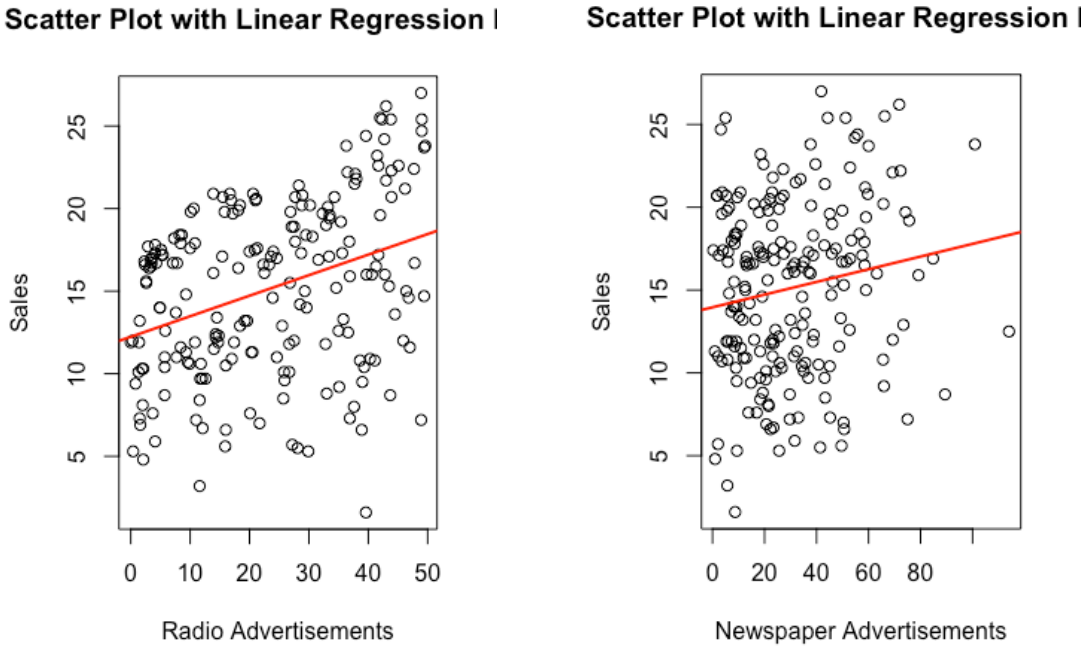



```
36  # Create a histogram
37  hist(TV_vs_Sales$TV, main = "Histogram", xlab = "TV Advertisements", ylab = "Sales")
38
39  # Create a box plot
40  boxplot(TV_vs_Sales$TV, main = "Box Plot", ylab = "Sales")
```

The histogram shows that the highest sales are more than $40K and belong to the $200K-$250K budget of TV advertisements and the lowest sales in the population are less than $25K and belong to the budget larger than $250K.

What the box plot shows about the population is that the mean of the sales is around $150K and the 1st quartile is between $50K and $100K, and the 3rd quartile is $200K and $250K.

Now, to see the overall relation between Radio and Sales, we repeat the same process except, we keep Radio column. The result is shown in the graph below. As you can see, linear regression cannot be applied in this case even with transformations as there is not a clear relationship between Radio ads and Sales. The same goes for Newspaper advertisements.



2.0 Data Collection Description

2.1 Sample Selection Method

In some of the statistical tests that will be conveyed to the data, two or more groups will be compared (e.g., t-test, chi-squared test, ANOVA). In these cases, it is necessary for the samples to be independent of each other. Else, there may be bias in the results of the tests or they may be inaccurate. For this purpose, the size of the sample must be at most 10% of the size of the population. As there are 200 observations in the data, sample sizes of 20 will be used to make sure all the samples are independent of each other. But, for normality test to be applied on the sample, the sample size has to be at least 30 for the central limit theorem (CLT) to be applied. This is why the first sample selected has a size of 30. Additionally, in all cases the process of the sample selection must be random.

To select a random sample of the TV advertisements, we extract the column from the main dataset and store it in a vector as shown in the code below:

```
81  # Extract a column as a vector using the $ operator
82  TVvector <- advertising_data$TV
83
84  # Extract a column as a vector using the [ ] operator
85  TVvector <- advertising_data["TV"]
```

Then, 30 random integers in the range [1, 200] is generated:

```
> rand_nums <- sample(1:200, 30, replace = FALSE)
> rand_nums
 [1] 105  29 118  91 101 155 199  52 152  72 126  57 171 140 151  24  26  71  35 115  73 131  47 100  46 129 166 123 116 177
```

Then, these integers are assigned to the indices of the vector and the data points are stored in the vector called SampleTV. After that, for a proper form, the data frame is transposed.

```
> # Create the data frame
> SampleTV <- data.frame(TV = c(44.7, 248.8, 76.4, 134.3, 222.4, 187.8, 283.6, 100.4, 121.0, 109.8, 87.2, 7.3, 50.0, 184.9, 280.7, 228.3, 262.9, 199.1, 95.7, 78.2, 26.
8, 0.7, 89.7, 135.2, 175.1, 220.3, 234.5, 224.0, 75.1, 248.4))
> # Transpose the data frame
> t(SampleTV)
   [,1]  [,2] [,3]  [,4]  [,5]  [,6]  [,7]  [,8] [,9] [,10] [,11] [,12] [,13] [,14] [,15] [,16] [,17] [,18] [,19] [,20] [,21] [,22] [,23] [,24] [,25] [,26]
TV 44.7 248.8 76.4 134.3 222.4 187.8 283.6 100.4  121 109.8  87.2   7.3    50 184.9 280.7 228.3 262.9 199.1  95.7  78.2  26.8   0.7  89.7 135.2 175.1 220.3
   [,27] [,28] [,29] [,30]
TV 234.5   224  75.1 248.4
```

A similar process is repeated for Sales and the corresponding indices are then assigned to the datapoints of the same column.

```
102
103 # Create the data frame for sales (the other dimension)
104 SampleSales <- data.frame(Sales = c(20.7, 18.9, 9.4,  14.0, 16.7, 20.6, 25.5, 10.7, 11.6, 12.4,  10.6,  5.5,
105 # Transpose the data frame
106 SampleSales <- t(SampleSales)
```

Also, A similar process is repeated for Radio and the corresponding indices of a new set of random integers in range [1,200] are then assigned to the datapoints of the same column.

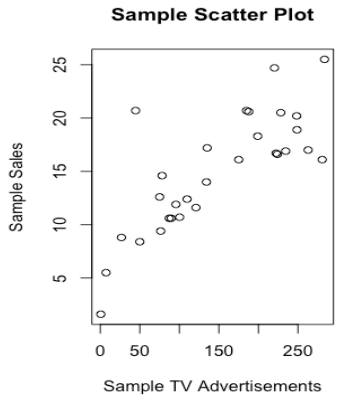
```
> #Now we can use the sample() function to pick 20 random integers from 1-200
> #Generate 20 random numbers between 1 and 200
> rand_nums <- sample(1:200, 20, replace = FALSE)
>
> # View the generated random numbers
> rand_nums
 [1] 146  28 143 200  76 193  72 141 173 134 124 108 123   5  87 102  19 136  81 119
```

```
252 # Create the data frame
253 SampleRadio <- data.frame(Radio = c(1.9, 16.7, 33.2, 8.6, 222.4, 43.7, 4.1, 14.3, 17.0, 20.1, 33.5, 34.6, 0.3
254 # Transpose the data frame
255 SampleRadio <- t(SampleRadio)
```

3 Diagrams, Charts (Sample Statistics)

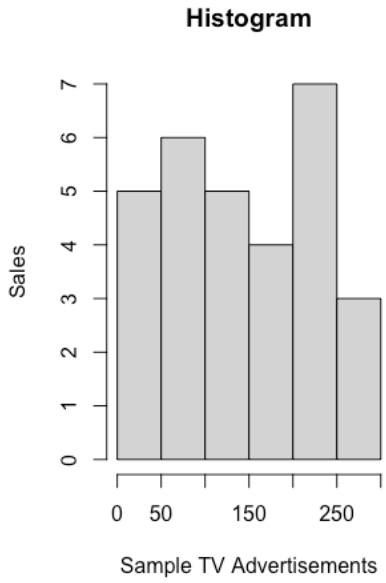
In the sample obtained, a scatter plot is created regarding the two dimensions of the sample through the code below.

```
# Create a scatter plot
plot(SampleTV, SampleSales,
     xlab = "Sample TV Advertisements", ylab = "Sample Sales",
     main = "Sample Scatter Plot")
```



As you can see, the linear relation that exists between the variables in the population, can also be detected in the sample scatter plot of the two variables. However, there are a few outliers in this case.

To study the sample further, we can use other methods of visualization. For instance, the histogram below shows that there are obvious similarities between the histogram of the population and the sample. In both of them, the highest sales belong to the $200K-$250K budget of TV advertisements and the lowest sales to the budget larger than $250K. The other bins relatively show a pattern similar to the population histogram.




```
# Create a histogram
hist(SampleTV, main = "Histogram", xlab = "Sample TV Advertisements", ylab = "Sales")
```

To be able to analyze the sample data better we also need to see the minimum, the 1st and 3rd quartile, the mean and the median and the maximum value. We could do this by using the summary function but as we transposed the sample before (to turn it into a column), we need to turn it into a row again, to be able to use this function.

```
> SampleTVForSummary <- t(SampleTV)
> summary(SampleTVForSummary)
       TV
 Min.   :  0.70
 1st Qu.: 80.45
 Median :134.75
 Mean   :147.78
 3rd Qu.:223.60
 Max.   :283.60
```

4.0 Normality Tests

4.1 Shapiro-Wilk Test

The normality test used to see whther or not the distribution of TV ads is normal, is the Shapiro-Wilk test. In this test:

The null hypothesis (H0): The data is normally distributed

The alternative hypothesis(H1): The data is not normally distributed.

The chosen significance level(alpha): 0.05

To determine whether to reject H0 a test statistic is calculated and compared to the critical value. If the calculated p-value is smaller than alpha, then H0 is rejected in favor of H1(The data is not normally distributed.). Else, there would not be enough evidence to reject H0 and we would consider the data to be normally distributed.

```
> shapiro.test(SampleTV)

        Shapiro-Wilk normality test

data:  SampleTV
W = 0.94488, p-value = 0.1232
```

As you can see, in this case the calculated p-value is larger than the significance level which is 0.05 in this case. Thus, the conclusion made would be:

There is not enough evidence to reject the null hypothesis.

Hence, we will consider that the sample TV data is normally distributed.

According to the Shapiro-Wilk test, neither the Radio population nor the TV population is normally distributed following a similar process. As their calculated p-value is much smaller than the significance level (alpha = 0.05). Thus, H0 is rejected in favor of H1 and the populations are not normally distributed.

```
> #Shapiro-Wilk test to see whether or not the Radio and the TV population is normally distributed.
> # Shapiro-Wilk normality test for the population(TV)
> TVvectorTransposed <- t(TVvector)
> shapiro.test(TVvectorTransposed)

        Shapiro-Wilk normality test

data:  TVvectorTransposed
W = 0.94951, p-value = 1.693e-06

> # Shapiro-Wilk normality test for the population(Radio)
> RadiovectorTransposed <- t(Radiovector)
> shapiro.test(RadiovectorTransposed)

        Shapiro-Wilk normality test

data:  RadiovectorTransposed
W = 0.94401, p-value = 5.198e-07
```

As shown by Shapiro-Wilk test, we assume that the Sales population is normally distributed as the calculated p-value is greater than alpha = 0.05. Because there is not enough evidence to reject the null hypothesis.

```
> # Shapiro-Wilk normality test for the population
> SalesvectorTransposed <- t(Salesvector)
> shapiro.test(SalesvectorTransposed)

        Shapiro-Wilk normality test

data:  SalesvectorTransposed
W = 0.98752, p-value = 0.07645
```

· 4.2 Examples

The significance level(alpha) by default is 0.05 which is smaller than the p-value calculated. Thus, there is not enough evidence to reject H0 (The null hypothesis is that the population is normally distributed.) so we can assume that the Sales population is normally distributed.

Through the following code we can calculate and store the standard deviation and the mean. We also had got the mean from the summary function.

```
> # Calculate and store the standard deviation
> sales_sd <- sd(SalesvectorTransposed)
>
> #View the standard deviation: 5.283892
> sales_sd
[1] 5.283892
>
> # Store the mean
> salesMean <- 15.13
```

The ggplot2 library is installed for depicting the graphs of the questions.

```
203 #ggplot library is installed for depicting the graphs of the questions.
204 install.packages("ggplot2")
205
206 library(ggplot2)
```

Question 1: Assume that x is our data (Sales population). Determine the value of x that solves

a. $P(X>x) = 0.6$

b. $P(X<x) = 0.85$

c. $P(Z<x) = 0.4$

(mu = 15.13, sigma = 5.283892)

Solution: Part a basically asks us to find the value of x for which the data points less than it would make up 0.6 of the whole data. Part b wants the value of x for which the data points bigger than it would make up 0.85 of the whole data. Part c wants the value of x for which the z-value of the data points below it (less than x) would make up 0.4 of the data. The solution is displayed on the next page.

```
> #Question 1
> # a. Calculate the value of x for P(X > x) = 0.6
> p1 <- 0.6
> x1 <- qnorm(1 - p1, salesMean, sales_sd)
> cat("x1 =", x1, "\n")
x1 = 13.79134
>
> # b. Calculate the value of x for P(X < x) = 0.85
> p2 <- 0.85
> x2 <- qnorm(p2, salesMean, sales_sd)
> cat("x2 =", x2, "\n")
x2 = 20.6064
>
> # c. Calculate the value of x for P(Z < x) = 0.04
> p3 <- 0.04
> x3 <- qnorm(p3, salesMean, sales_sd)
> x3 <- x3 * sales_sd + salesMean
> cat("x3 =", x3)
x3 = 46.19698
```

Question 2: As the SampleSales size is 30, according to the CLT the distribution is normal. What is the probability that the sale is less than 20?(mu = 15.13, sigma = 5.283892)

Solution: The probability of the sale being less than 20 is about 0.822 according to the solution.

```
> #Question 2
> # Calculate the probability P(X < 20)
> x <- 20
> p <- pnorm(x, salesMean, sales_sd)
> cat("P(X < 20) =", p)
P(X < 20) = 0.8216494
```

5. Point Estimations and Confidence Intervals

Point Estimations

Question 3: Estimate the mean of the sales population by using the sample sales.

Solution: In this problem, we can calculate the sample mean to estimate the population mean.

```
> #Question 3: Estimating the mean of the population using the sample
> # Calculate the sample mean
> sample_mean <- mean(SampleSales)
> # Print the sample mean
> cat("Sample Mean:", sample_mean)
Sample Mean: 14.98
```

The sample mean is 14.98 which must be a good estimation for the population mean. (The population mean 15.13 which was calculated before.)

Question 4: Estimate the standard deviation of the sales population by using the sample sales.

Solution: We can calculate the sample standard deviation (s = 5.507387) as a good estimation of the population standard deviation (sigma = 5.283892 as calculated before).

```
> #Question 4: Estimating the sd of the population using the sample
> # Calculate the sample standard deviation
> sample_std <- sd(SampleSales)
> # Print the estimated population standard deviation
> cat("Estimated population standard deviation:", sample_std, "\n")
Estimated population standard deviation: 5.507387
```

Confidence Intervals

Question 5: Find a confidence Interval for the population mean with 95% confidence level. [n = 30, xbar = 14.98, s = 5.507387)

Solution: First set the confidence level, then calculate the standard error by using the formula, then find the margin of error and add it to the sample mean for the upper bound and subtract it from the sample mean for the lower bound. The process and the answer is shown below.

```
> #Question 5: Confidence Interval for the Mean (95% confidence level)
> # Set the confidence level (e.g., 95%)
> confidence_level <- 0.95
> # Calculate the standard error
> standard_error <- sample_std / sqrt(length(SampleSales))
> # Calculate the margin of error
> margin_of_error <- qnorm(1 - (1 - confidence_level) / 2) * standard_error
> # Calculate the lower and upper bounds of the confidence interval
> lower_bound <- sample_mean - margin_of_error
> upper_bound <- sample_mean + margin_of_error
> # Print the confidence interval
> cat("Confidence Interval for the Mean (95% confidence level): [", lower_bound, ",", upper_bound, "]\n")
Confidence Interval for the Mean (95% confidence level): [ 13.00924 , 16.95076 ]
```

Question 6: Find a confidence interval for the standard deviation of the population (90% confidence level) using chi-square test. (n= 30, s = 5.507387)

Solution: First we set the confidence level then, we set the degrees of freedom and calculate the lower and upper bounds by using the formula and chi-squared test. The process and the final answer is shown below.

```
> #Question 6: Confidence Interval for the Standard Deviation (90% confidence level) using chi-square test
> # Set the confidence level
> confidence_level <- 0.90
> # Set the degrees of freedom
> df <- length(SampleSales) - 1
> # Calculate the lower and upper bounds of the confidence interval
> lower_bound <- sqrt(df * sample_std^2 / qchisq((1 + confidence_level) / 2, df))
> upper_bound <- sqrt(df * sample_std^2 / qchisq((1 - confidence_level) / 2, df))
> # Print the confidence interval
> cat("Confidence Interval for the Standard Deviation (95% confidence level): [", lower_bound, ",", upper_bound, "]
\n")
Confidence Interval for the Standard Deviation (95% confidence level): [ 4.546312 , 7.047829 ]
```

- 6. 2-3 Hypothesis Tests

.2 Hypothesis Test

In hypothesis tests, if the data is normally distributed the z test is used and if it is not normally distributed t test is used. For having more variety in the questions, a random sample from the Radio budget population will be used which is not normally distributed and requires the t test. 20 random integers from 1 to 200 was generated for the sample as explained in Data Collection Description section. Note that the sample size was picked as 20 because if it was 30, according to the CLT the sample would be normally distributed and we would not be able to use the t test. We would have to use the z test. Additionally, a new version of the TV sample will be stored which has a size of 20 for the same reason.

Question 7: A random sample of 20 data points from the Radio population claims that the population mean of Radio is greater than 23. If the significance level is 0.05, is the claim true? (mu = 23)
Solution: The null hypothesis (H0) is that the mean of the population is equal to 23.
The alternative hypothesis (Ha or the claim) is that the mean value of the population is greater than 23. This is suggested by the sample mean being larger than 23.
We solve the question by performing one-sample t-test and storing the p-value from the result of the t.test() function. If the p-value is equal to or less than alpha (0.05), we reject the null hypothesis in favor of the alternative hypothesis.

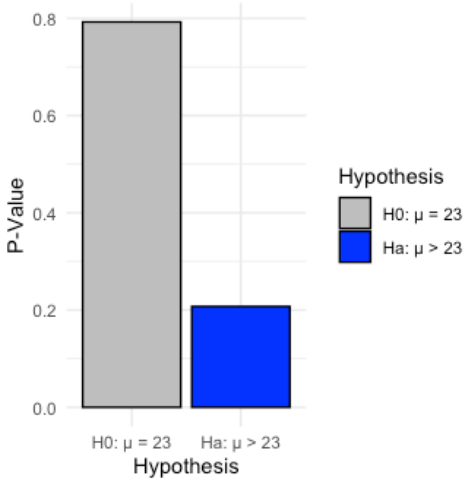
```r
#Question 7: Claim: The population mean of Radio is greater than 23.
# Set the significance level (alpha)
alpha <- 0.05
# Null hypothesis (H0): The mean value of Radio budgets is equal to 23
# Alternative hypothesis (Ha): The mean value of Radio budgets is greater than 23
# Perform one-sample t-test
t_test_result <- t.test(SampleRadio, mu = 23, alternative = "greater")
# Extract the p-value from the test result
p_value <- t_test_result$p.value
# Compare p-value with significance level to make a decision
if (p_value <= alpha) {
  cat("Reject the null hypothesis (H0), p-value = ", p_value)
} else {
  cat("Fail to reject the null hypothesis (H0), p-value = ", p_value)
}
```

Output:

```
Fail to reject the null hypothesis (H0), p-value =  0.2074528
```

As shown above, the p-value is calculated as 0.2074528 which is greater than the p-value (0.05). Thus, there is not enough evidence to reject the null hypothesis. It is then assumed that the population mean is 23.

In the bar plot you can see the null and the alternative hypothesis and the p-value. The code that generated the bar plot is below the graph.




```r
#Graph for Question 7
# Create a data frame for plotting
df <- data.frame(Hypothesis = c("H0: μ = 23", "Ha: μ > 23"),
                 P_Value = c(1 - p_value, p_value))
# Create a bar plot
ggplot(df, aes(x = Hypothesis, y = P_Value, fill = Hypothesis)) +
  geom_bar(stat = "identity", color = "black") +
  labs(x = "Hypothesis", y = "P-Value", fill = "Hypothesis") +
  scale_fill_manual(values = c("grey", "blue")) +
  theme_minimal()
```

Question 8: A random sample of 30 data points from the Sales population claims that the population mean of sales is not equal to 15.13. If the significance level is 0.05, is the claim true? (mu = 15.13, xbar = 14.98, sigma = 5.283892)

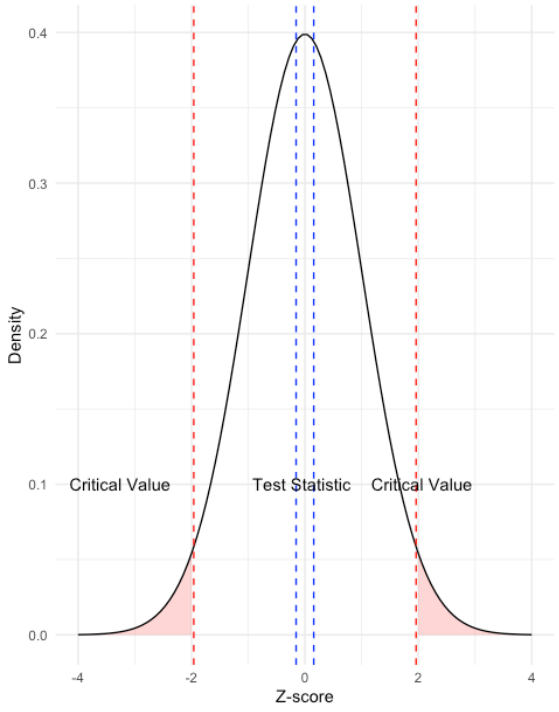
Solution: The null hypothesis (H0) is that the mean of the population is equal to 15.13.
The alternative hypothesis (Ha or the claim) is that the mean value of the population is not equal to 15.13. This problem is solved by finding the z-score of the sample mean and then calculating the critical value by using alpha in the qnorm() function. Then the p-value is calculated by the pnorm() function and then, the critical value and the p-value are compared to make a decision.
As the output of the code on the next page, we can see: p-value: 0.8764363 , critical value: 1.959964
As the p-value is smaller than the critical value, we fail to reject H0. There is not enough evidence to reject the null hypothesis. Thus, the population is assumed to have a mean of 15.13.

```
302 #Question 8: Claim: The mean weight of a population is not equal to 15.13
303 # Set the significance level (alpha)
304 alpha <- 0.05
305 # Null hypothesis (H0): The mean value of the Sales population is equal to 15.13
306 # Alternative hypothesis (Ha): The mean weight of a population is not equal to 15.13
307
308 # Calculate the test statistic (Z-score)
309 z_score <- (sample_mean - salesMean) / (sales_sd / sqrt(length(SampleSales)))
310 # Calculate the critical value for a two-tailed test
311 critical_value <- qnorm(1 - alpha/2)
312 # Calculate the p-value
313 p_value <- 2 * (1 - pnorm(abs(z_score)))
314 # Compare the test statistic with the critical value and p-value to make a decision
315 if (p_value > critical_value) {
316   cat("Reject the null hypothesis (H0)")
317 } else {
318   cat("Fail to reject the null hypothesis (H0)")
319 }
320 cat("\n")
321 cat("p-value:", p_value, ", critical value: ", critical_value)
322
322:1  (Top Level)
```

```
Console  Terminal  Background Jobs
R 4.3.0 · ~/Downloads/
Fail to reject the null hypothesis (H0)> cat("\n")

> cat("p-value:", p_value, ", critical value: ", critical_value)
p-value: 0.8764363 , critical value:  1.959964
```

The corresponding plot is on the next page. As you can see, a two-tailed graph is depicted and as calculated before, the absolute value of the critical value is very close to 2.
The p-value is approximately 0.9 and it is not in the critical region.
Thus, we fail to reject the null hypothesis and there is not enough evidence to reject H0.
The graph is obtained through the code on the next page.

```r
install.packages("ggplot2")

library(ggplot2)
# Set the significance level (alpha)
alpha <- 0.05
# Null hypothesis (H0): The mean value of the Sales population is equal to 15.13
# Alternative hypothesis (Ha): The mean weight of a population is not equal to 15.13
# Calculate the test statistic (Z-score)
z_score <- (sample_mean - salesMean) / (sales_sd / sqrt(length(SampleSales)))
# Calculate the critical value for a two-tailed test
critical_value <- qnorm(1 - alpha/2)
# Calculate the p-value
p_value <- 2 * (1 - pnorm(abs(z_score)))
# Create a data frame for plotting
df <- data.frame(x = c(-4, 4))
# Create a ggplot object and add layers for the two tails
ggplot(df, aes(x = x)) +
  stat_function(fun = dnorm, args = list(mean = 0, sd = 1), n = 100, geom = "area",
                xlim = c(-4, -2), fill = "red", alpha = 0.2) +
  stat_function(fun = dnorm, args = list(mean = 0, sd = 1), n = 100, geom = "area",
                xlim = c(2, 4), fill = "red", alpha = 0.2) +
  stat_function(fun = dnorm, args = list(mean = 0, sd = 1), n = 100, geom = "line") +
  geom_vline(xintercept = z_score, color = "blue", linetype = "dashed") +
  geom_vline(xintercept = -z_score, color = "blue", linetype = "dashed") +
  geom_vline(xintercept = critical_value, color = "red", linetype = "dashed") +
  geom_vline(xintercept = -critical_value, color = "red", linetype = "dashed") +
  annotate("text", x = z_score + 0.1, y = 0.1, label = "Test Statistic") +
  annotate("text", x = critical_value + 0.1, y = 0.1, label = "Critical Value") +
  annotate("text", x = -critical_value - 1.3, y = 0.1, label = "Critical Value") +
  labs(x = "Z-score", y = "Density") +
  theme_minimal()
```



.2 Hypothesis Test

Question 9: One claims that the means of the Radio budget and TV budget populations are significantly different. With alpha = 0.05, determine whether to accept or reject the claim. (n of both samples = 20).

Solution: By 2 hypothesis testing, we are able to determine whether the mean of the two populations are significantly different or not by using the t.test() function.
The null hypothesis (H0): The mean of the Radio budget population is not significantly different from the mean of the TV budget population.
The alternative hypothesis (Ha): The means of the two populations are significantly different.

```
398  #2 Hypothesis Test
399
400  #Question 9: Claim: The means of the two samples (Radio sample and TV sample) are significantly different.
401
402  # Hypothesis Test
403  # Step 1: Set the significance level
404  alpha <- 0.05
405  # Step 2: Calculate the test statistic (t-value) and p-value
406  result <- t.test(SampleRadio, SampleTVNew)
407  # Step 3: Compare the p-value to the significance level
408  if (result$p.value <= alpha) {
409    cat("Reject the null hypothesis. The means of the two samples are significantly different. , P-value = ", result$p.value)
410  } else {
411    cat("Fail to reject the null hypothesis. The means of the two samples are not significantly different.")
412  }
413
391:79  (Top Level)

Console  Terminal  Background Jobs

R 4.3.0 · ~/Downloads/

Reject the null hypothesis. The means of the two samples are significantly different. , P-value =  6.101197e-06
```

As shown in the output, the p-value is much smaller than the significance level which is 0.05. Thus, the null hypothesis is rejected in favor of the alternative hypothesis. Hence, the claim (Ha) is accepted and the means of the populations are significantly different.

Question 10: Perform 2 and 1 hypothesis testing on two random samples of the TV populations with the significance level of 0.05 and state the conclusion for each test.

Solution:

Hypothesis Test 1:
The null hypothesis (H0): The means of the two samples are not significantly different.

The alternative hypothesis (Ha): The means of the two samples are significantly different.

The p-value: 0.8202 alpha: 0.05 p-value > alpha

Conclusion: There is not enough evidence to reject H0.

We assume that the means of the two samples are not significantly different.

Hypothesis Test 2:
The null hypothesis (H0): The mean of sample 1 is not significantly less than the mean of sample 2.

The alternative hypothesis (Ha): The mean of sample 1 is significantly less than the mean of sample 2.

The p-value: 0.410103 alpha: 0.05 p-value > alpha

Conclusion: There is not enough evidence to reject H0.

We assume that the mean of sample 1 is not significantly less than the mean of sample 2.

The solution in R and the output of the code is provided on the next page.

```r
#Question 10

# Hypothesis Test 1
# Step 1: Set the significance level
alpha <- 0.05
# Step 2: Calculate the test statistic (t-value) and p-value
result_1 <- t.test(SampleTVNew, SampleTV2, paired = TRUE)
# Step 3: Compare the p-value to the significance level
if (result_1$p.value <= alpha) {
  cat("Reject the null hypothesis for Hypothesis Test 1. The means of the two samples are significantly different. , p-value = ", result_1$p.value)
} else {
  cat("Fail to reject the null hypothesis for Hypothesis Test 1. The means of the two samples are not significantly different. , p-value = ", result_1$p.value)
}
# Hypothesis Test 2
# Step 1: Set the significance level
alpha <- 0.05
# Step 2: Calculate the test statistic (t-value) and p-value
result_2 <- t.test(SampleTVNew, SampleTV2, alternative = "less", paired = TRUE)
# Step 3: Compare the p-value to the significance level
if (result_2$p.value <= alpha) {
  cat("Reject the null hypothesis for Hypothesis Test 2. The mean of sample 1 is significantly less than the mean of sample 2. , p-value = ", result_2$p.value)
} else {
  cat("Fail to reject the null hypothesis for Hypothesis Test 2. The mean of sample 1 is not significantly less than the mean of sample 2., p-value = ", result_2$p.value)
}

422:63  (Top Level)
```

Console | Terminal | Background Jobs

R 4.3.0 · ~/Downloads/

```
+ }
Fail to reject the null hypothesis for Hypothesis Test 1. The means of the two samples are not significantly different. , p-value =  0.8202061> # Hypothesis Test 2
> # Step 1: Set the significance level
> alpha <- 0.05
> # Step 2: Calculate the test statistic (t-value) and p-value
> result_2 <- t.test(SampleTVNew, SampleTV2, alternative = "less", paired = TRUE)
> # Step 3: Compare the p-value to the significance level
> if (result_2$p.value <= alpha) {
+     cat("Reject the null hypothesis for Hypothesis Test 2. The mean of sample 1 is significantly less than the mean of sample 2. , p-value = ", result_2$p.value)
+ } else {
+     cat("Fail to reject the null hypothesis for Hypothesis Test 2. The mean of sample 1 is not significantly less than the mean of sample 2., p-value = ", result_2$p.value)
+ }
Fail to reject the null hypothesis for Hypothesis Test 2. The mean of sample 1 is not significantly less than the mean of sample 2., p-value =  0.410103>
```

- 7. Goodness of Fits Tests and other Checks for Detecting the Distribution

1. Goodness-of-Fit Test (Chi-Squared Test)

Use the Chi-squared test to see whether the TV Budget population is normally distributed. (alpha = 0.05)

Solution:

The null hypothesis (H0): The distribution is normal.

The alternative hypothesis (Ha): The distribution is not normal.

```
> #Goodness of Fits Tests and other Checks for Detecting the Distribution
>
> #Question 11: Goodness-of-Fit Test (Chi-Squared Test):
> # Perform a chi-squared goodness-of-fit test
> # Assume you want to test if the observed variable follows a normal distribution
> # Perform the chi-squared test
> chi_squared_test <- chisq.test(TVvector)
> # Print the test result
> print(chi_squared_test)

        Chi-squared test for given probabilities

data:  TVvector
X-squared = 9975.5, df = 199, p-value < 2.2e-16
```

As you can see in the output of the code above, the p-value is calculated as a very small number which is less than the significance level.

Conclusion: Reject H0 in favor of Ha. Hence, the distribution of the TV Budget population is not normal.

Question 12: Use Chi-squared test to determine whether the Sales and Radio variables are independent. (alpha = 0.05)

Solution: Pearson's Chi-squared test is used for this purpose.

The null hypothesis (H0): The two variables are independent.
The alternative hypothesis (Ha): The two variables are dependent.

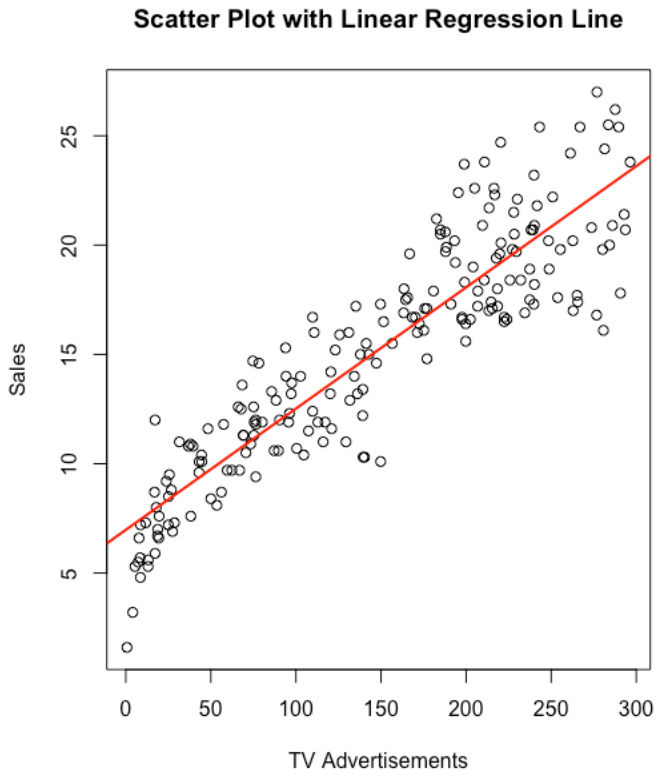
```
460  #Question 12: Use Chi-squared test to test two variables independence
461
462  # Create a contingency table of two categorical variables
463  # The two variables: Radio and Sales
464  # Create a data frame with the two variables
465  data <- data.frame(advertising_data$Sales, advertising_data$Radio)
466  # Create a contingency table
467  contingency_table <- table(data)
468  # Perform the chi-square test of independence
469  chi_square_test <- chisq.test(contingency_table)
470  # Print the test result
471  print(chi_square_test)
456:17  (Top Level)
Console  Terminal  Background Jobs
R 4.3.0 · ~/Downloads/

        Pearson's Chi-squared test

data:  contingency_table
X-squared = 19921, df = 19920, p-value = 0.4964
```

p-value > alpha

Conclusion: There is not enough evidence to reject H0.
The Sales and Radio variables are assumed to be independent.

In our data, all the values are numerical which means there cannot be the concept of proportions. Thus, we are not able to use the Chi-squared test for homogeneity as it involves proportions.

- 8. Linear Regression Model Design



As you can see the correlation of the TV advertisements and the sales is a positive linear relationship. For getting this graph, first we should fit the model.

```
#Linear Regression Model Design

# Fit a linear regression model
lm_model <- lm(Sales ~ TV, data = TV_vs_Sales)

# Create a scatter plot
plot(TV_vs_Sales$TV, TV_vs_Sales$Sales,
     xlab = "TV Advertisements", ylab = "Sales",
     main = "Scatter Plot with Linear Regression Line")
# Add the regression line to the scatter plot
abline(lm_model, col = "red", lwd = 2)
```

Question 13: Predict the Sales by using TV, Radio and Newspaper advertising budgets, calculate the residual error, p-value and F-statistic.
Solution: For obtaining all these values, we can use the summary function and use it with the fitted linear regression model. But first, we should fit the model again and this time, use all the variables.

```
472
473 #Linear Regression Model Design
474
475 # Fit a linear regression model
476 lm_model <- lm(Sales ~ TV, data = TV_vs_Sales)
477
478 # Create a scatter plot
479 plot(TV_vs_Sales$TV, TV_vs_Sales$Sales,
480      xlab = "TV Advertisements", ylab = "Sales",
481      main = "Scatter Plot with Linear Regression Line")
482 # Add the regression line to the scatter plot
483 abline(lm_model, col = "red", lwd = 2)
484
485 #Question 13: Predict the Sales by using TV, Radio and Newspaper advertising budgets, calculate the residual error, p-value and F-statistic
486
487 # Fit a linear regression model
488 lm_model <- lm(advertising_data$Sales ~ advertising_data$TV + advertising_data$Radio + advertising_data$Newspaper, data = advertising_data)
489 # Print the model summary
490 summary(lm_model)
476:47  (Top Level)
```

```
Console  Terminal  Background Jobs
R 4.3.0 · ~/Downloads/

Call:
lm(formula = advertising_data$Sales ~ advertising_data$TV + advertising_data$Radio +
    advertising_data$Newspaper, data = advertising_data)

Residuals:
    Min      1Q  Median      3Q     Max
-7.3034 -0.8244 -0.0008  0.8976  3.7473

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                4.6251241  0.3075012  15.041   <2e-16 ***
advertising_data$TV        0.0544458  0.0013752  39.592   <2e-16 ***
advertising_data$Radio     0.1070012  0.0084896  12.604   <2e-16 ***
advertising_data$Newspaper 0.0003357  0.0057881   0.058    0.954
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.662 on 196 degrees of freedom
Multiple R-squared:  0.9026,    Adjusted R-squared:  0.9011
F-statistic: 605.4 on 3 and 196 DF,  p-value: < 2.2e-16
```

Question 14: If the advertisement budgets for a company are TV: $230.1K, Radio: $37.8K, Newspaper: $69.2K what is the predicted sales and the residual error of the predicted sales?

Solution:

```
493 #Question 14: Print all the predicted values. Ad budgets, TV: $230.1K, Radio: $37.8K, Newspaper: $69.2K residual error of the predicted sales?
494
495 # Specify the values for the three variables
496 TV_value <- 230.1
497 Radio_value <- 37.8
498 Newspaper_value <- 69.2
499 # Create a new data frame with the values for prediction
500 new_data <- data.frame(TV = TV_value, Radio = Radio_value, Newspaper = Newspaper_value)
501 # Predict the outcome variable for the new data using the linear regression model
502 predicted_value <- predict(lm_model, newdata = new_data)#21.22097
503 # Print the predicted value
504 cat("Predicted value:", predicted_value, "\n")
505 # Calculate the residual error
506 observed_value <- 22.1 # Specify the observed value
507 residual <- observed_value - predicted_value
508 # Print the residual error
509 cat("Residual:", residual, "\n") #0.8790279
509:44 (Top Level)
Console Terminal Background Jobs
R 4.3.0 · ~/Downloads/
> # Print the predicted value
> cat("Predicted value:", predicted_value, "\n")
Predicted value: 21.22097 11.26825 10.49621 17.31245 15.64414 10.35634 11.27328 13.27062 5.318396 15.78871 8.85272 18.88401 9.698797 10.74921 19.27329
```

As you can see, the predicted value of the first element is 21.22097. (The requirements of the question belonged to the first row of advertising_data.)
For the residual error, we must subtract the predicted value from the observed value (22.1 as depicted in the original data). 0.8790279

```
> # Calculate the residual error
> observed_value <- 22.1 # Specify the observed value
> residual <- observed_value - predicted_value
> # Print the residual error
> cat("Residual:", residual, "\n") #0.8790279
Residual: 0.8790279 10.83175 11.60379 4.787553 6.455863 11.74366 10.82672 8.829375 16.7816 6.31129 13.24728 3.2
```

- 9. Analysis of Variance (ANOVA)

One-way ANOVA

Question 15: Is there a significant difference in the mean values of the three advertising methods?(alpha = 0.05)

Solution:
H0: There is no significant difference in the mean values of the three advertising methods.
Ha: There is a significant difference in the mean values of the three advertising methods.

```
> # Load the necessary packages
> library(dplyr)
> # Perform one-way ANOVA
> anova_result <- aov(advertising_data$Sales ~ advertising_data$TV * advertising_data$Radio * advertising_data$Newspaper, data = advertising_data )
> # Check the ANOVA table
> summary(anova_result)
                                                                Df Sum Sq Mean Sq  F value   Pr(>F)
advertising_data$TV                                              1   4512    4512 1834.699  < 2e-16 ***
advertising_data$Radio                                           1    502     502  204.244  < 2e-16 ***
advertising_data$Newspaper                                       1      0       0    0.004    0.951
advertising_data$TV:advertising_data$Radio                       1     64      64   25.924 8.44e-07 ***
advertising_data$TV:advertising_data$Newspaper                   1      2       2    0.649    0.421
advertising_data$Radio:advertising_data$Newspaper                1      1       1    0.514    0.474
advertising_data$TV:advertising_data$Radio:advertising_data$Newspaper  1      2       2    0.959    0.329
Residuals                                                      192    472       2
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> # Check the p-value
> p_value <- summary(anova_result)[[1]]$"Pr(>F)"[1]
> # Compare p-value with significance level to make a decision
> alpha <- 0.05
> if (p_value <= alpha) {
+     cat("There is a significant difference in the mean values of the groups.")
+ } else {
+     cat("There is no significant difference in the mean values of the groups.")
+ }
There is a significant difference in the mean values of the groups.
>
```

The p-value is smaller than alpha.
Conclusion: H0 is rejected in favor of Ha.
There is a significant difference in the mean values of the three advertising methods.

Two-way ANOVA

Question 16: Is there a significant effect of the TV advertisement Budget or the Radio advertisement Budget, or the Newspaper advertisement Budget or their interaction on sales? (alpha = 0.05)

Solution:

H0: There is not a significant effect of the TV advertisement Budget or the Radio advertisement Budget, or the Newspaper advertisement Budget or their interaction on sales.
Ha: There is a significant effect of the TV advertisement Budget or the Radio advertisement Budget, or the Newspaper advertisement Budget or their interaction on sales.

```
> # Perform two-way ANOVA
> anova_result <- aov(advertising_data$Sales ~ advertising_data$TV * advertising_data$Radio * advertising_data$Newspaper, data = advertising_data)
> # Check the ANOVA table
> summary(anova_result)
                                                                Df Sum Sq Mean Sq  F value   Pr(>F)
advertising_data$TV                                              1   4512    4512 1834.699  < 2e-16 ***
advertising_data$Radio                                           1    502     502  204.244  < 2e-16 ***
advertising_data$Newspaper                                       1      0       0    0.004    0.951
advertising_data$TV:advertising_data$Radio                       1     64      64   25.924 8.44e-07 ***
advertising_data$TV:advertising_data$Newspaper                   1      2       2    0.649    0.421
advertising_data$Radio:advertising_data$Newspaper                1      1       1    0.514    0.474
advertising_data$TV:advertising_data$Radio:advertising_data$Newspaper  1      2       2    0.959    0.329
Residuals                                                      192    472       2
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> # Check the p-value
> p_value <- summary(anova_result)[[1]]$"Pr(>F)"
> # Compare p-value with significance level to make a decision
> alpha <- 0.05
> if (any(p_value <= alpha)) {
+     cat("There is a significant effect of the TV advertisement Budget or the Radio advertisement Budget, or the Newspaper advertisement Budget or their interaction on sales.")
+ } else {
+     cat("There is no significant effect of the TV advertisement Budget or the Radio advertisement Budget, or the Newspaper advertisement Budget or their interaction on sales.")
+ }
There is a significant effect of the TV advertisement Budget or the Radio advertisement Budget, or the Newspaper advertisement Budget or their interaction on sales.
```

The p-value is smaller than alpha.

Conclusion: Reject H0 in favor of Ha.
There is a significant effect of the TV advertisement Budget or the Radio advertisement Budget, or the Newspaper advertisement Budget or their interaction on sales.

- 10. Applications of Nonparametric Tests

Question 17: By using Wilcoxon Signed-Rank Test Question state whether there is a significant difference in the sales rate before and after changing the method of advertising? (alpha = 0.05)
Solution:

H0: There is no significant difference in the sales before and after the intervention.
Ha: There is a significant difference in the sales before and after the intervention.

p-value < alpha

Conclusion: Reject H0 in favor of Ha.
There is a significant difference in the sales before and after the intervention.

```r
# Applications of Nonparametric Tests

#Question 17: Wilcoxon Signed-Rank Test Question: Is there a significant difference in the type of advertisements before and after a certain intervention?
#install stats library
#install.packages("stats")
# Load the necessary packages
library(stats)
# Perform Wilcoxon signed-rank test
wilcox_result <- wilcox.test(advertising_data$Radio, advertising_data$TV, paired = TRUE)
# Check the test statistic and p-value
test_statistic <- wilcox_result$statistic
p_value <- wilcox_result$p.value
# Compare p-value with significance level to make a decision
alpha <- 0.05
if (p_value <= alpha) {
  cat("There is a significant difference in the sales before and after the intervention.")
} else {
  cat("There is no significant difference in the sales before and after the intervention.")
}
```

572:39 (Top Level)

Console Terminal Background Jobs

R 4.3.0 · ~/Downloads/

```
There is a significant difference in the sales before and after the intervention.
```

Question 18: Single Sample Sign Test: Is there a significant difference in the sample median from the hypothesized value? (sample sales)

Solution:
H0: There is not a significant difference in the sample median from the hypothesized value.
Ha: There is a significant difference in the sample median from the hypothesized value.
Conclusion: Reject H0 in favor of Ha.
There is a significant difference in the sample median from the hypothesized value.

```r
#Question 18: Single Sample Sign Test: Is there a significant difference in the sample median from the hypothesized value?

# Load the necessary packages

library(stats)
# Perform Single Sample Sign Test
sign_test_result <- binom.test(sum(data$Signs == "+"), n = length(SampleSales), p = 0.5, alternative = "two.sided")
# Check the test statistic and p-value
test_statistic <- sign_test_result$statistic
p_value <- sign_test_result$p.value
# Compare p-value with significance level to make a decision
alpha <- 0.05
if (p_value <= alpha) {
  cat("There is a significant difference in the sample median from the hypothesized value.")
} else {
  cat("There is no significant difference in the sample median from the hypothesized value.")
}
```

579:9 (Top Level) R Script

Console Terminal Background Jobs

R 4.3.0 · ~/Downloads/

```
There is a significant difference in the sample median from the hypothesized value.
```

- 11. References and Image Citations

Page 1 image retrieved from:
https://www.vskills.in/certification/blog/techniques-and-importance-advertisement/

The data retrieved from:
https://www.kaggle.com/datasets/ashydv/advertising-dataset

1. "Nonparametric Statistical Methods" by Myles Hollander and Douglas A. Wolfe.
2. "Nonparametric Statistical Inference" by Jean Dickinson Gibbons and Subhabrata Chakraborti.
3. "Introductory Statistics with R" by Peter Dalgaard.
4. "Nonparametric Statistical Methods: Solutions Manual" by Myles Hollander and Douglas A. Wolfe.

http://www.statquest.org/

http://www.datacamp.com/

http://www.r-bloggers.com/

The reference of the questions:
The questions were written by the authors of this report and are inspired by the questions of the lectures.