PROJECT – The Statistical Analysis of Cold Storage Problem

Report:

Prepared by:

Chinedu H Obetta

Table of Content

Chapter 1. The Assignment -1

Problem 1: Cold Storage

Cold Storage started its operations in Jan 2016. They are in the business of storing Pasteurized Fresh Whole or Skimmed Milk, Sweet Cream, Flavored Milk Drinks. To ensure that there is no change of texture, body appearance, separation of fats the optimal temperature to be maintained is between 2 - 4 C.

In the first year of business, they outsourced the plant maintenance work to a professional company with stiff penalty clauses. It was agreed that if it was statistically proven that the probability of temperature going outside the 2 - 4 C during the one-year contract was above 2.5% and less than 5% then the penalty would be 10% of AMC (annual maintenance contract). In case it exceeded 5% then the penalty would be 25% of the AMC fee. The average temperature data at date level is given in the file "Cold_Storage_Temp_Data.csv"

1. Find mean cold storage temperature for Summer, Winter and Rainy Season (3 marks)
2. Find overall mean for the full year (3 marks)
3. Find Standard Deviation for the full year (3 marks)
4. Assume Normal distribution, what is the probability of temperature having fallen below 2 C? (6 marks)
5. Assume Normal distribution, what is the probability of temperature having gone above 4 C? (6 marks)
6. What will be the penalty for the AMC Company? (7 marks)
7. Perform a one-way ANOVA test to determine if there is a significant difference in Cold Storage temperature between rainy, summer and winter seasons and comment on the findings. (9 marks)

Table 1.- Excerpt of the data

```
>head(Cold_Storage_Temp_Data, 10)

##    Season Month Date Temperature
## 1  Winter   Jan    1         2.3
## 2  Winter   Jan    2         2.2
## 3  Winter   Jan    3         2.4
## 4  Winter   Jan    4         2.8
## 5  Winter   Jan    5         2.5
## 6  Winter   Jan    6         2.4
## 7  Winter   Jan    7         2.8
## 8  Winter   Jan    8         3.0
## 9  Winter   Jan    9         2.4
## 10 Winter   Jan   10         2.9
```

Chapter 2. The Solutions

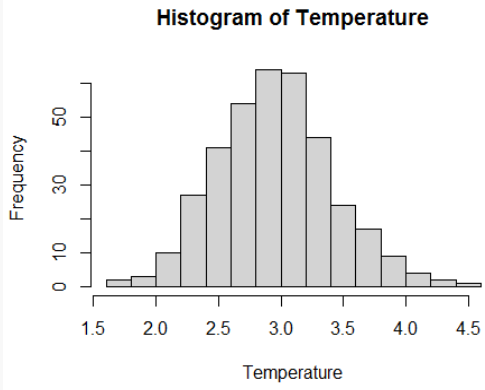
2.1. The Preliminary Analysis

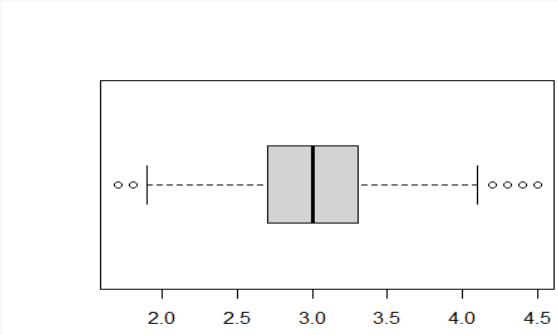
- The name of the dataset is Cold_Storage_Temp_Data and it contains 4 variables and 365 observations, thus, the sample size is 365.

The summary of the data is shown below

```
>summary (Cold_Storage_Temp_Data)

##     Season             Month                Date          Temperature
##  Length:365         Length:365         Min.   : 1.00   Min.   :1.700
##  Class :character   Class :character   1st Qu.: 8.00   1st Qu.:2.700
##  Mode  :character   Mode  :character   Median :16.00   Median :3.000
##                                        Mean   :15.72   Mean   :3.002
##                                        3rd Qu.:23.00   3rd Qu.:3.300
##                                        Max.   :31.00   Max.   :4.500
```

Histogram And Box Plot of the temperature for the year.





Observations:

- The temperature appears to be normally distributed and there is an existence of outliers as shown in the box plot above.
- The minimum and maximum temperature from 365 observations are 1.7c and 4.5c respectively.
- The difference between the values for mean and median is not much as depicted above.

greatlearning
Learning for Life

Question 1.1

1. Find mean cold storage temperature for Summer, Winter and Rainy Season (3 marks)

Using “by” function in “R console”, the mean cold storage temperature for summer, winter and Rainy Season are shown below:

Seasons	Rainy	Summer	Winter
Mean Temp(‘c)	3.087705	3.1475	2.776423

Observation: The Company’s storage temperature seems to be dependent on the season for the year. The mean temperature is highest during the summer season and lowest during the winter season as shown in the table above.

2. Find overall mean and Standard Deviation for the full year:

Solution:

The overall mean temperature and standard deviation for the full year are calculated as follows;

- Overall Mean: 3.002466C
- Overall Standard Deviation: 0.4658319

This means that the average distance between each of the temperature and the mean is 0.466.

```
Overall_mean <- mean(Temperature)
Overall_Std <- sd(Temperature)


Overall_Std              Overall_mean

## [1] 0.4658319        [1] 3.002466
```

3. Assume Normal distribution, what is the probability of temperature having fallen below 2 C? (6 marks)

Solution:

- If the assumption of normal distribution holds, then,
  Temp ~N (mean = Overall Mean, SD= Overall Standard Deviation).
  Recall that Overall Mean = 3.002466 and Overall Standard Deviation=0.4658319.

  Using “R”, the probability that the temperature is below 2C is determined as;
  Prob (tem< 2C) = pnorm (2, mean = Overall_mean, sd = Overall_Std)
  = 15.6%.

4. Assume Normal distribution, what is the probability of temperature having gone above 4 C? (6 marks)

Solution:

- If the assumption of normal distribution holds, then,
  Temp ~N (mean = Overall Mean, SD = Overall Standard Deviation).
  Recall that Overall Mean= 3.002466 and Overall Standard Deviation = 0.4658319.

  Using "R", the probability of that the temperature is above 4C is determined as;
  Prob (tem>4C) = 1-pnorm (4, mean = Overall_mean, sd = Overall_Std)
  = 16.1%.

5. Question: Perform a one-way ANOVA test to determine if there is a significant difference in Cold Storage temperature between rainy, summer and winter seasons and comment on the findings. (9 marks)

Solution: The purpose of the test is to determine whether there is a difference in the Cold Storage temperature between rainy, summer and winter seasons.

Descriptive Analysis of the Cold Storage Data

```
library(readr)
Cold_Storage_Temp_Data<-read.csv("Cold_Storage_Temp_Data.csv", heade
r = TRUE)
dim(Cold_Storage_Temp_Data)

## [1] 365   4

attach(Cold_Storage_Temp_Data)
head(Cold_Storage_Temp_Data, 4)

##   Season Month Date Temperature
## 1 Winter   Jan    1         2.3
## 2 Winter   Jan    2         2.2
## 3 Winter   Jan    3         2.4
## 4 Winter   Jan    4         2.8
```

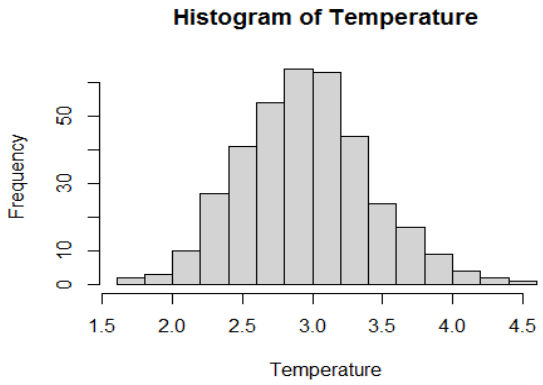
From the table above, "Temperature" is the response Y and "Season" is one factor at different levels.

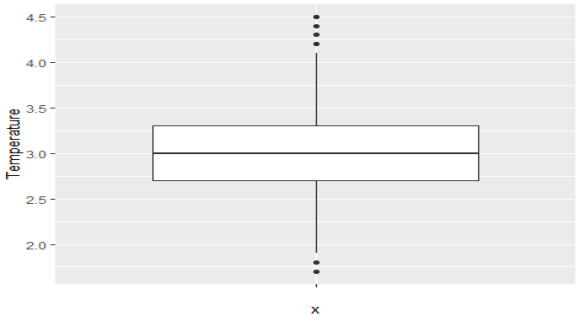
##Next Step: Summary of the response: Temperature

```
summary(Temperature)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.700   2.700   3.000   3.002   3.300   4.500

hist(Temperature)
```

```
library(ggplot2)

ggplot(data=Cold_Storage_Temp_Data, aes(x="", y= Temperature))+
  geom_boxplot()
```

The minimum value of temperature is 1.7C and maximum is 4.5C and the mean value is 3.002C. The pattern of the Cold Storage temperature can be visualized below through the Histogram and the Box Plot below.



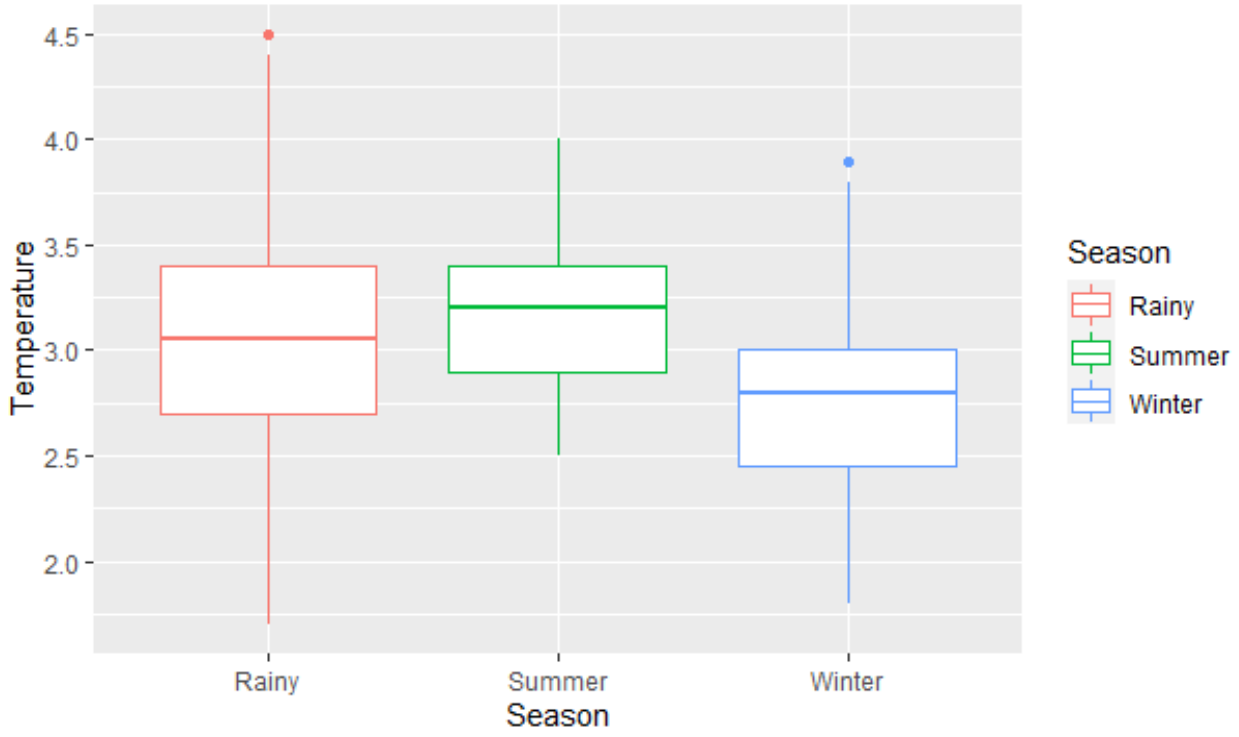


Frequency counts of the Cold Storage Temperature at different season are shown below:

```
> table(Season)
Season
 Rainy Summer Winter
   122    120    123
> round(tapply(Temperature, Season, mean),3)
 Rainy Summer Winter
 3.088  3.148  2.776
```

- Before one-way ANOVA procedure is applied to the data, visual analysis of the levels of factor is recommended. Moreover, the assumption of normality and equality of variance must be checked before we can proceed.

```
ggplot(Cold_Storage_Temp_Data, aes(x = Season, y = Temperature, color=Season)) +
geom_boxplot()
```



Observations: While the expected population mean for Rainy and Summer seasons are close to each other, the population of the factors(seasons) appear to be different from one another.

Test Of Normality for ANOVA

Shapiro-Wilk's test is to be applied to the response.

$H0$: Cold Storage temperature follows a normal distribution

-vs-

Ha: Cold Storage temperature does not follow a normal distribution.

Result of the test of normality

```
##  Shapiro-Wilk normality test
##
## data:  Temperature
## W = 0.99212, p-value = 0.05044
```

Reading of the result above:

- Noting that the p-value of the test is relatively small, we reject the null hypothesis that the Cold Storage temperature follows the normal distribution. This means that Cold Storage temperature does not follow a normal distribution as claimed.

Test of Homogeneity of Variance for ANOVA

- Next, we need to test the assumption that at all seasons, population variance is same. The problem is formulated as follow;

$H0$: $\sigma 1 = \sigma 2 = \sigma 3$ Population variances are same.
Ha: At least the variance of one season is different from others.

```
Levene's Test for Homogeneity of Variance (center = median)
##        Df F value    Pr(>F)
## group   2  8.0084 0.0003951 ***
##       362
```

Reading of the test result above:

- Given that the P-value = 0.003951 is very small, the null hypothesis that the variances are same is rejected, thus, at least the variance for one of the seasons is different from the rest.

Conclusion: Following the failure of the dataset, that is, Cold Storage temperature to comply with the conditions of the ANOVA test statistic, I am unable to establish if there is a significant difference in Cold Storage temperature between rainy, summer and winter seasons. Hence, my test is inconclusive.

Chapter 2. The Assignment 2

Problem 2-Complaints - Cold Storage

In Mar 2018, Cold Storage started getting complaints from their clients that they have been getting complaints from end consumers of the dairy products going sour and often smelling. On getting these complaints, the supervisor pulls out data of the last 35 days' temperatures. As a safety measure, the Supervisor decides to be vigilant to maintain the temperature at 3.9 C or below.

Assume 3.9 C as the upper acceptable value for mean temperature and at alpha = 0.1. Do you feel that there is a need for some corrective action in the Cold Storage Plant or is it that the problem is from the procurement side from where Cold Storage is getting the Dairy Products? The data of the last 35 days is in "Cold_Storage_Mar2018.csv"

1. Which Hypothesis test shall be performed to check if corrective action is needed at the cold storage plant? Justify your answer. (8 marks)
2. State the Hypothesis, perform hypothesis test and determine p-value (11 marks)
3. Give your inference (4 marks)

Table 2.

```
>head(Cold_Storage_Mar2018, 20)

##    Season Month Date Temperature
## 1  Summer   Feb   11         4.0
## 2  Summer   Feb   12         3.9
## 3  Summer   Feb   13         3.9
## 4  Summer   Feb   14         4.0
## 5  Summer   Feb   15         3.8
## 6  Summer   Feb   16         4.0
## 7  Summer   Feb   17         4.1
## 8  Summer   Feb   18         4.0
## 9  Summer   Feb   19         3.8
## 10 Summer   Feb   20         3.9
## 11 Summer   Feb   21         3.9
## 12 Summer   Feb   22         4.6
## 13 Summer   Feb   23         4.1
## 14 Summer   Feb   24         4.1
## 15 Summer   Feb   25         3.9
## 16 Summer   Feb   26         3.8
## 17 Summer   Feb   27         3.8
## 18 Summer   Feb   28         3.9
## 19 Summer   Mar    1         3.9
## 20 Summer   Mar    2         3.9
```

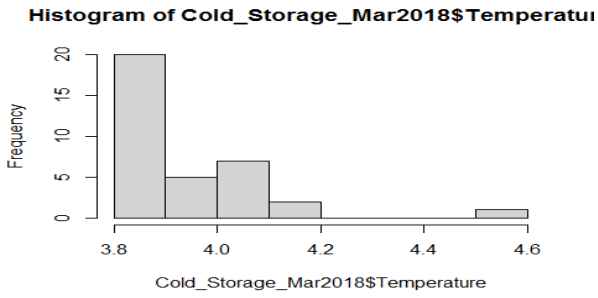
4.1 Preliminary Analysis

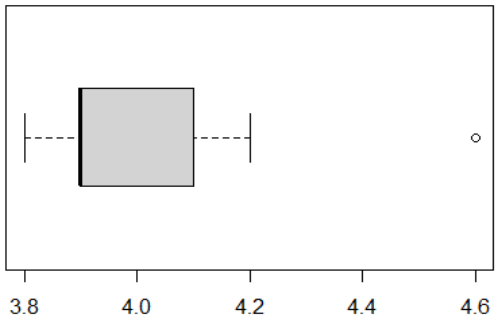
- The name of the dataset is Cold_Storage_Mar2018 and it contain 35 observation and 4 variables. Thus number of sample size is 35.
- The sample was selected daily for the last 35days of the data.

The summary of the data is shown below

```
##     Season             Month                Date       Temperature
##  Length:35          Length:35          Min.   : 1.0   Min.   :3.800
##  Class :character   Class :character   1st Qu.: 9.5   1st Qu.:3.900
##  Mode  :character   Mode  :character   Median :14.0   Median :3.900
##                                        Mean   :14.4   Mean   :3.974
##                                        3rd Qu.:19.5   3rd Qu.:4.100
##                                        Max.   :28.0   Max.   :4.600
```

Histogram And Box Plot of the temperature for the 35 days.





Observations:

- The dataset does not seem to follow normal distributions, it skewed to the right
- The minimum and maximum temperature from 35 days observations are 3.8c and 4.6c respectively.
- The Box Plot shows the presence of potential outlier and this could be the reason for the skewedness.

4.2. Question 2.1

2.1. Which Hypothesis test shall be performed to check if corrective action is needed at the cold storage plant? Justify your answer. (8 marks).

Solution:

- The given level of significance (Alpha) = 0.10.
- The sample size, N = 35, which is appropriately sizable for a Zstat Test.
- However, t- test statistc will be used to ascertain if corrective action is needed at the plant. This is because the population mean and standard deviation are not known.
- Degree of Freedom: Since the sample size is 35, the degree of freedom is 34. (N-1).
- The objective of the Hypothesis test is to check if corrective action is required at the Cold Storage Plant. This can only be achieved if we are able to prove that the temperature at the plant is greater than the upper acceptable value (3.9C). This is a one way test.
- Thus, our hypothesis is as follows;
  Ho: Temperature <= 3.9C.
  H1: Temperature > 3.9C.

4.3. Question 2.2

2.2 State the Hypothesis, perform hypothesis test and determine p-value (11 marks)

- The hypothesis is stated as follows;
  Ho: The Cold Storage Temperature <= 3.9C.
  H1: The Cold Storage Temperature > 3.9C.

  - Since the population mean and standard deviation are unknown, T- test statistic will be used for the hypothesis at 90% confidence interval.

  - **The T Test Result from R are displayed below:**

```
t.test(Cold_Storage_Mar2018$Temperature, alternative = "greater",
mu= 3.9, conf.level = .90)
```

```
##  One Sample t-test
##
## data:  Cold_Storage_Mar2018$Temperature
## t = 2.7524, df = 34, p-value = 0.004711
## alternative hypothesis: true mean is greater than 3.9
## 90 percent confidence interval:
##  3.939011      Inf
## sample estimates:
## mean of x
##  3.974286
```

4.4. Question 2.3

2.3 Give your inference (4 marks)

1. Noting that p-value which is 0.004711 is not greater than the 0.1 level of significance, the Null Hypothesis is rejected, thus the Alternative Hypothesis is accepted.
2. Thus, the dataset shows that we are 90% confidence that cold storage temperature at the factory is significantly greater than the upper acceptable value of 3.9C.
3. In summary, there is need some corrective action in the Cold Storage Plant as the problem seems from the plant instead of where the Diary Products are sourced.

Conclusion:

- While the analysis using test statistic shows that the Null Hypothesis should be rejected, there is a concern on the distribution of the data set.
- The histogram of the temperature shows that the dataset seems deviated from the assumed Normal Distribution.
- Thus, there is need for us to validate the assumption before the test result could be accepted, otherwise, the result will be inconclusive