Case Study: Statistical Methods of Decision Making - Cold Storage

Case Study: The Problem

Problem 1

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 the it was statistically proven that 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 case). 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_Temperature_Data.csv"

a) Find mean cold storage temperature for Summer, Winter and Rainy Season`
b) Find overall mean for the full year
c) Find Standard Deviation for the full year
d) Assume Normal distribution, what is the probability of temperature having fallen below 2 C?
e) Assume Normal distribution, what is the probability of temperature having gone above 4 C?
f) What will be the penalty for the AMC Company?

Problem 2

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 last 35 days' temperatures. As a safety measure, the Supervisor decides to be vigilant to maintain the temperature 3.9 C or below.

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

a) Which Hypothesis test shall be performed to check the if corrective action is needed at the cold storage plant? Justify your answer.
b) State the Hypothesis, perform hypothesis test and determine p-value
c) Give your inference

Solution – Problem 1

We have the following assumptions before starting the study of the population

- ✓ The temperature has been collected for complete one year when the plant maintenance was outsourced. So, we are assuming the **procedures has been stabilized**
- ✓ The temperature values recorded has **no influence on the outside temperature** as the outside temperature could vary over a period of 12 months.

Descriptive Statistics:

Let's import the data into R environment and do preliminary analysis using summary & descriptive statistics.

```
# Install necessary Packages and Invoke libraries

> setwd("Your Directory")


#Import and read the dataset

cold_temp = read.csv("Cold_Storage_Temp_Data.csv", header = TRUE)

> dim(cold_temp)

[1] 365  4

> str(cold_temp)

'data.frame': 365 obs. of 4 variables:
 $ Season     : Factor w/ 3 levels "Rainy","Summer",..: 3 3 3 3 3 3 3 3 3 3 ...
 $ Month      : Factor w/ 12 levels "Apr","Aug","Dec",..: 5 5 5 5 5 5 5 5 5 5 ...
 $ Date       : int 1 2 3 4 5 6 7 8 9 10 ...
 $ Temperature: num 2.3 2.2 2.4 2.8 2.5 2.4 2.8 3 2.4 2.9 ...

> summary(cold_temp)

  Season       Month          Date        Temperature
 Rainy :122   Aug    : 31 Min. : 1.00 Min.  :1.700
 Summer:120   Dec    : 31 1st Qu.: 8.00 1st Qu.:2.700
 Winter:123   Jan    : 31 Median :16.00 Median :3.000
              Jul    : 31 Mean  :15.72  Mean  :3.002
              Mar    : 31 3rd Qu.:23.00 3rd Qu.:3.300
              May    : 31 Max. :31.00 Max.  :4.500
              (Other):179
```

Information about few functions:

a) str() function tells the data type for each column. We have four variables in our dataset out of which Season, Month are of character type whereas Date, Temperature are of numerical type.

b) summary() function provides 5 number summary for continuous data. There are no missing values in the above data. For Date and Temperature mean and median will be fairly close together which means data set has a symmetrical distribution.

c) head() and tail() gives first n and last n rows respectively. Data is consistent.

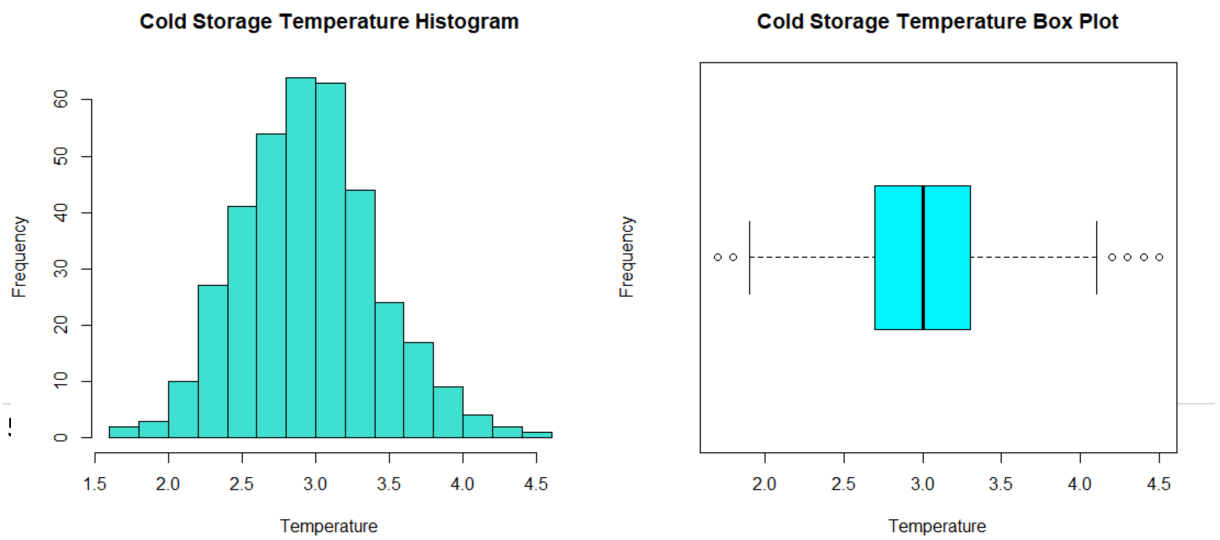
Conclusion

a) Dataset has records for 12 months and these months have been grouped and categorised in 3 groups as Rainy, Summer and Winter.
b) Summary has the minimum temperature as 1 Degree Celsius and maximum temperature as 4.5 Degree Celsius.

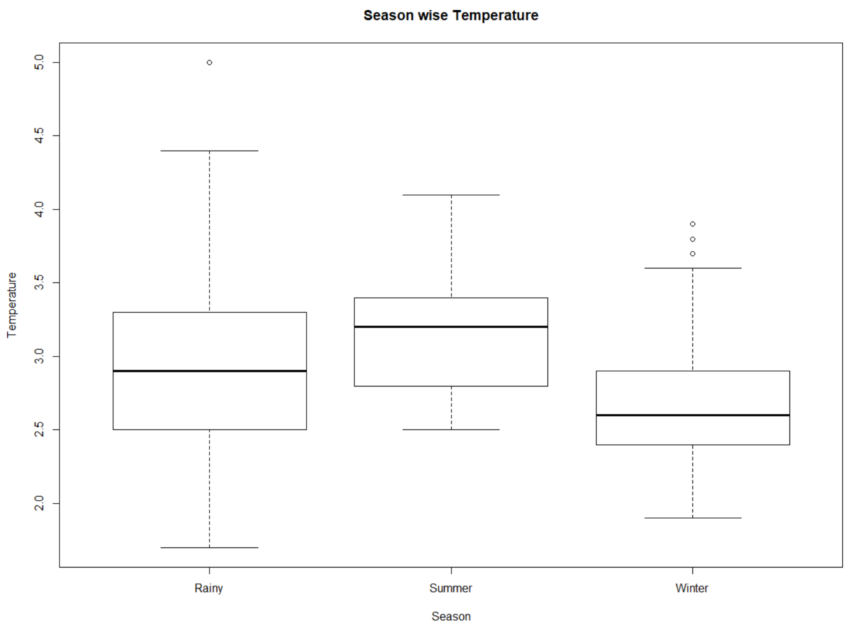
Graphical Analysis:

Box plots and histograms

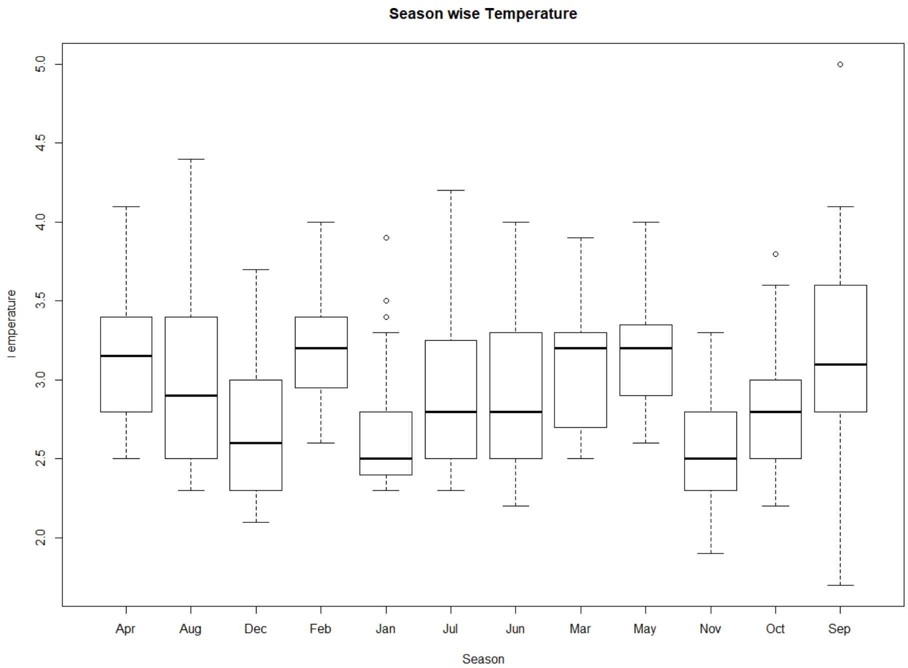
```
> par(mfrow=c(1,2))

> hist(Temperature,main='Cold Storage Temperature Histogram',xlab
= "Temperature",
ylab = "Frequency",col = "turquoise")

> boxplot(Temperature,main='Cold Storage Temperature Box Plot',xlab
= "Temperature",
ylab = "Frequency",col = "turquoise1",horizontal = TRUE)

> dev.off()
```



> boxplot(Temperature ~ Season, main='Cold Storage Temperature across Season')



> boxplot(Temperature~Month,main="Month wise Temperature",xlab = "Month",ylab="Temperature")



> sapply(my_data, function(x) sum(is.na(x)))

```
     Season       Month        Date Temperature
          0           0           0           0
```

Conclusion

Following is the conclusion from the graphical analysis.

a) Data set follows a normal distribution as skewness is not observed.
b) There are possible outliers/extreme values in the temperature
c) There is difference in minimum, maximum, 1st Quartile, 3rd Quartile and Median temperature value when the same was analysed across 3 seasons.
d) We can see that temperature is spread maximum in Sep month. There are few months like Jan, Oct, Sep which have outliers. Others don't have outlier. There is variation in temperature each month.

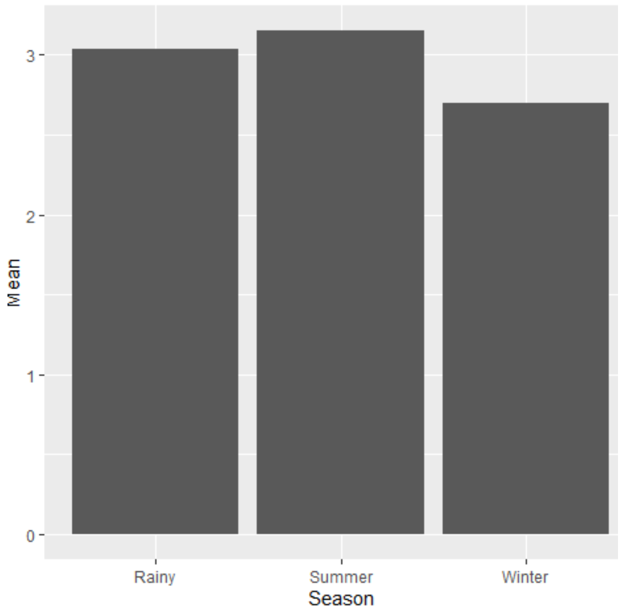
With descriptive statistics it seems, the ***temperature has not gone below 2 Degree and/or above 5 Degree*** However, inferential statistics need to be applied to draw any conclusion on the population.

a) Find mean cold storage temperature for Summer, Winter and Rainy Season

```
> mean_by_season <- my_data %>% group_by(Season) %>%
summarise(Mean=mean(Temperature))
> mean_by_season
```

```
# A tibble: 3 x 2
  Season  Mean
  <chr>  <dbl>
1 Rainy   3.04
2 Summer  3.15
3 Winter  2.70
```

```
> ggplot(mean_by_season,aes(x=Season,y=Mean))+geom_col()
```



b) Find overall mean for the full year

Overall mean is the mean value of all values in the dataset

```
> mean_full_year_temp <- mean(Temperature)
>mean_full_year_temp
```

[1] 2.9627

c) Find Standard Deviation for the full year

```
>sd_temp <- sd(cold_temp$Temperature)
>sd_temp
```

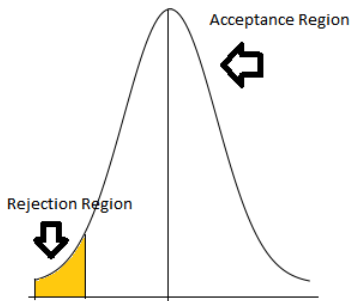
[1] 0.508589

d) Assume Normal distribution, what is the probability of temperature having fallen below 2 C?

As assumption is made that the distribution is Normal, we can use the function pnorm to find the probability. The probability needs to be calculated for values which have been fallen below 2 Degree C, this is a value which is shown in Yellow in the below diagram and hence lower tail is assumed to be true.

We have already calculated the mean and standard deviation value. Substituting these values, we get the following result

```
> prob_less_2 <- pnorm(q = 2,
                       mean = mean_full_year_temp,
                       sd = sd_temp, lower.tail = TRUE )
> prob_less_2
```



[1] 0.02918146

e) Assume Normal distribution, what is the probability of temperature having gone above 4 C?

Now in this problem we have to calculate the value greater than 4 Degree C hence the calculations would be similar to (d) section of the problem with an only difference that lower.tail=FALSE.

```
> prob_greater_4 <- pnorm(q = 4,
                          mean = mean_full_year_temp,
                          sd = sd_temp, lower.tail = FALSE)
> prob_greater_4
```

[1] 0.02070077

f) What will be the penalty for the AMC Company?

Let's understand the penalty condition in terms of probability of temperature going outside the 2 - 4 C:

(i) Above 2.5 % and less than 5 % then the penalty would be 10% of AMC (annual maintenance case).
(ii) In case it exceeded 5 % then the penalty would be 25% of the AMC fee.

So, let's calculate the total probability for condition less than 2 Degree or greater than 4 Degree, i.e. sum total of probability calculated in (d) and (e) part of the problem

```
> Total_Probability <- prob_greater_4+prob_less_2
>Total_Probability

[1] 0.04988223
```

Now use this calculated probability and check against the two conditions mentioned above to get the penalty

```
> if(Total_Probability>0.025 && Total_Probability <= 0.05) {
penalty <- '10%'
}else if(Total_Probability >0.05){
penalty <- '25%'
}else
penalty <- '0%'

penalty

[1] "10%"
```

Solution – Problem 2

We have the following assumptions before starting the study of the sample

a) The sample collection is random and not biased in nature
b) The sample is not affected by outside temperature

Descriptive Statistics:

Let's import the data into R environment and do preliminary analysis using summary & descriptive statistics.

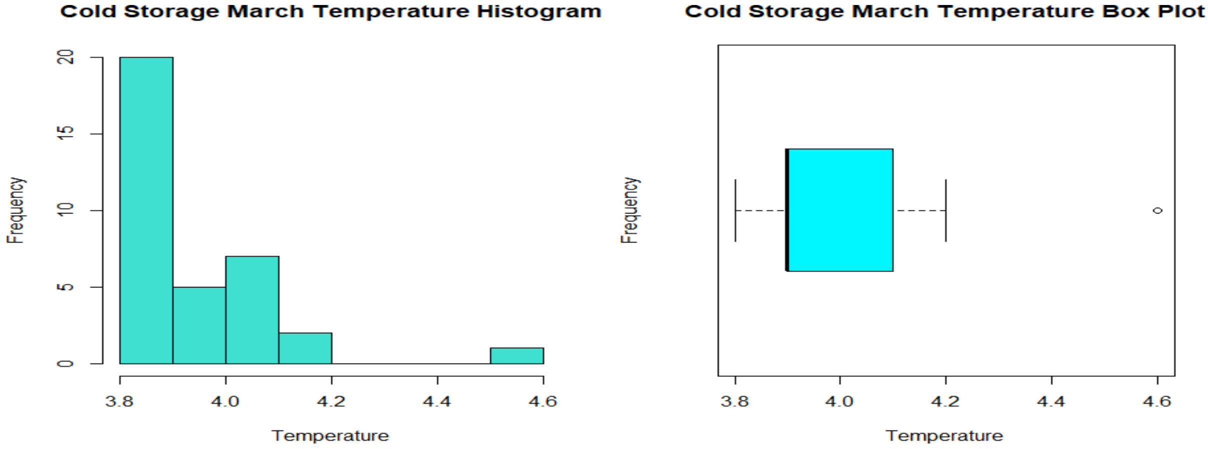
```
> cold_mar_data = read.csv("Cold_Storage_Mar2018.csv", header = TRUE)

> dim(cold_mar_data) [1]

35  4

> str(cold_mar_data)

'data.frame':   35 obs. of 4 variables:
 $ Season     : Factor w/ 1 level "Summer": 1 1 1 1 1 1 1 1 1 1 ...
 $ Month      : Factor w/ 2 levels "Feb","Mar": 1 1 1 1 1 1 1 1 1 1 ...
 $ Date       : int 11 12 13 14 15 16 17 18 19 20 ...
 $ Temperature: num 4 3.9 3.9 4 3.8 4 4.1 4 3.8 3.9 ...

> summary(cold_mar_data)


  Season     Month       Date         Temperature
 Summer:35 Feb:18 Min. : 1.0 Min.  :3.800
           Mar:17 1st Qu.: 9.5 1st Qu.:3.900
                  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
```

Conclusion

a) Sample has 35 records, hence as per central limit theorem, sample could be assumed to be normally distributed
b) Summary has the **minimum** temperature as **3.8** Degree Celsius and **maximum** temperature as **4.6** Degree Celsius.

Graphical Analysis:

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

hist(Temperature,main='Cold Storage March Temperature Histogram',xlab = "Temperature", ylab = "Frequency",col = "turquoise")

boxplot(Temperature,main='Cold Storage March Temperature Box Plot',xlab = "Temperature", ylab = "Frequency",col = "turquoise1",horizontal = TRUE)

dev.off()
```



Conclusion

Following is the conclusion from the graphical analysis.

a) There are possible outliers/extreme values in the sample
b) It seems, the ***temperature has gone above 3.9 Degree.*** However, inferential statistics need to be applied to draw any conclusion on the sample.

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

As the hypothesis testing requires to check if the men temperature of cold storage has gone above 3.9 Degree, hence two different type of tests could be performed

(i) T-test
(ii) Z-test

However, for performing Z test we require the Population Standard Deviation. .Also, Z test assumes that sample size is large enough and can be rightly compared with the population.

If we assume the Standard deviation calculated in Problem (1) to be the population values, Z test can be applied

Hence, we would perform both the test

As we are looking at only one sample, hence one sample test would be performed. Whether this would be a right tailed test or left tailed test, the same would be determined through alternate hypothesis.

b) State the Hypothesis, perform hypothesis test and determine p-value

Let's start the Z test

Hypothesis for Z- Testing

H_0: μ_O <= 3.9	[Temperature is maintained below or equal to 3.9 Degree and hence no corrective action is required.]
H_A: μ_O > 3.9	[Temperature is above 3.9 Degree and hence corrective action is required.]

Where
μ_O is mean of the sample data set

As alternate hypothesis is with a greater sign, it is a **Right tailed** test,

As mentioned in part (a) of Problem (2), standard deviation of the complete dataset calculated in Problem (1) would be used as a Population standard deviation hence **population SD (σ) = 00.508589**

#Z Score

z <- (mean - mu)/(sd_temp/(sqrt(n)))

mean of the sample is

```
> mean <- mean(Temperature) mean
  [1] 3.974286
```

Mu=3.9
n=35

Substituting these values in the above mentioned formula we get a Z score as

0.8641166

Given value of alpha =0.1 P value is

```
pValue = pnorm(z,lower.tail = FALSE)
pValue
```

[1] 0.1937

As P-value is > given alpha, hence we **Fail to Reject the Null Hypothesis**, which indicates that temperature of the cold storage was maintained below 3.9 Degree and hence no corrective action is required.

Let's perform the t test

Hypothesis for Z- Testing

H_0: μ_O <= 3.9	[Temperature is maintained below or equal to 3.9 Degree and hence no corrective action is required.]
H_A: μ_O > 3.9	[Temperature is above 3.9 Degree and hence corrective action is required.]

Where
μ_O is mean of the sample data set

As alternate hypothesis is with a greater sign, it is a **Right tailed** test,

With alpha=0.1, confidence interval is .9, mu=3.9

```
> t.test(Temperature, mu= mu,
     alternative = "greater",
     conf.level = conf)

        One Sample t-test

        data:  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
```

P value is < alpha and hence **Null Hypothesis is rejected** which indicates that temperature of the cold storage has gone above 3.9 and hence corrective action is required.

c) Give your inference

A test statistic is a numerical summary of the data that is compared to what would be expected under the null hypothesis. Test statistics can take on many forms such as the z-tests (usually used for large datasets) or t-tests (usually used when datasets are small).

Even though we have adequate sample size to perform Z test, we don't know the standard deviation of the population. So, Z test is not appropriate here.

Hence the output obtained through t-test would be appropriate and hence corrective action is recommended for cold storage.