**Mini Project - Cold Storage Case Study**

Ashit Debdas

BACP-2020

**Table of Contents**

1. [Project Objective 3](#_bookmark0)
   1. [Problem one 3](#_bookmark3)
   2. [Problem Two 3](#_bookmark3)
2. [Assumptions 4](#_bookmark2)
3. [Exploratory Data Analysis – Step by step approach 4](#_bookmark2)
   1. [Install Necessary Packages and Invoke Library 4](#_bookmark3)
   2. [Set up Working Directory 4](#_bookmark3)
   3. [Import and Read Data Set 4](#_bookmark3)
4. [Variable Identification………………………………………………………………………………………………………….…….4](#_bookmark7)
   1. [Variable Classes………………………………………………………………………………………………………………….4](#_bookmark9)
5. [Missing Value Treatment……………………………………………………………………………………………………….….4](#_bookmark7)
6. [Insights from Problem one ………………………………………………………………………………………………… …….4](#_bookmark7)
   1. [Mean cold storage temperature for Summer, Winter and Rainy Season ……………………………4](#_bookmark9)
   2. [Overall mean for the full year …………………………………………………………………………………………….4](#_bookmark9)
   3. [Standard Deviation for the full year …………………………………………………………………………….….….4](#_bookmark10)
   4. The [Probability of Temperature having fallen Below 2°C……………………………………………………4](#_bookmark9)
   5. [The probability of Temperature having gone Above 4°C …………………………………………………….4](#_bookmark9)
   6. [The Penalty for the AMC Company…………………………………………………………………………………….5](#_bookmark9)
7. [Insights from Problem Two……………………………………………………………………………………………….....…...5](#_bookmark7)
   1. Which Hypothesis Test shall be Performed to check the if corrective action needed…………5
   2. Perform Hypothesis Test and Determine P-Value……………………………………………………………….5
      1. T- Test…………………………………………………………………………………………………………………………5
      2. Z-Test………………………………………………………………………………………………………………………….6
   3. Inference After Two Tests…………………………………………………………………………………………………6
8. [Appendix A – Source code…………………………………………………………………………………………………………6](#_bookmark7)

# Project Objective

# 1.1 Problem one

The objective of the report is to explore in the year of January 2016 Cold Storage Case Study in R and generate insights. To ensure this storing Pasteurized Fresh Whole or Skimmed Milk, Sweet Cream, Flavored Milk Drinks that here is no change of texture, body appearance, separation of fats the optimal temperature to be maintained is between 2 - 4°C. As Per one-year agreement temperature going outside the 2 - 4°C above 2.5% and less than 5% then the penalty would be 10% of AMC and it exceeded 5% then the penalty would be 25% of the AMC fee. The average temperature data is in the file “Cold\_Storage\_Temp\_Data.csv”.

This exploration report will consist of the following:

* + Importing the dataset in R
  + Understanding the structure of dataset
  + Descriptive statistics
  + Find mean cold storage temperature for Summer, Winter and Rainy Season
  + Find out overall Mean, Standard Deviation for the full year
  + Find the probability of temperature having fallen below 2°C, gone above 4°C
  + Find out the penalty for the AMC Company

# Problem Two

# All of sudden In Mar 2018, Cold Storage started getting complaints from their Clients that consumers of the dairy products going sour and often smelling. The supervisor pulls out data of last 35 days’ temperatures to find out insights. Assume 3.9 C as upper acceptable value for mean temperature and at alpha = 0.1 do we need any kind of necessary action in the Cold Storage Plant. The data of the last 35 days is in “Cold\_Storage\_Mar2018.csv”.

This exploration report will consist of the following:

* + Importing the dataset in R
  + Understanding the structure of dataset
  + Hypothesis Test and Justify that
  + Perform hypothesis test and determine p-value
  + Inference

# Assumptions

# Problem one

Given the nature of the data provided in the dataset, it can be seen that this represent to the temperatures in the cold storage Over the entire year and temperature records each day the 365 rows of the dataset correspond to 365 unique days of the year and the temperatures recorded on each day.

To provide more insights in details of season-wise trends

* + The dataset has broken down to Three seasons: Summer, Rainy and winter
  + Summer corresponds to February to May
  + Rainy corresponds to Jun to September
  + Winter corresponds to January & October to December

Problem Two

All of sudden frequently receiving complaints from customers in 2018, the Cold Storage requested data from the maintenance company temperature data for last 35 days. The overall characteristics of this dataset is exactly same as the Problem one, except that the data pulled all correspond to the season of summer.

# Exploratory Data Analysis – Step by step approach

A Typical Data exploration activity consists of the following steup

* 1. Install necessary packages and invoke Library

Before start this section to install necessary packages and invoke associated libraries. Having all the packages at the same places increases code readability more efficiently. Please refer Appendix A for Source Code.

* 1. Set up Working Directory

Setting a working directory on starting of the R session makes importing and exporting data files and code files easier. Basically, working directory is the location/ folder on the PC where you have the data, codes related to the project which makes thinks more sophisticated. Please refer Appendix A for Source Code.

* 1. Import and Read the data

The given dataset is in .csv format. Hence, the command ‘read.csv’ is used for importing the file.

* + Problem One: Cold\_Storage\_Temp\_Data.csv
  + Problem Two: Cold\_Storage\_Mar2018.csv

Appendix A for Source Code

# Variable Identification

The dataset is analyzed for basic understanding of the features and data contained. it is usually an activity by which data is explored and organized.

* 1. Variable classes

# Problem one has 365 Rows and 4 columns

# Problem Two has 35 Rows and 4 columns

# Missing Value Treatment

# Missing value treatment is an important step in Exploratory Data Analysis, essentially missing data in the training data set can reduce the power of a model or can lead to a biased model because we have not analyzed the behavior and relationship with other variables correctly. It can lead to wrong prediction or classification. The datasets under scrutiny does not have any Missing values as we have already observed in the data summary

# Insights from Problem one

* 1. Mean cold storage temperate for summer, winter and Rainy season

As per my analysis Mean of summer season is 3.15, Winter season is 2.70 and Rainy season is 3.04. From the dataset we can observed that the highest average temperature is being clocked for the season of Summer whereas the lowest is for Winter. Although once can assume this to match the natural ambient temperatures of different seasons, statistically we cannot draw a conclusion due to the lack of weather and variability of temperature data across the year.

* 1. Overall mean for full year

Mean = 2.96274

* 1. Standard division for full year

Standard division = 0.508589

* 1. The probability of temperature having fallen below 2°C

Mean = 2.96(rounded), although it ideally should be populated in a to find out the probability value for Temp's less than 2, we have to find out Mean = 0.5086(rounded). variable while doing the calculations in R Using the pnorm function in R, the result is = 0.02918146 or 2.918146% probability.

* 1. The probability of temperature having gone above 4°C

Similarly, by using Pnorm function in R, we have to use the lower tail = FALSE, because its at the right half of graph. But I took same result directly by using pnorm function However, our calculate result is 0.02070077 or 2.070077%.

* 1. The Penalty for the AMC Company

There are two mutually exclusive occurred like "lower than 2" and then 4" at the same time, therefore probability would be P (AU B) = P(A) + P(B), by using R function we found total probability “P = P(Temp<2) + P(Temp>4)” = 0.04988223 or 4.988223% Therefore, penalty = 10% of AMC, since the probability of temperature out of the range of 2 - 4°C falls between the 2.5% and 5% bound mentioned on the problem statement.

# Insights from Problem Two

The Problem one and Two contains the temperatures throughout the year of 2016, the temperatures from a sample of 35 days that was procured on back of customer complaints from this, we can infer, the Problem one is not a sample of the Problem two. Therefore, we cannot get the population mean/standard deviation. As the problem two is independent, and we have not been provided with the population data for the same, cannot deduce the Population standard deviation. Therefore, we are assuming the population mean and standard deviation to be same as Sample mean Mu and Sample Standard deviation Sigma, thus assuming our Sample estimation will be reflective of the reality/population sampling

* 1. Which Hypothesis test shall be performed to check the if corrective action is needed?

The Problem one and two contains the temperatures throughout the year of 2016, the temperatures from a sample of 35 days that was procured on back of customer complaints from this, we can infer, the Problem one is not a sample of the Problem two. Therefore, we cannot get the population mean/standard deviation. As the problem two is independent, and we have not been provided with the population data for the same, cannot deduce the Population standard deviation.

Therefore, we are assuming the population mean and standard deviation to be same as Sample mean Mu and Sample Standard deviation Sigma, thus assuming our Sample estimation will be reflective of the reality/population sampling

Since the population standard deviation is unknown, the best Statistic test is student’s T-test then we will perform the Z test as well, and compare the outcomes. Finally, perform hypothesis test and determine p-value to conclude the both test result.

* 1. Perform Hypothesis Test and Determine P-Value

The supervisor has been tasked with maintaining the temperature at the cold storage to below 3.9°C this will be the Null Hypothesis Null Hypothesis

**Null Hypothesis, Ho: Mu <= 3.9**

On the other hand, since there have been complaints of product being degraded because the temperature is exceeding the upper higher acceptable limit of 3.9°C, this will be our Alternative Hypothesis.

**Alternative Hypothesis, Ha: Mu > 3.9**

* + 1. T-Test

State the Hypotheses: Ho: Mu <= 3.9 & Ha: Mu> 3.9 , Population mean(assumed), Mu = 3.9 Significance(given), alpha = 0.10, Hence conf.level = 0.90 ,Sample mean, s.mean (x) = 3.974286 , Sample standard deviation = s.sd = 0.159674, Sample size= 35,Degrees of freedom = df = (sample size - 1) = 34.

One Sample t-test

data: x

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

Even we can calculate in R by using pvalue= pt(z.test,df,lower.tail = FALSE), Therefore, we will get result pvalue = 0.004711198

After One sample **T-Test P value < alpha,** the Null Hypothesis is rejected, and Alternative Hypothesis is accepted, thus statistically concluding (via T Test) that the Temperature in the Cold Storage is greater than 3.9 C with 90% confidence (1 - 0.1), thus causing the products go sour or smelling, We will find the actual confidence by subtracting the Pvalue from 1 Actual Confidence = (1-Pvalue) \* 100 = 99.52888%

* + 1. Z-Test

The assumptions and calculation methods are same as already mentioned/ performed in the T-Test. As significance, alpha = 0.1, the critical values of Z.critical are +1.281552 We are using a probability value of 0.9 instead of 0.1 because MS Excel considers it cumulatively and it is a right tailed Test. =NORM.S.INV(0.9) 1.281552 The rejection region is Z < -1.28 or Z > +1.28 The non-rejection region is -1.28 <= Z <= 1.28

Since z-statistic value = 2.75 is greater than z-critical value of 1.28 at significance level alpha =0.1 Hence we Reject the null hypothesis We accept the Alternate hypothesis that temperature has indeed risen more than 3.9°C

* 1. Inference After Two Tests

Lastly, we have seen from the problem one that holds the values from the year 2016, the average temperature throughout the year is 2.96. However, as months went by, the working quality of the Cold Storage seems to have degraded. And from the samples taken in 2018, without even putting it through any Statistical analyses, we can see a mean temperature to be 3.97, which is 1 degree higher and going by the Working principle of Cold storages not good sight. However, we reserved our judgement before doing a root cause analysis through Statistical analysis and concluding the result. To find the actual cause we have done both T-statistic and Z-statistic tests, which gave us the following results

Hypothesis procedures both T and Z statistic tests, we have confirmed rejecting the Null Hypothesis of mean Temp <3.9°C With 90% confidence. We can conclude the temperature indeed crossed the with 99.53% actual confidence. other way, both tests proved that Temperature had been well over 3.9°C against the claims of Supervisor who had been vigilant about not breaching the 3.9°C, **t-statistic = z-statistic = 2.75** another striking similarity to reject the Null hypothesis. Hence, we can conclude the above statement

# Appendix A – Source Code

*setwd("F:/Data Science/BCAP/2.Fundamental of Business Statistic/Statistics and R/Week 3.Estimation and Hypothesis Testing/Project-1/")*

*getwd()*

*problem1= read.csv("Cold\_Storage\_Temp\_Data.csv", header = TRUE)*

*#library*

*library(ggplot2)*

*library(dplyr)*

*library(lattice)*

*#general oversvation*

*attach(problem1)*

*summary(problem1)*

*str(problem1)*

*anyNA(problem1)*

*anyDuplicated(problem1)*

*dim(problem1)*

*head(problem1)*

*tail(problem1)*

# 

*by(problem1,INDICES = Season,FUN = summary)*

*#2.Find overall mean for the full year*

*mean\_year=mean(problem1$Temperature)*

*print(mean\_year)*

*#3Find Standard Deviation for the full year*

*sd\_year=sd(problem1$Temperature)*

*print(sd\_year)*

*#the probability of temperature having fallen below 2 C?*

*temp\_below\_2C= pnorm(2, mean = mean\_year, sd = sd\_year)*

*print(temp\_below\_2C)*

*#the probability of temperature having gone above 4 C?*

*temp\_above\_4C= 1-pnorm(4, mean = mean\_year, sd = sd\_year)*

*print(temp\_above\_4C)*

*## Combined Probability for <2 and >4*

*TotalProb = temp\_above\_4C+temp\_below\_2C*

*print(TotalProb)*

*####the penalty for the AMC Company*

*if(TotalProb > 0.025 && TotalProb <= 0.05){ penalty = '10%'} else if(TotalProb > 0.5){penalty = '25%' }else penalty = '0%'*

*print(penalty)*

*#problem 2*

*problem2 =read.csv("Cold\_Storage\_Mar2018.CSV", header = TRUE)*

*summary(problem2)*

*str(problem2)*

*anyNA(problem2)*

*anyDuplicated(problem2)*

*dim(problem2)*

*head(problem2)*

*tail(problem2)*

*problem2$Season=as.character(problem2$Season)*

*problem2$Month=as.character(problem2$Month)*

*# Density plot for sample distribution of Temperatures*

*ggplot(data=problem2, aes(Temperature)) + geom\_density() + geom\_vline(xintercept = 3.9, col="blue")*

*##Z test*

*## sample mean*

*s.mean = mean(problem2$Temperature)*

*print(s.mean)*

*#sample Sd*

*s.sd= sd(problem2$Temperature)*

*print(s.sd)*

*##mu*

*mu=3.9*

*#sample*

*sample.size=35*

*##z test*

*z.test=(s.mean-mu)/(s.sd/(sqrt(sample.size)))*

*print(z.test)*

*#acceptable value for mean temperature and at alpha = 0.1*

*df=34*

*z.critical=qnorm(1-0.1)*

*print(z.critical)*

*z.test>=z.critical*

*pvalue= pt(z.test,df,lower.tail = FALSE)*

*print(pvalue)*

*Actualt.confint=(1-pvalue)\*100*

*print(Actualt.confint)*

*# t.test*

*x= problem2$Temperature*

*t.test(x,mu= 3.9,conf.level = 0.90,alternative = "greater")*

*THE END*