Great Learning
POWER AHEAD

Statistical method for Decision Making

-PRAJOTH
DSBA-2021

Table of Contents:

List of Tables:

List of Figures:

Problem 1:

Problem Statement:

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 2º - 4º 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 2º - 4º 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

Checking the head and tail of Dataset:

	Season	Month	Date	Temperature
0	Winter	Jan	1	2.3
1	Winter	Jan	2	2.2
2	Winter	Jan	3	2.4
3	Winter	Jan	4	2.8
4	Winter	Jan	5	2.5

	Season	Month	Date	Temperature
360	Winter	Dec	27	2.7
361	Winter	Dec	28	2.3
362	Winter	Dec	29	2.6
363	Winter	Dec	30	2.3
364	Winter	Dec	31	2.9

Table 1: Head and Tail of Dataset

Inference:

- Dataset has 365 rows and 4 columns.
- Dataset has both numerical and categorical variables.
- Numerical variables are date and Temperature.
- Categorical variable has Month and Season.
- There is no null value in dataset.
- There is no duplicate value in dataset.

Summary of Dataset:

	Temperature
count	365.00
mean	3.00
std	0.47
min	1.70
25%	2.70
50%	3.00
75%	3.30
max	4.50

Table 2: Summary of Dataset

Inference:

- Total count of days is 365.
- Mean of temperature is 3.00 and Standard deviation is 0.47.
- Mean, Median and Mode are equal so it seems to normal distribution.
- Minimum Temperature is 1.7 degree Celsius and maximum temperature is 4.50 degree Celsius

Univariate Analysis:

- Univariate analysis refers to the analysis of one variable. The purpose of univariate analysis is to understand the **distribution of values for a single variable**.

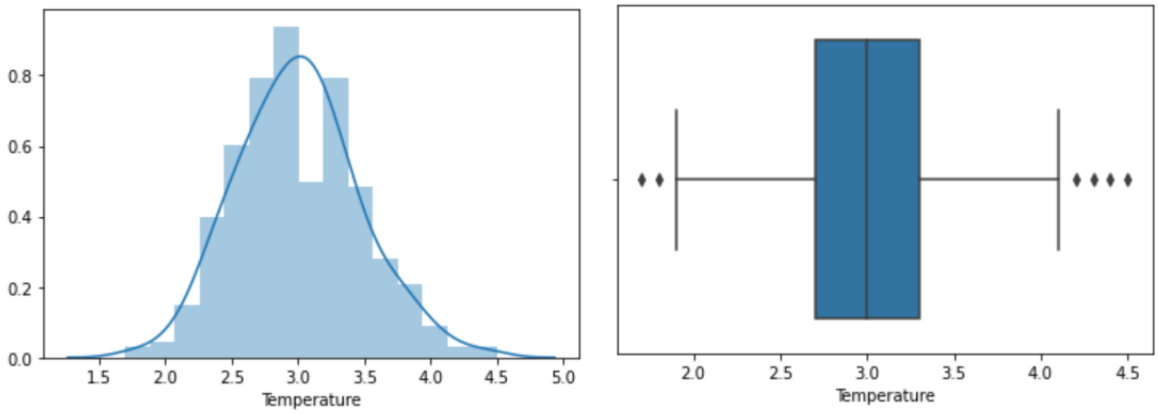

Figure 1: Univariate Analysis (Temperature)

Inference:

- Mean, Median and Mode are equal so it seems to normal distribution.
- Outliers are present in the data.

Bivariate Analysis:

- Bivariate analysis is **one of the simplest forms of quantitative (statistical) analysis**. It involves the **analysis of two variables** (often denoted as X, Y), for the purpose of determining the empirical relationship between them.

Temperature VS Month:

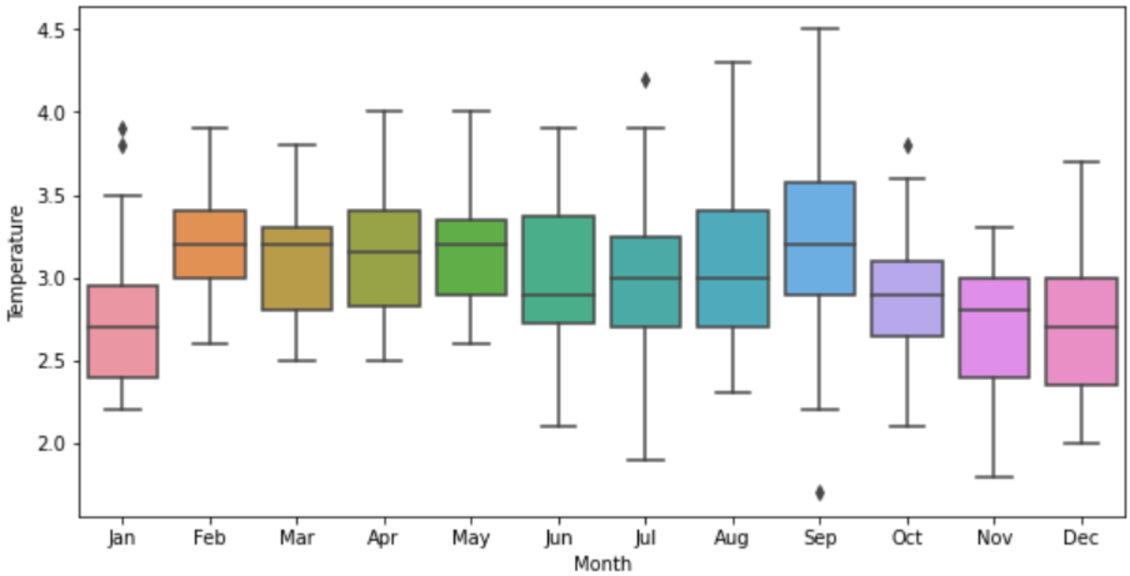

Figure 2: Bivariate Analysis (Temperature VS Month)

Inference:

- Jan, July, September and October has outliers.
- Sep, Nov, Jun, Jul, Jan, Dec, Aug are positively Skewed.

- Mar, May, Feb, Apr, Oct are negatively skewed.
- Normal Distribution is not observed in any month.

Temperature VS Season:

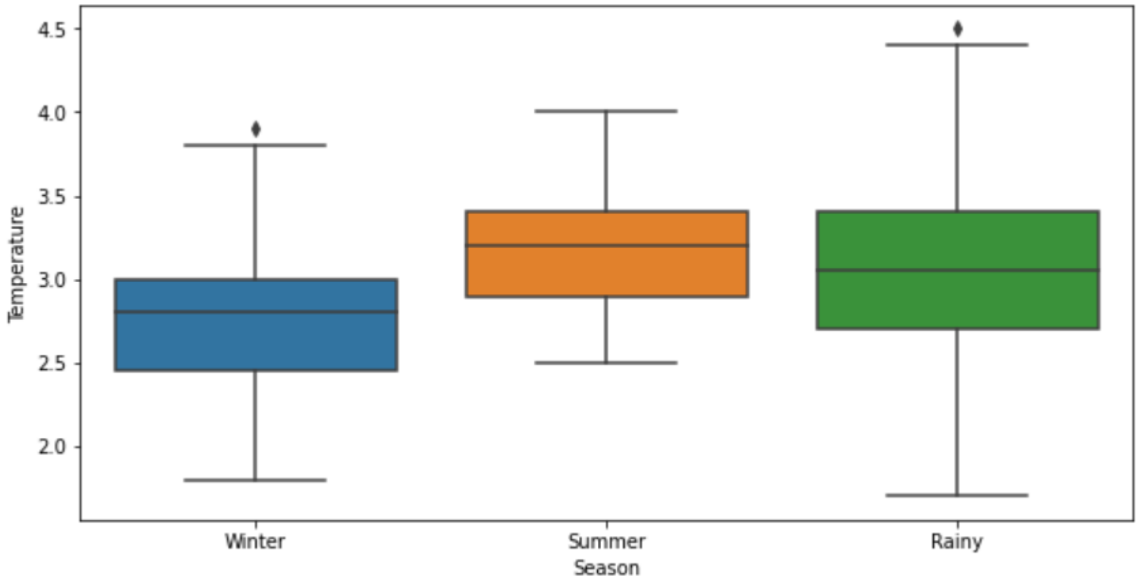

Figure 3: Bivariate Analysis (Temperature VS Season)

Inference:

- 1 Outliers in the winter season, 1 in Rainy, whereas no Outliers in summer.
- Rainy season has temperature higher than 4 Degree Celsius and lower than 2 Degree Celsius.
- Winter season has recorded temperature lower than 2 degree Celsius.

Month VS Temperature with Season:

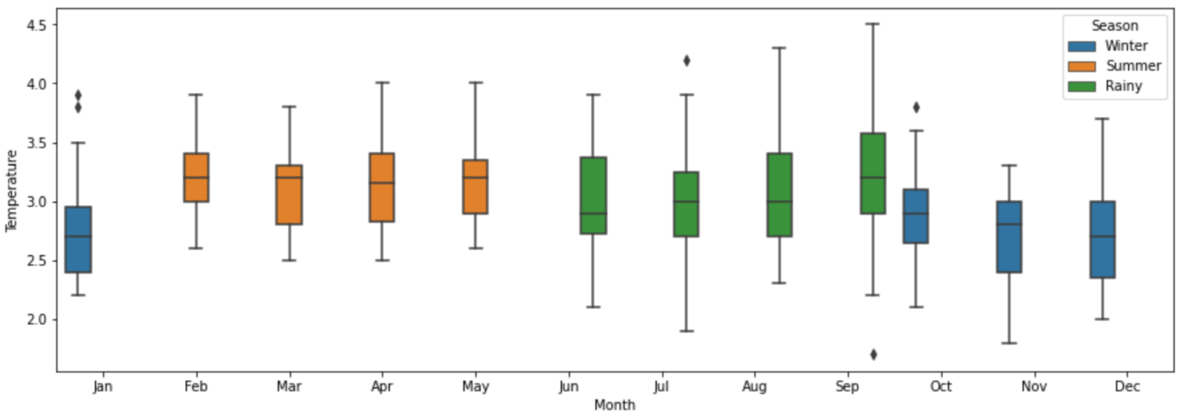

Figure 4: Bivariate Analysis (Temperature VS Month with season)

Inference:

- Summer corresponds to the months of Feb to May.
- Rainy corresponds to June to September.
- Winter corresponds to Jan & Oct to Dec.
- Going by the nature of these 2 variables, which being straight-forward,nothing substantial can be inferred statistically.

1.1 Find mean cold storage temperature for Summer, Winter, and Rainy Season.

Season	Mean
Rainy	3.09
Summer	3.15
Winter	2.78

Inference:

- From the dataset under analysis, it can be 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 data across the year.

1.2 Find the overall mean for the full year

Overall Mean for full Year	3.002

1.2 Find Standard Deviation for the full year.

Overall Standard Deviation of Year	0.465

Normal Distribution:

- Normal distribution, also known as the Gaussian distribution, is a probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean. In graph form, normal distribution will appear as a bell curve.

Importance of Normal Distribution:

- A normal distribution is the proper term for a probability bell curve.
- In a normal distribution the mean is zero and the standard deviation is 1. It has zero skew and a kurtosis of 3.
- Normal distributions are symmetrical, but not all symmetrical distributions are normal.
- In reality, most pricing distributions are not perfectly normal.

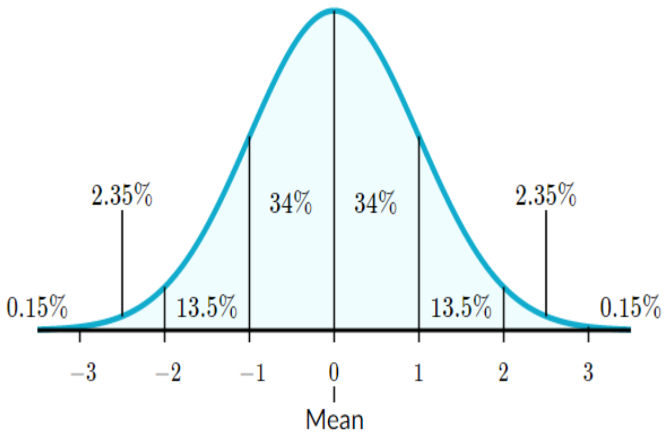

Figure 5: Normal Distribution

Features of Normal distributions:

68% (approximately), of the data falls within 1 standard deviation of the mean.
95% (approximately), of the data falls within 2 standard deviations of the mean.
99.7% (approximately), of the data falls within 3 standard deviations of the mean.

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

probability of temperature having fallen below 2 C	1.558

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

probability of temperature having gone above 4 C	1.592

1.6 What will be the penalty for the AMC Company

- A particular Temperature can never attain a value lower than 2 and higher than 4 at the same time, therefore these 2 are mutually exclusive events, thus P (A U B) = P(A) + P(B)Therefore,

P = P(Temp<2) + P(Temp>4) =3.15%.

Probability of Temperature	3.15

- **Therefore penalty = 10% of AMC, Since the probability of Temperature going outside of the range of 2–4 degree Celsius falls between the 2.5% and 5% boundary as mentioned on the problem statement.**

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 the last 35 days' temperatures. As a safety measure, the Supervisor has been vigilant to maintain the mean temperature 3.9º C or below.

Assume 3.9º C as the upper acceptable 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.

Checking the head and tail of Dataset:

	Season	Month	Date	Temperature
0	Summer	Feb	11	4.0
1	Summer	Feb	12	3.9
2	Summer	Feb	13	3.9
3	Summer	Feb	14	4.0
4	Summer	Feb	15	3.8

	Season	Month	Date	Temperature
30	Summer	Mar	13	4.2
31	Summer	Mar	14	4.2
32	Summer	Mar	15	3.8
33	Summer	Mar	16	3.9
34	Summer	Mar	17	3.9

Table 3: Head and Tail of Dataset

Inference:

- Dataset has 35 rows and 4 columns.
- Dataset has both numerical and categorical variables.

- Numerical variables are date and Temperature.
- Categorical variable has Month and Season.
- There is no null value in dataset.
- There is no duplicate values in dataset

Summary of Dataset:

	Temperature
count	35.00
mean	3.97
std	0.16
min	3.80
25%	3.90
50%	3.90
75%	4.10
max	4.60

Table 4: Summary of Dataset

Inference:

- Minimum temperature recorded is 3.8 degree Celsius and Minimum temperature recorded is 4.60 degree Celsius.
- Only summer season is taken for analysis.

Univariate Analysis:

- Univariate analysis refers to the analysis of one variable. The purpose of univariate analysis is to understand the **distribution of values for a single variable**.

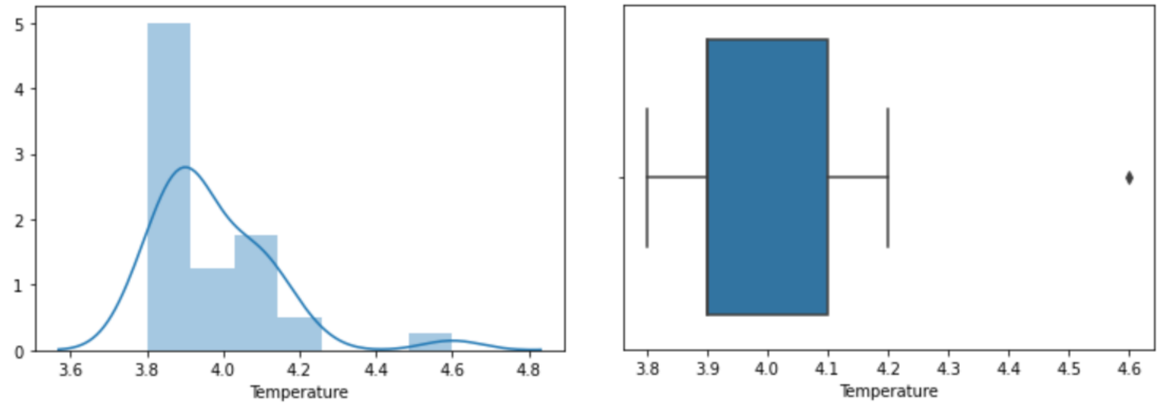

Figure 5: Univariate Analysis (Temperature)

Inference:

- Most of the temperature is distributed between in 3.8 to 4.2 degree Celsius.
- From the Temperature box plot, we can observe a very heavy Positive skewness in the data distribution.
- A single outlier seen that is quite off from the Max value.

Bivariate Analysis:

- Bivariate analysis is **one of the simplest forms of quantitative (statistical) analysis**. It involves the **analysis of two variables** (often denoted as X, Y), for the purpose of determining the empirical relationship between them.

Temperature VS Month:

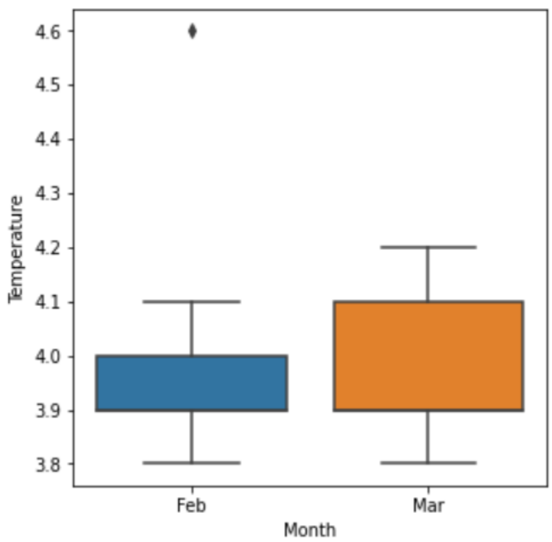

Figure 6: Bivariate Analysis (Temperature VS Month)

Temperature VS Season:

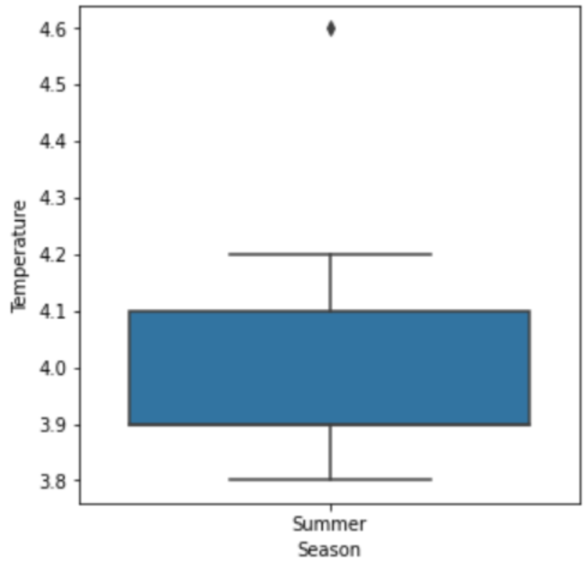

Figure 7: Bivariate Analysis (Temperature VS Month)

Month VS Temperature with Season:

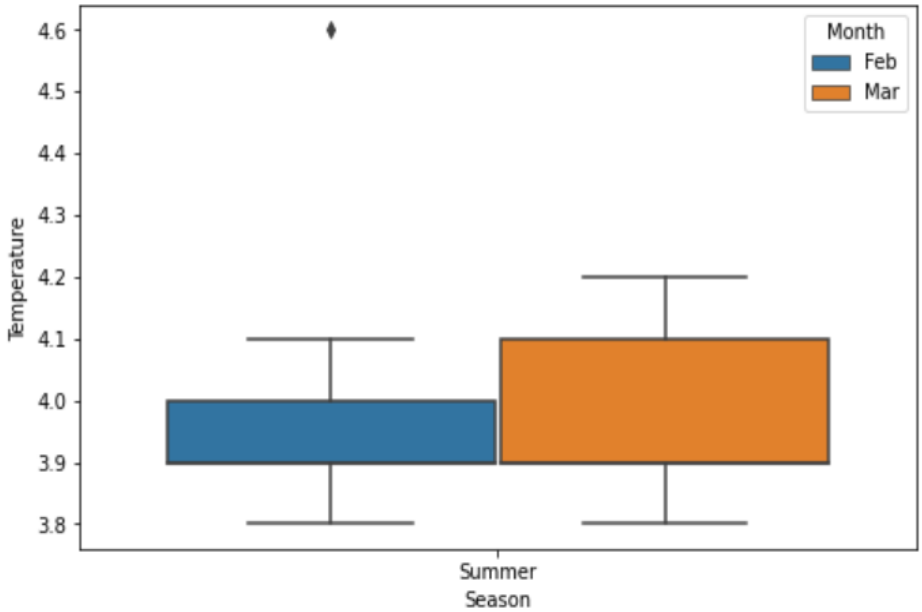

Figure 8: Bivariate Analysis (Temperature VS Month with Season)

Date VS Temperature with Month:

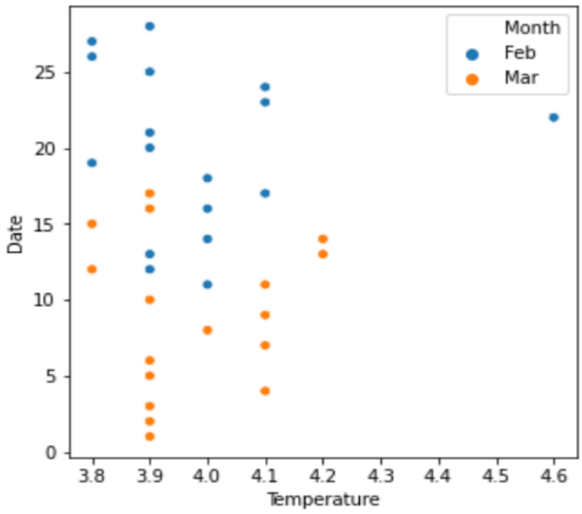

Figure 9: Bivariate Analysis (Date VS Temperature with Month)

Inference:

- Feb Month has recorded the highest temperature of 4.6.
- Temperature recorded in March is higher than the month Feb.

Hypothesis Testing:

- Hypothesis testing is the use of statistics to determine the probability that a given hypothesis is true.
- There are two types of statistical hypotheses.

Null hypothesis: The null hypothesis, denoted by H0, is usually the hypothesis that sample observations result purely from chance.

Alternative hypothesis: The alternative hypothesis, denoted by H1 or Ha, is the hypothesis that sample observations are influenced by some non-random cause.

Steps to perform Hypothesis Testing:

Step 1: Formulate the null hypothesis H0(commonly, that the observations are the result of pure chance) and the alternative hypothesis Ha(commonly, that the observations show a real effect combined with a component of chance variation).

Step 2: Identify a test statistic that can be used to assess the truth of the null hypothesis.

Step 3: Compute the P-value, which is the probability that a test statistic at least as significant as the one observed would be obtained assuming that the null hypothesis were true. The smaller the P-value, then stronger the evidence against the null hypothesis.

Step 4: Compare the P-value to an acceptable significance value (sometimes called an alpha value). If P<=Alpha, that the observed effect is statistically significant, the null hypothesis is ruled out, and the alternative hypothesis is valid.

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

- In a Hypothesis test scenario, if the sample mean is largely representative of the population mean, then we use a One-tailed T-Test.
- When the testing is done to show that the sample mean would be higher or lower than the population mean, it is referred to as a one-tailed test. (The basis assumption of Normal distribution of data however is necessary).
- For determining this, we require a Null & Alternate Hypothesis to be determined before carrying out such test. In this case, the supervisor needs to understand if the mean temperature of last 35 days falls below 3.9 C or not, hence in this case, a one tailed t-test would be the best suited, since the test is largely towards one side of the data or towards single tail.

2.2 State the Hypothesis and do the necessary calculations to accept or reject the corresponding null hypothesis

- We now have the data of 35 days temperatures of dairy products. Supervisor is setting an upper acceptable value of 3.9 C and alpha = 0.1, hence the confidence interval in such case is 90% or 0.90
- We now state the Null & Alternate Hypothesis for the case in question.

  Null Hypothesis, H0: Mu<=3.9 C
  Alternate Hypothesis, Ha: Mu>3.9 C

- In this case, we will reject the Null Hypothesis if the mean temperature is found to be above 3.9 C and the supervisor would then initiate an inquiry into the procurement side of the supply chain. At a confidence interval of 90%, we would perform the T-Test.
- Since 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 P-value from 1.

Actual Confidence = (1 – P-value) * 100 =99.52888%

It is proved with a statistical significance that the mean temperature is more than the expected 3.9 C and hence, a corrective action is needed at the Cold Storage plant.

2.3 Give your inference

- In other words, given the facts of the case, it is proved with a statistical significance that the mean temperature is more than the expected 3.9 C and hence, a corrective action is needed at the Cold Storage plant.

- Hence corrective measures are required by cold storage owner against the contracted company to avoid any such failures in future because such failures can lead in loss of customer on large scale.

- Statistically we can conclude that there needs to be corrective measures taken to keep the Cold Storage function properly, and there is no apparent problem (statistically speaking) from procurement side from where Cold Storage is getting the Dairy Products.

- We will submit the results to the owner of the Cold Storage and they need to figure out the resolution path, is it the lackadaisical approach in work by the Supervisor or some inherent problems with the machines being used. This we cannot conclude statistically due to lack of necessary data.

- Also, as we have seen earlier, there is ‘almost’ a 5% probability of the temperatures to be outside of the permissible range of 2-4 degree Celsius, thus attracting a hefty fine of 10% AMC, and until immediate necessary measures are taken, it could cross the 5 degree Celsius mark and attract even a heftier fine of 25% AMC.