**WHOLE SALE**

**CUSTOMER ANALYSIS**

**REPORT**

**Gaurav Sood**

**PGP-DSBA Online July 2021**

**Date: 12th September, 2021**

Table Of Contents

1. Problem Statement: ............................................................................................................4

1.1. Use methods of descriptive statistics to summarize data. Which Region and which Channel spent the most? Which Region and which Channel spent the least?........................................................................................................................................5

1.2. There are 6 different varieties of items that are considered. Describe and comment/explain all the varieties across Region and Channel? Provide a detailed justification for your answer......................................................................................................7

1.3. On the basis of the descriptive measure of variability, which item shows the most inconsistent behaviour? Which items shows the least inconsistent behaviour?...............................................................................................................................9

1.4. Are there any outliers in the data? Back up your answer with a suitable plot/technique with the help of detailed comments.................................................................................................................................9

1.5. On the basis of your analysis, what are your recommendations for the business? How can your analysis help the business to solve its problem? Answer from the business perspective.............................................................................................................................10

2. Problem Statement: ..........................................................................................................11

2.1. For this data, construct the following contingency tables (Keep Gender as row variable).................................................................................................................................12

2.2. Assume that the sample is a representative of the population of CMSU. Based on the data, answer the following questions:...............................................................................................................................13

2.3. Based on the above probabilities, do you think that the column variable in each case is independent of Gender? Justify your comment in each case........................................................................................................................................13

2.4. Assume that the sample is a representative of the population of CMSU. Based on the data, answer the following question:………………………………………………………………14

2.5. Assume that the sample is representative of the population of CMSU. Based on the data, answer the following question:………………………………………………………………15

2.6.  Construct a contingency table of Gender and Intent to Graduate at 2 levels (Yes/No). The Undecided students are not considered now and the table is a 2x2 table. Do you think the graduate intention and being female are independent events?.......................................15

2.7. Note that there are four numerical (continuous) variables in the data set, GPA, Salary, Spending, and Text Messages.

Answer the following questions based on the data…………………………………………......16

2.8. Note that there are four numerical (continuous) variables in the data set, GPA, Salary, Spending, and Text Messages. For each of them comment whether they follow a normal distribution. Write a note summarizing your conclusions………………………………………17

3. Problem Statement:…........................................................................................................20

3.1 Do you think that the population means for shingles A and B are equal? Form the hypothesis and conduct the test of the hypothesis. What assumption do you need to check before the test for equality of means is performed?.............................................................................................................................21

3.2 What assumption about the population distribution is needed in order to conduct the hypothesis tests above?...................................................................................................................................23

Introduction

The purpose of this whole exercise is to explore the dataset. Do the exploratory data analysis. Explore the dataset using central tendency and other parameters. This assignment should help the different section of people in exploring the summary statistics, contingency tables, conditional probabilities & hypothesis testing.

Problem 1

A wholesale distributor operating in different regions of Portugal has information on annual spending of several items in their stores across different regions and channels. The data (Wholesale Customer.csv) consists of 440 large retailers’ annual spending on 6 different varieties of products in 3 different regions (Lisbon, Oporto, Other) and across different sales channel (Hotel, Retail).

Data Description

1. Buyer/Spender: Index number of the buyer or spender.

2. Channel: Whether the product is purchased from a Hotel or a Retailer.

3. Region: Location the distributor is operating from (Lisbon, Oporto or Other).

4. Fresh: Annual spending on Fresh products.

5. Milk: Annual spending on Milk products.

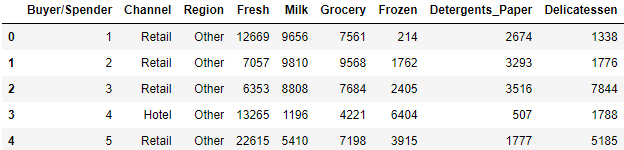
6. Grocery: Annual spending on Grocery products.

7. Frozen: Annual spending on Frozen products.

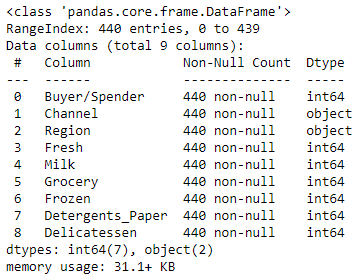
8. Detergents\_Paper: Annual spending on Detergents and Paper products.

9: Delicatessen: Annual spending on Delicatessen products.

Sample of the Dataset



Exploratory Data Analysis



There are total 440 rows and 10 columns, of which seven columns are integer type and two columns are object type.

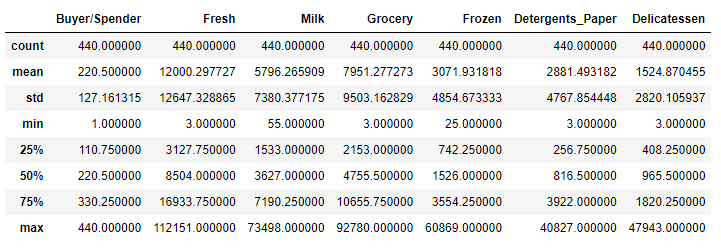
Checking for Missing Values in the Data Set



From the above Figure, we can see that there are no missing values in the given data set.

**1.1.1 Use methods of descriptive statistics to summarize data.**

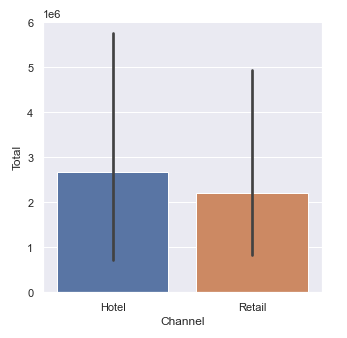
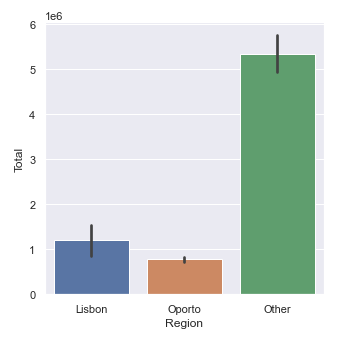
Descriptive statistics help describe and understand the features of a specific data set by giving short summaries about the sample and measures of the data. The most recognized types of descriptive statistics are measures of centre: the mean, median, and mode, which are used at almost all levels of math and statistics.

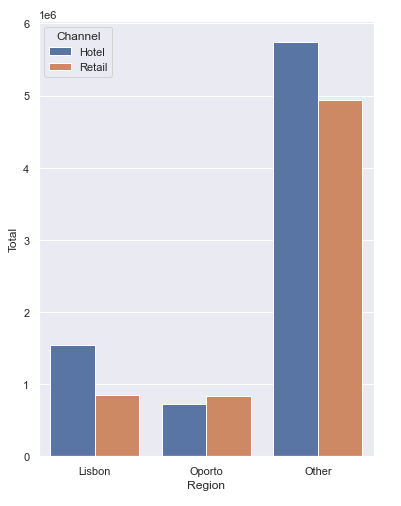


From the above describe function, we can see that there are 6 unique types of products sold by the whole seller. NaN shows that the values cannot be calculated for particular variables.

* Channel has two unique values, with "Hotel" as most frequent with 298 out of 440 entries. i.e. 67.7% of spending comes from "Hotel" channel.
* Region has three unique values, with "Other" as most frequent with 316 out of 440 entries. i.e. 71.8% of spending comes from "Other" region.

**1.1.2 Which Region and which Channel spent the most?**



From the above Figures, we can infer the following:

* Other is the region that spent the most with amount of **$ 10677599**
* Hotel is the channel that spent the most with amount of **$ 7999569**

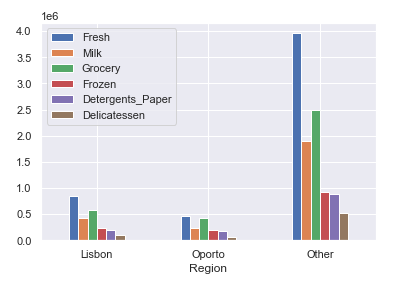
**1.1.3 Which Region and which Channel spent the least?**

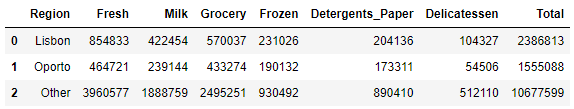
From the Figures illustrated in 1.1.2, we can infer the following:

* Oporto is the region that spent the least with amount of **$ 155508**
* Retail is the channel that spent the least with amount of **$ 6619931**

**1.2. There are 6 different varieties of items that are considered. Describe and comment/explain all the varieties across Region and Channel? Provide a detailed justification for your answer.**

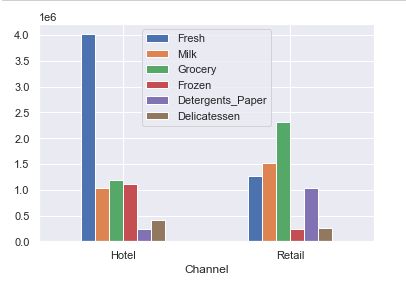
We have 6 different varieties of items. By plotting suitable graphs, we can get some insights on each of the 6 varieties, across Region and Channel.

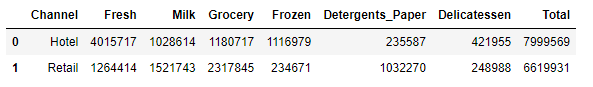
****



Some observations are given below:

* Other region is by far the highest spending region on all varieties of items, with Oporto being the lowest spender.
* Fresh and Groceries are found to be the most consumed items, with Delicatessen being the least.
* Spending pattern among items is consistent across regions.

****

****

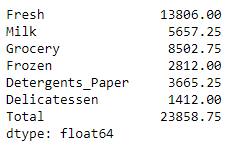
Some observations are given below:

* Hotel channel is found to have more spending than Retail channel.
* Fresh product is the most popular product in Hotel channel, with over 50% of Hotel revenue generated from its sales. Detergents and Paper are the least sold item in Hotel channel.
* In Retail channel, however, Fresh products are the third most consumed item, behind Groceries and Milk at one and two, respectively. Frozen and Delicatessen are the least consumed item in this channel.

**1.3 On the basis of the descriptive measure of variability, which item shows the most inconsistent behaviour? Which items shows the least inconsistent behaviour?**

We can use IQR method to find the most as well as the least consistent item.

IQR for all 6 varieties of items are as follows:

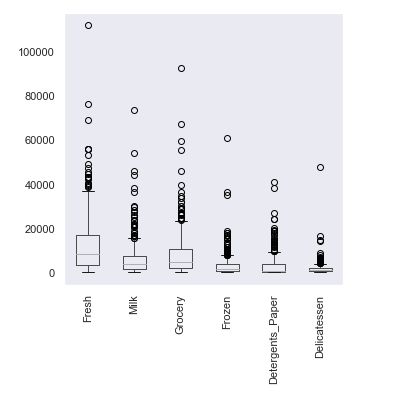


Statistics state that the one with highest IQR is the most inconsistent, and the one with the least IQR is the least inconsistent.

Therefore, we can say that Fresh items exhibit the most inconsistent behaviour, whereas Delicatessen items seem to be the least inconsistent.

**1.4 Are there any outliers in the data? Back up your answer with a suitable plot/technique with the help of detailed comments.**

To determine the presence of outliers across the products a box plot is drawn.



### As we can see from the above figure, it is safe to say that yes, there are indeed outliers present in all the items across the product range.

The item Fresh has the biggest outlier, with the item Delicatessen having the smallest outlier.

**1.5. On the basis of your analysis, what are your recommendations for the business? How can your analysis help the business to solve its problem? Answer from the business perspective.**

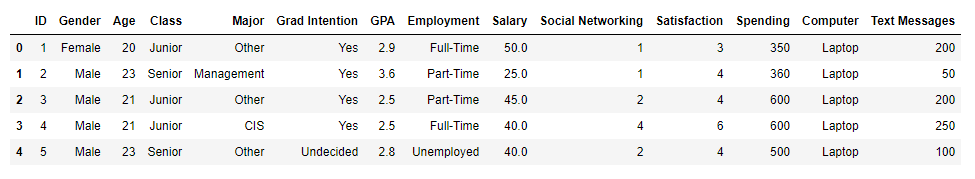
As per the analysis made of the data provided, following are some recommendations:

1. On observing the spending based on the region, Lisbon and Oporto are far lacking behind Others region. Whole Seller needs to find the grey areas causing less spending in these regions and try to ramp up market campaigns or promotions in order to increase spending.
2. There are great inconsistencies across some products. The Whole Seller has to focus more on those products (Frozen, Detergents & Paper and Delicatessen) to make them more aligned with each other.
3. There are many outliers present in all the products, indicating uneven spending. There is a need to identify a reason for outliers.
4. There is a discrepancy in total spending between Hotel and Retail channels, when they should be more or less equal. The Whole Seller ought to work out on pricing/discounts offered to increase the revenue generated.
5. Whole Seller must also focus more on products other than Fresh and Grocery as mean spending is comparatively more on those other products.
6. There are only two modes of sales, either by Hotel or Retail. Expanding channels of sales in order to reach customers would be very beneficial.

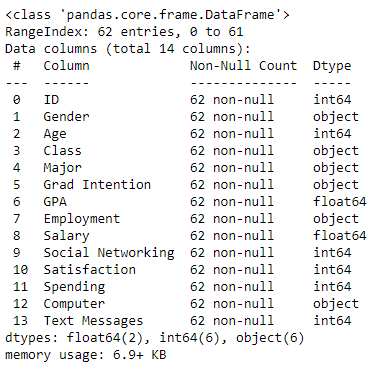
Problem 2

The Student News Service at Clear Mountain State University (CMSU) has decided to gather data about the undergraduate students that attend CMSU. CMSU creates and distributes a survey of 14 questions and receives responses from 62 undergraduates (stored in the **Survey** data set).

Sample of the Data Set



Exploratory Data Analysis



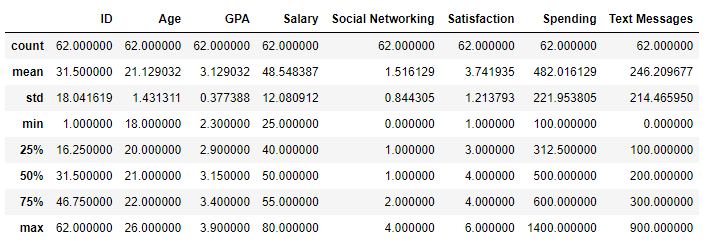
There are 62 entries across 14 columns, of which six columns each are of integer type and object type, while two columns are float type.

Checking for Missing Values in the Data Set



There are no missing values in the given data set.

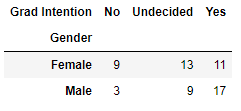
Descriptive Statistics for the Data Set



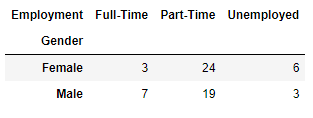
**2.1. For this data, construct the following contingency tables (Keep Gender as row variable)**

**2.1.1. Gender and Major**

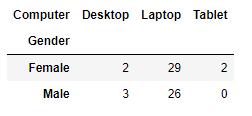
**2.1.2. Gender and Grad Intention**



**2.1.3. Gender and Employment**

****

**2.1.4. Gender and Computer**



**2.2. Assume that the sample is representative of the population of CMSU. Based on the data, answer the following question:**

**2.2.1. What is the probability that a randomly selected CMSU student will be male?**

Probability that a randomly selected CMSU student will be male is (29/62) = **46.77%**

**2.2.2. What is the probability that a randomly selected CMSU student will be female?**

Probability that a randomly selected CMSU student will be female is (33/62) = **53.23%**

**2.3. Assume that the sample is representative of the population of CMSU. Based on the data, answer the following question:**

****

**2.3.1. Find the conditional probability of different majors among the male students in CMSU.**

Probability of Accounting being a male student is (4/29) = **13.79%**

Probability of CIS being a male student is (1/29) = **3.45%**

Probability of Economics/Finance being a male student is (4/29) = **13.79%**

Probability of International Business being a male student is (2/29) = **6.9%**

Probability of Management being a male student is (6/29) = **20.69%**

Probability of Other being a male student is (4/29) = **13.79%**

Probability of Retailing/Marketing being a male student is (5/29) = **17.24**%

Probability of Undecided being a male student is (3/29) = **10.3%**

**2.3.2 Find the conditional probability of different majors among the female students of CMSU.**

Probability of Accounting being a female student is (3/33) = **9.09%**

Probability of CIS being a female student is (3/33) = **9.09%**

Probability of Economics/Finance being a female student is (7/33) = **21.21%**

Probability of International Business being a female student is (4/33) = **12.12%**

Probability of Management being a female student is (4/33) = **12.12%**

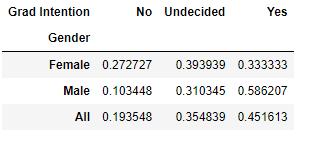
Probability of Other being a female student is (3/33) = **9.09%**

Probability of Retailing/Marketing being a female student is (9/33) = **27.27%**

Probability of Undecided being a female student is (0/33) = **0%**

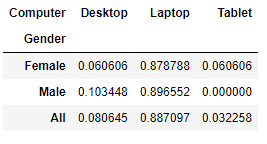
**2.4. Assume that the sample is a representative of the population of CMSU. Based on the data, answer the following question:**

**2.4.1. Find the probability that a randomly chosen student is a male and intends to graduate.**



Probability of a student being male and intending to graduate is **58.62%**

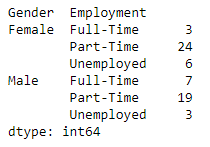
**2.4.2 Find the probability that a randomly selected student is a female and does NOT have a laptop.**

****

Probability of a student being female and not having a laptop is **12%**

**2.5. Assume that the sample is representative of the population of CMSU. Based on the data, answer the following question:**

**2.5.1. Find the probability that a randomly chosen student is a male or has full-time employment?**

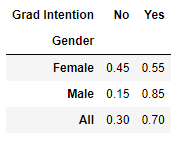
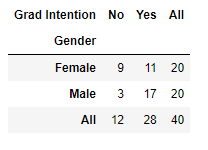


Probability that a randomly chosen student is a male or has a full-time employment is **51.61%**

**2.5.2. Find the conditional probability that given a female student is randomly chosen, she is majoring in international business or management.**

Probability that when a female student is randomly chosen, she is majoring in international business or management is **24.24%**

**2.6. Construct a contingency table of Gender and Intent to Graduate at 2 levels (Yes/No). The Undecided students are not considered now and the table is a 2x2 table. Do you think the graduate intention and being female are independent events?**

****

If two events are independent then it should satisfy the below rule:

P (A ∩ B) = P(A). P(B)

Probability of student being a female is P(F) = 0.5

Probability of student intending to graduate is P(G) = 0.7

Probability of a student being Female and Intent to graduate is P(F∩G) = 0.275

Since, P(F∩G) ≠ P(F).P(G) Therefore, graduate intention and being a female are not independent events.

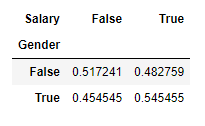
**2.7. Note that there are four numerical (continuous) variables in the data set, GPA, Salary, Spending, and Text Messages.**

**Answer the following questions based on the data**

**2.7.1. If a student is chosen randomly, what is the probability that his/her GPA is less than 3?**

Probability of a randomly chosen student’s GPA being less than 3 is **27.42%**

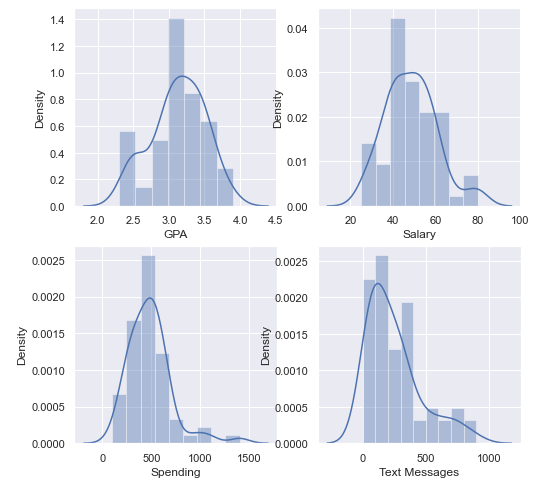
**2.7.2. Find the conditional probability that a randomly selected male earns 50 or more. Find the conditional probability that a randomly selected female earns 50 or more.**

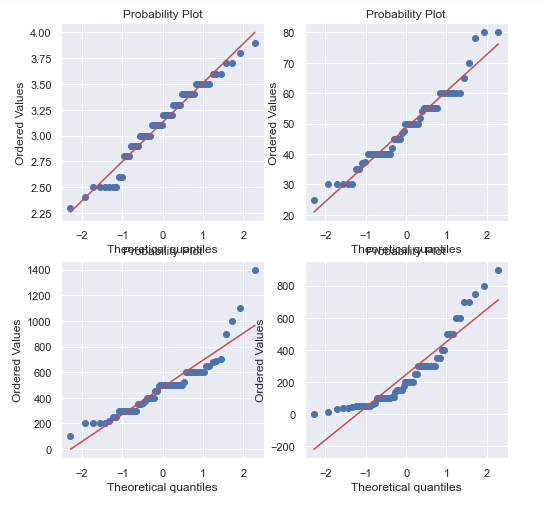
****

The probability that a randomly selected male earns 50 or more is **48.28%**

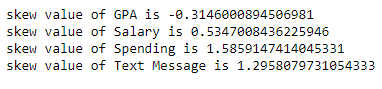
The probability that a randomly selected female earns 50 or more is **54.55%**

**2.8. Note that there are four numerical (continuous) variables in the data set, GPA, Salary, Spending, and Text Messages. For each of them comment whether they follow a normal distribution. Write a note summarizing your conclusions.**





From the graphs it is clear that the continuous variables more or less fall under normal distribution.



Taking a look at the skew value, if the value is zero it is said to be symmetric. A negative skew value indicates that the data is left skewed and on the other hand, a positive skew value indicates that the data is right skewed.

**Empirical Relation for GPA:**

1. mu + 1 sigma = 3.129 + 0.377 = 3.506 mu

2. Almost 70% of the data points lies in between this range

**Empirical Relation for Salary:**

1. mu + 1 sigma = 48.548 + 12.080 = 60.628 mu

2. Almost 83% of the data points lies in between this range

**Empirical Relation for Spending:**

1. mu + 1 sigma = 48.548 + 12.080 = 60.628 mu

2. Almost 83% of the data points lies in between this range and slightly left skewed

**Empirical Relation for Text Messages:**

1. mu + 1 sigma = 246.21 + 214.466 = 460.676 mu

2. Almost 82% of the data points lies in between this range

Conclusion

The dataset of students answering to the survey resulted in 62 responses, from both male and female students. A number of observations made are listed below:

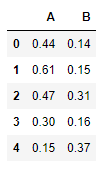
* There are slightly more female students than male students.
* Many students intend to graduate.
* The vast majority of students own a laptop.
* Retailing/Marketing is the most popular Major pursued by students.
* 2/3rd of students are looking for a part-time job.
* The mean salary is about 50.

Problem 3

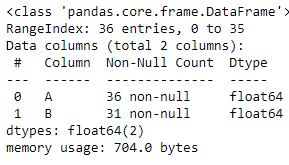
An important quality characteristic used by the manufacturers of ABC asphalt shingles is the amount of moisture the shingles contain when they are packaged. Customers may feel that they have purchased a product lacking in quality if they find moisture and wet shingles inside the packaging.   In some cases, excessive moisture can cause the granules attached to the shingles for texture and coloring purposes to fall off the shingles resulting in appearance problems. To monitor the amount of moisture present, the company conducts moisture tests. A shingle is weighed and then dried. The shingle is then reweighed, and based on the amount of moisture taken out of the product, the pounds of moisture per 100 square feet are calculated. The company would like to show that the mean moisture content is less than 0.35 pounds per 100 square feet.

The file ([A & B shingles.csv](https://olympus.greatlearning.in/courses/53595/files/3060221/download?verifier=SfvK8Oq4wTCI0Pq2VGzne5WoIQ5pfy1avNAnysjJ&wrap=1)) includes 36 measurements (in pounds per 100 square feet) for A shingles and 31 for B shingles.

Sample of the Data Set

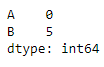


Exploratory Data Analysis



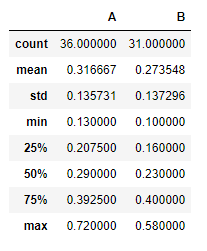
The given data set contains 36 entries across 2 columns, with both of them being of float type.

Checking for Missing Values in the Data Set



There are five missing values in the given data set, all belonging to Sample B.

Descriptive Statistics for the Data Set



**3.1 Do you think there is evidence that means moisture contents in both types of shingles are within the permissible limits? State your conclusions clearly showing all steps.**

We are given two independent samples of shingles A and B.

Since the population Standard Deviation is unknown, we cannot perform Z test. So we move forward with T-test.

We must find if the mean moisture level is less than the permissible limit for both samples.

So, one sample T-test for sample A and sample B is required.

SAMPLE A

1. DEFINE NULL & ALTERNATE HYPOTHESIS

The null hypothesis states that the moisture content of sample A is greater than or equal to the permissible limit, 𝜇 ≥ 0.35

The alternative hypothesis states that the moisture content of sample A is less than permissible limit, 𝜇 < 0.35

𝐻0 : 𝜇 ≥ 0.35 𝐻𝐴 : 𝜇 < 0.35

1. DETERMINE THE SIGNIFICANCE LIMIT

Since alpha value is not given in the question, we have to assume 𝛼 = 0.05.

1. IDENTIFY THE TEST STATISTIC

We have samples A and we do not know the population standard deviation. The sample size is n=36. So, we will use the t distribution and the 𝑡𝑆𝑇𝐴𝑇 test statistic for one sample t-test.

1. CALCULATE THE P-VALUE & TEST STATISTIC

****

1. DECIDE TO REJECT OR ACCEPT NULL HYPOTHESIS

Since tstat > P Value, we fail to reject the null hypothesis.

We conclude that the moisture content is greater than the permissible limit in sample A.

SAMPLE B

1. DEFINE NULL & ALTERNATE HYPOTHESIS

The null hypothesis states that the moisture content of sample B is greater than or equal to the permissible limit, 𝜇 ≥ 0.35

The alternative hypothesis states that the moisture content of sample B is less than the permissible limit, 𝜇 < 0.35

𝐻0 : 𝜇 ≥ 0.35 𝐻A : 𝜇 < 0.35

1. DETERMINE THE SIGNIFICANCE LIMIT

Since alpha value is not given in the question, we have to assume 𝛼 = 0.05.

1. IDENTIFY THE TEST STATISTIC

We have sample B and we do not know the population standard deviation. The sample size is n=31. So, we will use the t distribution and the 𝑡𝑆𝑇𝐴𝑇 test statistic for one sample t-test.

1. CALCULATE THE P-VALUE & TEST STATISTIC

****

1. DECIDE TO REJECT OR ACCEPT NULL HYPOTHESIS

Since tstat < P Value, we reject the null hypothesis.

We conclude that the moisture content is less than permissible limit in sample B.

**3.2 Do you think that the population mean for shingles A and B are equal? Form the hypothesis and conduct the test of the hypothesis. What assumption do you need to check before the test for equality of means is performed?**

A. DEFINE NULL & ALTERNATE HYPOTHESIS

In testing whether the mean for shingles A and Shingles B are the same, the null hypothesis states that the mean of shingle A to mean of shingle B are the same,   equals . The alternative hypothesis states that the mean are different,   is not equal to

* 𝐻0 : 𝜇 A - 𝜇B =/= 0 i.e 𝜇A IS NOT EQUAL TO 𝜇B
* 𝐻A : 𝜇A - 𝜇B = 0 i.e  𝜇A IS EQUAL TO 𝜇B

B. DETERMINE THE SIGNIFICANCE LIMIT

Since alpha value is not given in the question, we have to assume 𝛼 = 0.05.

C. IDENTIFY THE TEST STATISTIC

We have two samples and we do not know the population standard deviation. Sample sizes for both samples are not the same. The sample size is n > 30. So, we will use the t distribution and the 𝑡𝑆𝑇𝐴𝑇 test statistic for two sample test.

D. CALCULATE THE P-VALUE & TEST STATISTIC

****

E. DECIDE TO REJECT OR ACCEPT NULL HYPOTHESIS

Since tstat > P Value, we fail to reject the null hypothesis.

We conclude that the mean for shingles A and singles B is not equal.

THE END