Thakur Arun Singh

This Business Report shall provide detailed explanation of how we approached each problem given in the assignment. It shall also provide relative resolution and explanation with regards to the problems

Statistica l Method of Decision Making– Project

Business Report | March 2021

Contents

[Wholesale Customer Analysis 2](#_Toc66053622)

[Problem Statement 1: 2](#_Toc66053623)

[Problem 1.1 2](#_Toc66053624)

[Problem 1.2 3](#_Toc66053625)

[Problem 1.3 4](#_Toc66053626)

[Problem 1.4 4](#_Toc66053627)

[Problem 1.5 5](#_Toc66053628)

[Problem Statement 2: 6](#_Toc66053629)

[Problem 2.1 6](#_Toc66053630)

[Problem 2.1.1 6](#_Toc66053631)

[Problem 2.1.2 6](#_Toc66053632)

[Problem 2.1.3 6](#_Toc66053633)

[Problem 2.1.4 7](#_Toc66053634)

[Problem 2.2 7](#_Toc66053635)

[Problem 2.2.1 7](#_Toc66053636)

[Problem 2.2.2 7](#_Toc66053637)

[Problem 2.3 7](#_Toc66053638)

[Problem 2.3.1 7](#_Toc66053639)

[Problem 2.3.2 8](#_Toc66053640)

[Problem 2.4 8](#_Toc66053641)

[Problem 2.4.1 8](#_Toc66053642)

[Problem 2.4.2 8](#_Toc66053643)

[Problem 2.5 9](#_Toc66053644)

[Problem 2.5.1 9](#_Toc66053645)

[Problem 2.5.2 9](#_Toc66053646)

[Problem 2.6 9](#_Toc66053647)

[Problem 2.7 10](#_Toc66053648)

[Problem 2.7.1 10](#_Toc66053649)

[Problem 2.7.2 10](#_Toc66053650)

[Problem 2.8 10](#_Toc66053651)

[Problem 3 12](#_Toc66053652)

[Problem 3.1 12](#_Toc66053653)

[Problem 3.2 13](#_Toc66053654)

# Wholesale Customer Analysis

## Problem Statement 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 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).

### Problem 1.1

Use methods of descriptive statistics to summarize data. Which Region and which Channel seems to spend more? Which Region and which Channel seems to spend less?

**Resolution:**

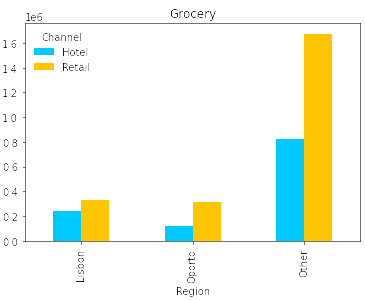
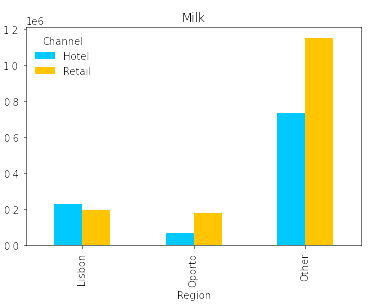
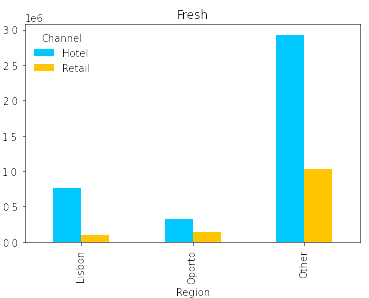
First we imported the ‘Wholesale Customer data’ dataset in python to analyze the ‘Spend’ under each store items across all regions and then find the solutions to each problem. Below is the detailed approach and resolutions.

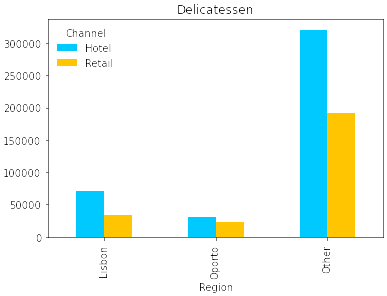
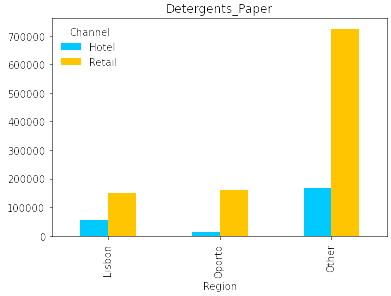
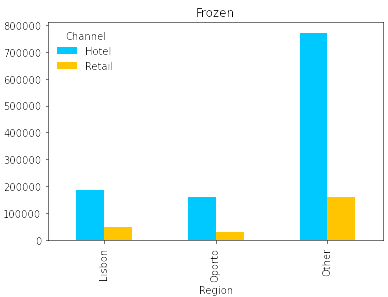
|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **count** | **mean** | **std** | **min** | **25%** | **50%** | **75%** | **max** |
| **Buyer/Spender** | 440 | 220.5 | 127.161315 | 1 | 110.75 | 220.5 | 330.25 | 440 |
| **Fresh** | 440 | 12000.2977 | 12647.3289 | 3 | 3127.75 | 8504 | 16933.75 | 112151 |
| **Milk** | 440 | 5796.26591 | 7380.37718 | 55 | 1533 | 3627 | 7190.25 | 73498 |
| **Grocery** | 440 | 7951.27727 | 9503.16283 | 3 | 2153 | 4755.5 | 10655.75 | 92780 |
| **Frozen** | 440 | 3071.93182 | 4854.67333 | 25 | 742.25 | 1526 | 3554.25 | 60869 |
| **Detergents\_Paper** | 440 | 2881.49318 | 4767.85445 | 3 | 256.75 | 816.5 | 3922 | 40827 |
| **Delicatessen** | 440 | 1524.87046 | 2820.10594 | 3 | 408.25 | 965.5 | 1820.25 | 47943 |

Which region and channel *spend most* & *least*?

Using describe function in python we can first looked at the basic descriptive statistics of the data set. Using bar graph with Region and Channel we were able to identify region with most spend and least spend. Below is the bar graphical representations – Looking at the bar graphs, **Hotel** Channel spends more and **Retail** spends least.

* Hotel channel spent amount is **$8070603** with the highest spend amount
* Retail spent amount **$6645917** with least spend amount





Similarly I have grouped totals by region

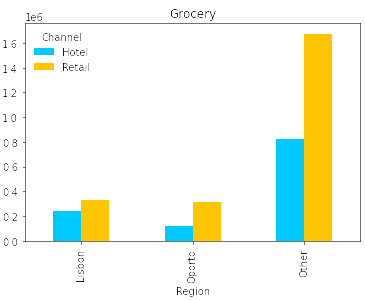
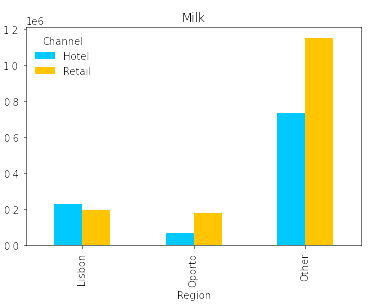
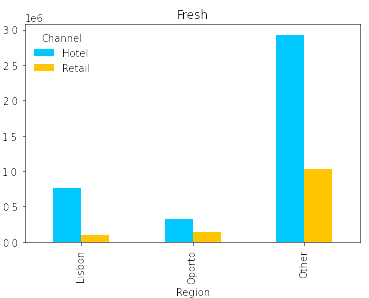
* **Other regions** spent amount is **$10741625** with the highest spend amount
* **Oporto region** spent amount is **$1569987** has least spend amount

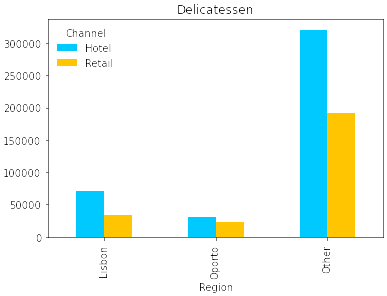
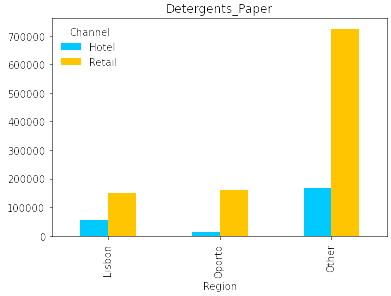
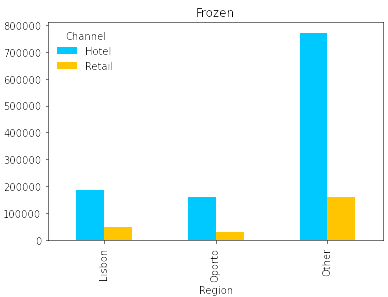
### Problem 1.2

There are 6 different varieties of items are considered. Do all varieties show similar behavior across Region and Channel?  Provide justification for your answer

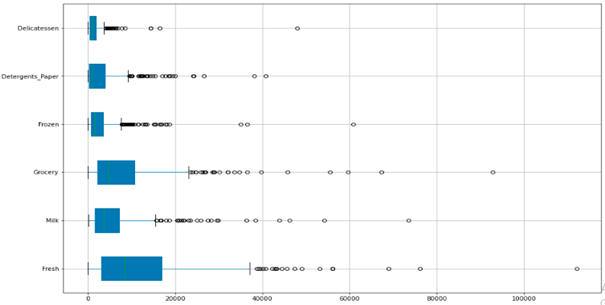
**Resolution:**

We have used pivot tables for each category for checking spent across Region and Channel we get the following outputs –





From the above graphs, we can infer that some categories such as **Milk, Grocery & Detergents\_Paperhave** has **higher spent** in the Retail channel Vs. Hotel, across all regions. On the other hand, **Fresh and Frozen** has **higher consumption** in the Hotel channel Vs. Retail, across all regions. Additionally, if we plot a box plot we can infer that the spent for **Fresh and Groceries** is the **maximum** across region and channel while for **Delicatessen** it is the **least** across region and channel. Below is the output of boxplot



### Problem 1.3

On the basis of a descriptive measure of variability, which item shows the most inconsistent behavior? Which items show the least inconsistent behavior?

**Resolution:**

Using Coefficient of Variation we can find the **least** value is of Category **Fresh** (1.05) and **highest** value is of Category **Delicatessen** (1.84)

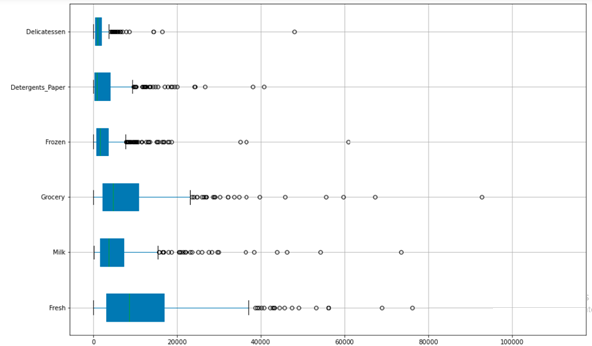
Therefore, with the given data we can infer that most inconsistent behavior shown by item **Delicatessen** and least **inconsistent** behavior shown by item **Fresh**

### Problem 1.4

Are there any outliers in the data?

**Resolution:**

Here to find the outliers we can plot a boxplot and the output shall give us the details for the given data and their outliers



### Problem 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

**Resolution:**

From the above analysis, we can say that there are **inconsistencies in spending** of different items after calculating Coefficient of Variation, so **spending can be minimized**. The **spending of Hotel and Retail** channel are **different** which **should be balanced**. Additionally, **spent should be equal distributed for different regions**. Business needs to **pay attention on other items** rather than just focusing on Fresh and Grocery.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

## Problem Statement 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).

### Problem 2.1

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

**Resolution:**

### Problem 2.1.1

**Gender and Major**

|  |  |  |
| --- | --- | --- |
| **Major** | **Female** | **Male** |
| Accounting | 3 | 4 |
| CIS | 3 | 1 |
| Economics/Finance | 7 | 4 |
| International Business | 4 | 2 |
| Management | 4 | 6 |
| Other | 3 | 4 |
| Retailing/Marketing | 9 | 5 |
| Undecided |  | 3 |

### Problem 2.1.2

**Gender and Grade Intention**

|  |  |  |
| --- | --- | --- |
| **Grade Intention** | **Female** | **Male** |
| No | 9 | 3 |
| Undecided | 13 | 9 |
| Yes | 11 | 17 |

### Problem 2.1.3

**Gender and Employment**

|  |  |  |
| --- | --- | --- |
| **Employment** | **Female** | **Male** |
| Full-Time | 3 | 7 |
| Part-Time | 24 | 19 |
| Unemployed | 6 | 3 |

### Problem 2.1.4

**Gender and Computer**

|  |  |  |
| --- | --- | --- |
| **Computer** | **Female** | **Male** |
| Desktop | 2 | 3 |
| Laptop | 29 | 26 |
| Tablet | 2 |  |

### Problem 2.2

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

**Resolution:**

### Problem 2.2.1

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

**Resolution:**

First we will find out total number of male students from all students from the given data. The probability of male in CMSU if randomly selected students is **46.77%**

### Problem 2.2.2

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

**Resolution:**

Here we will find out total number of female students from all students from the given data. The probability of female in CMSU if randomly selected students is **53.23%**

### Problem 2.3

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

### Problem 2.3.1

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

**Resolution:**

**By taking the help of contingency tables, we got** Gender and Majors total number of males and of which who are opting for different majors

**Probabilities of male students opting for different majors are**

**Accounting 13.79%**

**CIS 3.45%**

**Economics/Finance 13.79%**

**International Business 6.90%**

**Management 20.69%**

**Other 13.79%**

**Retailing/Marketing 17.24%**

**Undecided 10.34%**

From the above table we can infer that most of the male students prefer **Management** **as Majors** and **CIS** **is the least preferred**.

### Problem 2.3.2

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

**Resolution:**

**By taking the help of contingency tables, we got** Gender and Majors total number of females and of which who are opting for different majors

**Probabilities of female students opting for different majors are**

**Accounting 9.09%**

**CIS 9.09%**

**Economics/Finance 21.21%**

**International Business 12.12%**

**Management 12.12%**

**Other 9.09%**

**Retailing/Marketing 27.27%**

**Undecided 0.00%**

From the above table we can infer that most of the female students prefer **Retailing/Marketing as Majors** and none of the female students are undecided.

### Problem 2.4

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

### Problem 2.4.1

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

**Resolution:**

Again with the help of contingency tables of Gender and Grad Intention we get the total numbers of males and number of males who intends to be graduated. After calculating we can say that - Probability of male and intends to be Graduate is **58.62%**

### Problem 2.4.2

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

**Resolution:**

With the help of contingency tables of Gender and Computer we get the total numbers of females who do not have laptops. After calculating we can say that - Probability of female and do not have laptop is **13.79%**

### Problem 2.5

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

### Problem 2.5.1

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

**Resolution:**

With the help of contingency tables of Gender and Employment we get the total numbers of males who are full time employed. After calculating we can say that - Probability of male and who has full time employment is **74.19%**

### Problem 2.5.2

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

**Resolution:**

With the help of contingency tables of Gender and Major we get the total numbers of females who took International Business or Management. After calculating we can say that - Probability of female and who are majoring in International Business or Management is **24.24%**

### Problem 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?**

**Resolution:**

|  |  |  |
| --- | --- | --- |
| **Grade Intentions / Gender** | **Female** | **Male** |
| No | 9 | 3 |
| Yes | 11 | 17 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Grad Intention** | **No** | **Yes** | **Total** |
| Gender |  |  |  |
| Female | 9 | 11 | 20 |
| Male | 3 | 17 | 20 |
| Total | 12 | 0 | 40 |

The Probability that a randomly selected student being female

The Probability that a randomly selected student the graduate intention and being female

P(Grad Intention Yes) = 28/40 = 0.7

P(Grad Intention Yes | female) = 11 / 20 = 0.55

These probabilities are **not equal**. This suggests that the **two events are independent**.

### Problem 2.7

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

### Problem 2.7.1

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

**Resolution:**

Using contingency tables of Gender and GPA we get the total number of students and number of students who’s GPA less than 3. After calculating we can say that - Probability of randomly students and his/her GPA is less than 3 is **22.58%**

### Problem 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.

**Resolution:**

Using contingency tables of Gender and Salary we get the total number of students’ who are male and female and earning more than 50. After calculating we can say that - Probability of randomly selected **male** earns 50 or more is **34.48%** andfor female **30.30%**

### Problem 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.

**Resolution:**

Here with the help of using ‘distplot’ we get the normal distribution of these four numerical (continuous) variables in the data set – GPA, Salary, Spending and Text Messages

We can confirm whether these four data sets are following normal distribution or not, we done the Shapiro–Wilk test and the output from Python

ShapiroResult(statistic=0.953252375125885, pvalue=0.09815297275781631)

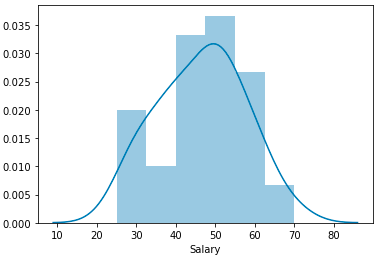
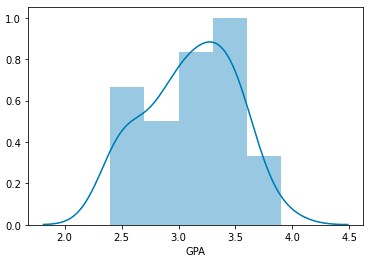
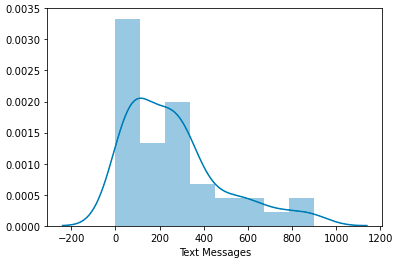
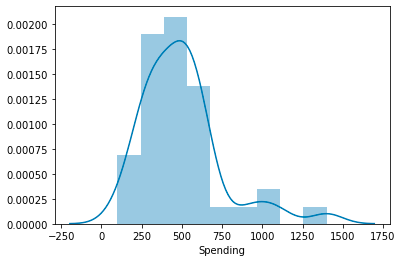
ShapiroResult(statistic=0.9689891934394836, pvalue=0.33416980504989624)

ShapiroResult(statistic=0.8724251985549927, pvalue=0.00033097428968176246)

ShapiroResult(statistic=0.8824034929275513, pvalue=0.0006114590214565396)

From the above details we can say that out of the given four data sets ‘GPA’ and ‘Salary’ are following normal distribution whereas other two ‘Spending’ and ‘Text Messages’ are not following the normal distribution

Below are the graphs

****

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

### 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 colouring 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 is calculated. The company would like to show that the mean moisture content is less than 0.35 pound per 100 square feet.

### Problem 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.

**Resolution:**

**Input** from Python

t\_statistic, p\_value = ttest\_1samp(df.A, 0.35)

print('One sample t test \nt statistic: {0} p value: {1} '.format(t\_statistic, p\_value/2))

**Output** from Python

One sample t test

t statistic: -1.4735046253382782 p value: 0.07477633144907513

Since P-value > 0.05, do not reject H0 . There is not enough evidence to conclude that the mean moisture content for Sample A shingles is less than 0.35 pounds per 100 square feet. P-value = 0.0748.

If the population mean moisture content is in fact no less than 0.35 pounds per 100 square feet, the probability of observing a sample of 36 shingles that will result in sample mean moisture content of 0.3167 pounds per 100 square feet or less is .0748.

**Input** from Python

t\_statistic, p\_value = ttest\_1samp(df.B, 0.35,nan\_policy='omit' )

print('One sample t test \nt statistic: {0} p value: {1} '.format(t\_statistic, p\_value/2))

**Output** from Python

One sample t test

t statistic: -3.1003313069986995 p value: 0.0020904774003191826

Since P-value < 0.05, reject H0 . There is enough evidence to conclude that the mean moisture content for Sample B shingles is not less than 0.35 pounds per 100 square feet. P-value = 0.0021.

If the population mean moisture content is in fact no less than 0.35pounds per 100 square feet, the probability of observing a sample of 31 shingles that will result in sample mean moisture content of 0.2735 pounds per 100 square feet or less is .0021.

### Problem 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?

**Resolution:**

H0 : μ(A)= μ(B)

Ha : μ(A)!= μ(B)

α = 0.05

**Input** fromPython

t\_statistic,p\_value=ttest\_ind(df['A'],df['B'],equal\_var=True ,nan\_policy='omit')

print("t\_statistic={} and pvalue={}".format(round(t\_statistic,3),round(p\_value,3)))

**Output** from Python

t\_statistic=1.29 and pvalue=0.202

As the P-value > α , do not reject H0; and we can say that population mean for shingles A and B are equal Test Assumptions When running a two-sample t-test, the basic assumptions are that the distributions of the two populations are normal, and that the variances of the two distributions are the same. If those assumptions are not likely to be met, another testing procedure could be used.

The End

Thakur Arun Singh

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*