**Contents**

Problem 1

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?

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

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?

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

the help of detailed comments

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

Problem 2

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

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 a representative of the population of CMSU. Based on the data,

answer the following questions:

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

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.

2.4. Note that there are three numerical (continuous) variables in the data set, Salary, Spending and Text Messages. For each of them comment whether they follow a normal distribution. Write a note summarizing your conclusions. [Recall that symmetric histogram does not necessarily mean that the underlying distribution is symmetric]

Problem 3

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?

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

**PROBLEM 1**

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

Measure of Central Tendency - Mean, Median, mode Measure of Dispersion - Range, IQR, Standard Deviation

From the above two describe function, we can infer the following

Channel has two unique values, with "Hotel" as most frequent with 298 out of 440 transactions. i.e 67.7 percentage of spending comes from "Hotel" channel.

Retail has three unique values, with "Other" as most frequent with 316 out of 440 transactions. i.e.71.8 percentage of spending comes from "Other" region.

Fresh item (440 records),has a mean of 12000.3, standard deviation of 12647.3, with min value of 3 and max value of 112151 .

The other aspect is Q1(25%) is 3127.75, Q3(75%) is 16933.8, with Q2(50%) 8504

range = max-min =112151-3=112,148 & IQR = Q3-Q1 = 16933.8-3127.75 = 13,806.05 (this helpful in calculating the outlier(1.5 IQR Lower/Upper limit))

Milk item (440 records),has a mean of 5796.27, standard deviation of 7380.38, with min value of 55 and max value of 73498.

The other aspect is Q1(25%) is 1533, Q3(75%) is 7190.25, with Q2(50%) 3627

range = max-min =73498-55=73443 & IQR = Q3-Q1 = 7190.25-1533 = 5657.25 (this helpful in calculating the outlier(1.5 IQR Lower/Upper limit))

Grocery item (440 records),has a mean of 7951.28, standard deviation of 9503.16, with min value of 3 and max value of 92780.

The other aspect is Q1(25%) is 2153, Q3(75%) is 10655.8, with Q2(50%) 4755.5

range = max-min =92780-3=92777 & IQR = Q3-Q1 = 10655.8-2153 = 8502.8 (this helpful in calculating the outlier(1.5 IQR Lower/Upper limit))

Frozen (440 records),has a mean of 3071.93, standard deviation of 4854.67, with min value of 25 and max value of 60869.

The other aspect is Q1(25%) is 742.25, Q3(75%) is 3554.25, with Q2(50%) 1526

range = max-min =60869-25=60844 & IQR = Q3-Q1 = 3554.25-742.25 = 2812 (this helpful in calculating the outlier(1.5 IQR Lower/Upper limit))

Detergents\_Paper (440 records),has a mean of 2881.49, standard deviation of 4767.85, with min value of 3 and max value of 40827.

The other aspect is Q1(25%) is 256.75, Q3(75%) is 3922, with Q2(50%) 816.5

range = max-min =40827-3=40824 & IQR = Q3-Q1 = 3922-256.75 = 3665.25 (this helpful in calculating the outlier(1.5 IQR Lower/Upper limit)).

Delicatessen (440 records),has a mean of 1524.87, standard deviation of 2820.11, with min value of 3 and max value of 47943.

The other aspect is Q1(25%) is 408.25, Q3(75%) is 1820.25, with Q2(50%) 965.5

range = max-min =47943-3=47940 & IQR = Q3-Q1 = 1820.25-408.25 = 1412 (this helpful in calculating the outlier(1.5 IQR Lower/Upper limit))

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

First looking at the basic descriptive statistics of the data set. Using

bar graph with Region and Channel we were able to identify region with maximum spend and minimum spend.Looking at the bar graph, Hotel Channel spends more and Retail spends least.

Hotel channel spend amounts to 8070603$ with the highest spend amount

Retail spend amounts to 6645917$ has least spend amount based on Channel.

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

Lowest spendings in the Region/Channel is from Oporto/Hotel

**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.**

Some categories like Milk, Grocery & Detergents\_Paper have higher spend in the Retail channel versus Hotel, across all regions. On the other hand, Fresh and Frozen

have higher consumption in the Hotel channel versus Retail, across all regions. Also, if we plot a box plot

So, we can say that the spend for Fresh and groceries is the maximum across region and channel while for Delicatessen it is the least across region and 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?**

Using Coefficient of Variation we find out the least value is of Category “Fresh” and highest value is of Category “Delicatessen”

So from the given data it is clear that most inconsistent behavior shown by item – Delicatessen

and least inconsistent behavior shown by item – Fresh

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

We use boxplot to and see that there are outliers in all the data present.

**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?**

I found out that there are inconsistencies in spending of different items (by calculating each Coefficient of Variation), which should be minimized. The spending of Hotel and Retail channel are different which should be more or less equal. And the spent should equal for different regions. There is a need to focus on other items also than “Fresh” and “Grocery”.

**PROBLEM 2**

****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****

**We make contingency tables using python for these.**

****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?****

**After calculation we got the result that probability of 46.77% student will be male in CMSU if randomly selected.**

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

**After calculation we got the result that probability of 53.23% student will be female in CMSU if randomly selected.**

****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.****

**By Contingency tables of Gender and Majors we got the total numbers of males and number of males**

**opting for different majors**

**Below is the output from Python –**

**Probability of Males opting for Accounting. is 13.79%**

**Probability of Males opting for CIS. is 3.45%**

**Probability of Males opting for Economics/Finance. is 13.79%**

**Probability of Males opting for InternationalBusiness. is 6.90%**

**Probability of Males opting for Management. is 20.69%**

**Probability of Males opting for Other. is 13.79%**

**Probability of Males opting for Retailing/Marketing. is 17.24%**

**Probability of Males opting for Undecided. is 10.34%**

**Most of the males students prefer Management as Majors and CIS**

**is the least preferred one.**

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

Probability of Females opting for Accounting. is 9.09%

Probability of Females opting for CIS. is 9.09%

Probability of Females opting for Economics/Finance. is 21.21%

Probability of Females opting for International Business. is 12.12%

Probability of Females opting for Management. is 12.12%

Probability of Females opting for Other. is 9.09%

Probability of Females opting for Retailing/Marketing. is 27.27%

Probability of Females opting for Undecided. is 0.00%

Most of the females students prefer Retailing/Marketing as Majors.

****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.****

**Using contingency tables of Gender and Grad Intention we got the total numbers of males and number of males intends to be graduate and post calculation we find out that - Probability of Males and intends to be Graduate. is 58.62%.**

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

Using contingency tables of Gender and Computer we got the total numbers of females and number of females does not have a laptop And post calculation we find out that - Probability of randomly selected student is a Female and does NOT have a laptop. is 13.79%.

****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?****

**Using contingency tables of Gender and Employment we got the total numbers of males and number of males who are full time employed**

**And post calculation we find out that - Probability of randomly chosen student is either Male or has full time employment. is 74.19%**

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

**Using contingency tables of Gender and Major we got the total numbers of females and number of females majoring in international business or management.**

**And post calculation we find out that - Probability that given 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?****

Grad Intention No Yes

Gender

Female 9 11

Male 3 17

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.

**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?**

Using contingency tables of Gender and GPA we got the total numbers of students and number of students GPA less than 3

And post calculation we find out that - Probability that student is chosen randomly and that his/her GPA is

less than 3 is 22.58%

**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.**

Probability that randomly selected male earns 50 or more is: 48.28 %

Probability that 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.**

By looking at histogram details we confirm 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.