

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 Customers Data.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/Restaurant/Café HoReCa, Retail).

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?

1.2. There are 6 different varieties of items are considered.

Do all varieties show similar behaviour 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?

1.4. Are there any outliers in the data?

1.5. On the basis of this report, what are the recommendations?

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.csv file).

Part I

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 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?

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

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

2.2.3. Find the conditional probability of intent to graduate, given that the student is a male.

Find the conditional probability of intent to graduate, given that the student is a female.

2.2.4. Find the conditional probability of employment status for the male students as well as for the female students.

2.2.5. Find the conditional probability of laptop preference among the male students as well as among the female students.

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.

Part II

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

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.

The file (A & B shingles.csv) includes 36 measurements (in pounds per 100 square feet) for A shingles and 31 for B shingles.

3.1. For the A shingles, form the null and alternative hypothesis to test whether the population mean moisture content is less than 0.35 pound per 100 square feet.

3.2. For the A shingles, conduct the test of hypothesis and find the p-value.

Interpret the p-value.

Is there evidence at the 0.05 level of significance that the population mean moisture content is less than 0.35 pound per 100 square feet?

3.3. For the B shingles, form the null and alternative hypothesis to test whether the population mean moisture content is less than 0.35 pound per 100 square feet.

3.4. For the B shingles, conduct the test of the hypothesis and find the p-value. Interpret the p-value.

Is there evidence at the 0.05 level of significance that the population mean moisture content is less than 0.35 pound per 100 square feet?

3.5. 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?

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

3.7. Check the assumptions made with histograms, boxplots, normal probability plots or empirical rule.

3.8. Do you think that the assumption needed in order to conduct the hypothesis tests above is valid? Explain.

Scoring guide (Rubric) - SMDM Project

Criteria	Points
----------	--------

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?	3
---	---

1.2. There are 6 different varieties of items are considered. Do all varieties show similar behaviour across Region and Channel?	3
--	---

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?	3
--	---

1.4. Are there any outliers in the data?	2
--	---

1.5. On the basis of this report, what are the recommendations?	4
---	---

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
---	---

2.2.1. Assume that the sample is a representative of the population of CMSU. Based on the data, answer the following question: 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?	3
--	---

2.2.2. Assume that the sample is a representative of the population of CMSU. Based on the data, answer the following question: Find the conditional probability of different majors among the male students in CMSU. Find the conditional probability	
---	--

of different majors among the female students of CMSU. 3

2.2.3. Assume that the sample is a representative of the population of CMSU. Based on the data, answer the following question: Find the conditional probability of intent to graduate, given that the student is a male. Find the conditional probability of intent to graduate, given that the student is a female. 3

2.2.4. Assume that the sample is a representative of the population of CMSU. Based on the data, answer the following question: Find the conditional probability of employment status for the male students as well as for the female students. 3

2.2.5. Assume that the sample is a representative of the population of CMSU. Based on the data, answer the following question: Find the conditional probability of laptop preference among the male students as well as among the female students. 3

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. 3

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] 4

3.1. For the A shingles, form the null and alternative hypothesis to test whether the population mean moisture content is less than 0.35 pound per 100 square feet. 2

3.2. For the A shingles, conduct the test of hypothesis and find the p-value. Interpret the p-value. Is there evidence at the 0.05 level of significance that the population mean moisture content is less than 0.35 pound per 100 square feet? 3

3.3. For the B shingles, form the null and alternative hypothesis to test whether the population mean moisture content is less than 0.35 pound per 100 square feet. 2

3.4. For the B shingles, conduct the test of the hypothesis and find the p-value. Interpret the p-value. Is there evidence at the 0.05 level of significance that the population mean moisture content is less than 0.35 pound per 100 square feet? 3

3.5. 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

3.6. What assumption about the population distribution is needed in order to conduct the hypothesis tests above? 2

3.7. Check the assumptions made with histograms, boxplots, normal probability plots or empirical rule. 3

3.8. Do you think that the assumption needed in order to conduct the hypothesis tests above is valid? Explain.