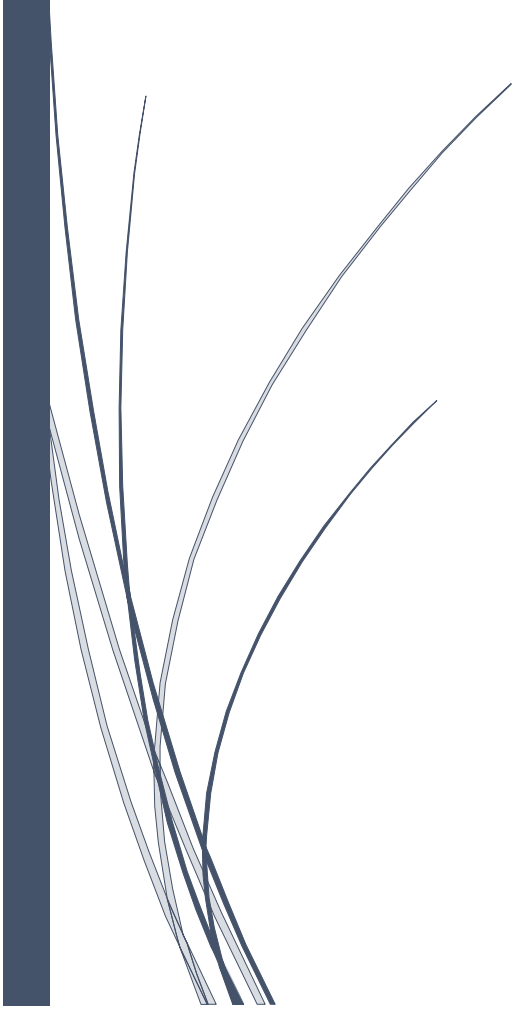
greatlearning
Learning for Life

10/30/2022

Advanced Statistics Project



Contents:

Problem 1:

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:

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 pounds 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.

Problem 3:

Salary is hypothesized to depend on educational qualification and occupation. To understand the dependency, the salaries of 40 individuals **[SalaryData.csv]** are collected and each person's educational qualification and occupation are noted. Educational qualification is at three levels, High school graduate, Bachelor's, and Doctorate. Occupation is at four levels, Administrative and clerical, Sales, Professional or specialty, and Executive or managerial. A different number of observations are in each level of education – occupation combination.

[Assume that the data follows a normal distribution. In reality, the normality assumption may not always hold if the sample size is small.]

Solutions:

Problem 1:

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).

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

1.1.1. Gender and Major

Ans.:

Major Gender	Accounting	CIS	Economics/Finance	International Business	Management	Other	Retailing/Marketing	Undecided
Female	3	3	7	4	4	3	9	0
Male	4	1	4	2	6	4	5	3

1.1.2. Gender and Grad Intention

Ans.:

Grad Intention Gender	No	Undecided	Yes
Female	9	13	11
Male	3	9	17

1.1.3. Gender and Employment

Ans.:

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

1.1.4. Gender and Computer

Ans.:

Computer Gender	Desktop	Laptop	Tablet
Female	2	29	2
Male	3	26	0

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

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

Ans.:
```
Female    33
Male      29
Name: Gender, dtype: int64
```

Total no. of students: 62
Probability that a randomly selected CMSU student will be male = 46.774193548387096 %

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

Ans.: Total no. of students: 62
Probability that a randomly selected CMSU student will be male = 53.2258064516129 %

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

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

Ans.:
Probability of Males opting for accounting. is: 13.793103448275861 %
Probability of Males opting for CIS. is: 3.4482758620689653 %
Probability of Males opting for Economics/Finance. is: 13.793103448275861 %
Probability of Males opting for International Business. is: 6.896551724137931 %
Probability of Males opting for Management. is: 20.689655172413794 %
Probability of Males opting for Other. is: 13.793103448275861 %
Probability of Males opting for Retailing/Marketing. is: 17.24137931034483 %
Probability of Males opting for Undecided. is: 10.344827586206897 %

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

Ans.:
Probability of Females opting for accounting. is: 9.090909090909092 %
Probability of Females opting for CIS. is: 9.090909090909092 %
Probability of Females opting for Economics/Finance. is: 21.21212121212121 %
Probability of Females opting for International Business. is: 12.121212121212121 %
Probability of Females opting for Management. is: 12.121212121212121 %
Probability of Females opting for Other. is: 9.090909090909092 %
Probability of Females opting for Retailing/Marketing. is: 27.27272727272727 %
Probability of Females opting for Undecided. is: 0.0 %

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

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

Ans.: The probability That a randomly chosen student is a male and intends to graduate. is: 58.620689655172406 %

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

Ans.: The probability that a randomly selected student is a female and does NOT have a laptop. is: 12.121212121212121 %

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

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

Ans.:
Total no. of employed people: 62
Total no. of Males: 29
Total no. of fulltime employees: 10
Total no. of fulltime employees: 7

Probability of male students: 46.774193548387096 %
Probability of fulltime employees: 16.129032258064516 %
Probability of male fulltime employees: 11.29032258064516 %

The probability that a randomly chosen student is either a male or has full-time employment 51.61290322580645 %

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

Ans.: Total Female count: 33

The conditional probability that given a female student is randomly chosen, she is majoring in international business or management: 24.242424242424242 %

1.6. 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?

Ans.:

Grad Intention	No	Yes	All
Gender			
Female	9	11	20
Male	3	17	20
All	12	28	40

- The probability that a randomly selected Student being Female is: 50.0 %
- The probability that a randomly selected student is female and intends to graduate 55.00000000000001 %

#Probability value for the graduate intention and being female are not equal, they are not independent events.

1.7. 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.

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

Ans.: **The probability that his/her GPA is less than 3 is 27.419354838709676 %**

1.7.2. 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.

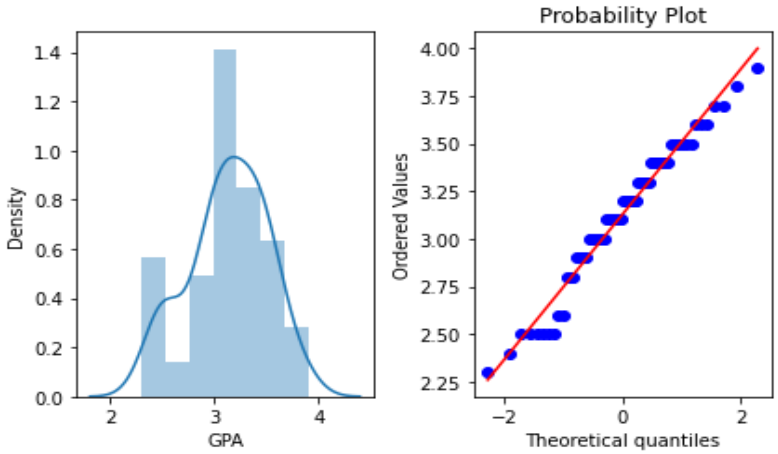
Ans.: **the conditional probability that a randomly selected male earns 50 or more is: 48.275862 %**

the conditional probability that a randomly selected female earns 50 or more is: 54.545454 %

1.8. 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.

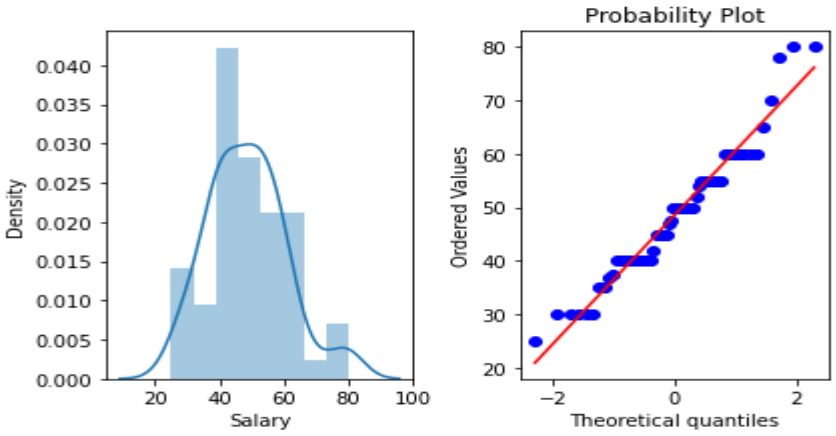
- Null Hypotheses(H0) :(P-value>0.05): Sample follows the normal distributions.
- Alternative Hypotheses (HA):(P-value<0.05): Sample does not follow the normal distributions.

Ans.: *Skew value for GPA is: -0.3146000894506981
**Kurtosis value for GPA is: -0.5040435381579838



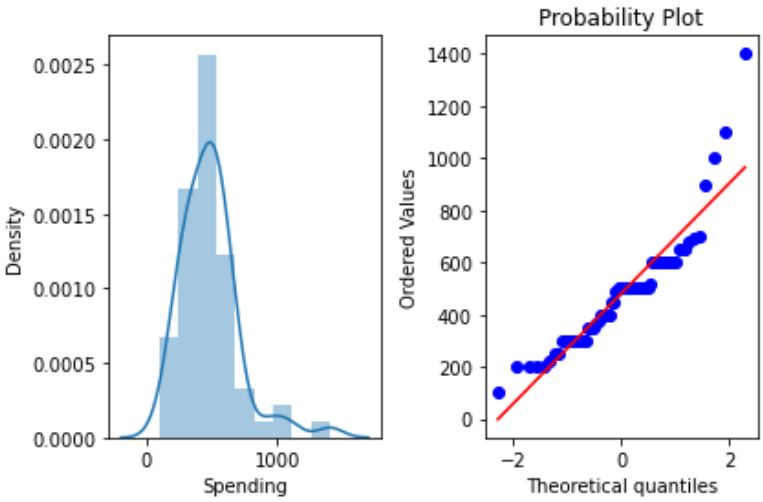
- **stat=0.9685, p=0.1120**
- **GPA follows Normal Distribution.**

*Skew value for Salary is: 0.5347008436225946
**Kurtosis value for Salary is: 0.4242636177584149



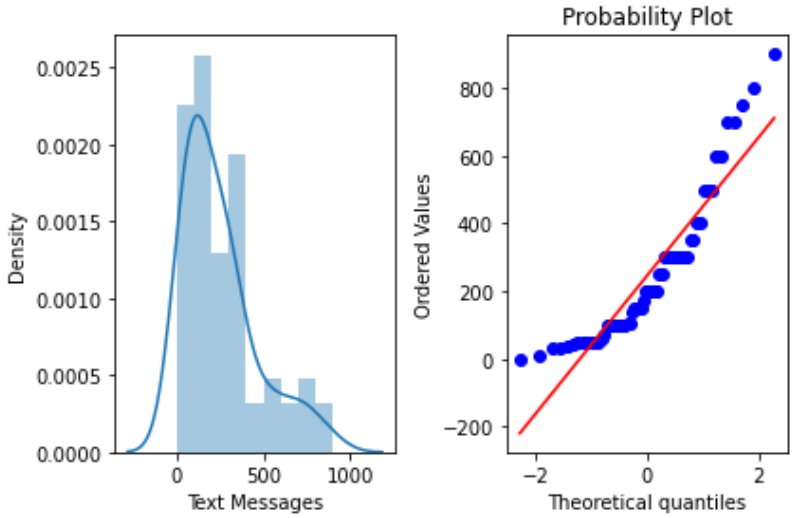
- **stat=0.9566, p=0.0280**
- **Salary does not follow Normal Distribution.**

*Skew value for Spending is: 1.5859147414045331
**Kurtosis value for Spending is: 4.559914423727916



- **stat=0.8777, p=0.0000**
- **Spending does not follow Normal Distribution.**

*Skew value for Text Messages is: 1.2958079731054333
**Kurtosis value for Text Messages is: 1.1356852071694052



- **stat=0.8594, p=0.0000**
- **Text Messages does not follow Normal Distribution.**

- *Write a note summarizing your conclusions:*

Ans.: GPA follows Normal Distribution.

Salary does not follow Normal Distribution.

Spending does not follow Normal Distribution.

Text Messages does not follow Normal Distribution.

Problem 2:

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 pounds 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.

2.1 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.

Ans.: The null hypothesis states that the moisture content of sample A and sample B is greater or than equal to the permissible limit, $\mu \geq 0.35$, and the alternative hypothesis states that the moisture content of sample A and sample B is less than permissible limit, $\mu < 0.35$.

Null Hypotheses (HA): $\mu \geq 0.35$

Alternative Hypotheses (HA): $\mu < 0.35$

##Here we need to consider significance level(α) = 0.05 as given in the question.

*t_stat value for sample A: -1.4735046253382782
**One-sample t-test p-value= 0.07477633144907513
We have enough evidence to reject the alternative hypothesis in favour of null hypothesis,
Hence, can conclude that the moisture content is greater than permissible limit in sample A.

*t_stat value for sample B: -3.1003313069986995
**one-sample t-test p-value= 0.0020904774003191826
We have enough evidence to reject the null hypothesis in favour of alternative hypothesis,
Hence, can conclude that the moisture content is less than permissible limit in sample B.

2.2 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?

Ans.: We have two samples and the sizes for both samples are not the same. The sample size is, n > 30. So, we use the t distribution and the t-test statistic for two sample tests.

To check the population, mean(s) for shingles A and B 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, μa equals μb. The alternative hypothesis states that the mean is different, μa is not equal to μb.

We can frame the Hypotheses as:

Null Hypotheses(H0): μa = μb
Alternative Hypotheses (HA): μa ≠ μb

*t_stat value: 1.2896282719661123
*Two-sample t-test p-value= 0.2017496571835306

** Here p- value is greater than level of significance so we have enough evidence to reject the alternative hypothesis in favour of null hypothesis. **Hence, we can conclude that mean for shingles-A and singles-B are same.**

Problem 3:

Salary is hypothesized to depend on educational qualification and occupation. To understand the dependency, the salaries of 40 individuals **[SalaryData.csv]** are collected and each person's educational qualification and occupation are noted. Educational qualification is at three levels, High school graduate, Bachelor's, and Doctorate. Occupation is at four levels, Administrative and clerical, Sales, Professional or specialty, and Executive or managerial. A different number of observations are in each level of education – occupation combination.

[Assume that the data follows a normal distribution. In reality, the normality assumption may not always hold if the sample size is small.]

3.1 State the null and the alternate hypothesis for conducting one-way ANOVA for both Education and Occupation individually.

Ans.: **Null Hypotheses(H0):** The mean salary is the same with respect to all the 4 categories of Education.
Alternative Hypotheses (HA): The mean salary is different in at least one category of Education.

Null Hypotheses(H0): The mean salary is the same with respect to all the 4 categories of occupation.
Alternative Hypotheses (HA): The mean salary is different in at least one category of occupation.

3.2 Perform a one-way ANOVA on Salary with respect to Education. State whether the null hypothesis is accepted or rejected based on the ANOVA results.

Ans.:

	df	sum_sq	mean_sq	F	PR(>F)
C(Education)	2.0	1.026955e+11	5.134773e+10	30.95628	1.257709e-08
Residual	37.0	6.137256e+10	1.658718e+09	NaN	NaN

#Since the p-value is less than the significance level (0.05), we can reject the null hypothesis and conclude that there is a difference in the mean salaries for at-least one category of education.

3.3 Perform a one-way ANOVA on Salary with respect to Occupation. State whether the null hypothesis is accepted or rejected based on the ANOVA results.

Ans.:

	df	sum_sq	mean_sq	F	PR(>F)
C(Occupation)	3.0	1.125878e+10	3.752928e+09	0.884144	0.458508
Residual	36.0	1.528092e+11	4.244701e+09	NaN	NaN

#Since the p-value is greater than the significance level (0.05), we can reject the Alternative Hypothesis and conclude that there is no difference in the mean salaries across the 4 categories of occupation.

3.4 If the null hypothesis is rejected in either (2) or in (3), find out which class means are significantly different. Interpret the result. **(Non-Graded)**

3.5 Perform a two-way ANOVA based on Salary with respect to both Education and Occupation (along with their interaction Education*Occupation). State the null and alternative hypotheses and state your results. How will you interpret this result?

Ans.: **#Null Hypotheses (H0):** The mean salary with respect to each education category and occupation is equal.

#Alternative Hypotheses (HA): At least one of the means of salary with respect to each education category and occupation is unequal.

	df	sum_sq	mean_sq	F	\
C(Education)	2.0	1.026955e+11	5.134773e+10	72.211958	
C(Occupation)	3.0	5.519946e+09	1.839982e+09	2.587626	
C(Education):C(Occupation)	6.0	3.634909e+10	6.058182e+09	8.519815	
Residual	29.0	2.062102e+10	7.110697e+08	NaN	

	PR(>F)
C(Education)	5.466264e-12
C(Occupation)	7.211580e-02
C(Education):C(Occupation)	2.232500e-05
Residual	NaN

#Since the p-value is lesser than the significance level (0.05), so we will reject the null hypothesis and conclude that there is significant amount of interaction between the variables (Education and Occupation) and at least one of the means of the salary with respect to each education category and occupation is unequal.

3.6 Explain the business implications of performing ANOVA for this particular case study.

Ans.: From the ANOVA table we can implies:

- From the ANOVA table we can see the p-value for Education is = 5.466264e-12, which is less than 0.05, hence we can reject null-hypothesis and can conclude that the mean salary for different level of education is different. There is a significant effect due to Education.
- From the ANOVA table we can see the p-value for Occupation is = 7.211580e-02, which is greater than 0.05, hence we fail to reject null-hypothesis and can conclude that the mean salary for different level of Occupation is same. There is no significant effect due to Occupation.
- From the ANOVA table we can see the p-value for Education & Occupation is = 2.232500e-05, which is less than 0.05, hence we reject null-hypothesis and can conclude that There is an interaction effect between Education & Occupation.