SMDM PROJECT REPORT

DSBA

greatlearning
Learning for Life

CONTENTS:

Problem 1:

Problem 2:

Problem 3:

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 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).

1.1.1 Use methods of descriptive statistics to summarize data

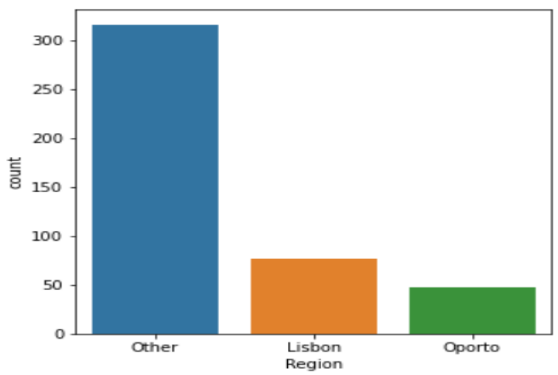
Summarized Data:

	Buyer/Spender	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
count	440.000000	440.000000	440.000000	440.000000	440.000000	440.000000	440.000000
mean	220.500000	12000.297727	5796.265909	7951.277273	3071.931818	2881.493182	1524.870455
std	127.161315	12647.328865	7380.377175	9503.162829	4854.673333	4767.854448	2820.105937
min	1.000000	3.000000	55.000000	3.000000	25.000000	3.000000	3.000000
25%	110.750000	3127.750000	1533.000000	2153.000000	742.250000	256.750000	408.250000
50%	220.500000	8504.000000	3627.000000	4755.500000	1526.000000	816.500000	965.500000
75%	330.250000	16933.750000	7190.250000	10655.750000	3554.250000	3922.000000	1820.250000
max	440.000000	112151.000000	73498.000000	92780.000000	60869.000000	40827.000000	47943.000000

1.1.2 Which Region and which Channel spent the most?

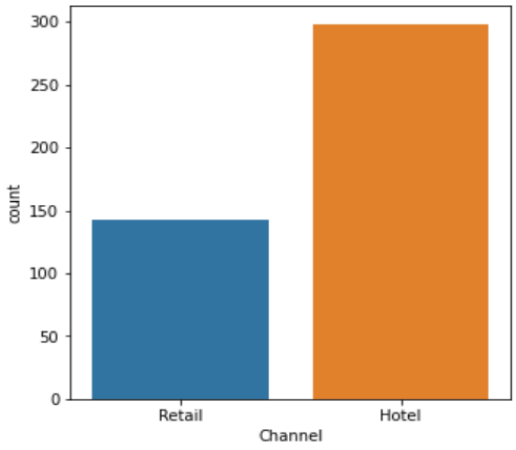
1.1.3 Which Region and which Channel spent the least?

Region:



Region	Buyer/Spender	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen	Total
Lisbon	18095	854833	422454	570037	231026	204136	104327	2386813
Oporto	14899	464721	239144	433274	190132	173311	54506	1555088
Other	64026	3960577	1888759	2495251	930492	890410	512110	10677599

Channel:



Channel	Buyer/Spender	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen	Total
Hotel	71034	4015717	1028614	1180717	1116979	235587	421955	7999569
Retail	25986	1264414	1521743	2317845	234671	1032270	248988	6619931

Conclusion:

- Region and Channel spent the most: Other (10677599) and Hotel (7999569)
- Region and Channel spent the least: Oporto (1555088) and Retail (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.

Region: Mean

Region	Buyer/Spender	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
Lisbon	235.000000	11101.727273	5486.415584	7403.077922	3000.337662	2651.116883	1354.896104
Oporto	317.000000	9887.680851	5088.170213	9218.595745	4045.361702	3687.468085	1159.702128
Other	202.613924	12533.471519	5977.085443	7896.363924	2944.594937	2817.753165	1620.601266

Region: Standard deviation

Region	Buyer/Spender	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
Lisbon	22.371857	11557.438575	5704.856079	8496.287728	3092.143894	4208.462708	1345.423340
Oporto	13.711309	8387.899211	5826.343145	10842.745314	9151.784954	6514.717668	1050.739841
Other	143.615303	13389.213115	7935.463443	9537.287778	4260.126243	4593.051613	3232.581660

Region: Coefficient of variation

Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
Lisbon	1.041049	1.039815	1.147670	1.030599	1.587430	0.993008
Oporto	0.848318	1.145076	1.176182	2.262291	1.766718	0.906043
Other	1.068277	1.327648	1.207808	1.446761	1.630040	1.994680

Channel: Mean

Channel	Buyer/Spender	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
Hotel	238.369128	13475.560403	3451.724832	3962.137584	3748.251678	790.560403	1415.956376
Retail	183.000000	8904.323944	10716.500000	16322.852113	1652.612676	7269.507042	1753.436620

Channel: Standard deviation

Channel	Buyer/Spender	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
Hotel	120.910343	13831.687502	4352.165571	3545.513391	5643.912500	1104.093673	3147.426922
Retail	132.136132	8987.714750	9679.631351	12267.318094	1812.803662	6291.089697	1953.797047

Channel: Coefficient of variation

Channel	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
Hotel	1.026428	1.260867	0.894849	1.505745	1.396596	2.222828
Retail	1.009365	0.903246	0.751543	1.096932	0.865408	1.114267

Skewness:

	Skewness
Fresh	2.552583
Milk	4.039922
Grocery	3.575187
Frozen	5.887826
Detergents_Paper	3.619458
Delicatessen	11.113534

Conclusion:

From the above observations by considering mean, std, CV, skewness we can see that the behaviour in all items (Channel, Region) are different.

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

Finding CV for each item:

Fresh	1.0527196084948245
Milk	1.2718508307424503
Grocery	1.193815447749267
Frozen	1.5785355298607762
Detergents paper	1.6527657881041729
Delicatessen	1.8473041039189306

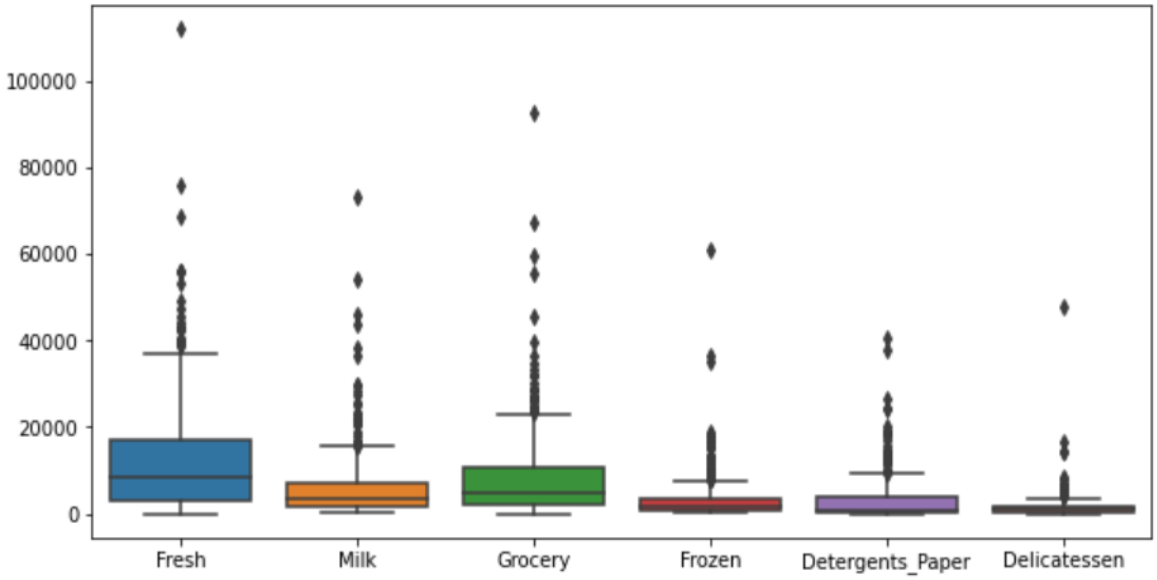
Conclusion:

While comparing among variables whose means showed wide differences. Taken the coefficient of variation as a descriptive measure, because it normalizes the standard deviation with respect to the mean.

- **Fresh items** have the lowest coefficient of variation. Hence, it is **least** inconsistent behaviour.

- **Delicatessen** has the highest coefficient of variation. Hence, it is the **most** inconsistent behaviour.

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



Conclusion:

There are outliers within a data set in all the items. we can see that a data point that is located outside the whiskers of the box plot.

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

Correlation table:

	Buyer/Spender	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
Buyer/Spender	1.000000	-0.061151	-0.162290	-0.140509	0.053802	-0.134365	-0.101845
Fresh	-0.061151	1.000000	0.100510	-0.011854	0.345881	-0.101953	0.244690
Milk	-0.162290	0.100510	1.000000	0.728335	0.123994	0.661816	0.406368
Grocery	-0.140509	-0.011854	0.728335	1.000000	-0.040193	0.924641	0.205497
Frozen	0.053802	0.345881	0.123994	-0.040193	1.000000	-0.131525	0.390947
Detergents_Paper	-0.134365	-0.101953	0.661816	0.924641	-0.131525	1.000000	0.069291
Delicatessen	-0.101845	0.244690	0.406368	0.205497	0.390947	0.069291	1.000000

Analysis:

- Considering above data, we can say that there is a strong correlation between Detergents_paper and Grocery (0.924641).
- There are inconsistencies in spending of different items, which should be minimized.
- Hotel and other gets maximum expenditure.
- Hotel channel expenditure on the Fresh items is higher in almost every region.
- Frozen and Delicatessen are the least contributors of expenditure in retail channel.
- Other regions have maximum expenditures in both the channels.
- Grocery items are more in demand at retail channel
- Behaviour in all items (Channel, Region) are different.
- Customer spends more money on Detergents_paper and Grocery products. Spent should be equal for different regions
- Hotel Channels must be established in Lisbon, Oporto region.

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).

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

2.1.1. Gender and Major

Major Gender	Accounting	CIS	Economics/Finance	International Business	Management	Other	Retailing/Marketing	Undecided	All
Female	3	3	7	4	4	3	9	0	33
Male	4	1	4	2	6	4	5	3	29
All	7	4	11	6	10	7	14	3	62

2.1.2. Gender and Grad Intention

Grad Intention Gender	No	Undecided	Yes	All
Female	9	13	11	33
Male	3	9	17	29
All	12	22	28	62

2.1.3. Gender and Employment

Employment Gender	Full-Time	Part-Time	Unemployed	All
Female	3	24	6	33
Male	7	19	3	29
All	10	43	9	62

2.1.4. Gender and Computer

Computer	Desktop	Laptop	Tablet	All
Gender				
Female	2	29	2	33
Male	3	26	0	29
All	5	55	2	62

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?

Here the sum of male and female is 62.
Total number of males is 29 and the probability that a randomly selected CMSU student will be male: P(Male)=29/62 **i.e. 46.77 %**

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

Here the sum of male and female is 62.
Total number of females is 33 and the probability that a randomly selected CMSU student will be female: P(Female)=33/62 **i.e. 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.

P(Accounting/male) = 4/29 = 13.79%
P(CIS/male) = 1/29 = 3.45%
P(Economics/Finance/male) = 4/29 = 13.79%
P(International Business/male) = 2/29 = 6.9%
P(Management/male) = 6/29 = 20.69%
P(Other/male) = 4/29 = 13.79%
P(Retailing/Marketing/male) = 5/29 = 17.24%
P(undecided/male) = 3/29 = 10.34%

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

P(Accounting/ female) = 3/33 = 9.09%
P(CIS/ female) = 3/33 = 9.09%
P(Economics/Finance/ female) = 7/33 = 21.21%
P(International Business/ female) = 4/33 = 12.12%
P(Management/ female) = 4/33 = 12.12%
P(Other/ female) = 3/33 = 9.09%
P(Retailing/Marketing/ female) = 9/33 = 27.27%
P(undecided/ female) = 0/33 = 0.0%

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.

P(males intends to graduate/total students) = 17/62 = 27.42%

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

P(females doesn't have a laptop/total students) = 4/62 = 6.45%

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?

Total number of students = 62

No. of male students = 29

No. of fulltime students = 10

No. of male fulltime students = 7

(male full time/total no. of students) => p [(A U B / C)] =p(A/C) + p(B/C) – p [(A n B)/(C)]

p(C)=total number of students

p(A)= no. of male students

p(B)= no. of fulltime students

p (A n B) = no. of male fulltime students

(29/62) + (10/62) - (7/62) = **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.

As every student is majoring in one subject all these are independent events

So, probability that given a female student who has majored in international business or management is:

P [(IB U M / F)] = p(IB/F) +p(M/F)-p [(IB n M)/(F)]

As they are independent events, we can consider p [(IB n M)/(F)] as "0"

P(IB/F) + P(M/F) = 4/33+4/33 = **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

Do you think graduate intention and being female are independent events?

Ans: No

Explanation: (The Events are dependent)

Total number of students = 40

No. of female students = 20

No. of students who intend to graduate = 28

No. of female students who intend to graduate = 11

(male full time/total males) => p [(A U B / C)] =p(A/C) + p(B/C) – p [(A n B)/(C)]

p(C) = Total number of students

p(A) = No. of female students

p(B) = No. of graduate intent students

p (A n B) = No. of female students who intend to graduate

(20/40) + (28/40) - (11/40) = **92.5%**

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

GPA

GPA	2.3	2.4	2.5	2.6	2.8	2.9	3.0	3.1	3.2	3.3	3.4	3.5	3.6	3.7	3.8	3.9	All
Gender																	
Female	1	1	2	0	1	3	5	2	4	3	2	4	1	2	1	1	33
Male	0	0	4	2	2	1	2	5	2	2	5	2	2	0	0	0	29
All	1	1	6	2	3	4	7	7	6	5	7	6	3	2	1	1	62

Salary

Salary	25.0	30.0	35.0	37.0	37.5	40.0	42.0	45.0	47.0	47.5	50.0	52.0	54.0	55.0	60.0	65.0	70.0	78.0	80.0	All
Gender																				
Female	0	5	1	0	1	5	1	1	0	1	5	0	0	5	5	0	1	1	1	33
Male	1	0	1	1	0	7	0	4	1	0	4	1	1	3	3	1	0	0	1	29
All	1	5	2	1	1	12	1	5	1	1	9	1	1	8	8	1	1	1	2	62

Spending

Spending	100	200	220	250	300	350	360	375	400	450	...	600	650	680	690	700	900	1000	1100	1400	All
Gender																					
Female	1	3	1	1	5	3	0	0	1	2	...	3	2	0	1	1	1	1	0	0	33
Male	0	1	0	1	3	0	1	1	4	0	...	6	0	1	0	0	0	0	1	1	29
All	1	4	1	2	8	3	1	1	5	2	...	9	2	1	1	1	1	1	1	1	62

Text Messages.

Text Messages	0	10	30	35	40	45	50	60	70	100	...	300	350	400	500	600	700	750	800	900	All
Gender																					
Female	1	1	1	1	0	0	4	0	0	4	...	5	2	1	2	1	1	0	0	1	33
Male	0	0	0	0	1	1	2	1	1	4	...	5	0	1	1	1	1	1	1	0	29
All	1	1	1	1	1	1	6	1	1	8	...	10	2	2	3	2	2	1	1	1	62

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

Number of students having < 3 = (1+1+6+2+3+4) i.e 17

Total number of students = 62

Probability that his/her GPA is less than 3 = 17/62 i.e. **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.

Number of **male** students earns >= 50 => (4+1+1+3+3+1+1) i.e 14

Total number of male students = 29

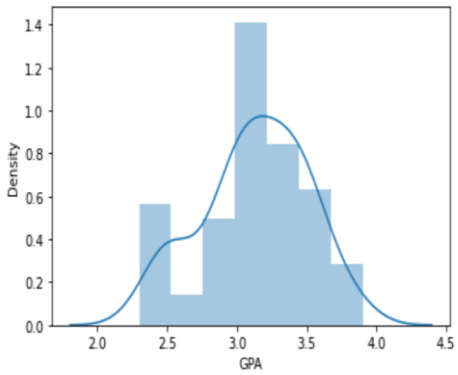
Probability that a randomly selected male earns 50 or more = 14/29 i.e **48.28 %**

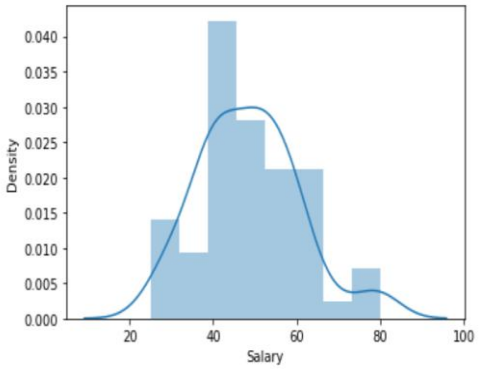
Number of **female** students earns >= 50 => (5+5+5+1+1+1) i.e 18

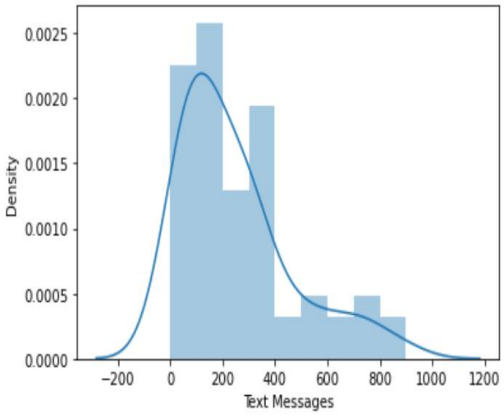
Total number of female students = 33

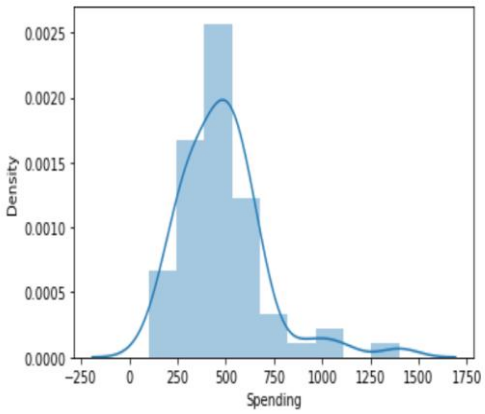
Probability that a randomly selected female earns 50 or more = 18/33 i.e **54.55 %**

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 for this whole Problem 2.









	count	mean	std	min	25%	50%	75%	max
GPA	62.0	3.129032	0.377388	2.3	2.9	3.15	3.4	3.9
Salary	62.0	48.548387	12.080912	25.0	40.0	50.00	55.0	80.0
Spending	62.0	482.016129	221.953805	100.0	312.5	500.00	600.0	1400.0
Text Messages	62.0	246.209677	214.465950	0.0	100.0	200.00	300.0	900.0

All the numerical continuous variable seems to be almost normally distributed.

But, GPA having Normal distribution, salary (slightly right skewed), spending (right skewed), text messages (right skewed)

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.

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.

SAMPLE A: Population std is unknown. Sample size is > 30.Opting for T-Test

Step 1: Defining null and alternative hypothesis

H0: μA <= 0.35

HA: μA > 0.35

Step 2: Level of significance

Alpha is not given in the question. Taking 0.05 as significance level

α = 0.05

Step 3: calculating p-value and test statistic

The sample size for this problem is 36

- One sample t-test :

  t = (Xbar-mu)/(s/sqrt n)

  Xbar = Avg of given sample

  mu = 0.35

  s = std of given sample

  n = 36

- test statistic = -1.4735046253382782

- P-value = TDIST (test statistic,dof,2); {dof is n-1 i.e 35}

final P-value = 0.07477633144907513

- ttest_1samp (Gives t statistic and 2 tailed p value .So, we are dividing by 2 to convert into one tail)

Step 4: Accepting or rejecting null hypothesis

We have no evidence to reject the null hypothesis since p value > Level of significance.Our one-sample t-test p-value= [0.07477633144907513]

Sample B: Population std is unknown. Sample size is > 30. Opting for T-Test

Step 1: Defining null and alternative hypothesis

H0: μB <= 0.35

HA: μB > 0.35

Step 2: Level of significance

Alpha is not given in the question. Taking 0.05 as significance level

α = 0.05

Step 3: calculating p-value and test statistic

The sample size for this problem is 31

- One sample t-test :

t = (Xbar-mu)/(s/sqrt n)

Xbar = Avg of given sample

mu = 0.35

s = std of given sample

n = 31

- test statistic = -3.1003313069986995

- P-value = TDIST (test statistic,dof,2); {dof is n-1 i.e 30}

final P-value = 0.0020904774003191826

- ttest_1samp (Gives t statistic and 2 tailed p value .So, we are dividing by 2 to convert into one tail)

Step 4: Accepting or rejecting null hypothesis

We have evidence to reject the null hypothesis since p value < Level of significance.Our one-sample t-test p-value= [0.0020904774003191826]

- Shingles of type A are with in permissible limits
- Shingles of type B are not with in permissible limits

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?

Assumption before the test for equality of means:

Computing difference between two group means we use t-test

Here we have two samples (unpaired independent)

Check whether the two samples having equal or unequal variance.

1st we calculate variance for both groups by using F-statistic formula to check whether the variances are equal or not.

F = Larger variance/Smaller variance

If the resulted values are closer to 1 will indicate equal variance between two samples and values that deviate from 1 indicate unequal variance.

In our case F-statistic values is 1.02 which indicates equal variance between two samples.

Using t-Test: Two-Sample Assuming Equal Variances:

	Variable 1	*Variable 2*
Mean	0.316666667	0.273548387
Variance	0.018422857	0.018850323
Observations	36	31
Pooled Variance	0.018620149	
Hypothesized Mean Difference	0	
df	65	
t Stat	1.289628272	
P(T<=t) one-tail	0.100874829	
t Critical one-tail	1.668635976	
P(T<=t) two-tail	0.201749657	
t Critical two-tail	1.997137908	

Step 1: Defining null and alternative hypothesis

H0: μA - μB = 0

HA: μA - μB != 0

Step 2: Level of significance

Alpha is not given in the question. Taking 0.05 as significance level

α = 0.05

Step 3: calculating p-value and test statistic

- two sample t-test :

t = (X1bar-X2bar)/(sqrt s1^2/n1)+(sqrt s2^2/n2)

X1bar = Avg of given sample 1 : 0.316666667

X2bar = Avg of given sample 2 : 0.273548387

s1 = variance of given sample 1 : 0.018422857

s2 = variance of given sample 2 : 0.018850323

n1 = 36, n2=31

df = 65

- using ttest_ind(for means of two independent samples from observations)

Step 4: Accepting or rejecting null hypothesis

We have no evidence to reject the null hypothesis since p value > Level of significance

Our two-sample t-test p-value= [0.2017496571835306].

Hence, it is clear that the population mean for shingles A and B are equal