STATISTICS FOR COMPUTER SCIENCE PROJECT

GROUP-15:-

S.Ajay Kumar 19BCS101
Bokka Srikar 19BCS024
Bhanupriya T P 19BCS022
Dhyan M G 19BCS038

About Data :

The data set marketing_data.csv consists of 2,240 customers of XYZ company with data on:

- Customer profiles
- Product preferences
- Campaign successes/failures
- Channel performance

Columns Details:

ID :Customer's unique identifier

Year_Birth: Customer's birth year

Education: Customer's education level

Marital_Status: Customer's marital status.

Income: Customer's yearly household income

Kidhome: Number of children in customer's household

Teenhome: Number of teenagers in customer's household

Dt_Customer: Date of customer's enrollment with the company.

Recency: Number of days since customer's last purchase.

MntWines: Amount spent on wine in the last 2 years.

MntFruits: Amount spent on fruits in the last 2 years

MntMeatProducts: Amount spent on meat in the last 2 years

MntFishProducts: Amount spent on fish in the last 2 years

MntSweetProducts: Amount spent on sweets in the last 2 years

MntGoldProds: Amount spent on gold in the last 2 years.

NumDealsPurchases: Number of purchases made with a discount.

NumWebPurchases: Number of purchases made through the company's website.

NumCatalogPurchases: Number of purchases made using a catalogue.

NumStorePurchases: Number of purchases made directly in stores.

NumWebVisitsMonth: Number of visits to the company's web site in the last month.

AcceptedCmp1: 1 if customer accepted the offer in the 1st campaign, 0 otherwise

AcceptedCmp2: 1 if customer accepted the offer in the 2nd campaign, 0 otherwise.

AcceptedCmp3: 1 if customer accepted the offer in the 3rd campaign, 0 otherwise.

AcceptedCmp4: 1 if customer accepted the offer in the 4th campaign, 0 otherwise.

AcceptedCmp5: 1 if customer accepted the offer in the 5th campaign, 0 otherwise

Response: 1 if customer accepted the offer in the last campaign, 0 otherwise

Complain: 1 if customer complained in the last 2 years, 0 otherwise

Country: Customer's location

The Data is cleaned i.e the outliers are removed from the data and null values are removed orreplaced with median. So that data will not affect the statistical values.

Categorical data analysis:

Categorical data analysis is the analysis of data where the response variable has been grouped into a set of mutually exclusive ordered (such as age group) or unordered (such as eye color) categories.

Chi-Square goodness of fit:

The Chi-square goodness of fit test is a statistical hypothesis test used to determine whether a variable is likely to come from a specifieddistribution or not. It is often used to evaluate whether sample data is representative of the full population. Goodness of fit refers to how close the observed data are to those predicted from a hypothesis.

Question) Checking whether the campaigns are successful thenumber of people purchased after 5 campaigns are collected from data. The company thinks thecampaign is successful only if the there are atleast 140 customers purchase the products fromtheir company in every campaign.

Hypothesis:

H0 = "The campaign is Successful. It is reached to expectedcustomers."
H1 = "The campaign is not Successful. It isn't reached to expectedcustomers."

CONCLUSION :

P-value is = 1.617168209072961e-20
Chi-Square value is = 98.98571428571428
Degrees of Freedom DF = 4
The campaign is not Successful. It isn't reached to expected customers.

Chi-Square contingency test of independence:

The Chi-square test of independence is a statistical hypothesis test used to determine whether two categorical or nominal variables are likely to be related or not. The test statistic for the Chi-Square Test of chosen confidence level. If the calculated X^2 value > critical X^2 value, then we reject the null hypothesis.

Q) Do customers single and married spent different ways in purchasing products like (wine, fruits, meat, fish, sweet, Gold)

Hypothesis

H0 = " The Amount spent on different products is Independent ofMarital Status."
H1 = "The Amount spent on different products is Dependent ofMarital Status."

CONCLUSION :

Statistic value is = 0.1735704803722132P-
Value is = 0.9993723879228041
Degrees of Freedom = 5
The Amount spent on differnet products is Independent of MaritalStatus.

PARAMETRIC HYPOTHESIS TESTS:

The One-Sample t-Test:

The one-sample t-test is a statistical hypothesis test used to determine whether an unknown population mean is different from aspecific value.
The one sample t test compares the mean of your sample data to a known value. For example, you might want to know how your samplemean compares to the population mean.
The average age of Company customers is 52 at alpha = 0.05.

Hypothesis

H0 = " The average age of Customers is 52."
H1 = " The Average age of customers is not equal to 52 "

Conclusion:

Avg age in a sample is = 50.13793103448276
Statistic test value is = -0.9465015739114689
P-value is = 0.3519901616647836
The Average age of customers is not equal to 52

Independent two sample t-Test:

The independent t-test, also called the two sample t-test, independent-samples t-test or student's t-test, is an inferential statistical test that determines whether there is a statistically significant difference between the means in two unrelated groups.

Q) Does the product type hinders the purchasing of products between individuals?

Hypothesis

H0 = "μ1 > μ2 Mean Amount spent on FISH products is less than orequal to MEAT."

H1 = "μ1 <= μ2 Mean Amount spent on Fish products is greaterthan or equal to MEAT."

Result :

Sp = 40.92167314924126

Test Statistic Value is = 33.45141976637732

CONCLUSION :

μ1 > μ2 Mean Amount spent on FISH products is less than or equal toMEAT.

Dependent two sample t-Test:

The dependent t-test (also called the paired t-test or paired-samples t-test) compares the means of two related groups to determine whether there is a statistically significant difference between these means.

The Vegeterian people are Spending equal amount on Sweet and Fruitswhich are contibuting more for total purchases.

Does the amount spent on Fruits and Sweets are dependent.

Hypothesis

H0 = "μD = 0 No differnece in Amount spent on Sweet products and Fruits."

H1 = "μD != 0 There is difference in Amount spent on Sweetproducts and Fruits.

CONCLUSION :

Sp = 40.92167314924126

Test Statistic Value is = -0.24801785651577193

μD != 0 There is difference in Amount spent on Sweet products andFuits.

One-sample z-test:

The one-sample z-test is used to test whether the mean of apopulation is greater than, less than, or not equal to a
specific value.

Q) Avg number of web purchases is greater than3.

Hypothesis:

H0 = "μ <= 3 Average number of web purchases is LESS than or equalto 3."
H1 = "μ > 3 Average number of web purchases is GREATER than orequal to 3."

CONCLUSION :

Z-Value is = 18.4999262909087
Reject μ <= 3 Average number of web purchases is LESS than or equal to 3.

Two-Sample Z-Test:

A Z-test is a type of hypothesis test—a way for you to figure out ifresults from a test are valid or repeatable.
The Two-Sample Z-test is used to compare the means of two samplesto see if it is feasible that they come from the same population. The null hypothesis is: the population means are equal.

Average amount spent on wines is greater for Singles than Married.

Hypothesis

H0 = "μ1 <= μ2 Average amount spent on wines is less than AverageAmount spent on Gold."
H1 = "μ1 > μ2 Average amount spent on wines is greater than AverageAmount spent on Gold." (Claim)

CONCLUSION :

Z-Value is = 0.35559057941777744
14.882854294937603
μ1 <= μ2 Average amount spent on wines is less than Average Amount spent on Gold.

One sample z-test for proportions:

The One proportion Z-test is used to compare an observed proportion to a theoretical one, when there are only two categories. Itcompares the proportion to a target or reference value and also calculates a range of values that is likely to include the population proportion. This is also called hypothesis of inequality.

There are 20% of customers who accepted the offer atleastonce.Checking if it is true

Hypothesis

H0 = "p = 0.2 There are 20% of customers who accepted the offeratleast once"
H1 = "p != 0.2 The customers who accepted the offer atleast once isnot equal to 0.2."

CONCLUSION :

Z-Value is = 0.771720207508393
p = 0.2 There are 20% of customers who accepted the offer atleastonce

Two Sample z-test for proportions:

The purpose of two sample Z test is to compare the random samples of two populations. Use two sample z test of proportion for large sample size and Fisher exact probability test is an excellent non-parametric test for small sample sizes.
There is difference in proportions of accepting Campaign offersbetween singles and married

Hypothesis

H0 = "P1-P2 = 0 There is no difference in proportions of acceptingCampaign offers between singles and married"
H1 = "P1-P2 != 0 There is difference in proportions of acceptingCampaign offers between singles and married"

CONCLUSION :

Z-Value is = -5.661516235110674

P1-P2 != 0 There is difference in proportions of accepting Campaign offers between singles and married

NON PARAMETRIC HYPOTHESIS TESTS:

Runs test:

A runs test is a statistical procedure that examines whether a string of data is occurring randomly from a specific distribution. The runs test analyzes the occurrence of similar events that are separatedby events that are different.

Hypothesis

H0 : The data is produced in a random manner based on Income

H1 : The data is not produced in a random manner based on Income

CONCLUSION :

----> As elements n>=30 So we are using Z -test for comparison

----> A z-score of less than 0 represents an element less than themedian.

----> Z stat value doesnot fall in rejection region Z-value < Z - tablevalue

----> Reject H0

Z- value = -5.853544528527859

P- value = 4.8120533902742105e-09

Sign Test One Sample:

The sign test is a statistical method to test for consistent differences between pairs of observations, such as the weight of subjects before and after treatment. Given pairs of observations (such as weight pre- and post-treatment) for each subject, the sign test determines if one member of the pair (such as pre-treatment) tends tobe greater than (or less than) the other member of the pair (such as post-treatment).

Q) Number of median purchases assumed by company is 15

Hypothesis

H0 : median = 15

H1 : Median != 15 alpha =0.05

Condition: Sample size< 26 Critical value for two tailed test at n=25 is6 Give critical value as Input 6

Test Value is = 7

Test Value is greater than critical Value. 7 > 6

CONCLUSION :

--->Failed to reject Null Hypothesis Median is equal to 15

Mann Whitney U Test:

The Mann-Whitney U test is used to compare differences between two independent groups when the dependent variable iseither ordinal or continuous, but not normally distributed.

Checking Whether there is differnece between amount spent onproducts among singles and married

Hypothesis

H0: There is no difference on amount spent on purchasing productsbetween singles and Married

H1: There is difference on amount spent on purchasing productsbetween singles and Married

Result:

Test Statistic U1 is = 563156.5Test

Statistic U2 is = 582585.5

manwhitneyu p value is = 0.2531744435010702P-value is = 0.5063504962214381

Z- Value is = -0.6645310290084294

Z-table value = 1.6448536269514722Test

Value is greater than critical Value.

-0.6645310290084294 > -1.6448536269514722

CONCLUSION :

--->Failed to reject Null Hypothesis

There is no difference on amount spent on purchasing productsbetween singles and Married

Wilcoxon Signed Rank Test:

The Wilcoxon signed rank test should be used if the differencesbetween pairs of data are non-normally distributed.
Check whether the customer who purchases meat also purchases fishNon vegeteraian

Hypothesis:

H0: There is no difference among customers purchasing fish and meat.
H1: There is difference among customers purchasing fish and meat.

Result:

Test statistic value is = 5.0
P-Value is = 2.8716584471854804e-06
Interfer the result or Give critical value for n=30 at alpha =0.05 fromthe table 136
Test Value is less than critical Value.5.0
<= 136

CONCLUSION :

--->Null Hypothesis is Rejected
There is difference among customers purchasing fish and meat.

NOTE:

Some of the Samples in code are drawn randomly by the system so Output may change if yourun the code. Statistic values may change when you run.