**Name: - UPENDRA DAMA**

**Batch: - 25/01/2020 (Weekend)**

**Module: - 5**

A F&B manager wants to determine whether there is any significant difference in the diameter of the cutlet between two units. A randomly selected sample of cutlets was collected from both units and measured? Analyze the data and draw inferences at 5% significance level. Please state the assumptions and tests that you carried out to check validity of the assumptions.

File : **Cutlets.csv**

**Solution: -**

1. **Problem Statement**

Is there any difference in diameter of cutlet between two units?

1. **Business Problem**

Should two units produce cutlets with same diameter?

1. **Data**

Here Y is diameter which is continuous.

X is 2 samples (unit A and unit B)

1. **Normality Test**

***Let’s create Hypothesis test for Unit A***

H0 = Data are normally distributed

Ha = Data are not normally distributed

P value = 0.32.

Alpha = 5% = 0.05

P value > Alpha

P high H0 fly - > Fail to reject H0. i.e. we will go with H0.

H0 = Data are normal.

***Let’s create Hypothesis test for Unit B***

H0 = Data are normally distributed

Ha = Data are not normally distributed

P value = 0.5225.

Alpha = 5% = 0.05

P value > Alpha

P high H0 fly - > Fail to reject H0. i.e. we will go with H0.

H0 = Data are normal.

1. **External conditions**

Not same

1. **Variance Test**

***Let’s create Hypothesis for two units***

H0 = Variance of diameters of Unit A is equal to the variance of diameters of Unit B

Ha = Variance of diameters of Unit A is not equal to the variance of diameters of Unit B

P value = 0.3136.

Alpha = 5% = 0.05

P value > Alpha

P high H0 fly - > Fail to reject H0. i.e. we will go with H0.

H0 = Variance of diameters of Unit A is equal to the variance of diameters of Unit B

1. **2-sample T Test**

Since we are assuming variances are equal we will go with 2-sample T test.

H0 = Averages of diameters of Unit A is equal to Averages of diameters of unit B -----> There is no significance difference between diameters of cutlets from unit A and Unit B.

Ha = Averages of diameters of Unit A is not equal to Averages of diameters of unit B --🡪 There is significance difference between diameters of cutlets from unit A and Unit B.

P value = 0.4723

P value > Alpha (0.05)

P high H0 fly - > Fail to reject H0. i.e. we will go with H0.

H0 = Averages of diameters of Unit A is equal to Averages of diameters of unit B

Hence, Inference is that **there is no significant difference in the diameters of cutlets from Unit A and Unit B**

|  |
| --- |
| **R Code** |
| > library(readr)  > Cutlets <- read\_csv("Desktop/Digi 360/Module 5/Datasets-2/Cutlets.csv")  Parsed with column specification:  cols(  `Unit A` = col\_double(),  `Unit B` = col\_double()  )  > View(Cutlets)  > attach(Cutlets)  The following objects are masked from Cutlets (pos = 3):  Unit A, Unit B  ###Renaming the columns  > colnames(Cutlets) <- c("unit\_A","unit\_B")  > attach(Cutlets)  > attach(Cutlets)  The following objects are masked from Cutlets (pos = 3):  unit\_A, unit\_B  ###Normality Test  > shapiro.test(unit\_A)  Shapiro-Wilk normality test  data: unit\_A  W = 0.96495, p-value = 0.32  > shapiro.test(unit\_B)  Shapiro-Wilk normality test  data: unit\_B  W = 0.97273, p-value = 0.5225  ####Variance Test  > var.test(unit\_A, unit\_B)  F test to compare two variances  data: unit\_A and unit\_B  F = 0.70536, num df = 34, denom df = 34, p-value = 0.3136  alternative hypothesis: true ratio of variances is not equal to 1  95 percent confidence interval:  0.3560436 1.3974120  sample estimates:  ratio of variances  0.7053649  > # since variances are equal, will go with 2 sample T test  > t.test(unit\_A, unit\_B, alternative = "two.sided", conf.level = 0.95)  Welch Two Sample t-test  data: unit\_A and unit\_B  t = 0.72287, df = 66.029, p-value = 0.4723  alternative hypothesis: true difference in means is not equal to 0  95 percent confidence interval:  -0.09654633 0.20613490  sample estimates:  mean of x mean of y  7.019091 6.964297  **#### Since p value is greater than 0.05, fail to reject H0. There is no significant difference in the diameters of cutlets from Unit A and Unit B** |

|  |
| --- |
| **Python Code** |
| ####A F&B manager wants to determine whether there is any significant difference in the ##diameter of the cutlet between two units.  import pandas as pd  import numpy as np  import scipy  from scipy import stats  import statsmodels.api as sm  cutlet = pd.read\_csv ("~/desktop/Digi 360/Module 5/Datasets-2/Cutlets.csv")  cutlet.head()   |  | **Unit A** | **Unit B** | | --- | --- | --- | | 0 | 6.8090 | 6.7703 | | 1 | 6.4376 | 7.5093 | | 2 | 6.9157 | 6.7300 | | 3 | 7.3012 | 6.7878 | | 4 | 7.4488 | 7.1522 |   ###Rename Columns#######  cutlet1 = cutlet.rename(columns = {'Unit A':'unita', 'Unit B':'unitb'})  cutlet1.head()   |  | **unita** | **unitb** | | --- | --- | --- | | 0 | 6.8090 | 6.7703 | | 1 | 6.4376 | 7.5093 | | 2 | 6.9157 | 6.7300 | | 3 | 7.3012 | 6.7878 | | 4 | 7.4488 | 7.1522 |   ###Normality Test#############  a = pd.Series(cutlet1.unita).dropna()  print("p value for unit A:", stats.shapiro(a))  b = pd.Series(cutlet1.unitb).dropna()  print("p value for unit B:", stats.shapiro(b))  p value for unit A: (0.9649458527565002, 0.3199819028377533)  p value for unit B: (0.9727300405502319, 0.5224985480308533)  ####Variance test######  scipy.stats.levene(a,b)  LeveneResult(statistic=0.665089763863238, pvalue=0.4176162212502553)  # Here p value is greater than 0.05. So, fail to reject H0. That means variances are ##equal. Since variances are equal, will go with 2 sample T test  stats.ttest\_ind(a,b)  Ttest\_indResult(statistic=0.7228688704678061, pvalue=0.4722394724599501)  **#### Since p value is greater than 0.05, fail to reject H0. There is no significant difference in the diameters of cutlets from Unit A and Unit B** |

A hospital wants to determine whether there is any difference in the average Turn Around Time (TAT) of reports of the laboratories on their preferred list. They collected a random sample and recorded TAT for reports of 4 laboratories. TAT is defined as sample collected to report dispatch.

Analyze the data and determine whether there is any difference in average TAT among the different laboratories at 5% significance level.

File: **LabTAT.csv**

**Solution: -**

1. **Problem Statement**

Is there any difference in average turnaround time of reports of 4 labs?

1. **Business Problem**

Should all 4 labs dispatch reports within same turnaround time?

1. **Data**

Here Y is time which is continuous.

X is 4 samples (4 laboratories)

1. **Normality Test**

***Let’s create Hypothesis test for Lab1***

H0 = Data are normally distributed

Ha = Data are not normally distributed

P value = 0.5508.

Alpha = 5% = 0.05

P value > Alpha

P high H0 fly - > Fail to reject H0. i.e. we will go with H0.

H0 = Data are normal.

***Let’s create Hypothesis test for Lab2***

H0 = Data are normally distributed

Ha = Data are not normally distributed

P value = 0.8637.

Alpha = 5% = 0.05

P value > Alpha

P high H0 fly - > Fail to reject H0. i.e. we will go with H0.

H0 = Data are normal.

***Let’s create Hypothesis test for Lab3***

H0 = Data are normally distributed

Ha = Data are not normally distributed

P value = 0.4205.

Alpha = 5% = 0.05

P value > Alpha

P high H0 fly - > Fail to reject H0. i.e. we will go with H0.

H0 = Data are normal.

***Let’s create Hypothesis test for Lab4***

H0 = Data are normally distributed

Ha = Data are not normally distributed

P value = 0.6619.

Alpha = 5% = 0.05

P value > Alpha

P high H0 fly - > Fail to reject H0. i.e. we will go with H0.

H0 = Data are normal.

1. **External conditions**

Not same

1. **Variance Test**

***Let’s create Hypothesis for lab 1 and lab 2***

H0 = Variance of TAT of Lab 1 is equal to variance of TAT of Lab 2

Ha = Variance of TAT of Lab 1 is not equal to variance of TAT of Lab 2

P value = 0.1675

Alpha = 5% = 0.05

P value > Alpha

P high H0 fly - > Fail to reject H0. i.e. we will go with H0.

H0 = Variance of TAT of Lab 1 is equal to variance of TAT of Lab 2

***Let’s create Hypothesis for lab 2 and lab 3***

H0 = Variance of TAT of Lab 2 is equal to variance of TAT of Lab 3

Ha = Variance of TAT of Lab 2 is not equal to variance of TAT of Lab 3

P value = 0.2742.

Alpha = 5% = 0.05

P value > Alpha

P high H0 fly - > Fail to reject H0. i.e. we will go with H0.

H0 = Variance of TAT of Lab 2 is equal to variance of TAT of Lab 3

***Let’s create Hypothesis for lab 3 and lab 4***

H0 = Variance of TAT of Lab 3 is equal to variance of TAT of Lab 4

Ha = Variance of TAT of Lab 3 is not equal to variance of TAT of Lab 4

P value = 0.3168.

Alpha = 5% = 0.05

P value > Alpha

P high H0 fly - > Fail to reject H0. i.e. we will go with H0.

H0 = Variance of TAT of Lab 3 is equal to variance of TAT of Lab 4

***Let’s create Hypothesis for lab 1 and lab 4***

H0 = Variance of TAT of Lab 1 is equal to variance of TAT of Lab 4

Ha = Variance of TAT of Lab 1 is not equal to variance of TAT of Lab 4

P value = 0.1408.

Alpha = 5% = 0.05

P value > Alpha

P high H0 fly - > Fail to reject H0. i.e. we will go with H0.

H0 = Variance of TAT of Lab 1 is equal to variance of TAT of Lab 4

1. **One way ANOVA Test**

Since we are assuming variances are equal we will go with One way ANOVA test.

H0 = Average TAT for all the samples is same -----> There is no significance difference in average turnaround time of reports of 4 labs

Ha = Average TAT for all the samples is same --🡪 There is significance difference between in average turnaround time of reports of 4 labs

P value of lab 2 wrt to lab1 = 0.166299

P value of lab 3 wrt to lab1 = 0.277335

P value of lab 4 wrt to lab1 = 0.215323

P value > Alpha (0.05)

P high H0 fly - > Fail to reject H0. i.e. we will go with H0.

H0 = Average TAT for all the samples is same.

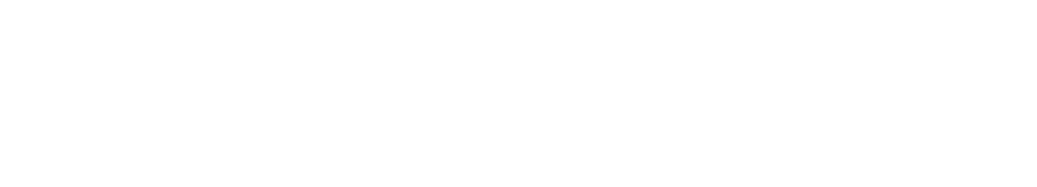
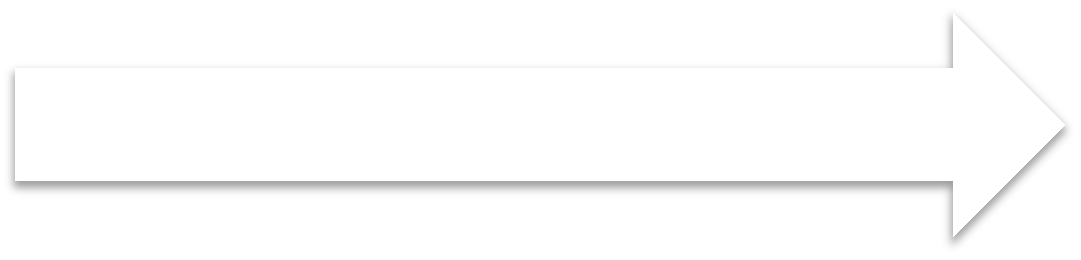
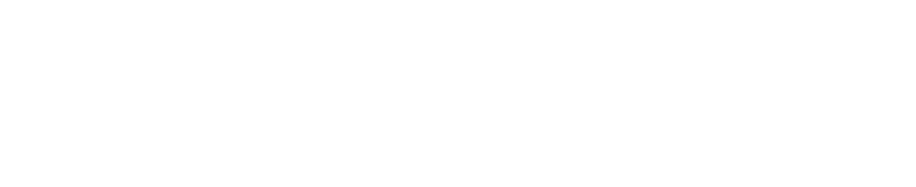
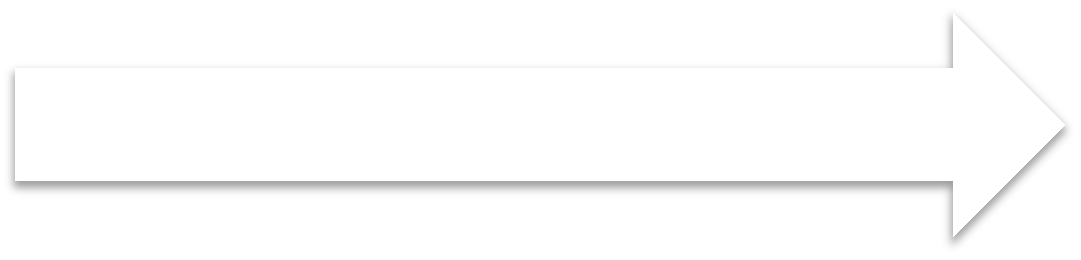
Hence, Inference is that **There is no significance difference in average turnaround time to dispatch the reports of 4 labs.**

|  |
| --- |
| **Python Code** |
| import pandas as pd  import scipy  from scipy import stats  import statsmodels.api as sm  from statsmodels.formula.api import ols  labs = pd.read\_csv("~/Desktop/Digi 360/Module 5/Datasets-2/LabTAT.csv")  labs.head()   |  | **Laboratory 1** | **Laboratory 2** | **Laboratory 3** | **Laboratory 4** | | --- | --- | --- | --- | --- | | 0 | 185.35 | 165.53 | 176.70 | 166.13 | | 1 | 170.49 | 185.91 | 198.45 | 160.79 | | 2 | 192.77 | 194.92 | 201.23 | 185.18 | | 3 | 177.33 | 183.00 | 199.61 | 176.42 | | 4 | 193.41 | 169.57 | 204.63 | 152.60 |   ###Rename Columns#######  labs.rename(columns = {'Laboratory 1':'lab1', 'Laboratory 2':'lab2',  'Laboratory 3':'lab3', 'Laboratory 4':'lab4'}, inplace = True)  labs.head()   | **lab1** | **lab2** | **lab3** | **lab4** | | --- | --- | --- | --- | | 0 | 185.35 | 165.53 | 176.70 | 166.13 | | 1 | 170.49 | 185.91 | 198.45 | 160.79 | | 2 | 192.77 | 194.92 | 201.23 | 185.18 | | 3 | 177.33 | 183.00 | 199.61 | 176.42 | | 4 | 193.41 | 169.57 | 204.63 | 152.60 |   ###Normality Test#############  a = pd.Series(labs.lab1).dropna()  print("p value for lab 1:", stats.shapiro(a))  b = pd.Series(labs.lab2).dropna()  print("p value for lab 2:", stats.shapiro(b))  c = pd.Series(labs.lab3).dropna()  print("p value for lab 3:", stats.shapiro(c))  d = pd.Series(labs.lab4).dropna()  print("p value for lab 4:", stats.shapiro(d))  p value for lab 1: (0.9901824593544006, 0.5506953597068787)  p value for lab 2: (0.9936322569847107, 0.8637524843215942)  p value for lab 3: (0.9886345267295837, 0.4205053448677063)  p value for lab 4: (0.9913753271102905, 0.6618951559066772)  ####variance Test#######  print(scipy.stats.levene(a,b))  print(scipy.stats.levene(b,c))  print(scipy.stats.levene(c,d))  print(scipy.stats.levene(d,a))  LeveneResult(statistic=3.5495027780905763, pvalue=0.06078228171776711)  LeveneResult(statistic=0.9441465124387124, pvalue=0.33220021420602397)  LeveneResult(statistic=2.037958464521512, pvalue=0.15472618294425391)  LeveneResult(statistic=1.5000140718506723, pvalue=0.22188001348277267)  ####one way ANOVA Test#######  ### Since we are assuming variances are equal we will go with One way ANOVA test.  mod = ols('lab1 ~ lab2+lab3+lab4',data=labs).fit()  aov\_table=sm.stats.anova\_lm(mod, type=2)  print(aov\_table)  df sum\_sq mean\_sq F PR(>F)  lab2 1.0 332.030416 332.030416 1.940311 0.166299  lab3 1.0 203.853111 203.853111 1.191271 0.277335  lab4 1.0 265.614707 265.614707 1.552192 0.215323  Residual 116.0 19850.186366 171.122296 NaN NaN  **Here P value is greater than 0.05 so we will fail to reject H0. So, there is no significance difference in average turnaround time to dispatch the reports of 4 labs.** |

|  |
| --- |
| **R Code** |
| > library(readr)  > LabTAT <- read\_csv("Desktop/Digi 360/Module 5/Datasets-2/LabTAT.csv")  Parsed with column specification:  cols(  `Laboratory 1` = col\_double(),  `Laboratory 2` = col\_double(),  `Laboratory 3` = col\_double(),  `Laboratory 4` = col\_double()  )  > View(LabTAT)  > ###Renaming columns  > colnames(labs) <- c("lab1","lab2","lab3","lab4")  Error in colnames(labs) <- c("lab1", "lab2", "lab3", "lab4") :  object 'labs' not found  > colnames(LabTAT) <- c("lab1","lab2","lab3","lab4")  > attach(LabTAT)  > attach(LabTAT)  The following objects are masked from LabTAT (pos = 3):  lab1, lab2, lab3, lab4  > ###Normality Test  > shapiro.test(lab1)  Shapiro-Wilk normality test  data: lab1  W = 0.99018, p-value = 0.5508  > shapiro.test(lab2)  Shapiro-Wilk normality test  data: lab2  W = 0.99363, p-value = 0.8637  > shapiro.test(lab3)  Shapiro-Wilk normality test  data: lab3  W = 0.98863, p-value = 0.4205  > shapiro.test(lab4)  Shapiro-Wilk normality test  data: lab4  W = 0.99138, p-value = 0.6619  > ####Variance Test for lab1 and lab2  > var.test(lab1, lab2)  F test to compare two variances  data: lab1 and lab2  F = 0.77573, num df = 119, denom df = 119, p-value = 0.1675  alternative hypothesis: true ratio of variances is not equal to 1  95 percent confidence interval:  0.5406345 1.1130690  sample estimates:  ratio of variances  0.7757342  > ####Variance Test for lab2 and lab3  > var.test(lab2, lab3)  F test to compare two variances  data: lab2 and lab3  F = 0.81785, num df = 119, denom df = 119, p-value = 0.2742  alternative hypothesis: true ratio of variances is not equal to 1  95 percent confidence interval:  0.5699887 1.1735038  sample estimates:  ratio of variances  0.8178532  > ####Variance Test for lab3 and lab4  > var.test(lab3, lab4)  F test to compare two variances  data: lab3 and lab4  F = 1.2021, num df = 119, denom df = 119, p-value = 0.3168  alternative hypothesis: true ratio of variances is not equal to 1  95 percent confidence interval:  0.8377527 1.7247817  sample estimates:  ratio of variances  1.202057  > ####Variance Test for lab4 and lab1  > var.test(lab4, lab1)  F test to compare two variances  data: lab4 and lab1  F = 1.3113, num df = 119, denom df = 119, p-value = 0.1408  alternative hypothesis: true ratio of variances is not equal to 1  95 percent confidence interval:  0.9138556 1.8814637  sample estimates:  ratio of variances  1.311254  > # since p value is greater than 0.05, we go with H0 i.e. variances are equal, so, will go with  **One way ANOVA Test**  > res.aov <- aov(lab1 ~ lab2+lab3+lab4, data = LabTAT)  > summary(res.aov)  Df Sum Sq Mean Sq F value Pr(>F)  lab2 1 332 332.0 1.940 0.166  lab3 1 204 203.9 1.191 0.277  lab4 1 266 265.6 1.552 0.215  Residuals 116 19850 171.1  21 observations deleted due to missingness  >  **Here P value is greater than 0.05 so we will fail to reject H0. So, there is no significance difference in average turnaround time to dispatch the reports of 4 labs.** |

Sales of products in four different regions is tabulated for males and females. Find if male-female buyer rations are similar across regions.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **East** | **West** | **North** | **South** |
| Males | 50 | 142 | 131 | 70 |
| Females | 550 | 351 | 480 | 350 |



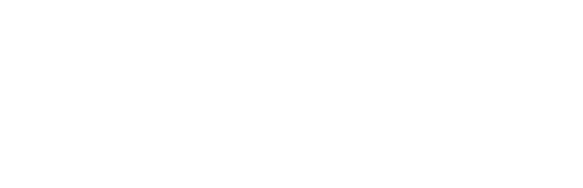
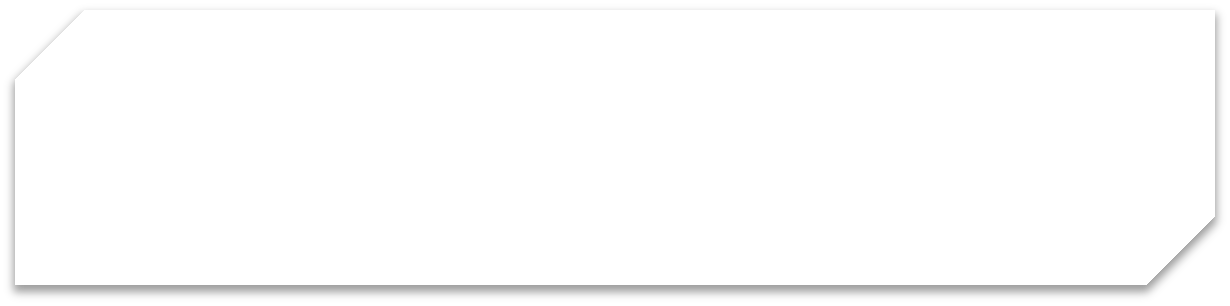
H0

* All proportions are equal

Ha

* Not all Proportions are equal

1. Check p-value
2. If p-Value < alpha, we reject Null Hypothesis



Buyer Ratio.csv

**Solution: -**

1. **Problem Statement**

Are proportion of males versus females same across four regions?

1. **Business Problem**

Should proportion of males and females same across all 4 regions?

1. **Data**

Here Y is proportion which is discrete.

X is 4 samples (4 regions)

1. **chi-square Test**

H0 = Proportions of Male and Female are same

Ha = Proportions of Male and Female are not same

P value = 0.297

P value > Alpha (0.05)

P High H0 Fly - > Fail to Reject H0. i.e. we will go with H0.

H0 = Proportions of Male are equal to Proportions of Female

Hence, Inference is that **Proportions of Male are equal to Proportions of Female across all 4 regions.**

|  |
| --- |
| **R Code** |
| > library(readr)  > BuyerRatio <- read\_csv("Desktop/Digi 360/Module 5/Datasets-2/BuyerRatio.csv")  Parsed with column specification:  cols(  `Observed Values` = col\_character(),  East = col\_double(),  West = col\_double(),  North = col\_double(),  South = col\_double()  )  > View(BuyerRatio)  > attach(BuyerRatio)  The following objects are masked from BuyerRatio (pos = 3):  East, North, Observed Values, South, West  The following objects are masked from BuyerRatio (pos = 4):  East, North, Observed Values, South, West  ##Creating matrix for Chi Square Test  > x <- matrix(c(50,435,142,1523,131,1356,70,750),nrow=2)  > x  [,1] [,2] [,3] [,4]  [1,] 50 142 131 70  [2,] 435 1523 1356 750  > chisq.test(x)  Pearson's Chi-squared test  data: x  X-squared = 1.5959, df = 3, p-value = **0.6603**  **##Here p value is greater than 0.05 so we go with H0 that means Proportions of Male are equal to Proportions of Female across all 4 regions.** |

|  |
| --- |
| **Python Code** |
| ####Sales of products in four different regions is tabulated for males and females.  import pandas as pd  import numpy as np  import scipy  from scipy import stats  from scipy.stats import chisquare  br = pd.read\_csv ("~/desktop/Digi 360/Module 5/Datasets-2/BuyerRatio.csv")  br.head()   |  | **Observed Values** | **East** | **West** | **North** | **South** | | --- | --- | --- | --- | --- | --- | | 0 | Males | 50 | 142 | 131 | 70 | | 1 | Females | 435 | 1523 | 1356 | 750 |   ##Renaming columns  br1 = br.rename(columns = {'Observed Values':'obv'})  br1.head()   |  | **obv** | **East** | **West** | **North** | **South** | | --- | --- | --- | --- | --- | --- | | 0 | Males | 50 | 142 | 131 | 70 | | 1 | Females | 435 | 1523 | 1356 | 750 |   ###Creating matrix for chi square test  br1 = br1.set\_index('obv').rename\_axis('Region', axis=1)  br1   | **Region** | **East** | **West** | **North** | **South** | | --- | --- | --- | --- | --- | | **obv** |  |  |  |  | | Males | 50 | 142 | 131 | 70 | | Females | 435 | 1523 | 1356 | 750 |   ###Getting p value by Chi Square Test. Since the output is discrete and four samples.  from scipy.stats import chi2\_contingency  chi2\_results = scipy.stats.chi2\_contingency(br1)  Chi\_square=[['', 'Test statistic','p-value'],['sample data',chi2\_results[0], chi2\_results[1]]]  Chi\_square  [['', 'Test statistic', 'p-value'],  ['sample data', 1.595945538661058, **0.6603094907091882**]]  **##Here p value is greater than 0.05 so we go with H0 that means Proportions of Male are equal to Proportions of Female across all 4 regions.** |

Telecall uses 4 centers around the globe to process customer order forms. They audit a certain % of the customer order forms. Any error in order form renders it defective and must be reworked before processing. The manager wants to check whether the defective % varies by center. Please analyze the data at *5%* significance level and help the manager draw appropriate inferences

File: **Customer OrderForm.csv**

**Solution: -**

1. **Problem Statement**

Is defective% varies by center?

1. **Business Problem**

Should defective % same or vary across all centers?

1. **Data**

Here Y is % which is discrete.

X is 4 samples

1. **chi-square Test**

H0 = Defective % is same across all centers.

Ha = Defective % is not same across all centers.

P value = 0.2771

P value > Alpha (0.05)

P High H0 Fly - > Fail to Reject H0. i.e. we will go with H0.

H0 = Defective % is same across all centers.

Hence, Inference is that **Defective % is same across all centers.**

|  |
| --- |
| R Code |
| > library(readr)  > CustomerOrderform <- read\_csv("Desktop/Digi 360/Module 5/Datasets-2/CustomerOrderform.csv")  Parsed with column specification:  cols(  Phillippines = col\_character(),  Indonesia = col\_character(),  Malta = col\_character(),  India = col\_character()  )  > View(CustomerOrderform)  > attach(CustomerOrderform)  > attach(CustomerOrderform)  The following objects are masked from CustomerOrderform (pos = 3):  India, Indonesia, Malta, Phillippines  ###Since our output is discrete and four samples. Let’s apply chi-Square test  ###Stacking the data  > stacked2 <- stack(CustomerOrderform)  > table3 <- table(stacked2)  > table3  ind  values Phillippines Indonesia Malta India  Defective 29 33 31 20  Error Free 271 267 269 280  ###Chi-Square Test  > chisq.test(table(stacked2$ind,stacked2$values))  Pearson's Chi-squared test  data: table(stacked2$ind, stacked2$values)  X-squared = 3.859, df = 3, p-value = 0.2771  **##Here p value is greater than 0.05 so we go with H0 that means Defective % is same across all centers.** |

|  |
| --- |
| **Python Code** |
| ####Defect% varies by center out of 4 centers  import pandas as pd  import numpy as np  import scipy  from scipy import stats  from scipy.stats import chisquare  cof = pd.read\_csv ("~/desktop/Digi 360/Module 5/Datasets-2/CustomerOrderform.csv")  cof.head()   |  | **Phillippines** | **Indonesia** | **Malta** | **India** | | --- | --- | --- | --- | --- | | 0 | Error Free | Error Free | Defective | Error Free | | 1 | Error Free | Error Free | Error Free | Defective | | 2 | Error Free | Defective | Defective | Error Free | | 3 | Error Free | Error Free | Error Free | Error Free | | 4 | Error Free | Error Free | Defective | Error Free |   ###Stacking the dataframe  stack = cof.stack().rename\_axis(('sno', 'country')).reset\_index(name='val')  print(stack)  sno country val  0 0 Phillippines Error Free  1 0 Indonesia Error Free  2 0 Malta Defective  3 0 India Error Free  4 1 Phillippines Error Free  ... ... ... ...  1195 298 India Error Free  1196 299 Phillippines Error Free  1197 299 Indonesia Defective  1198 299 Malta Defective  1199 299 India Error Free  [1200 rows x 3 columns]  ###Getting p value by Chi Square Test. ###Since the output is discrete and four samples.  table = pd.crosstab(stack.val,stack.country)  from scipy.stats import chi2\_contingency  chi2\_results = scipy.stats.chi2\_contingency(table)  Chi\_square=[['', 'Test statistic','p-value'],['sample data',chi2\_results[0], chi2\_results[1]]]  Chi\_square  [['', 'Test statistic', 'p-value'],  ['sample data', 3.858960685820355, 0.2771020991233135]]  **##Here p value is greater than 0.05 so we go with H0 that means Defective % is same across all centers.** |

Fantaloons Sales managers commented that *%* of males versus females walking into the store differ based on day of the week. Analyze the data and determine whether there is evidence at *5 %* significance level to support this hypothesis.

File: **Fantaloons.csv**

**Solution: -**

1. **Problem Statement**

Is proportion of males versus females walking into the store differ or not?

1. **Business Problem**

Should proportion of males and females walking into the store is same?

1. **Data**

Here Y is proportion which is discrete.

X is 2 samples (males and females)

1. **2 Proportion Test**

H0 = Proportions of Male and Female are same

Ha = Proportions of Male and Female are not same

P value = 2.2e-16

P value < Alpha (0.05)

P Low H0 go - > Reject H0. i.e. we will go with Ha.

Ha = Proportions of Male and Female are not same

Ho -> Proportions of Male greater than Proportions of Female

Ha -> Proportions of Male not greater than Proportions of Female

P value = 1

P value > Alpha (0.05)

P High H0 Fly - > Fail to Reject H0. i.e. we will go with H0.

H0 = Proportions of Male greater than Proportions of Female

Hence, Inference is that **Proportions of Male greater than Proportions of Female**

|  |
| --- |
| **R Code** |
| > library(readr)  > Fantaloons <- read\_csv("Desktop/Digi 360/Module 5/Datasets-2/Fantaloons.csv")  Parsed with column specification:  cols(  Weekdays = col\_character(),  Weekend = col\_character()  )  > View(Fantaloons)  > attach(Fantaloons)  The following objects are masked from Fantaloons (pos = 3):  Weekdays, Weekend  ##Stack the columns for 2 sample proportion test  > stacked <- stack(Fantaloons)  > attach(stacked)  > attach(stacked)  The following objects are masked from stacked (pos = 3):  ind, values  > a <- table(stacked)  > a  ind  values Weekdays Weekend  Female 287 233  Male 113 167  ###Here Y is discrete and proportion. Also we have 2 samples here. So, we will go for 2 proportion test.  > prop.test(x=c(113,167),n=c(400,400),conf.level = 0.95,alternative = "two.sided")  2-sample test for equality of proportions with continuity correction  data: c(113, 167) out of c(400, 400)  X-squared = 15.434, df = 1, p-value = 8.543e-05  alternative hypothesis: two.sided  95 percent confidence interval:  -0.20293811 -0.06706189  sample estimates:  prop 1 prop 2  0.2825 0.4175  ###Here p value is less than 0.05, so we reject null hypothesis and go with Ha. That means proportions of male and female are not same.  > prop.test(x=c(113,167),n=c(400,400),conf.level = 0.95,alternative = "greater")  2-sample test for equality of proportions with continuity correction  data: c(113, 167) out of c(400, 400)  X-squared = 15.434, df = 1, p-value = 1  alternative hypothesis: greater  95 percent confidence interval:  -0.1924174 1.0000000  sample estimates:  prop 1 prop 2  0.2825 0.4175  **##Here p value is greater than 0.05 so we will go with H0. That means Proportions of Male greater than Proportions of Female** |

|  |
| --- |
| **Python Code** |
| ####Pantaloons Male vs Female %  import pandas as pd  import numpy as np  import scipy  from scipy import stats  import statsmodels.api as sm  fant = pd.read\_csv ("~/desktop/Digi 360/Module 5/Datasets-2/Fantaloons.csv")  fant.head()   |  | **Weekdays** | **Weekend** | | --- | --- | --- | | 0 | Male | Female | | 1 | Female | Male | | 2 | Female | Male | | 3 | Male | Female | | 4 | Female | Female |   ##Stack the columns for 2 sample proportion test  ##Stack the columns for 2 sample proportion test  stack = fant.stack().rename\_axis(('sno', 'day')).reset\_index(name='val')  print(stack)  sno day val  0 0 Weekdays Male  1 0 Weekend Female  2 1 Weekdays Female  3 1 Weekend Male  4 2 Weekdays Female  .. ... ... ...  795 397 Weekend Female  796 398 Weekdays Female  797 398 Weekend Male  798 399 Weekdays Female  799 399 Weekend Female  [800 rows x 3 columns]  ###Here Y is discrete and proportion. Also we have 2 samples here. So, we will go for 2 proportion test.  table = pd.crosstab(stack.val,stack.day)  print(table)  day Weekdays Weekend  val  Female 287 233  Male 113 167  from statsmodels.stats.proportion import proportions\_ztest  count1 = stack.day.value\_counts()  count2 = stack.val.value\_counts()  print(count1, count2)  Weekdays 400  Weekend 400  Name: day, dtype: int64 Female 520  Male 280  Name: val, dtype: int64  count = np.array([113,167])  observ = np.array([400,400])  stats,pval1 = proportions\_ztest(count,observ,alternative='two-sided')  print(pval1)  6.26114287794602e-05  ###Here p value is less than 0.05, so we reject null hypothesis and go with Ha. That means proportions of male and female are not same.  stats,pval1 = proportions\_ztest(count,observ,alternative='larger')  print(pval1)  0.9999686942856103  **##Here p value is greater than 0.05 so we will go with H0. That means Proportions of Male greater than Proportions of Female** |