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)  > attach(Cutlest)  Error in attach(Cutlest) : object 'Cutlest' not found  > attach(Cutlets)  The following objects are masked from Cutlets (pos = 3):  Unit A, Unit B  > colnames(Cutlets) <- c("unit\_A","unit\_B")  > attach(Cutlets)  > attach(Cutlets)  The following objects are masked from Cutlets (pos = 3):  unit\_A, unit\_B  > 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  > 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 |

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.

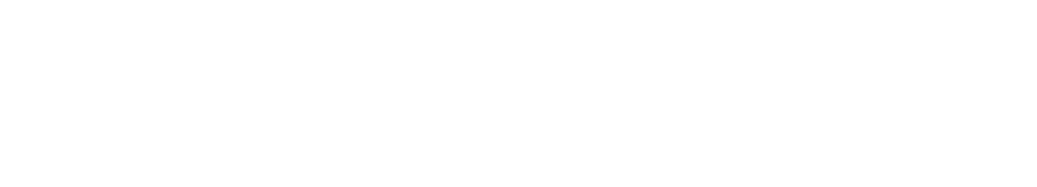
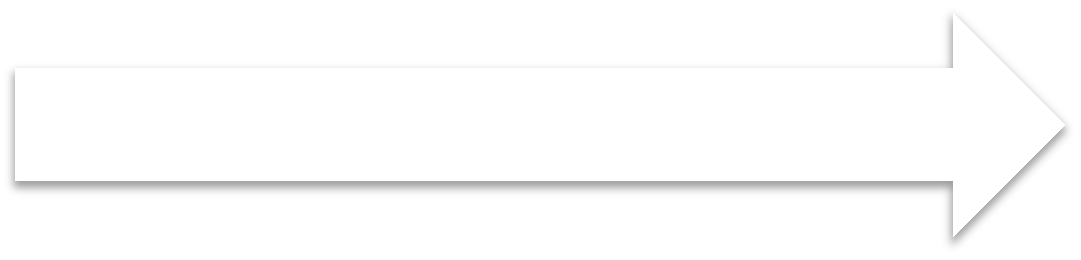
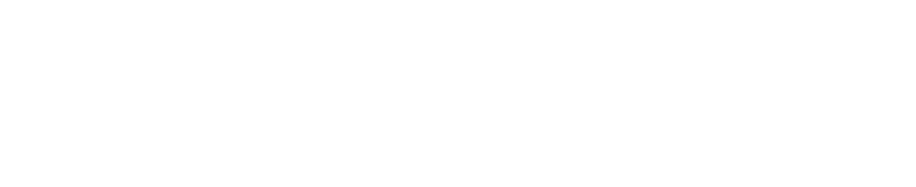
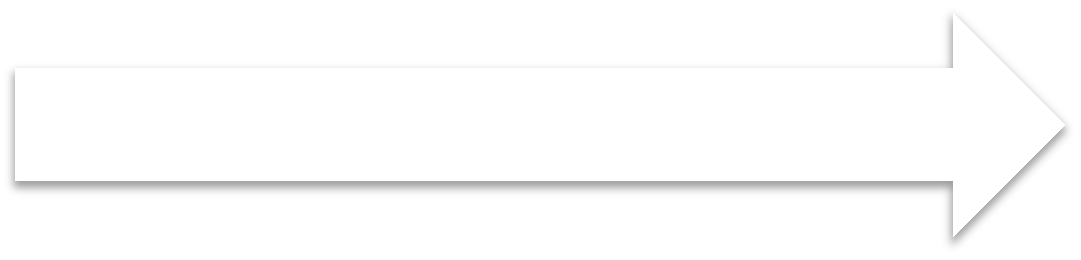
H0 = Average TAT for all the samples is same.

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  ###One way ANOVA Test##########  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)  ####one way ANOVA Test#######  mod = ols('lab1 ~ lab2+lab3+lab4',data=labs).fit()  aov\_table=sm.stats.anova\_lm(mod, type=2)  help(sm.stats.anova\_lm)  print(aov\_table)  Help on function anova\_lm in module statsmodels.stats.anova:  anova\_lm(\*args, \*\*kwargs)  Anova table for one or more fitted linear models.    Parameters  ----------  args : fitted linear model results instance  One or more fitted linear models  scale : float  Estimate of variance, If None, will be estimated from the largest  model. Default is None.  test : str {"F", "Chisq", "Cp"} or None  Test statistics to provide. Default is "F".  typ : str or int {"I","II","III"} or {1,2,3}  The type of Anova test to perform. See notes.  robust : {None, "hc0", "hc1", "hc2", "hc3"}  Use heteroscedasticity-corrected coefficient covariance matrix.  If robust covariance is desired, it is recommended to use `hc3`.    Returns  -------  anova : DataFrame  When args is a single model, return is DataFrame with columns:    sum\_sq : float64  Sum of squares for model terms.  df : float64  Degrees of freedom for model terms.  F : float64  F statistic value for significance of adding model terms.  PR(>F) : float64  P-value for significance of adding model terms.    When args is multiple models, return is DataFrame with columns:    df\_resid : float64  Degrees of freedom of residuals in models.  ssr : float64  Sum of squares of residuals in models.  df\_diff : float64  Degrees of freedom difference from previous model in args  ss\_dff : float64  Difference in ssr from previous model in args  F : float64  F statistic comparing to previous model in args  PR(>F): float64  P-value for significance comparing to previous model in args    Notes  -----  Model statistics are given in the order of args. Models must have been fit  using the formula api.    See Also  --------  model\_results.compare\_f\_test, model\_results.compare\_lm\_test    Examples  --------  >>> import statsmodels.api as sm  >>> from statsmodels.formula.api import ols  >>> moore = sm.datasets.get\_rdataset("Moore", "carData", cache=True) # load  >>> data = moore.data  >>> data = data.rename(columns={"partner.status" :  ... "partner\_status"}) # make name pythonic  >>> moore\_lm = ols('conformity ~ C(fcategory, Sum)\*C(partner\_status, Sum)',  ... data=data).fit()  >>> table = sm.stats.anova\_lm(moore\_lm, typ=2) # Type 2 Anova DataFrame  >>> print(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 |

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 |



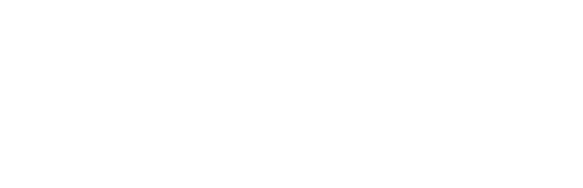
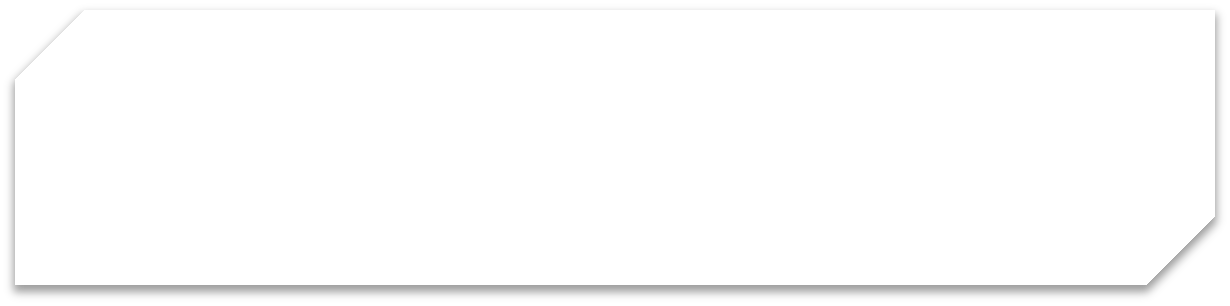
H0

* 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 (males and females)

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)  > attach(BuyerRatio)  The following objects are masked from BuyerRatio (pos = 3):  East, North, Observed Values, South, West  > #########Chi Square#################  > stacked1 <- stack(BuyerRatio)  > table(stacked1$ind,stacked1$values)    131 1356 142 1523 435 50 70 750 Females Males  Observed Values 0 0 0 0 0 0 0 0 1 1  East 0 0 0 0 1 1 0 0 0 0  West 0 0 1 1 0 0 0 0 0 0  North 1 1 0 0 0 0 0 0 0 0  South 0 0 0 0 0 0 1 1 0 0  > table(stacked1$ind,stacked1$values)    131 1356 142 1523 435 50 70 750 Females Males  Observed Values 0 0 0 0 0 0 0 0 1 1  East 0 0 0 0 1 1 0 0 0 0  West 0 0 1 1 0 0 0 0 0 0  North 1 1 0 0 0 0 0 0 0 0  South 0 0 0 0 0 0 1 1 0 0  > chisq.test(table(stacked1$ind,stacked1$values))  Pearson's Chi-squared test  data: table(stacked1$ind, stacked1$values)  X-squared = 40, df = 36, p-value = 0.297 |

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 (males and females)

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.**

|  |
| --- |
|  |
| > 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  > stacked2 <- stack(CustomerOrderform)  > table3 <- table(stacked2)  > table3  ind  values Phillippines Indonesia Malta India  Defective 29 33 31 20  Error Free 271 267 269 280  > chisq.test(table(stacked3$ind,stacked3$values))  Error in table(stacked3$ind, stacked3$values) :  object 'stacked3' not found  > 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 |

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  > stacked <- stack(Fantaloons)  > table 1 <- table(male)  Error: unexpected numeric constant in "table 1"  > table 1 <- table(Male)  Error: unexpected numeric constant in "table 1"  > 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  > prop.test(x=c(280,520),n=c(800,800),conf.level = 0.95,alternative = "two.sided")  2-sample test for equality of proportions with continuity correction  data: c(280, 520) out of c(800, 800)  X-squared = 142.8, df = 1, p-value < 2.2e-16  alternative hypothesis: two.sided  95 percent confidence interval:  -0.3479922 -0.2520078  sample estimates:  prop 1 prop 2  0.35 0.65  > prop.test(x=c(280,520),n=c(800,800),conf.level = 0.95,alternative = "greater")  data: c(280, 520) out of c(800, 800)  X-squared = 142.8, df = 1, p-value = 1  alternative hypothesis: greater  95 percent confidence interval:  -0.3404773 1.0000000  sample estimates:  prop 1 prop 2  0.35 0.65 |