Non_Parametric_Statistical_Testing.R

kriti

Thu Sep 20 16:32:39 2018

```
#q1
#A company wants to learn if sales income is
#equaly distributed among the stores. In order to test it, 8 stores were
#randomly selected. The sales figures are: 102, 300, 102, 100, 205, 105, 71 and 92
#units of product.

#Are the sales equaly distributed among the stores, on the level of significance
#of 95%?


#here we compare the observed freq to expected freq
#example if obs is 51,49 then expected is 50,50

## Hence we use chi square goodness of fit test

chisq.test(c(102,300,102,205,105,71,92))
```

```
##
##  Chi-squared test for given probabilities
##
## data:  c(102, 300, 102, 205, 105, 71, 92)
## X-squared = 293.77, df = 6, p-value < 2.2e-16
```

```
# we see the pvalue is less than 0.05 hence the sales income is not
#uniformly dist



#q2
#A company sells the same product in two types of stores: classical
#and self-service stores. The data about income earned in each type of
#store are as follows:

# Classical stores: 50, 50, 60, 70, 75, 80, 90, 85
#Self-service: 55, 75, 80, 90, 105, 65

#On the level of significance of 95%, is there a difference in
#income among different types of stores?


income<-c(50,50,60,70,75,80,90,85)
store_type=rep("Classical",length(income))
x=cbind(store_type,income)

income<-c(55,75,80,90,105,65)
store_type=rep("Self_Service",length(income))
y=cbind(store_type,income)

data<-rbind(x,y)
data<-data.frame(data)
data
```

```
##      store_type income
## 1     Classical     50
## 2     Classical     50
## 3     Classical     60
## 4     Classical     70
## 5     Classical     75
## 6     Classical     80
## 7     Classical     90
## 8     Classical     85
## 9  Self_Service     55
## 10 Self_Service     75
## 11 Self_Service     80
## 12 Self_Service     90
## 13 Self_Service    105
## 14 Self_Service     65
```

```
data$income<-as.numeric(data$income)
#now performing one sample t test

var.test(data$income~data$store_type)
```

```
##
##  F test to compare two variances
##
## data:  data$income by data$store_type
## F = 0.85198, num df = 7, denom df = 5, p-value = 0.8157
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.1243207 4.5029126
## sample estimates:
## ratio of variances
##          0.8519793
```

```
# we see the variances are equal

aggregate(data$income,by=list(data$store_type),FUN=function(x) shapiro.test(x)$p.value)
```

```
##       Group.1         x
## 1    Classical 0.4678184
## 2 Self_Service 0.9553720
```

```
t.test(data$income~data$store_type,var.equal=TRUE)
```

```
##
##  Two Sample t-test
##
## data:  data$income by data$store_type
## t = 0.1941, df = 12, p-value = 0.8493
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.408313  4.074980
## sample estimates:
##    mean in group Classical mean in group Self_Service
##                   6.000000                   5.666667
```

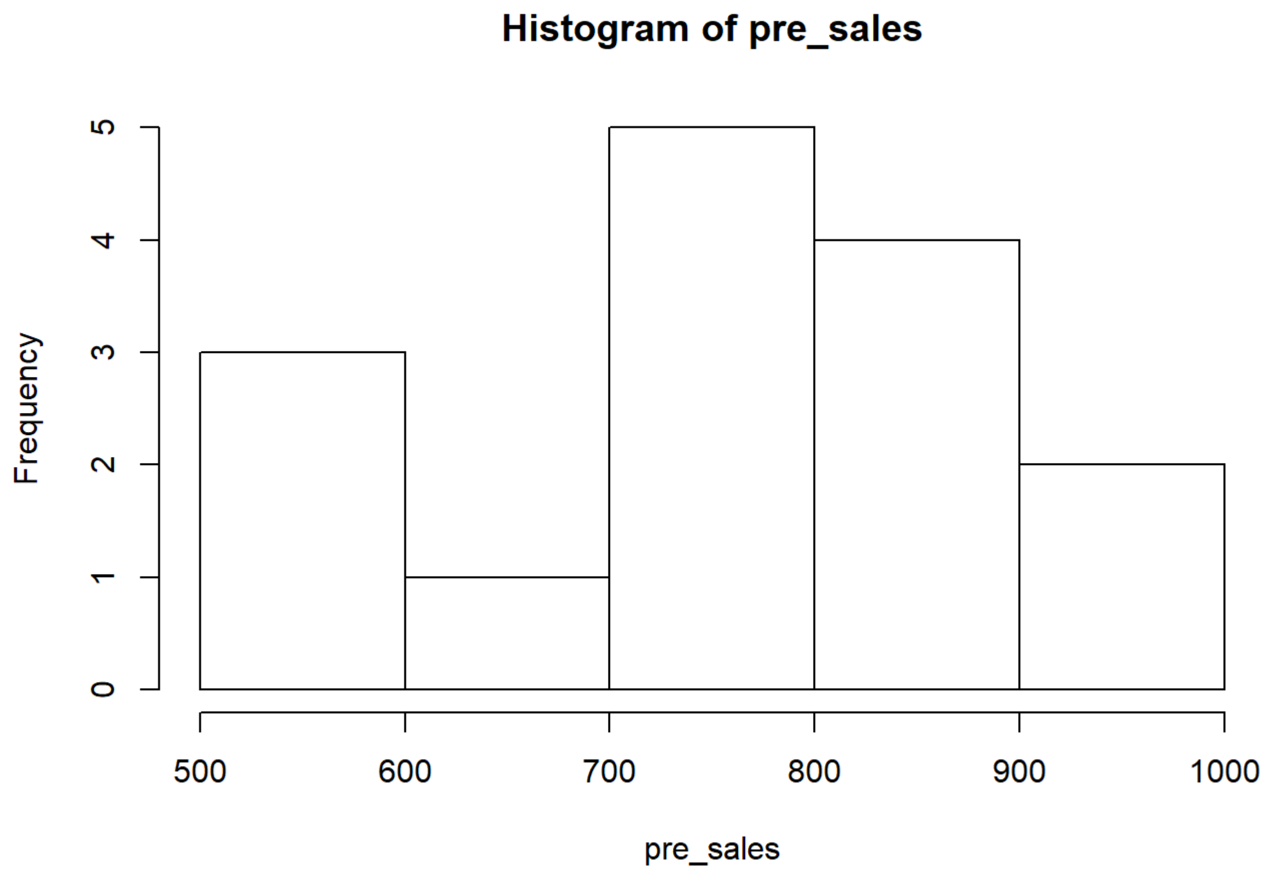
```
#we accept the null hypothesis and hence there is no diff between the income of the
#two types of stores





#q3

#Exercise 3
#Accounting data for sales showed that in randomly selected 15 stores
#the quantities of products sold are:
  #509, 517, 502, 629, 830, 911, 847, 803, 727, 853, 757, 730, 774, 718, 904
#Unsatisfied with those results, a company decided to start advertising campaign. After the camp
aign finished, the amount of products sold in these same stores were:
 #517, 508, 523, 730, 821, 940, 818, 821, 842, 842, 709, 688, 787, 780, 901
#Did the advertizing campaign produce statistically significant results?


##solution

#since we are comparing the results before and after analysis we should use
#paired testing


pre_sales<-c(509, 517, 502, 629, 830, 911, 847, 803, 727, 853, 757, 730, 774, 718, 904)
post_sales<-c(517, 508, 523, 730, 821, 940, 818, 821, 842, 842, 709, 688, 787, 780, 901)

#checking if the distribution is normal
shapiro.test(pre_sales)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  pre_sales
## W = 0.90113, p-value = 0.09902
```

```
shapiro.test(post_sales)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  post_sales
## W = 0.88582, p-value = 0.05799
```

```
hist(pre_sales)
```




```
hist(post_sales)
##both are close to a normal distribution but not exactly normally dist
##so we try t test and wilcox test

##checking if they have equal variance
var.test(pre_sales,post_sales)
```

```
##
##  F test to compare two variances
##
## data:  pre_sales and post_sales
## F = 1.0117, num df = 14, denom df = 14, p-value = 0.983
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.3396515 3.0133828
## sample estimates:
## ratio of variances
##           1.011682
```

```
###by seeing the results of the f test we see that the variance is equal

##computing paired t test
t.test(pre_sales,post_sales,var.equal = TRUE,paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  pre_sales and post_sales
## t = -1.1814, df = 14, p-value = 0.2571
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -40.54261  11.74261
## sample estimates:
## mean of the differences
##                   -14.4
```

```
wilcox.test(pre_sales,post_sales,correct = FALSE,paired = TRUE)
```

```
## Warning in wilcox.test.default(pre_sales, post_sales, correct = FALSE,
## paired = TRUE): cannot compute exact p-value with ties
```

```
##
##  Wilcoxon signed rank test
##
## data:  pre_sales and post_sales
## V = 45.5, p-value = 0.41
## alternative hypothesis: true location shift is not equal to 0
```

```r
##we accept the null hypothesis.No diff between the means of both the samples
##hence the campaign did not have any effect


#Exercise 4

#One product is produced in white, blue and red color.
#Five stores were randomly selected in order to test, with the 5% risk of
#error, if the color influences the number of products sold. Data about
#sales are given in the following table:

#Store  White   Blue    Red
#1. 510 925 730
#2. 720 735 745
#3. 930 753 875
#4. 754 685 610
#5. 105


##SOLUTION
##here we use Kruskal test since the distributions are not normal.Else we could
#use anova

white<-c(510,720,930,754,105)
Blue<-c(925,735,753,685)
Red<-c(730,745,875,610)

##performing Kruskal Wallis test
kruskal.test(list(white,Blue,Red))
```

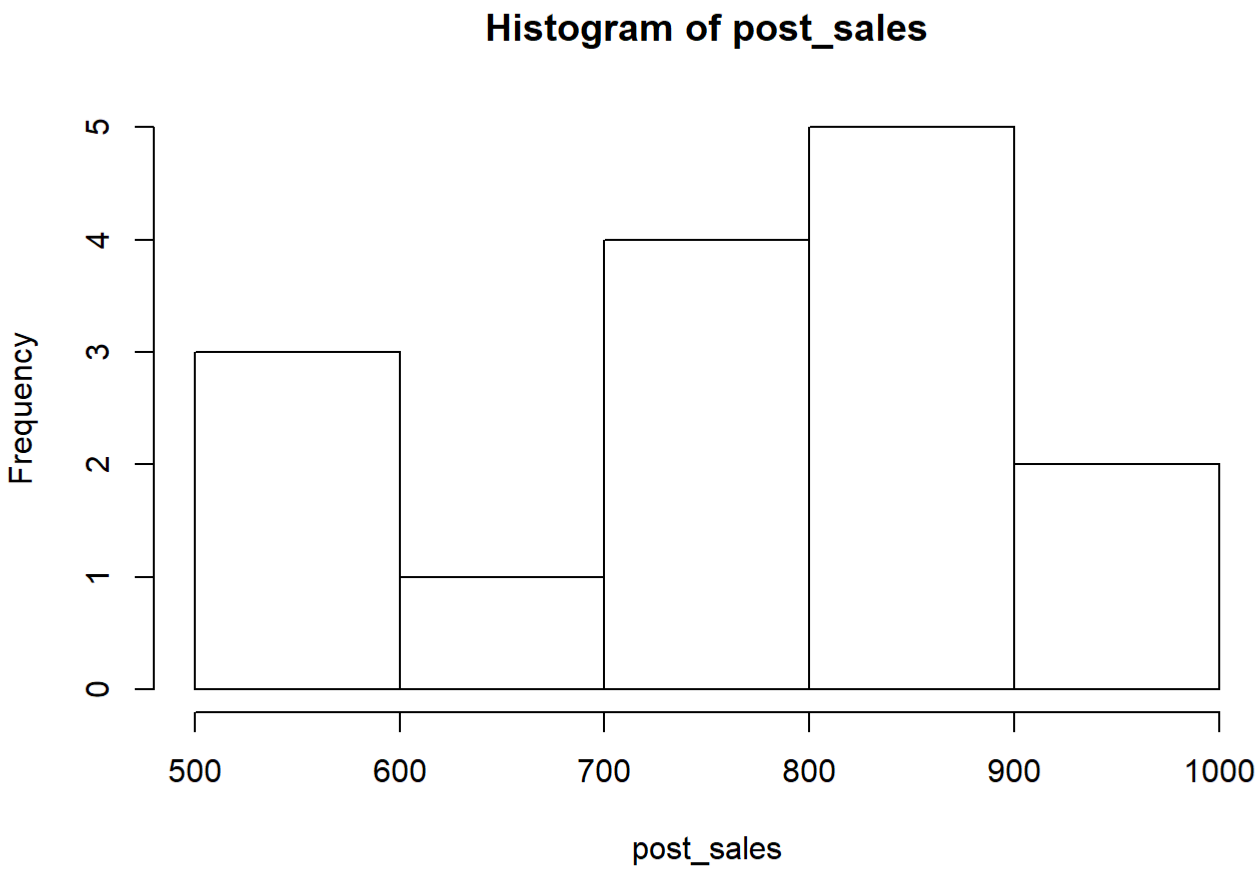
```
##
##  Kruskal-Wallis rank sum test
##
## data:  list(white, Blue, Red)
## Kruskal-Wallis chi-squared = 0.47473, df = 2, p-value = 0.7887
```

```r
##here the null hypothesis is that white=blue=red
##since p value is greater than 0.05 we accept the null hypothesis and
#color doesnt accept the sales




#Exercise 5

#A TV station conducted surveys in March, April, May and June asking a
#number of it's viewers about their satisfaction with the program in the
#previous month. The same viewers participated in all four surveys. You can
#download survey data here

#Did the viewer's satisfaction change during four months?

##solution

#Tip: in order to conduct this test, you'll need to install and use CVST
#library.

March<-c(1,0,0,1,1)
April<-c(0,0,0,1,0)
May<-c(1,1,1,1,1)
June<-c(0,1,0,1,0)
Data<-cbind(March,April,May,June)

##since we have the same set of viewers this is paired testing
library("CVST")
```

```
## Warning: package 'CVST' was built under R version 3.4.4
```

```
## Loading required package: kernlab
```

```
## Loading required package: Matrix
```




```
cochranq.test(Data)
```

```
##
##  Cochran's Q Test (monte-carlo)
##
## data:  mat[index, ]
## Cochran's Q = 7, df = 3, p-value = 0.088
```

```
##since p value is greater than 0.05 we accept the null hypothesis.viewer
#satisfaction did not change

#Exercise 6

#A company conducted survey in order to learn about customer satisfaction
#with company's service. Then, after improvement of the service, company
#conducted another survey on the same customers. The summary of two surveys is
#given in the following table:

#Survey Satisfied   Not satisfied
#Before improvement 32  68
#After improvement  48  52


##solution
#we could run a chi square test of independence but here before and after are related
##hence we run a McNemar test

#McNemar test only works on 2*2 matrix

mat=matrix(data=c(32,68,48,52),nrow=2,ncol=2)
rownames(mat)<-c("before","after")
colnames(mat)<-c("Satisfied","Not Satisfied")
mat
```

```
##        Satisfied Not Satisfied
## before        32            48
## after         68            52
```

```
mcnemar.test(mat)
```

```
##
##  McNemar's Chi-squared test with continuity correction
##
## data:  mat
## McNemar's chi-squared = 3.1121, df = 1, p-value = 0.07771
```

```
#in this case the null hypothesis is that both are equal
#we accept the null hypothesis and conclude that there is no change in before and after


#Exercise 7

#A company conducted a survey in order to examine if the frequency of usage of
#company's service depends on the size of the city where it's clients live.
#The summary of survey is given in the following table:

#City size  Frequency of service usage
      #Always|Sometime|Never
#Small      151 252 603
#Medium   802   603 405
#Large      753 55  408
#Does the frequency of usage of company's service depend on the size of the city?


#Solution

#here the two categorical variables are independent and not 2*2 so we use

# Chi-square test for homogeneity

# H0: oij=eij for all cells

mat=matrix(data=c(151,252,603,802,603,405,753,55,408),nrow=3,ncol=3)

chisq.test(mat)
```

```
##
##  Pearson's Chi-squared test
##
## data:  mat
## X-squared = 821.31, df = 4, p-value < 2.2e-16
```

```
##since p value is less than 0.05 we reject the null hypothesis and hence the
#categorical variables are associated with each other
```