STAT 231 Assignment 5

The purpose of this assignment is to use the software R to perform Goodness of Fit tests and test for independence in two way tables.

The code for this assignment is posted both as a text file called RCodeAssignment5.txt and an R file called RCodeAssignment5R.R which are posted in the Assignment 5 folder in the Assignments folder under Content on Learn.

Problem 1: Run the following R code.

```
##########################################################################################
# run this code only once
library(MASS)
##########################################################################################

##########################################################################################
# Problem 1: Testing the Multinomial Model with Equal Probabilities
id<-20456458
set.seed(id)
k<-sample(5:9,1)   # randomly choose number of categories for Multinomial data
p<-sample(1:9,k,replace=TRUE)
p<-p/sum(p)       # choose random probabilities which must sum to one
y<-rmultinom(1,150,p)    # generate random data
e<- rep(150/k, k)  # calculate expected frequencies assuming equal probabilities for each category
# print table of observed and expected frequencies
cat("Table of Observed and Expected Frequencies ")
print(data.frame("Category" = rbind(y[,1],e), row.names = c("Observed", "Expected")),digits=4)
# observed values of likelihood ratio test statistic and Goodness of Fit test statistic
# and corresponding p-values
df<-k-1      # degrees of freedom for the Chi-squared distribution
lambda<-2*sum(y*log(y/e))
pvalue<-1-pchisq(lambda,df)
cat("Observed value of likelihood ratio statistic = ", lambda)
cat("with p-value = ",pvalue, "and degrees of freedom = ",df)
pearson<-sum(((y-e)^2)/e)
pvalue<-1-pchisq(pearson,df)
cat("Observed value of Goodness of Fit statistic = ", pearson)
cat("with p-value = ", pvalue, "and degrees of freedom = ",df)
##########################################################################################
```

Verify that you obtain the following output:

```
         Category.1 Category.2 Category.3 Category.4 Category.5
Observed          8         23         43         63         13
Expected         30         30         30         30         30

Observed value of likelihood ratio statistic =  69.33145
with p-value =  3.141931e-14 and degrees of freedom =  4


Observed value of Goodness of Fit statistic =  69.33333
with p-value =  3.141931e-14 and degrees of freedom =  4
```

Problem 2: Run the following R code.

```
###########################################################################################
# Problem 2: Testing the Goodness of Fit of a Poisson Model
set.seed(id)
model<-sample(c(1:4),1)
cat("Model = ", model)
# Data are randomly generated from one of four different models all with mean 4
# Model=1: Poisson(4) distribution
# Model=2:  Negative Binomial(3,3/7)
# Model=3:  G(4,1) distribution and discretized
# Model=4:  Gamma(3,4/3) distribution and discretized
if (model==1) {
  y<-rpois(150,4)    # 150 observations from Poisson(4)
} else if (model==2) {
  y<-rnbinom(150,3,3/7)   # 150 observations from NB(3,3/7)
} else if (model==3) {
  y<-round( rnorm(150,4,1))   # 150 observations from G(4,1) rounded
  y[y<0]<-0   # convert any negative observations to 0
} else if (model==4) {
  y<-round(rgamma(150,3,3/4))    # 150 observations from Gamma(3,4/3) rounded
}
ymin<-min(y)
ymax<-max(y)
# determine categories and frequencies for the data
data<-table(c(y, ymin:ymax))-1      # Done to ensure all categories are accounted for
f<-as.numeric(data)     # frequencies
cat<-as.numeric(names(data))
# determine the maximum likelihood estimate of theta which is the sample mean calculated
# from the frequency table
```

```
thetahat<-sum(cat*f)/150
# determine the expected frequencies
e<-dpois(cat,thetahat)*150   #expected frequencies for Poisson data
#frequency for ymin must be sum for y<=ymin
e[1]<-ppois(ymin,thetahat)*150
ncat<-length(e)
# frequency for ymax must be sum of frequencies for y>=ymax
e[ncat]<- ppois(ymax- 1,thetahat, lower = F)*150
# Table of Observed and expected frequencies
data<-rbind("y" = ymin:ymax, "observed" = f, "expected" = e)
# print table of observed and expected frequencies
cat("Table of Observed and Expected Frequencies ")
print(data,digits=4)
# Expected frequencies must all be at least 5 to apply tests. Collapse categories if necessary.
nbins<-ncol(data)
while(data[3, nbins] < 5){
  data[2:3, nbins - 1]<-data[2:3, nbins - 1] + data[2:3, nbins]
  data<-data[, -nbins]
  nbins<-nbins - 1
}
nbins<-1
while(data[3, nbins] < 5){
  data[2:3, nbins + 1]<-data[2:3, nbins + 1] + data[2:3, nbins]
  data<-data[, -nbins]
}
cat("Table of Observed and Expected Frequencies ")
# print table of observed and expected frequencies
print(data,digits=4)
# observed values of likelihood ratio test statistic and Goodness of Fit test statistic
# and corresponding p-values
df = ncol(data)-2      # degress of freedom for the Chi-squared distribution
f<-data[2,]
e<-data[3,]
lambda<-2*sum(f*log(f/e))
pvalue<-1-pchisq(lambda,df)
cat("Observed value of likelihood ratio statistic = ", lambda)
cat("with p-value = ",pvalue, "and degrees of freedom = ",df)
pearson<-sum(((f-e)^2)/e)
pvalue<-1-pchisq(pearson,df)
cat("Observed value of Goodness of Fit statistic = ", pearson)
cat("with p-value = ", pvalue, "and degrees of freedom = ",df)
##########################################################################################
```

Verify that you obtain the following output:

```
Model =  1

Table of Observed and Expected Frequencies

          [,1]  [,2]  [,3] [,4] [,5]  [,6]  [,7]  [,8]  [,9] [,10]
y        0.000  1.00  2.00  3.0  4.0  5.00  6.00 7.000 8.000 9.000
observed 3.000  8.00 20.00 35.0 28.0 28.00 15.00 6.000 4.000 3.000
expected 2.822 11.21 22.27 29.5 29.3 23.29 15.42 8.753 4.347 3.087


Table of Observed and Expected Frequencies

          [,1]  [,2] [,3] [,4]  [,5]  [,6]  [,7]  [,8]
y         1.00  2.00  3.0  4.0  5.00  6.00 7.000 8.000
observed 11.00 20.00 35.0 28.0 28.00 15.00 6.000 7.000
expected 14.03 22.27 29.5 29.3 23.29 15.42 8.753 7.434

Observed value of likelihood ratio statistic =  4.212926
with p-value =  0.6478865 and degrees of freedom =  6

Observed value of Goodness of Fit statistic =  4.108531
with p-value =  0.6619919 and degrees of freedom =  6
```

Problem 3: Run the following R code.

```
##########################################################################################
# Problem 3: Testing for Independence in Two Way Tables
set.seed(id)
# generate data for a two way table by first simulating bivariate data
# from the Bivariate Normal distribution and then discretize the data
# Random uniform between -0.75 and 0.75
corrCoef<-runif(1, -0.75, 0.75)
sigma<-max(id %% 10, 1)
# Last digit of UWID using modulo, minimum value of 1.
mu1<-max(id %% 100 - id %% 10, 20)
# (Second last digit*10) is extracted here, minimum value of 20
mu2<-max(id %% 1000 - id %% 100, 30)
# (Third last digit*100) is extracted here, minimum value of 30
VarCovar<-cbind(c(sigma^2, corrCoef*sigma^2), c(corrCoef*sigma^2, sigma^2))
# Simulate data from a bivariate Normal
n<-sample(c(100:200),1)    # n =  sample size
cat("Number of observations = ",n)
data2<-mvrnorm(n, mu = c(mu1, mu2), Sigma = VarCovar)
# Create smoker/non-smoker variable by mapping 1 to smoker and 2 to non-smoker
```

```
data3 = as.data.frame(data2)
data3[, 1]<-ifelse(data3[,1] < median(data3[,1]), 1, 2)
data3[, 1]<-c("Smoker", "Non-smoker") [data3[,1]]
# Create tall/avg/short height variable by mapping 1 to tall, 2 to average and 3 to short
data3[, 2]<-floor((rank(data3[, 2])-0.1)/nrow(data3)*3) + 1
data3[, 2]<-c("Tall", "Average", "Short")[data3[, 2]]
data3[, 1]<-factor(data3[, 1])
data3[, 2]<-factor(data3[, 2])
colnames(data3)<-c("Smoker Indicator", "Height Indicator")
f<-table(data3)
cat("Table of Observed Frequencies:")
f
r<-margin.table(f,1)     # row totals
c<-margin.table(f,2)     # column totals
e<-outer(r,c)/sum(f)   # matrix of expected frequencies
cat("Table of Expected Frequencies:")
print(e,digits=4)
lambda<-2*sum(f*log(f/e))  # observed value of likelihood ratio statistic
df<-(length(r)-1)*(length(c)-1)  # degrees of freedom
pvalue<-1-pchisq(lambda,df)
cat("Observed value of likelihood ratio statistic = ", lambda)
cat("with p-value = ",pvalue, "and degrees of freedom = ",df)
pearson<-sum(((f-e)^2)/e)
pvalue<-1-pchisq(pearson,df)
cat("Observed value of Goodness of Fit statistic = ", pearson)
cat("with p-value = ", pvalue, "and degrees of freedom = ",df)
##########################################################################################
```

Verify that you obtain the following output:

```
Number of observations =  109

Table of Observed Frequencies:

                Height Indicator
Smoker Indicator Average Short Tall
      Non-smoker      23     6   26
      Smoker          13    31   10

Table of Expected Frequencies:

                Height Indicator
Smoker Indicator Average Short  Tall
      Non-smoker   18.17 18.67 18.17
      Smoker       17.83 18.33 17.83
```

```
Observed value of likelihood ratio statistic =  28.66472
with p-value =  5.963973e-07 and degrees of freedom =  2

Observed value of Goodness of Fit statistic =  26.77386
with p-value =  1.535077e-06 and degrees of freedom =  2
```

Run the R code above again except modify the line

"id<-20456458"

in Problem 1 by replacing the number 20456458 with your UWaterloo ID number.

Download the Assignment 5 Template which is posted as a Word document on Learn. Fill in the required information and plots based on the output for the data generated using your ID number. <u>Your assignment must follow the template exactly.</u> See Assignment 5 Example posted on Learn.

Create a .pdf file for the answer to EACH problem.

Upload your assignment to Crowdmark using the link which was emailed to you.