21bcs022-ex6

Nagajothi

2024-03-13

## R Markdown

#Individual Questions:  
#1.Chi-Squared Test:  
set.seed(0)  
x=sample(c("Male","Female"),1000,replace=TRUE)  
y=sample(c("BJP","CONGRESS","DMK"),1000,replace=TRUE)  
observed=table(x,y)  
observed

## y  
## x BJP CONGRESS DMK  
## Female 159 172 165  
## Male 165 166 173

#Built-in function calculation:  
chi\_test=chisq.test(observed)  
names(chi\_test)

## [1] "statistic" "parameter" "p.value" "method" "data.name" "observed"   
## [7] "expected" "residuals" "stdres"

print(chi\_test$p.value)

## [1] 0.842404

if(chi\_test$p.value<0.05){  
 print("Gender and Political preferences are dependent.\n")  
} else{  
 print("Gender and Political preferences are not dependent.\n")  
}

## [1] "Gender and Political preferences are not dependent.\n"

# Chi-square calculation function  
calculate\_chi\_square <- function(observed) {  
 # Calculate row and column sums  
 row\_sums <- rowSums(observed)  
 col\_sums <- colSums(observed)  
 total\_sum <- sum(observed)  
   
 # Calculate expected frequencies for each attribute pair  
 expected <- matrix(NA, nrow = nrow(observed), ncol = ncol(observed))  
 for (i in 1:nrow(observed)) {  
 for (j in 1:ncol(observed)) {  
 expected[i, j] <- (row\_sums[i] \* col\_sums[j]) / total\_sum  
 }  
 }  
   
 # Calculate chi-square statistic  
 chi\_squared <- sum((observed - expected)^2 / expected)  
   
 # Determine degrees of freedom  
 rows <- nrow(observed)  
 cols <- ncol(observed)  
 df <- (rows - 1) \* (cols - 1)  
   
 # Perform chi-square test  
 p\_value <- 1 - pchisq(chi\_squared, df)  
   
 # Return chi-square statistic and p-value  
 return(list(chi\_squared = chi\_squared, p\_value = p\_value))  
}  
  
# Calculate chi-square statistic  
chi\_square\_results <- calculate\_chi\_square(observed)  
chi\_squared <- chi\_square\_results$chi\_squared  
p\_value <- chi\_square\_results$p\_value  
  
# Compare with default error rate  
alpha <- 0.001 # You can adjust this value based on your desired level of significance  
print(p\_value)

## [1] 0.842404

# Compare with significance level  
if (p\_value < alpha) {  
 print("Gender and Political preferences are dependent.\n")  
} else {  
 print("Gender and Political preferences are dependent.\n")  
}

## [1] "Gender and Political preferences are dependent.\n"

#2.Correlation coefficient:  
x <- sample(1:100,100,replace=TRUE) # Age in years  
y <- sample(1:1000,100,replace=TRUE) # Salary in Rupees  
print(cor(x,y))

## [1] 0.04398303

# Calculate means  
mean\_x <- mean(x)  
mean\_y <- mean(y)  
  
# Calculate standard deviations  
sd\_x <- sd(x)  
sd\_y <- sd(y)  
  
# Number of data points  
n <- length(x)  
  
# Compute correlation coefficient manually  
correlation <- sum((x - mean\_x) \* (y - mean\_y)) /   
 (sqrt(sum((x - mean\_x)^2)) \* sqrt(sum((y - mean\_y)^2)))  
correlation

## [1] 0.04398303

if (correlation==0){  
 cat("Age and salary are completely independent.\n")  
} else {  
 if(correlation>0){  
 cat("Age and salary are positively correlated.\n")  
} else {  
 cat("Age and salary are negatively correlated.\n")  
}  
}

## Age and salary are positively correlated.

#3.Covariance Test:  
x <- sample(1:100,100,replace=TRUE) # Age in years  
y <- sample(1:1000,100,replace=TRUE) # Salary in Rupees  
print(cov(x,y))

## [1] 43.9901

# Calculate means  
mean\_x <- mean(x)  
mean\_y <- mean(y)  
  
# Number of data points  
n <- length(x)  
  
# Determine if attributes are independent or not  
covariance <- sum((x - mean\_x) \* (y - mean\_y)) / n  
covariance

## [1] 43.5502

# Check if covariance is 0  
if (covariance == 0) {  
 cat("Age and salary are independent.\n")  
} else {  
 cat("Age and salary are dependent.\n")  
}

## Age and salary are dependent.

#Common Questions:  
#1.Find out if the ‘cyl’ and ‘carb’ variables are in ‘mtcars’ dataset and whether it is dependent or not.   
data(mtcars)  
names(mtcars)

## [1] "mpg" "cyl" "disp" "hp" "drat" "wt" "qsec" "vs" "am" "gear"  
## [11] "carb"

"cyl" %in% colnames(mtcars)

## [1] TRUE

"carb" %in% colnames(mtcars)

## [1] TRUE

cor\_test=cor.test(mtcars$carb,mtcars$cyl,method="pearson")  
names(cor\_test)

## [1] "statistic" "parameter" "p.value" "estimate" "null.value"   
## [6] "alternative" "method" "data.name" "conf.int"

if(cor\_test$p.value<0.05){  
 print("Carb and Cyl are dependent.\n")  
} else{  
 print("Carb and Cyl are independent.\n")  
}

## [1] "Carb and Cyl are dependent.\n"

#2.Use chi-squared test for the same.  
tab<-table(mtcars$carb,mtcars$cyl)  
chi\_test=chisq.test(tab)

## Warning in chisq.test(tab): Chi-squared approximation may be incorrect

names(chi\_test)

## [1] "statistic" "parameter" "p.value" "method" "data.name" "observed"   
## [7] "expected" "residuals" "stdres"

if(chi\_test$p.value<0.05){  
 print("Carb and Cyl are dependent.\n")  
} else{  
 print("Carb and Cyl are not dependent.\n")  
}

## [1] "Carb and Cyl are dependent.\n"

#3.Test the hypothesis whether the students smoking habit is independent   
#of their exercise level at 0.05 significance level using survey dataset.   
library(MASS)  
t<-table(survey$Smoke,survey$Exer)  
c<-chisq.test(t)

## Warning in chisq.test(t): Chi-squared approximation may be incorrect

c

##   
## Pearson's Chi-squared test  
##   
## data: t  
## X-squared = 5.4885, df = 6, p-value = 0.4828

if(c$p.value<0.05){  
 print("Smoking Habit and exercise level are dependent.\n")  
}else{  
 print("Smoking Habit and exercise level are not dependent.\n")  
}

## [1] "Smoking Habit and exercise level are not dependent.\n"

#4.Conduct the Chi-squared independence test of the smoking and   
#exercise survey by computing the p-value with the textbook formula.   
obs<-c$observed  
exp<-round(c$observed,3)  
x<-((obs-exp)^2)/exp  
row<-nrow(t)  
col<-ncol(t)  
d<-(row-1)\*(col-1)  
pchisq(q=sum(x),df=d,lower.tail=FALSE)

## [1] 1