Compupter Practical 3

Statistics 2

Dom Hutchinson

```
options(warn=-1)
diabetes<-read.csv("data/diabetes_data.csv",header=T)
head(diabetes)
```

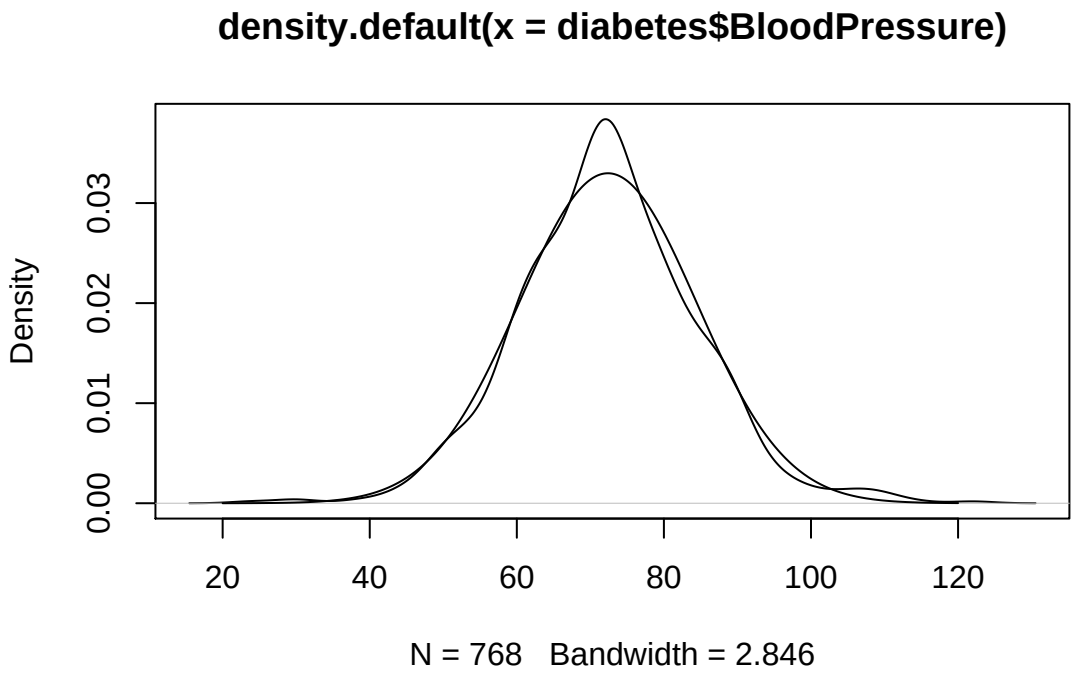
```
##   Pregnancies Glucose BloodPressure SkinThickness Insulin  BMI
## 1           6     148            72            35       0 33.6
## 2           1      85            66            29       0 26.6
## 3           8     183            64             0       0 23.3
## 4           1      89            66            23      94 28.1
## 5           0     137            40            35     168 43.1
## 6           5     116            74             0       0 25.6
##   DiabetesPedigreeFunction Age Outcome
## 1                    0.627  50       1
## 2                    0.351  31       0
## 3                    0.672  32       1
## 4                    0.167  21       0
## 5                    2.288  33       1
## 6                    0.201  30       0
```

```
missing<-function(var) { # Map missing values to median
  med<-median(var[var>0])
  var[var==0]<-med
  return(var)
}
diabetes$Glucose<-missing(diabetes$Glucose)
diabetes$BloodPressure<-missing(diabetes$BloodPressure)
diabetes$Insulin<-missing(diabetes$Insulin)
diabetes$SkinThickness<-missing(diabetes$SkinThickness)
diabetes$BMI<-missing(diabetes$BMI)

head(diabetes)
```

```
##   Pregnancies Glucose BloodPressure SkinThickness Insulin  BMI
## 1           6     148            72            35     125 33.6
## 2           1      85            66            29     125 26.6
## 3           8     183            64            29     125 23.3
## 4           1      89            66            23      94 28.1
## 5           0     137            40            35     168 43.1
## 6           5     116            74            29     125 25.6
##   DiabetesPedigreeFunction Age Outcome
## 1                    0.627  50       1
## 2                    0.351  31       0
## 3                    0.672  32       1
## 4                    0.167  21       0
## 5                    2.288  33       1
## 6                    0.201  30       0
```

Question 1

```
breaks<-c(-Inf,seq(45,95,by=5),Inf) # Quantise data
obs<-table(cut(diabetes$BloodPressure,breaks))

# Perform Pearson's Goodness of Fit Test
n<-length(diabetes$BloodPressure)
m<-length(breaks)-1-3
mu<-mean(diabetes$BloodPressure) # MLE
sigma<-sd(diabetes$BloodPressure) # MLE

x<-seq(20,120,0.1)
plot(density(diabetes$BloodPressure))
lines(x,dnorm(x,mu,sigma))
```




```
exp<-n*(pnorm(breaks[-1],mean=mu,sd=sigma)-pnorm(breaks[-length(breaks)],mean=mu,sd=sigma)) # calculate
round(cbind(obs,exp),1) # Display observed & expected values
```

```
##           obs   exp
## (-Inf,45]   9   9.1
## (45,50]    20  15.6
## (50,55]    24  33.2
## (55,60]    70  59.6
## (60,65]    85  90.5
## (65,70]   132 116.0
## (70,75]   139 125.7
## (75,80]   124 115.1
## (80,85]    59  89.1
## (85,90]    68  58.3
```

```
## (90,95]    15  32.2
## (95, Inf]  23  23.6
```

```
t_obs<-sum((obs-exp)^2/exp) # observered test statistic
p_val<-1-pchisq(t_obs,df=m) # p-value
cat("mu=",mu,"\nsigma=",sigma,"\ndf=",m,"\nt_obs=",t_obs,"\np-val=",p_val,sep="")
```

```
## mu=72.38672
## sigma=12.09664
## df=9
## t_obs=31.18803
## p-val=0.0002748274
```

Question 2

Let $X_1, \dots, X_n \overset{\text{iid}}{\sim} \text{Normal}(\mu, 12^2)$ model the blood pressume of members of the study.
Here I shall test the hypotheses

$$H_0 : \mu = 70 \text{ against } H_1 : \mu > 70$$

I shall use Pearson's test statistic

$$T(\mathbf{X}) := \sum_{i=1}^{m} \frac{(o_i - e_i)^2}{e_i} = \sum_{i=1}^{m} \frac{(o_i - np_i)^2}{np_i} \to_{\mathcal{D}} \chi^2_{m-1}$$

where o_i is the number of observations in inteval i, e_i is the expected number of observations of interval i & p_i is the probability of an observation belonging to interval i given the null-hypothes is true.
Due to the breaks chosen in Question 1 $m =' rlength(obs)'$.

```
mu<-70; sigma<-12
exp<-n*(pnorm(breaks[-1],mean=mu,sd=sigma)-pnorm(breaks[-length(breaks)],mean=mu,sd=sigma))
round(cbind(obs,exp),1)
```

```
##           obs   exp
## (-Inf,45]   9  14.3
## (45,50]    20  22.4
## (50,55]    24  44.4
## (55,60]    70  74.2
## (60,65]    85 104.5
## (65,70]   132 124.1
## (70,75]   139 124.1
## (75,80]   124 104.5
## (80,85]    59  74.2
## (85,90]    68  44.4
## (90,95]    15  22.4
## (95, Inf]  23  14.3
```

```
t_obs<-sum((obs-exp)^2/exp) # observered test statistic
p_val<-1-pchisq(t_obs,df=m) # p-value
cat("mu=",mu,"\nsigma=",sigma,"\ndf=",m,"\nt_obs=",t_obs,"\np-val=",p_val,sep="")
```

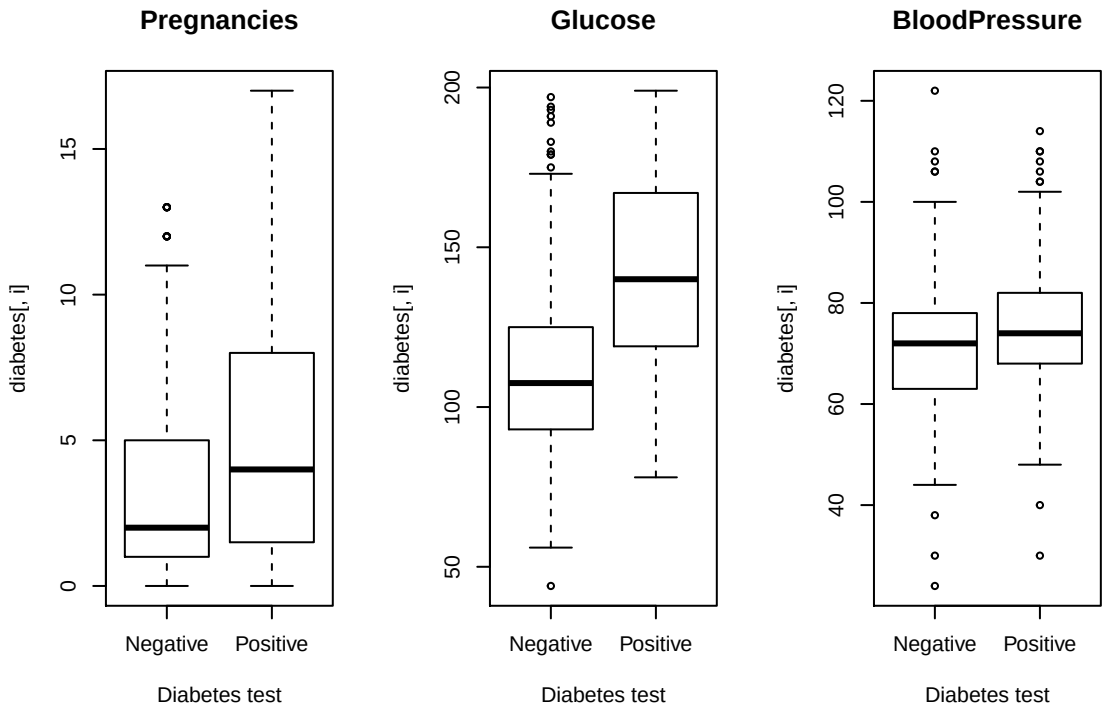
```
## mu=70
## sigma=12
## df=9
## t_obs=44.82382
## p-val=9.945303e-07
```

Question 3

Let $Y_i \overset{\text{ind}}{\sim} \text{Bernoulli}(\sigma(\theta^T x_i))$ for $i \in [1, n]$ $\pi_i = \mathbb{P}(Y_i = 1)$ where $\sigma(z) := \frac{1}{1+e^{-z}}$.
Then

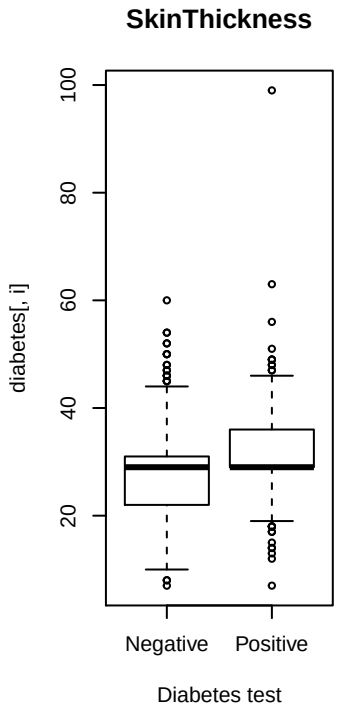
$$
\begin{aligned}
\pi_i &:= \mathbb{P}(Y_i = 1) \\
&= \sigma(\theta^T x_i) \\
&:= \frac{1}{1 + e^{-\theta^T x_i}} \\
&= \frac{1}{1 + e^{-\sum_{j=1}^{d} \theta_j x_{ij}}} \\
\implies \ln \pi_i &= \ln\left(\frac{1}{1 + e^{-\sum_{j=1}^{d} \theta_j x_{ij}}}\right) \\
&= \ln 1 - \ln(1 + e^{-\sum_{j=1}^{d} \theta_j x_{ij}}) \\
&= -\ln(1 + e^{-\sum_{j=1}^{d} \theta_j x_{ij}}) \\
\text{and} \quad \ln(1 - \pi_i) &= \ln\left(1 - \frac{1}{1 + e^{-\sum_{j=1}^{d} \theta_j x_{ij}}}\right) \\
&= \ln\left(\frac{e^{-\sum_{j=1}^{d} \theta_j x_{ij}}}{1 + e^{-\sum_{j=1}^{d} \theta_j x_{ij}}}\right) \\
&= \ln(e^{-\sum_{j=1}^{d} \theta_j x_{ij}}) - \ln(1 + e^{-\sum_{j=1}^{d} \theta_j x_{ij}}) \\
&= -\left(\sum_{j=1}^{d} \theta_j x_{ij}\right) - \ln(1 + e^{-\sum_{j=1}^{d} \theta_j x_{ij}}) \\
\implies \ln \frac{\pi_i}{1 - \pi_i} &= \ln(\pi_i) - \ln(1 - \pi) \\
&= -\ln(1 + e^{-\sum_{j=1}^{d} \theta_j x_{ij}}) + \left(\sum_{j=1}^{d} \theta_j x_{ij}\right) + \ln(1 + e^{-\sum_{j=1}^{d} \theta_j x_{ij}}) \\
&= \sum_{j=1}^{d} \theta_j x_{ij}
\end{aligned}
$$

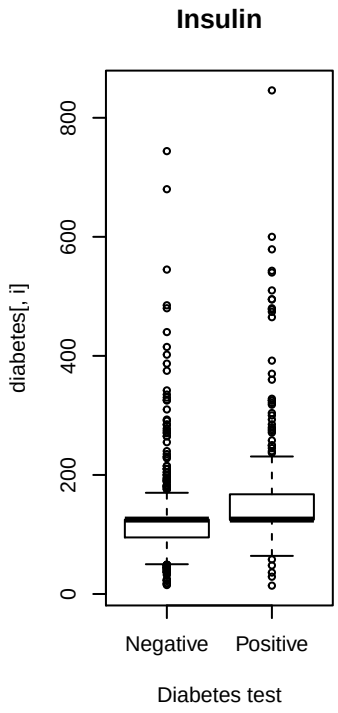
```
Xnames <- colnames(diabetes[, -9]) #get names of explanatory variables
par(mfrow=c(1,3))
for (i in 1:8) {
boxplot(diabetes[,i]~diabetes$Outcome,main=paste(Xnames[i]),
names=c("Negative","Positive"),xlab="Diabetes test")
}
```

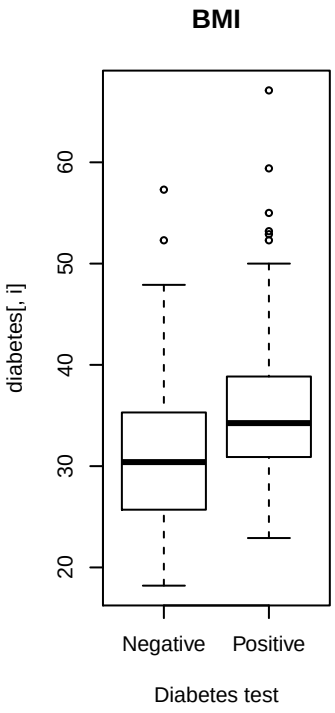

Pregnancies
diabetes[, i]
0
5
10
15
Negative
Positive
Diabetes test

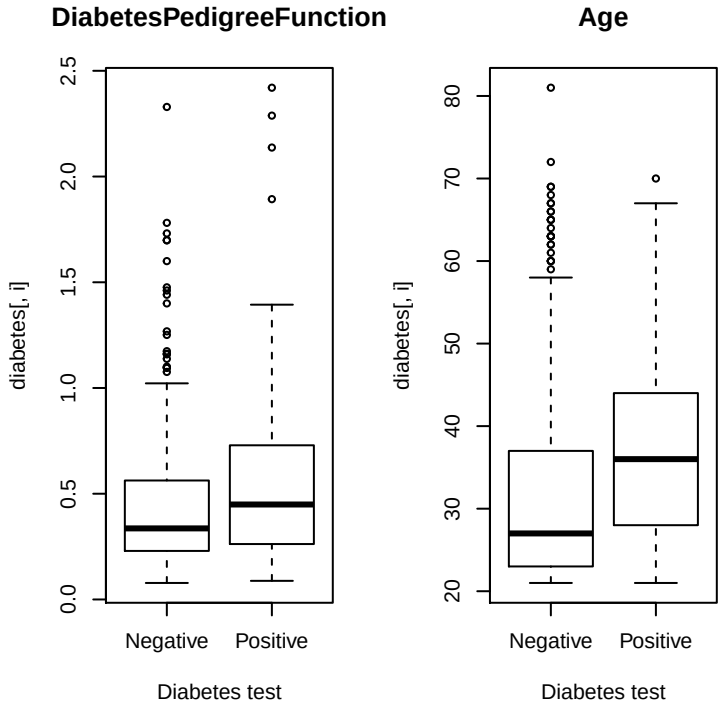
Glucose
diabetes[, i]
50
100
150
200
Negative
Positive
Diabetes test

BloodPressure
diabetes[, i]
40
60
80
100
120
Negative
Positive
Diabetes test


SkinThickness
diabetes[, i]
20
40
60
80
100
Negative
Positive
Diabetes test


Insulin
diabetes[, i]
0
200
400
600
800
Negative
Positive
Diabetes test


BMI
diabetes[, i]
20
30
40
50
60
Negative
Positive
Diabetes test



Question 4

Here I shall test whether the variables *BloodPressure, SkinThichness, Insulin* and *Age* are statistically significant to the development of diabetes.
To do so I shall test the hypotheses

$$H_0 : \boldsymbol{\theta} := (\theta_3, \theta_4, \theta_5, \theta_8) = \mathbf{0} \text{ against } H_1 : \boldsymbol{\theta} \neq \mathbf{0}$$

Consider the likelihood ratio statistic

$$\Lambda_n := \frac{L(\hat{\boldsymbol{\theta}}_0; \boldsymbol{x})}{L(\hat{\boldsymbol{\theta}}_{\text{MLE}}; \boldsymbol{x})}$$

and define test statistic

$$T_n(\mathbf{X}) := -2\Lambda_n = -2[\ell(\hat{\boldsymbol{\theta}}_0; \boldsymbol{X}) - \ell(\hat{\boldsymbol{\theta}}_{\text{MLE}}; \boldsymbol{X})] \sim \chi^2_r$$

where $r = 4$ since we only have four restrictions under the null hypothesis.

```
X_rest<-cbind(1,as.matrix(diabetes[,c(1,2,6,7)])) # Variables we are not testing (ie assuming others=0)
X_full<-cbind(1,as.matrix(diabetes[,1:8])) # all variables
Y<-diabetes[,9] # outcomes

# sigmoid function
sigma<-function(z) {
   1/(1+exp(-z))
}


# Log likelihood
ell<-function(theta,X,y) {
```

```
   p<-as.vector(sigma(X%*%theta))
   sum(y*log(p)+(1-y)*log(1-p))
}

# score function
score<-function(theta,X,y) {
   p<-as.vector(sigma(X%*%theta))
   as.vector(t(X)%*%(y-p))
}

# MLE
maximise.ell<-function(ell,score,X,y,theta0) {
   optim.out<-optim(theta0, fn=ell, gr=score, X=X, y=y, method="BFGS", control=list(fnscale=-1, maxit=1
   return(list(theta=optim.out$par, value=optim.out$value))
}

theta_hat_0.value<-maximise.ell(ell,score,X_rest,Y,rep(0,5))$value
theta_hat_mle.value<-maximise.ell(ell,score,X_full,Y,rep(0,9))$value
cat("ell(theta_hat_0): ",theta_hat_0.value,"\nell(theta_hat_mle): ",theta_hat_mle.value,sep="")
```

```
## ell(theta_hat_0): -358.1828
## ell(theta_hat_mle): -356.4209
```

Using these results we can calculate an observed test statistic

$$T_n(\mathbf{x}) = -2[\ell(\hat{\boldsymbol{\theta}}_0;\boldsymbol{x}) - \ell(\hat{\boldsymbol{\theta}}_{\text{MLE}};\boldsymbol{x})] = -2[-358.182779 - -356.4209352] = 3.5236876$$

Since $T_n(\mathbf{X}) \sim \chi^2_4$ we have an observered p-value of

$$p(\mathbf{x}) := \mathbb{P}(T_n(\mathbf{X}) \geq T_n(\mathbf{x}); H_0) = \mathbb{P}(\chi^2_4 \geq 3.5236876) = 0.4742857$$

Using the code described in the epilogue we can confirm this value

```
model1<-glm(Y~X_full,family=binomial) #full model
model2<-glm(Y~X_rest,family=binomial) #restricted model
suppressMessages(library(lmtest))
lrtest(model1, model2)
```

```
## Likelihood ratio test
##
## Model 1: Y ~ X_full
## Model 2: Y ~ X_rest
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   9 -356.42
## 2   5 -358.18 -4 3.5237     0.4743
```

This is not statistically significant enough to suggest rejecting H_0.
Thus we accept that the variables *BloodPressure, SkinThichness, Insulin* and *Age* are not statistically significant for the development of diabetes.

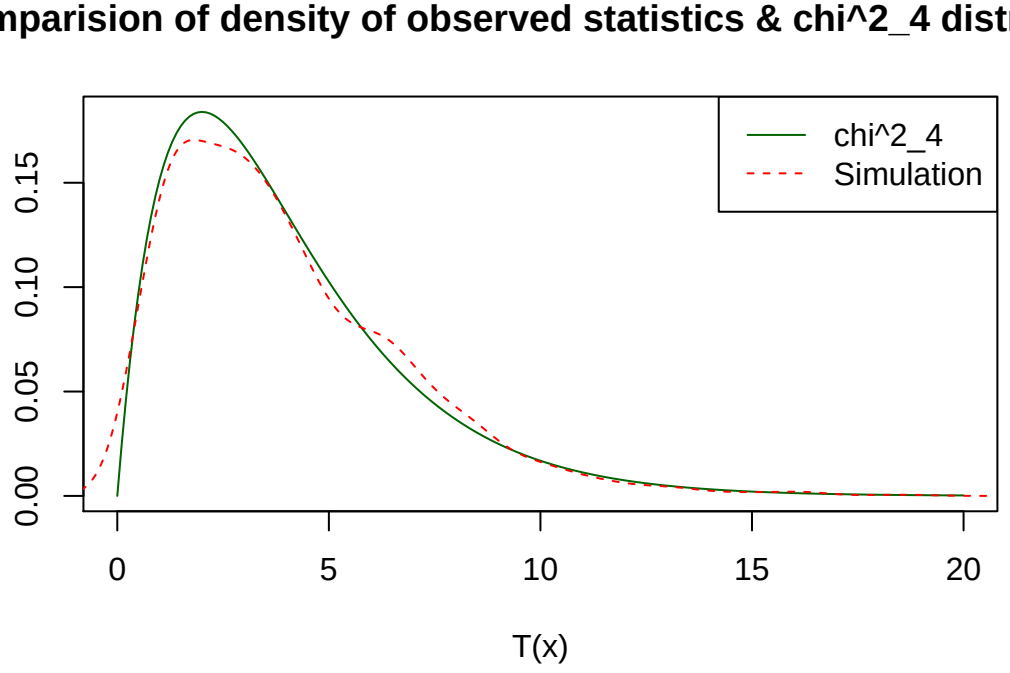
Question 5

```
set.seed(16111998)
generate.ys<-function(X,theta) {
   n<-dim(X)[1]
```

```
   rbinom(n,size=1,prob=sigma(X%*%theta))
}

simulate<-function(theta_hat_mle) {
   new_Y<-generate.ys(X_rest,theta_hat_mle)

   theta_hat_0.value<-maximise.ell(ell,score,X_rest,new_Y,rep(0,5))$value
   theta_hat_mle.value<-maximise.ell(ell,score,X_full,new_Y,rep(0,9))$value

   t_obs<--2*(theta_hat_0.value-theta_hat_mle.value)
}

n_trials=2000; m<-4
theta_hat_mle<-maximise.ell(ell,score,X_rest,Y,rep(0,5))$theta
simulation.raw<-sapply(1:n_trials, function(i) simulate(theta_hat_mle))
```

a)

```
x<-seq(0,20,0.1)
plot(x,dchisq(x,m),type="l",col="darkgreen",xlab="T(x)",ylab="",main="Comparision of density of observe
lines(density(simulation.raw),col="red",lty=2)
legend("topright",legend=c("chi^2_4","Simulation"),lty=1:2,col=c("darkgreen","red"))
```



b)

```
breaks<-c(-Inf,seq(1,13,by=1),Inf)
obs<-table(cut(simulation.raw,breaks))
```

```r
exp<-n_trials*(pchisq(breaks[-1],4)-pchisq(breaks[-length(breaks)],4))
round(cbind(obs,exp),1)
```

```
##           obs   exp
## (-Inf,1]  177 180.4
## (1,2]     365 348.1
## (2,3]     338 355.9
## (3,4]     298 303.6
## (4,5]     231 237.4
## (5,6]     157 176.3
## (6,7]     163 126.5
## (7,8]      91  88.6
## (8,9]      70  61.0
## (9,10]     40  41.3
## (10,11]    25  27.7
## (11,12]    15  18.4
## (12,13]    10  12.2
## (13, Inf]  20  22.6
```

```r
t_obs<-sum((obs-exp)^2/exp) # observered test statistic
p_val<-1-pchisq(t_obs,df=m) # p-value
cat("df=",m,"\nt_obs=",t_obs,"\np-val=",p_val,sep="")
```

```
## df=4
## t_obs=17.71702
## p-val=0.001401555
```