

Computer Practical 3

Statistics 2

Dom Hutchinson

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

missing<-function(v){ # Replace missing values with median of non-zero values
  med<-median(v[v>0])
  v[v==0]<-med
  return(v)
}

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)
```

Question 1

Here I shall test the hypotheses

H_0 : BloodPressure data can be modelled by a $\mathcal{N}(\mu, \sigma^2)$ distribution

H_1 : BloodPressure data can **not** be modelled by a $\mathcal{N}(\mu, \sigma^2)$ distribution

I shall use the maximum likelihood estimates for μ and σ^2 .

$$\hat{\mu}_{\text{MLE}} = \frac{1}{n} \sum_{i=1}^n x_i \text{ and } \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

Define the test statistic

$$T_{\text{pearson}}(\mathbf{X}) = \sum_{j=1}^m \frac{(o_j - e_j)^2}{e_j} \rightarrow_{\mathcal{D}} \chi_r^2$$

where o_j is the number of observations of class j and e_j is the expected number of observations of class j , assuming H_0 is true.

In this case $r = m - 3 = 7 - 3 = 4$ since we have 7 classes of observations but 3 constraints due to the normal distribution.

```
breaks<-c(-Inf,seq(45,95,by=10),Inf) # Quantise data
obs<-table(cut(diabetes$BloodPressure,breaks))

# Perform Pearson's Goodness of Fit Test
n<-length(diabetes$BloodPressure) # number of data points
r<-length(obs)-3 # degrees of freedom
mu<-sum(diabetes$BloodPressure)/n # MLE mean
sigma<-sqrt(sum((diabetes$BloodPressure-mu)^2)/n) # MLE variance

exp<-n*(pnorm(breaks[-1],mean=mu,sd=sigma)-pnorm(breaks[-length(breaks)],mean=mu,sd=sigma)) # expected
round(cbind(obs,exp),1) # return data table
```

```
##          obs    exp
## (-Inf,45]    9    9.0
## (45,55]     44   48.7
## (55,65]    155  150.1
## (65,75]    271  241.9
## (75,85]    183  204.3
## (85,95]     83   90.4
## (95, Inf]   23   23.6

t_obs<-sum((obs-exp)^2/exp) # observed test statistic
p_val<-1-pchisq(t_obs,df=r) # p-value
cat("mu=",mu,"\nsigma=",sigma,"\ndf=",r,"\nt_obs=",t_obs,"\np_val=",p_val,sep="") # return test values

## mu=72.38672
## sigma=12.08876
## df=4
## t_obs=6.959031
## p_val=0.1380692
```

Here the p -value (0.1380692) is not statistically significant enough to reject H_0 , for a reasonable significance level.

Thus we accept that *BloodPressure* can be modelled by a Normal distribution.

Question 2

Let $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \text{Normal}(\mu, 12^2)$ model the blood pressure 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

$$T_{\text{pearson}}(\mathbf{X}) = \sum_{j=1}^m \frac{(o_j - e_j)^2}{e_j} \rightarrow_{\mathcal{D}} \chi_r^2$$

where o_j is the number of observations of class j and e_j is the expected number of observations of class j , assuming H_0 is true.

In this case $r = 7 - 1 = 6$ due to the breaks chosen in Question 1.

```
mu<-70; sigma<-12
m<-length(obs) # number of classes

exp<-n*(pnorm(breaks[-1],mean=mu,sd=sigma)-pnorm(breaks[-length(breaks)],mean=mu,sd=sigma)) # expected
round(cbind(obs,exp),1) # display data table
```

```
##          obs    exp
## (-Inf,45]    9   14.3
## (45,55]     44   66.8
## (55,65]    155  178.8
## (65,75]    271  248.1
## (75,85]    183  178.8
## (85,95]     83   66.8
## (95, Inf]   23   14.3

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

```
## mu=70
## sigma=12
## df=6
## t_obs=24.35191
## p_val=0.000449861
```

Here the p -value (4.4986098×10^{-4}) is statistically significant enough to reject H_0 , for a reasonable significance level.

Thus we accept the alternative hypothesis, that $\mu > 70$.

Question 3

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

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

```
# 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=1000))
  return(list(theta=optim.out$par, value=optim.out$value))
}

```

Question 4

Here I shall test whether the variables *BloodPressure*, *SkinThickness*, *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

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

where $\hat{\boldsymbol{\theta}}_0$ is the maximum likelihood estimator under the null hypothesis and $\hat{\boldsymbol{\theta}}_{\text{MLE}}$ is the maximum likelihood estimator for the full model.

Define test statistic

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

where $r = 4$ since the null hypothesis specifies restrictions on four variables.

```

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

```

```

theta_hat_0.value<-maximise.ell(ell,score,X_rest,Y,rep(0,5))$value # MLE under H0
theta_hat_mle.value<-maximise.ell(ell,score,X_full,Y,rep(0,9))$value # MLE for full model
cat("ell(theta_hat_0): ",theta_hat_0.value,"\nell(theta_hat_mle): ",theta_hat_mle.value,sep="") # output

```

```

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

```

Using these results we can calculate an observed test statistic

$$T_n(\mathbf{x}) = -2[\ell(\hat{\boldsymbol{\theta}}_0; \mathbf{x}) - \ell(\hat{\boldsymbol{\theta}}_{\text{MLE}}; \mathbf{x})] = -2[(-358.18) - (-356.42)] = 3.52$$

Since $T_n(\mathbf{X}) \sim \chi_4^2$ we have an observed p -value of

$$p(\mathbf{x}) := \mathbb{P}(T_n(\mathbf{X}) \geq T_n(\mathbf{x}); H_0) = \mathbb{P}(\chi_4^2 \geq 3.52) = 0.4743$$

Using the code described in the epilogue we can confirm this calculation.

```

model1<-glm(Y~X_full,family=binomial) #full model
model2<-glm(Y~X_rest,family=binomial) #restricted model
suppressMessages(library(lmtest)) # load library
lrtest(model1, model2) # perform linear regression test

```

```

## 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
```

Here the p -value (0.4743) is not statistically significant enough to reject H_0 , for a reasonable significance level. Thus we accept that the variables *BloodPressure*, *SkinThickness*, *Insulin* and *Age* are not statistically significant for the development of diabetes and thus $(\theta_3, \theta_4, \theta_5, \theta_8) = \mathbf{0}$.

Question 5

```
set.seed(795430357) # Set RNG seed

generate.y<-function(X,theta) { # Generate new outcomes from data
  n<-dim(X)[1]
  rbinom(n,size=1,prob=sigma(X%%theta))
}

simulate<-function(theta_hat_mle) {
  new_Y<-generate.y(X_rest,theta_hat_mle) # Generate new outcomes

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

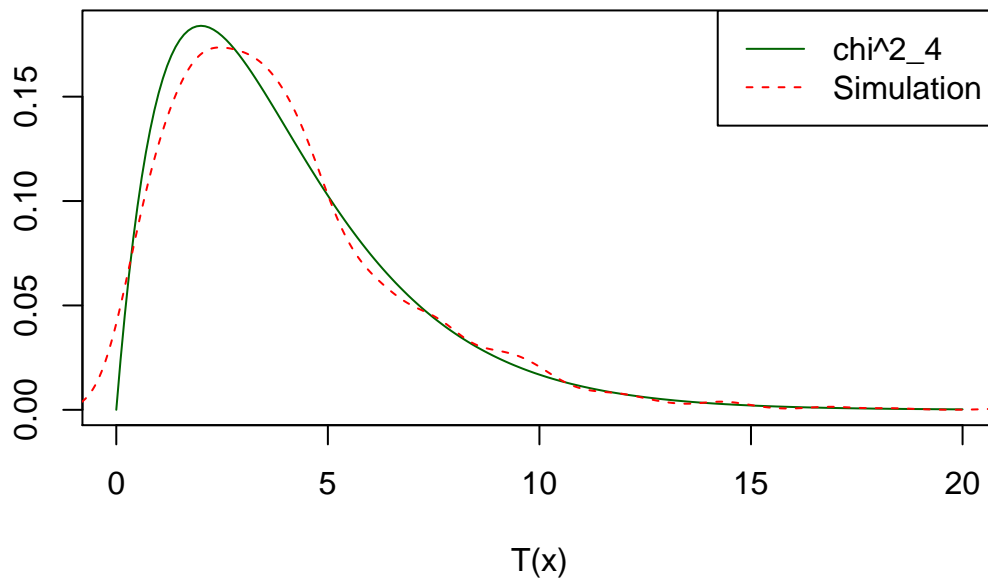
  t_obs<-2*(theta_hat_0.value-theta_hat_mle.value) # Calculate observed test statistic
}

n_trials<-1000
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)) # Run simulation
```

a)

```
x<-seq(0,20,0.1)
plot(x,dchisq(x,4),type="l",col="darkgreen",xlab="T(x)",ylab="",
      ,main="Comparision of density of observed statistics & chi^2_4 distribution") # Plot chi^2_m distr
lines(density(simulation.raw),col="red",lty=2) # Plot distribution of obsered test statistics
legend("topright",legend=c("chi^2_4","Simulation"),lty=1:2,col=c("darkgreen","red"))
```

Comparison of density of observed statistics & chi^2_4 distribu



b)

Here I shall test the hypotheses

$$H_0 : -2 \ln \Lambda_n \sim \chi_4^2 \text{ against } H_1 : -2 \ln \Lambda_n \not\sim \chi_4^2$$

This shall be done using Pearson's Goodness-of-Fit test. Usint test statistic

$$T_{\text{pearson}}(\mathbf{X}) = \sum_{j=1}^m \frac{(o_j - e_j)^2}{e_j} \rightarrow_{\mathcal{D}} \chi_r^2$$

In this case $r = 12 - 1 = 11$ since I split the observed data into twelve classes and there is one constraint on these classes.

```
breaks<-c(-Inf,seq(1,11,by=1),Inf) # quantise data
obs<-table(cut(simulation.raw,breaks))
```

```
exp<-n_trials*(pchisq(breaks[-1],4)-pchisq(breaks[-length(breaks)],4)) # Expected number of observation.
round(cbind(obs,exp),1) # Output data table
```

```
##      obs  exp
## (-Inf,1]   91  90.2
## (1,2]     159 174.0
## (2,3]     176 177.9
## (3,4]     171 151.8
## (4,5]     133 118.7
## (5,6]      74  88.1
## (6,7]      52  63.3
## (7,8]      52  44.3
## (8,9]      26  30.5
```

```
## (9,10]      30  20.7
## (10,11]     11  13.9
## (11, Inf]   25  26.6
```

```
r<-length(obs)-1
t_obs<-sum((obs-exp)^2/exp) # observed test statistic
p_val<-1-pchisq(t_obs,df=r) # p-value
cat("df=",r,"\\nt_obs=",t_obs,"\\np_val=",p_val,sep="") # output test values
```

```
## df=11
## t_obs=16.63217
## p_val=0.1192364
```

Here the p -value (0.1192364) is not statistically significant enough to reject H_0 , for a reasonable significance level.

Thus we accept that the test statistic $-2 \ln \Lambda_n$ is distributed according to χ_4^2 .