Statistics 2: Computer Practical 3

Joshua Acton 2100807

1 Data description

For this computer practical you will need to load the following data:

```
diabetes<-read.csv("diabetes_data.csv",header=T)
```

This dataset was originally created by the National Institute of Diabetes and Digestive and Kidney Diseases and was downloaded from https://www.kaggle.com/datasets. It includes medical records for 768 female patients of Pima Indian heritage, aged at least 21 years old who were tested for diabetes. Analytically, the dataset consists of the following measurements:

- `Pregnancies` : Number of pregnancies
- `Glucose` : Plasma glucose concentration a 2 hours in an oral glucose tolerance test (mg/dL)
- `BloodPressure`: Diastolic blood pressure (mm Hg)
- `SkinThickness`: Triceps skin fold thickness (mm)
- `Insulin` : 2-Hour serum insulin (mu U/ml)
- `BMI` : Body mass index (kg/m^2)
- `DiabetesPedigreeFunction`: a function indicating the likelihood of diabetes based on family history
- `Age` : Age of patient (>21)
- `Outcome` : Outcome of diabetes test (1: Test positive, 0:Test negative)

Observing the dataset carefully, it becomes obvious that for some of the variables (Glucose, BloodPressure, SkinThickness, Insulin and BMI) the missing values were recorded as 0's. This could lead to misleading results in the analysis, therefore we need to adjust these entries before we proceed. In general, we need to be very careful with missing data. Deleting observations is not usually a good practice. By doing so, we may loose a lot of information and in addition we may introduce bias in the data, e.g. if the missingness is related to the outcome of interest (diabetes in this case).

For these data, Pregnancies, DiabetesPedigreeFunction, Age and of course Outcome have no missing values and hence we prefer not to delete any of the patients. Instead, we replace all missing values with the median observation of the corresponding variable (we could have used the mean or even more complicated methods for *imputation*). To do this we run the following code:

```
missing<-function(var){
  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)
```

2 Testing the mean of normal data

We now focus on the BloodPressure measurements. According to the American Heart Association (AHA) a normal diastolic pressure should be in the range $60 - 80$mm Hg (millimeters of mercury), so one could say that the blood pressure for a 'healthy' person should be 70mm Hg. We want to test whether the diastolic pressure for the patients in the diabetes dataset is higher than 70mm Hg.

If the distribution is known, the test could be simplified significantly.

Question 1.[2 marks] Use the intervals

$$(-\infty, 45],\ (45, 55],\ (55, 65],\ (65, 75],\ (75, 85],\ (85, 95],\ (95, \infty)$$

to quantize the data. You can do this using the following code:

```
BP<-diabetes$BloodPressure
breaks<-c(-Inf,seq(45,95,by=10),Inf)
Obs<-table(cut(BP,breaks))
```

Hence, using Pearson's goodness of fit test, confirm that a Normal distribution is a valid assumption for the BloodPressure data, i.e. they are derived from $\mathcal{N}(\mu, \sigma^2)$ for an appropriate choice of the parameters μ and σ^2. You may find the handout on *Goodness-of-fit for continuous distributions* and the relevant case study helpful.

Solution:
We will conduct a Pearson's goodness of fit test using the following hypotheses:

$$H_0 : \text{The distribution of BloodPressure data is consistent with a } \mathcal{N}(\theta) \text{ distribution.}$$
$$H_1 : \text{The distribution of BloodPressure data is not consistent with a } \mathcal{N}(\theta) \text{ distribution.}$$

for some $\theta = (\mu, \sigma^2) \in \Theta = \mathbb{R} \times \mathbb{R}_+$.
Firstly, in order to compute the Pearson test statistic $T_{Pearson}(\mathbf{x})$ we need to compute the expected number of observations that fall in to each of the quantiles

$$\{I_1, ..., I_m\} = (-\infty, 45],\ (45, 55],\ (55, 65],\ (65, 75],\ (75, 85],\ (85, 95],\ (95, \infty)$$

under the null hypothesis.
To do this we will estimate

$$\theta = \hat{\theta}_{mle} = (\hat{\mu}_{mle}, \widehat{\sigma^2}_{mle}) = (\bar{x}, S^2)$$

from the BloodPressure data, and compute each probability

$$\mathbb{P}(B_P \in I_j),\ B_P \sim \mathcal{N}(\bar{x}, S^2),\ j \in \{1, ..., m\},$$

storing these in the vector $\mathbf{p}_0$.
We will then compute the pearson test statistic

$$T_{Pearson}(\mathbf{x}) = \sum_{j=1}^{m} \frac{(O_j - E_j)^2}{E_j}$$

where O_j is the number of observations in the j^{th} quantile given by the table Obs computed above, and $E_j = np_{0,j}$ is the expected number of observations in the j^{th} quantile.
Since we have modified our test statistic by using $\theta = \hat{\theta}_{mle}$ instead of $\theta = \theta_{opt}$ for the parameters under the null hypothesis, we know that

$$\mathbb{P}(W_{m-1-\dim(\Theta)} \geq T(\mathbf{x})) \leq \mathbb{P}(T_{modified}(\mathbf{X}) \geq T(\mathbf{x}); \theta) \leq \mathbb{P}(W_{m-1} \geq T(\mathbf{x}))$$

where $W_r \sim \chi^2_r$. Hence we can compute lower and upper bounds for our p-value as $\mathbb{P}(W_{m-1-\dim(\Theta)} \geq T(\mathbf{x}))$ and $\mathbb{P}(W_{m-1} \geq T(\mathbf{x}))$ respectively, where $m = 7$ and $\dim(\Theta) = 2$.

```r
# find the maximum likelihood estimate of the population mean
mu = mean(diabetes$BloodPressure)
# find the maximum likelihood estimate of the population variance
sigma2 = var(diabetes$BloodPressure)
# obtain the total number of samples in the original data
n = length(diabetes$BloodPressure)
# obtain the number of quantiles data will be "binned" into
m = length(breaks) - 1
# create a vector to store probabilities of observations falling into each quantile
p_0 = rep(0, m)

# loop through each quantile
for(i in 0:m) {
  # set the probability equal to the CDF of a Normal R.V. evaluated at the upper
  # end-point of the quantile
  p_0[i] = pnorm(breaks[i+1], mean=mu, sd=sqrt(sigma2))
  # if not the first quantile, subtract the probabilities of the data falling
  # into all previous quantiles
  if(i > 0) {
    p_0[i] = p_0[i] - sum(p_0[0:(i-1)])
  }
}

# multiply the total number of samples by the probability for each quantile
# to obtain a vector of the expected number of observations per quantile
E = n*p_0
E
```

```
## [1]   9.052613  48.788427 150.070593 241.767134 204.241387  90.437250  23.642597
```

Before proceeding, we verify that the expected values $\mathbf{E}$ have been quantised such that $E_j \geq 5 \; \forall \; j \in \{1, ..., m\}$.

```r
# convert the table containing observations in each quantile to a vector for
# ease-of-use
O <- as.vector(as.matrix(Obs[1:(length(Obs))]))

# compute the pearson test-statistic
T_x <- sum(((E-O)^2)/E)

# compute the lower bound on the p-value
pchisq(T_x, df=m-1-2, lower.tail=FALSE)
```

```
## [1] 0.1356217
```

```r
# compute the upper bound on the p-value
pchisq(T_x, df=m-1, lower.tail=FALSE)
```

```
## [1] 0.3203806
```

Hence performing an $\alpha-$level test with any reasonable α, (e.g. $\alpha \leq 0.1$), we have insufficient evidence to reject H_0 since the lower bound on our p-value is $> \alpha$, so we can reasonably assume that the data does come from a $\mathcal{N}(\mu, \sigma^2)$ distribution.

Question 2 [2 marks] Assuming that the BloodPessure measurements are realisations of i.i.d. random variables from $\mathcal{N}(\mu, 12)$, perform the following test for an appropriate (UMP) test statistic

$$H_0 : \mu = 70 \quad vs \quad H_1 : \mu > 70$$

reporting the p-value of the test. State clearly the test statistic used and justify your choice appropriately.

Solution:
The Neyman-Pearson test statistic for the mean of a normal distribution with known variance is

$$\begin{aligned} T_{\mathrm{NP}}(\mathbf{x}) &= \frac{f_n(\mathbf{x};\mu_1,\sigma^2)}{f_n(\mathbf{x};\mu_0,\sigma^2)} \\ &= \prod_{i=1}^{n} \frac{\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_i-\mu_1}{\sigma}\right)^2}}{\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_i-\mu_0}{\sigma}\right)^2}} \\ &= \prod_{i=1}^{n} \frac{e^{-\frac{1}{2}\left(\frac{x_i-\mu_1}{\sigma}\right)^2}}{e^{-\frac{1}{2}\left(\frac{x_i-\mu_0}{\sigma}\right)^2}} \\ &= e^{\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[(x_i-\mu_0)^2-(x_i-\mu_1)^2\right]} \\ &= e^{\frac{1}{2\sigma^2}\sum_{i=1}^{n}[x_i^2-2x_i\mu_0+\mu_0^2-x_i^2+2x_i\mu_1-\mu_1^2]} \\ &= e^{\frac{1}{2\sigma^2}n[(\mu_0^2-\mu_1^2)+2(\mu_1-\mu_0)\bar{x}]} \\ &= e^{\frac{1}{2\sigma^2}n[(\mu_0^2-\mu_1^2)]} \cdot e^{\frac{1}{2\sigma^2}n[2(\mu_1-\mu_0)\bar{x}]} \\ &= Mf(\bar{x}) \end{aligned}$$

where $M = e^{\frac{1}{2\sigma^2}n[(\mu_0^2-\mu_1^2)]}$ is independent of $\mathbf{x}$ and $f(\bar{x}) = e^{\frac{1}{2\sigma^2}n[2(\mu_1-\mu_0)\bar{x}]}$ is bijective and increasing in $\bar{x}$ since $\mu_1 > \mu_0$.
Hence

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

is a UMP test statistic for μ.
Thus we can compute the p-value directly as

$$p(\mathbf{x}) = \mathbb{P}(T(\mathbf{X}) \geq T(\mathbf{x}); \mu_0, \sigma^2)$$

where

$$T(\mathbf{x}) = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

is the sample mean, and the test statistic has distribution

$$T(\mathbf{X}) = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right),$$

with the following R code:

```r
# obtain the sample size
n <- length(diabetes$BloodPressure)
# obtain the population mean under the null hypothesis
mu_0 <- 70
# obtain the value of the observed test statistic
T_x <- mean(diabetes$BloodPressure)
# obtain the assumed known value of population variance
var <- 12

# compute the p-value
pnorm(T_x, mu_0, sqrt(var/n), lower.tail=FALSE)
```

```
## [1] 1.422938e-81
```

This p-value is minuscule, which can be understood as $\frac{\sigma^2}{n} = \frac{12}{768} \approx 0.0156$ shows that the variance of the test statistic $T(\mathbf{X})$ is tiny under the null hypothesis, so the probability of seeing a sample mean that far away from the population mean is correspondingly small. Thus, for any reasonable α-level test, we have sufficient evidence to reject H_0 and conclude that most likely $\mu > 70$.

3 Logistic regression (again!)

Recall the logistic model from the previous computer practicals,

$$Y_i \overset{\text{ind}}{\sim} \text{Bernoulli}(\sigma(\theta^T x_i)), \qquad i \in \{1, \ldots, n\},$$

where $x_1, \ldots, x_n$ are d-dimensional real vectors of explanatory variables, and σ is the standard logistic function

$$\sigma(z) = \frac{1}{1 + \exp(-z)}.$$

Question 3 [1 mark] Show that for $\pi_i = \mathbb{P}(Y_i = 1)$[1],

$$\log \frac{\pi_i}{1 - \pi_i} = \sum_{j=1}^{d} \theta_j x_{ij}.$$

Solution:
Since $Y_i \overset{\text{ind}}{\sim} \text{Bernoulli}(\sigma(\theta^T x_i))$ we have that $\pi_i = \mathbb{P}(Y_i = 1) = \sigma(\theta^T x_i)$. Hence,

$$\begin{aligned}
\log \frac{\pi_i}{1 - \pi_i} &= \log \frac{\sigma(\theta^T x_i)}{1 - \sigma(\theta^T x_i)} \\
&= \log \frac{\frac{1}{1+e^{-\theta^T x_i}}}{1 - \frac{1}{1+e^{-\theta^T x_i}}} \\
&= \log \frac{1}{1 + e^{-\theta^T x_i}} \cdot \frac{1}{1 - \frac{1}{1+e^{-\theta^T x_i}}} \\
&= \log \frac{1}{1 + e^{-\theta^T x_i}} \cdot \frac{1}{\frac{e^{-\theta^T x_i}}{1+e^{-\theta^T x_i}}} \\
&= \log \frac{1}{1 + e^{-\theta^T x_i}} \cdot \frac{1 + e^{-\theta^T x_i}}{e^{-\theta^T x_i}} \\
&= \log \frac{1}{e^{-\theta^T x_i}} \\
&= \log e^{\theta^T x_i} \\
&= \theta^T x_i \\
&= \sum_{j=1}^{d} \theta_j x_{ij}
\end{aligned}$$

as required.

4 Hypothesis testing in logistic regression

In the second computer practical it was mentioned that if an explanatory variable has no effect on the probability of the response variable then we expect the corresponding coefficient to be equal to 0. We will

[1] The quantity $\frac{\pi_i}{1-\pi_i}$ is called the *odds* for individual i and the logarithm of this is usually referred as *logit*. By calculating the ratio of the odds for two different individuals (*odds ratio*) we can compare the odds of these two individuals.

examine this in a more formal way through hypothesis testing. Consider the logistic model described above and assume we want to test:

$$H_0 : \theta_{i_1} = \theta_{i_2} = ... = \theta_{i_r} = 0$$
$$H_1 : \text{at least one of } \theta_{i_1}, \theta_{i_2}, ..., \theta_{i_r} \text{ not equal to } 0,$$

where $r \leq d$ and $i_1, i_2, ..., i_r \in \{1, 2, ..., d\}$.

Consider the generalised likelihood ratio statistic for this *nested* test,

$$\Lambda_n = \frac{\sup_{H_0} L(\boldsymbol{\theta}; \boldsymbol{y})}{\sup_{\boldsymbol{\theta}} L(\boldsymbol{\theta}; \boldsymbol{y})} = \frac{L(\hat{\boldsymbol{\theta}}_0; \boldsymbol{y})}{L(\hat{\boldsymbol{\theta}}_{MLE}; \boldsymbol{y})}$$

where $\hat{\boldsymbol{\theta}}_0$ is the maximum likelihood estimator under the null hypothesis, and $\hat{\boldsymbol{\theta}}_{MLE}$ is the maximum likelihood estimator for the full model (i.e. all $\theta_i \neq 0$).

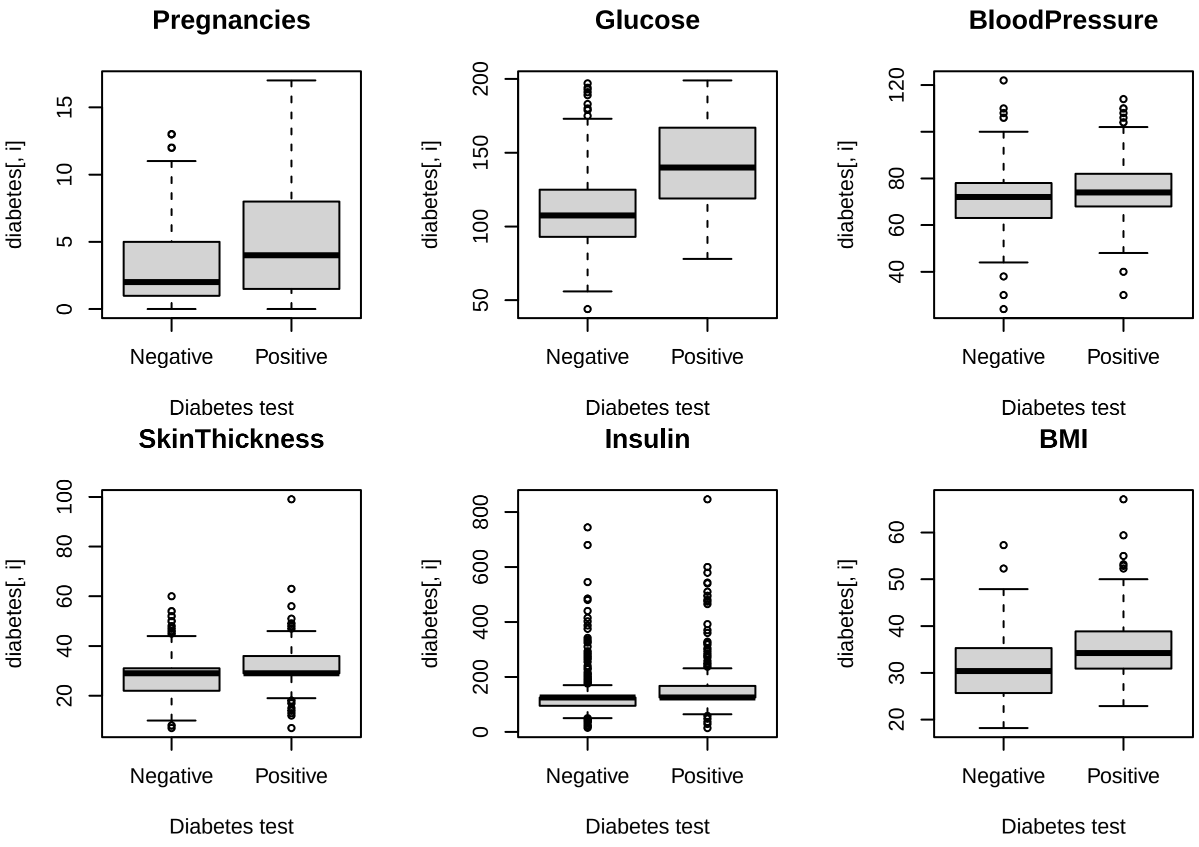
Then,

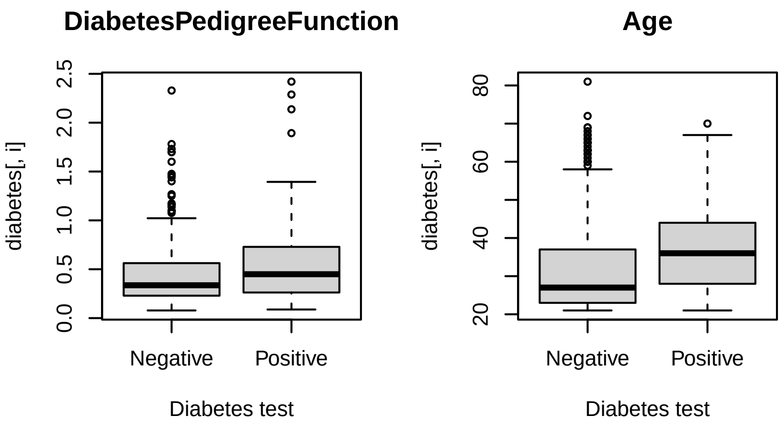
$$-2 \log \Lambda_n = -2\{l(\hat{\boldsymbol{\theta}}_0; \boldsymbol{y}) - l(\hat{\boldsymbol{\theta}}_{MLE}; \boldsymbol{y})\}$$

has a $\mathcal{X}^2_r$ distribution **under the null hypothesis** (notice that r is the number of restrictions under the null hypothesis).

Returning to the diabetes data, we plot all explanatory variables in the data against the Outcome (patient diabetic or not) to get a first impression of possible effects.

```r
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")
}
```





Question 4 [2 marks] Use the generalised likelihood ratio test to decide whether the variables BloodPressure, SkinThickness, Insulin and Age are statistically significant for the development of diabetes. To do this start by forming the matrices that correspond to the restricted model under the null hypothesis (X_rest) and to the full model (X_full).

```
X_full<-cbind(1,as.matrix(diabetes[,1:8]))
X_rest<-cbind(1,as.matrix(diabetes[,c(1,2,6,7)]))
Y<-diabetes[,9]
```

Notice that, as in computer practical 2, we add a constant term in the model, and hence the vector of parameters becomes $\boldsymbol{\theta} = (\theta_0, \theta_1, ..., \theta_8)$. You will need the following functions from computer practical 2 (notice the modification in the return vector of ell.maximize):

```
sigma <- function(v) {
  1/(1+exp(-v))
}

ell <- function(theta, X, y) {
  p <- as.vector(sigma(X%*%theta))
  sum(y*log(p) + (1-y)*log(1-p))
}

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

maximize.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, reltol=1e-16))
return(list(theta=optim.out$par, value=optim.out$value))
}
```

Solution:

```
# total number of parameters in the model (+1 constant)
parameters_total <- dim(X_full)[2]
# number of parameters we are interested in (+1 constant)
parameters_reduced <- dim(X_rest)[2]
# degrees of freedom
df <- parameters_total - parameters_reduced
```

```r
# value of log-likelihood function under ML estimate of parameters of full model
ell_mle <- maximize.ell(ell, score, X_full, Y, rep(0,parameters_total))$value
# value of log-likelihood function under ML estimate of parameters under null hypothesis
ell_mle0 <- maximize.ell(ell, score, X_rest, Y, rep(0, df+1))$value

# value of the GLR test statistic
T_x <- -2*(ell_mle0-ell_mle)

# computation of the p-value from the assumed chisq distribution under the null hypothesis
pchisq(T_x, df=df, lower.tail=FALSE)
```

```
## [1] 0.4742857
```

This p-value is extremely large, hence we do not have sufficient evidence to reject H_0 and can conclude that the variables BloodPressure, SkinThickness, Insulin and Age are most likely not statistically significant for the development of diabetes.

5 Verifying the distribution of $-2\log\Lambda_n$.

Recall the following function from computer practical 2, which simulates values of the response variable for a given matrix of explanatory variables, X, and a given vector of parameters, theta.

```r
generate.ys <- function(X, theta) {
  n <- dim(X)[1]
  rbinom(n, size = 1, prob=sigma(X%*%theta))
}
```

Question 5 [3 marks] Consider the test in Question 4. By repeated experiments, simulate the test statistic $-2\log\Lambda_n$ *under the null hypothesis* using the appropriate maximum likelihood estimate for $\boldsymbol{\theta}$. Hence, verify that it has a $\mathcal{X}^2$ distribution with the underlying degrees of freedom. To do this,

(a) Plot the density of the simulated $-2\log\Lambda_n$ values against the density of the corresponding chi-squared distribution

(b) Perform the Pearson's goodness-of-fit test. [Note that the intervals chosen to quantize the observed data should meet the criterion of each interval having expected counts ≥ 5.]

Solution:
(a)
Under the null hypothesis, when we simulate values of the test statistic $T(\mathbf{x}) = -2\log\Lambda_n$, we do so using simulated values of the response variable, generated from the ML estimate of θ_0 and the restricted model X_rest, where we have assumed $\theta_{BloodPressure} = \theta_{SkinThickness} = \theta_{Insulin} = \theta_{Age} = 0$.

To do this, we obtain $\widehat{\theta}_{0mle}$ and use this to generate 1000 simulated values of the response variable, each a vector of length 768 corresponding to the original sample size. We then use these simulated values to generate 1000 values of the test statistic using the same procedure as Q4, just with the actual response variable values replaced with each of our simulated values.

Finally, we plot the density of our 1000 simulated test statistics against the density of a χ^2_4 distribution, as we assume $T(\mathbf{x})$ follows this distribution under the null hypothesis.

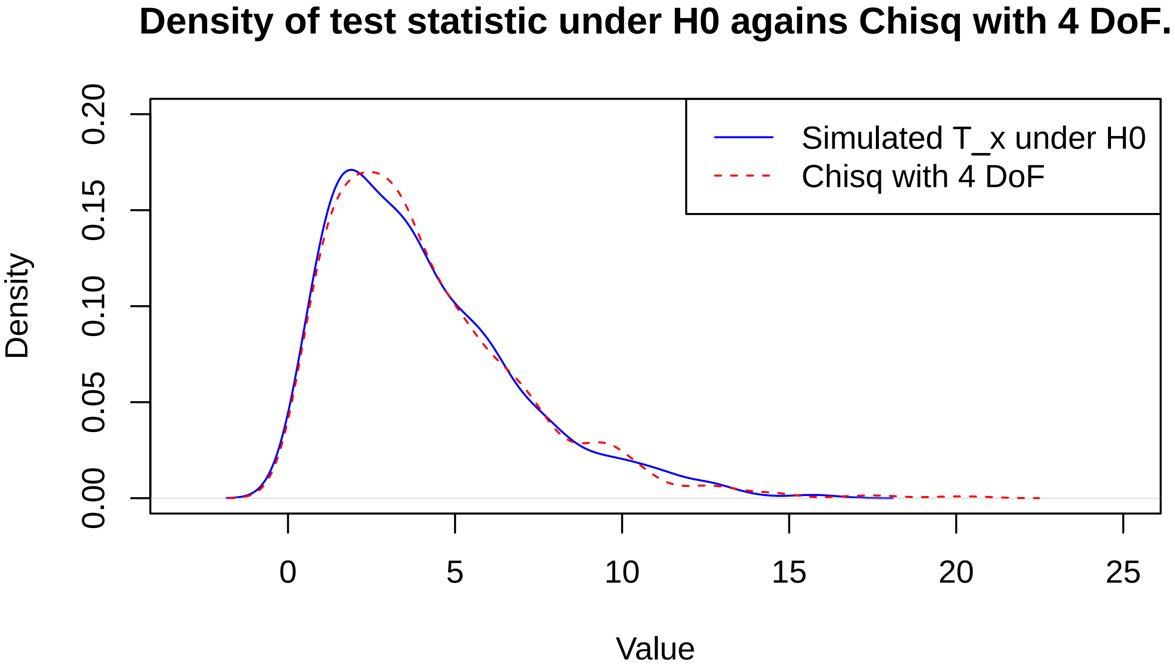
```r
# obtain the ML estimate of the parameters under the null hypothesis
theta_0 <- maximize.ell(ell, score, X_rest, Y, rep(0, df+1))$theta

# set the number of simulated test statistics to obtain
trials <- 1000
# create a vector to store the simulated test statistics
```

```r
test_statistics <- rep(0, trials)

for(i in 1:trials) {
  # generate a set of simulated response variable values
  y <- generate.ys(X_rest, theta_0)
  # obtain the value of the log-likelihood function for these simulated values under the
  # full model
  ell_mle <- maximize.ell(ell, score, X_full, y, rep(0,parameters_total))$value
  # obtain the value of the log-likelihood function for these simulated values under the
  # null hypothesis
  ell_mle0 <- maximize.ell(ell, score, X_rest, y, rep(0, df+1))$value
  # obtain the value of the simulated test statistic
  test_statistics[i] <- -2*(ell_mle0-ell_mle)
}

# plot the density of the simulated test statistics
plot(density(test_statistics), col="blue", xlim=c(-3, 25), ylim=c(0, 0.2), xlab="Value",
     main="Density of test statistic under H0 agains Chisq with 4 DoF.")
# plot the density of 1000 samples from a chisq distribution with df=4
lines(density(rchisq(n=1000, df=df)), col="red", lty=2)
ticks = seq(from=0, to=0.2, by=0.02)
legend = c("Simulated T_x under H0", "Chisq with 4 DoF")
legend("topright", legend=legend, lty=c(1, 2), col=c("blue", "red"))
```



We can see that the density of our simulated test statistic under H_0 matches the density of a χ^2_4 distribution very closely.

Next we verify statistically that $T(\mathbf{x})$ does approximate a χ^2_4 distribution H_0 using Pearson's goodness-of-fit test.

(b)

We will conduct a Pearson's goodness of fit test using the following hypotheses:

$$H_0 : \text{The distribution of } T(\mathbf{x}) \text{ is consistent with a } \ \chi^2_4 \text{ distribution.}$$
$$H_1 : \text{The distribution of } T(\mathbf{x}) \text{ is not consistent with a } \ \chi^2_4 \text{ distribution.}$$

To conduct this test, firstly we must quantise our simulated test statistics into j intervals with $e_j \geq 5$. When $W \sim \chi^2_4$ as under H_0,

$$\mathbb{P}(W \in (-\infty, 0]) = 0,$$

so we will only consider quantising into intervals in the range $(0, \infty)$.

In particular, we will construct our m quantiles such that

$$\forall\, j \in \{1, ..., m\},\ p_{0,j} = \mathbb{P}(W \in I_j) = \frac{1}{m}$$

where $W \sim \chi^2_4$.

Once we have quantised the data, we will follow much the same testing procedure as in Q1, with the exception that here we are testing a specified parameter of the distribution, not the family of distributions as a whole. Hence, since we know that

$$T_{Pearson}(\mathbf{x}) = \sum_{i=1}^{m} \frac{(O_j - E_j)^2}{E_j} \longrightarrow_D W \sim \chi^2_{m-1},$$

we can derive an explicit p-value for our test.

```
# derive m equiprobable quantiles from the chisq distribution with df=4, each with
# probability 1/m
m <- 10
breaks <- rep(0, m)

for(i in 0:m) {
  breaks[i+1] = qchisq(i/m, df=df)
}

# create a 0-vector to store the probabilities of observations falling into each quantile
p_0 = rep(0, m)

# Loop through each quantile
for(i in 0:m) {
  # Set the probability equal to the CDF of a Chi-squared R.V. evaluated at the upper
  # end-point of the quantile
  p_0[i] = pchisq(breaks[i+1], df=df)
  # If not the first quantile, subtract the probabilities of the data falling
  # into all previous quantiles
  if(i > 0) {
    p_0[i] = p_0[i] - sum(p_0[0:(i-1)])
  }
}

# compute the expected number of observations in each quantile
E = trials*p_0

E
```

```
##  [1] 100 100 100 100 100 100 100 100 100 100
```

Here we can verify that the expected values $\mathbf{E}$ have been quantised such that $E_j \geq 5 \; \forall \; j \in \{1, ..., m\}$.

```
# quantise the observed data into the intervals derived above
T_Obs<-table(cut(test_statistics,breaks))

# convert the table containing observations in each quantile to a vector for
# ease-of-use
O <- as.vector(as.matrix(T_Obs[1:(length(T_Obs))]))

# computing the pearson test statistic
T_x <- sum(((E-O)^2)/E)

# computing the p-value
pchisq(T_x, df=m-1, lower.tail=FALSE)
```

```
## [1] 0.4226379
```

This p-value is extremely large, hence we do not have sufficient evidence to reject H_0 and can conclude that the test statistic is consistent with a χ^2_4 distribution.

6 Epilogue

Throughout this computer practical, we applied various forms of hypothesis testing to make decisions about the parameters of a distribution (using generalised likelihood ratio test) or even for the distribution of some observed data (goodness-of-fit test).

We also saw how one can decide between two nested models in logistic regression analysis using an appropriate form of the generalised likelihood ratio. As mentioned in the epilogue of Computer Practical 2, we can easily obtain the estimated values for the parameters in logistic regression using the *glm* function. One can then use the *lrtest* (need to load *lmtest* package) to perform hypothesis testing for nested models. For example, if we have two explanatory variables x_1, x_2 and we want to test if x_1 is significant for the response y, we can test this in R as follows:

```
model1<-glm(y~x1+x2,family=binomial) #full model

model2<-glm(y~x2,family=binomial) #restricted model

library(lmtest)

lrtest(model1, model2)
```

[You may use this (only!) to confirm your result in question 4.]