Q1: Use the gene data in the below link

http://web.stanford.edu/~hastie/CASI_files/DATA/leukemia_big.csv

(a) Perform 1000 nonparametric bootstrap replications of $\overline{ALL} = 0.752$. You can use program bcanon from the CRAN library "bootstrap"

~~~~
```{r}

library(ggplot2)

library(ggpubr)

theme_set(theme_pubr())

library(dplyr)

library(tidyverse)

all<-read.csv("C:/Users/Alejandro/Documents/1 ITESM/MCI/2do semestre/Nueva carpeta/Final/ALL.csv",  header = TRUE, stringsAsFactors = FALSE, na.strings = "NA")

```

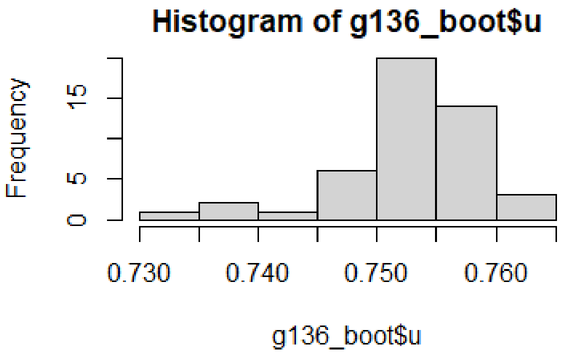
```{r}

all<-select(all, -c(Columna1))



g136 <- select(all, X136)

```

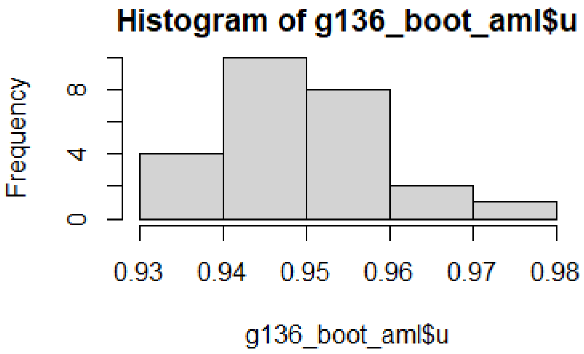
```{r}

library(bootstrap)

g136<-as.numeric(unlist(g136))

theta<- function(g136){mean(g136)}

g136_boot <- bcanon(g136,1000, theta = theta)

g136_boot$u

hist(g136_boot$u)

```
~~~~

(b) Do the same for $\overline{AML} = 0.95$

~~~
```{r}

aml<-read.csv("C:/Users/Alejandro/Documents/1 ITESM/MCI/2do semestre/Nueva carpeta/Final/AML.csv",  header = TRUE, stringsAsFactors = FALSE, na.strings = "NA")

```

```{r}

aml<-select(aml, -c(Columna1))

g136_aml <- select(aml, X136)

```

```{r}

library(bootstrap)

g136_aml<-as.numeric(unlist(g136_aml))

theta<- function(g136_aml){mean(g136_aml)}

g136_boot_aml <- bcanon(g136_aml,1000, theta = theta)

g136_boot_aml$u

hist(g136_boot_aml$u)

```
~~~

(c) Plot histograms of the results, and suggest an inference.





Suppose that there were no differences between AML and ALL patients for any gene, so that

$$t = \frac{\overline{ALL} - \overline{AML}}{\widehat{sd}}$$

exactly followed a student-t distribution with 70 degrees of freedom in all 7128 cases. About how big might you expect the largest observed t value to be? Hint: $1/7128$ = 0.00014.

If there where no differences between AML an ALL then the highest value of t would be 0

(3 points)

****
Q2: Consider your own choice of distribution, and the normality distribution of parameter and score function; Draw a schematic graph of $\dot{l}_x(\theta)$ (score function) versus θ. Use it to justify the

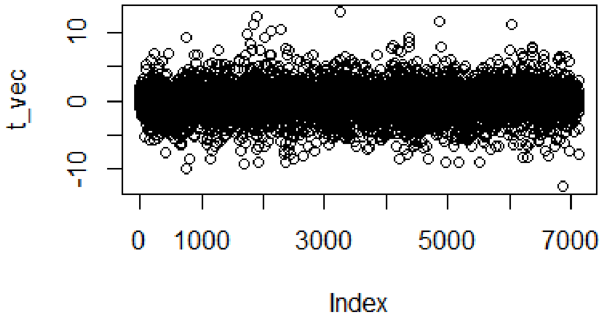
$$\hat{\theta} = \theta + \frac{\dot{l}_x(\theta)/n}{-\ddot{l}_x(\theta)/n}$$

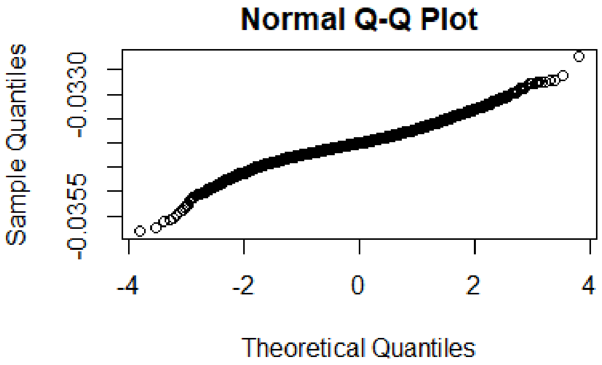
where $\ddot{l}x(\theta)$ is the second derivative of log likelihood function.
(3 points)

****
Q3: Consider data in Q1, Fisher suggested using permutations of the 72 data points (randomly selected either by bootstrapping or jackknifing). The 72 values are *randomly* divided into disjoint sets of size 47 and 25, and the two-sample t-statistic with equal variance is recomputed. This is done some large number (i.e., 1000), yielding permutation t -values t_1 ; t_2 ; ... ; t_{1000} . Compare these permutation t-values with the observed t-statistics using whole data. Plot a histogram of permutation t-values and find significant level of permutation.

```{r}

library(stats)

library(bootstrap)

t_vec<-vector()
```

```
for (j in 1:7128){
 t<-t.test(aml[,j], all[,j])
 t_vec[j]<-t$statistic
}
plot(t_vec)
t_vec_boot <- bcanon(t_vec,1000, theta = theta)
qqnorm(t_vec_boot$u)
```





(3 points)

****

Q4: Draw a sample of 1000 bivariate normal vectors x = $(x_1, x_2)'$, with

(a) Regress x_2 on x_1, and numerically check conditional distribution of $x_2|x_1$.

$$N1\left(\frac{x1}{1000}+\frac{x12}{x22}\left(\frac{999*x2}{1000}\right),\frac{x2}{1000}\left(\frac{x12}{x22}\right)\right)$$

$$N1\left(\left(\frac{x1*x22+999x2*x12}{1000x22}\right),\left(\frac{x2*x12}{1000x22}\right)\right)$$

(b) Do the same for x_1 on x_2.

$$N2\left(\frac{x2}{1000}+\frac{x22}{x12}\left(\frac{999*x1}{1000}\right),\frac{x1}{1000}\left(\frac{x21}{x11}\right)\right)$$

$$N2\left(\left(\frac{x1*x11+999x1*x21}{1000x11}\right),\left(\frac{x2*x21}{1000x11}\right)\right)$$

(c) What is your conclusion/observation?

What we can see is that one is the inverse of the other so they are compementary

(2point)

****

Q5: If x ~ $\text{Mult}_L(n, \boldsymbol{\pi})$ (multinomial with L parameters), use the Poisson trick as follow

$\text{Mult}_L(N, \boldsymbol{\pi})$~ Poi($n\boldsymbol{\pi}$) if N ~Poi(n)

to approximate the mean and variance of x_1/x_2, consider that for $S_i \sim Poi(\mu_i)$

$$\frac{S_1}{S_2}=\frac{\mu_1}{\mu_2}\left(1+\frac{S_1-\mu_1}{\mu_1}-\frac{S_2-\mu_2}{\mu_2}\right)$$

(3 points)

****

Q6: An experimental new anti-cancer drug called **Xilathon** is under development. Before human testing can begin, animal studies are needed to de- termine safe dosages. To this end, a *bioassay* or dose–response experiment was carried out: 11 groups of n D 10 mice each were injected with increasing amounts of **Xilathon**, dosages coded 1; 2; ... ; 11. the numbers of mice dying in the 11 groups were 0,0,0,3,6,6,5,9,9, 10, 10. Use the R package glm to calculate the logistic regression curve. What were the regression curve values at x = 0,1,2,...,10?

```{r}

med<-read.csv("C:/Users/Alejandro/Documents/1 ITESM/MCI/2do semestre/Nueva carpeta/Final/med.csv", header = TRUE, stringsAsFactors = FALSE, na.strings = "NA")

med

obs<-med

```

~~~
```{r}

#Parametrize data to obtain vales between 0 and 1

observed = med$y

maximo = max(obs$y)

med$y<-med$y/maximo


str(med)

sum(is.na(med$y))

logita<-glm(y~.,data=med, family = "binomial")

logita

summary(logita)


prediction = predict(object = logita, type = "response")


response = prediction*maximo

res <- trunc(response)


res

```
~~~

The regression values are

X	1	2	3	4	5	6	7	8	9	10	11
Y	0	0	0	1	3	5	7	8	9	9	9

(3 points)

****

Q7:

(a) Fit the Poisson regression model to below data:

https://web.stanford.edu/~hastie/CASI_files/DATA/galaxy.txt

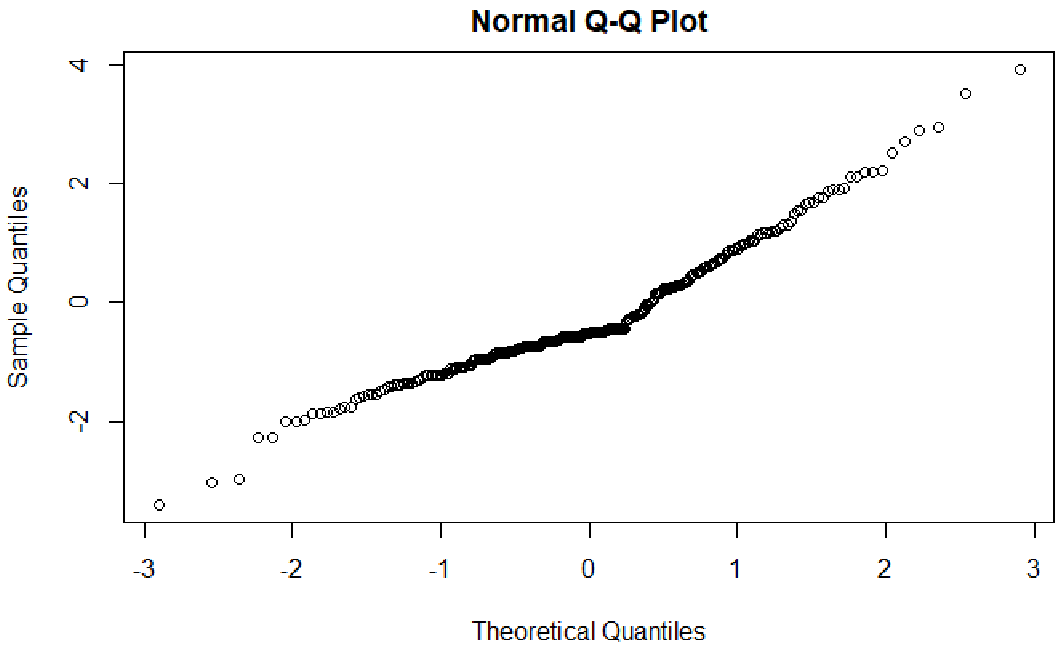
```{r}

ref<-read.csv("https://web.stanford.edu/~hastie/CASI_files/DATA/galaxy.txt", header = FALSE, stringsAsFactors = FALSE, na.strings = "NA")

gx<-read.csv("C:/Users/Alejandro/Documents/1 ITESM/MCI/2do semestre/Nueva carpeta/Final/gx.csv",  header = TRUE, stringsAsFactors = FALSE, na.strings = "NA")

gx



```

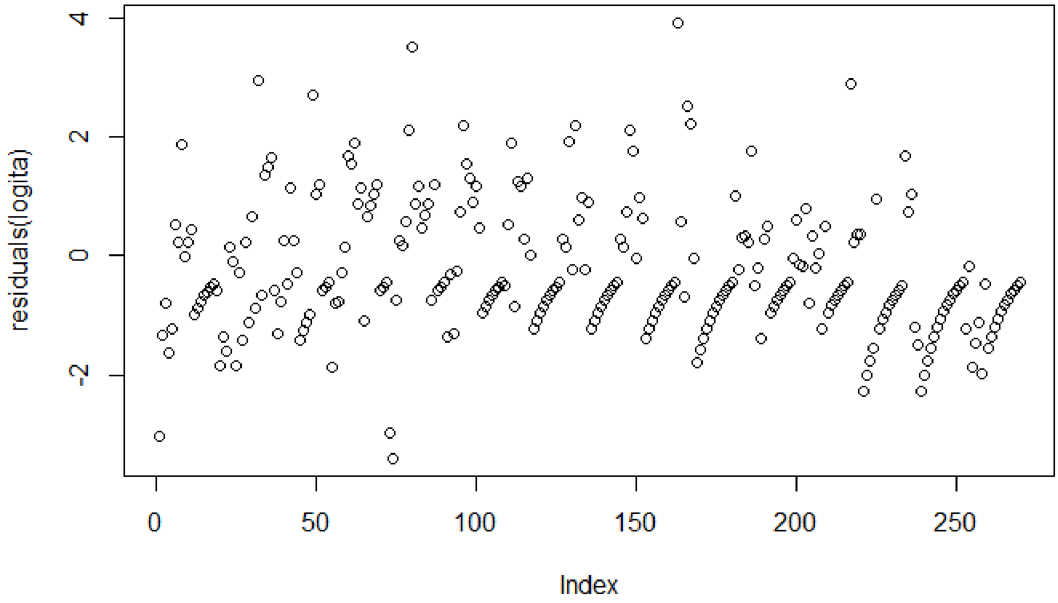
```{r}

#Parametrize data to obtain vales between 0 and 1

observed = gx$gx

# maximo = max(data_tool$CP)

# data_tool$CP<-data_tool$CP/maximo

str(gx)

sum(is.na(gx$gx))

logita<-glm(gx~.,data=gx, family = "poisson")

logita

summary(logita)

c<-sort(logita$coefficients)

c

prediction = predict(object = logita, type = "response")

prediction = trunc(prediction)

pred_mat <- matrix(, nrow = 18, ncol = 15)

rownames(pred_mat)<- c("21.38", "21.13", "20.88", "20.63", "20.38", "20.13", "19.88", "19.63", "19.38", "19.13", "18.88", "18.63", "18.38", "18.13", "17.88", "17.63", "17.38", "17.13")
```

~~~
colnames(pred_mat)<- c("-1.56", "-1.44", "-1.31", "-1.19", "-1.06", "-0.94", "-0.81", "-0.69", "-0.56",
"-0.44", "-0.31", "-0.19", "-0.06", "0.06", "0.19")


c=1


for (j in 1:15) {

 for (i in 1:18){

    pred_mat[i,j]<-prediction[c]

    c <- c+1

 }

 }

pred_mat

```
~~~

	-1.56	**-1.44**	**-1.31**	**-1.19**	**-1.06**	**-0.94**	**-0.81**	**-0.69**	**-0.56**	**-0.44**	**-0.31**	**-0.19**	**-0.06**	**0.06**	**0.19**
21.38	7	7	7	7	7	7	7	7	7	7	7	7	7	6	6
21.13	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
20.88	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
20.63	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
20.38	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
20.13	2	2	2	2	2	2	2	2	2	2	2	2	2	1	1
19.88	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
19.63	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
19.38	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0
19.13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18.88	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18.63	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18.38	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18.13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17.88	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17.63	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17.38	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17.13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

(b) Plot the Poisson deviance residuals.





qqnorm(residuals(logita))

plot(residuals(logita))

(d) Where does the fit seem poor?

The fit seems poor from 19.38 to 17.13

(d) How might you add to model to get a better fit?

What we could do to get a better fit is to observe the data behavior and select a distribution according to it, or to transform the data in order to fit another distribution.

Other thing that could be done is to add more observations in the lower region.

(3 points)

Note: Include all of your code, plots, numerical result, derivation, …

Explain your interpretation clearly and for mathematical derivations use WORD equation editor

THIS IS AN INDIVIDUAL-WORK EXAM, ANY EXACT SIMILARITY (copy-paste) BETWEEN YOUR WORK AND OTHER STUDENTS WILL CONSIDER AS VIOLATING THE COURSE RULES.