

**Solutions:**

1. Find E(X), E(Y), sd(X) and sd(Y).

> # Finding E(X)

> X.Range<- c(1,2,3)

> f.x<- function(x) 2.469862\*(x\*exp(-x^2))

> f\_x<- function(x) f.x(x)\*(x %in% X.Range)

> E.X<- sum(X.Range\*f\_x(X.Range))

> print("The E(X) is ")

[1] "The E(X) is "

> print(E.X)

[1] 1.092303

* The value of E(X) is 1.092303

> # Finding E(Y)

> f.Y<- function(y) 2\*(y\*exp(-y^2))\*(0<=y & y<=Inf)

> E.Y<- integrate(function(y) y\*f.Y(y), lower=0, upper=Inf)$value

> print("The E(Y) is")

[1] "The E(Y) is"

> print(E.Y)

[1] 0.8862269

* The value of E(Y) is 0.8862269

> # Finding sd(X)

> X.Range<-c(1,2,3)

> f.x<- function(x) 2.469862\*(x\*exp(-x^2))

> f\_x<- function(x) f.x(x)\*(x %in% X.Range)

> E.X<- sum(X.Range\*f\_x(X.Range))

> # findind variance

> Var.X<- sum((X.Range-E.X)^2 \* f\_x(X.Range))

> # finding standard deviatioin

> sd.X<- sqrt(Var.X)

> print("the sd(x) is")

[1] "the sd(x) is"

> print (sd.X)

[1] 0.2925953

* The value of sd(x) is 0.2925953

> # Finding sd(Y)

> f.Y<- function(y) 2\*(y\*exp(-y^2))\*(0<=y & y<=Inf)

> E.Y<- integrate(function(y) y\*f.Y(y), lower=0, upper=Inf)$value

> Var.Y<-integrate(function(y) (y-E.Y)^2\*f.Y(y), lower=0, upper=Inf)$value

> sd.Y<- sqrt(Var.Y)

> print("The sd(Y) is")

[1] "The sd(Y) is"

> print(sd.Y)

[1] 0.4632514

* The value of sd(Y) is 0.4632514

b) If X and Y are independent, find E(2X-3Y) and sd(2X-3Y).

> # Finding E(2X-3Y)

> (2\*(E.X) - 3\*(E.Y))

[1] -0.4740746

* The value of E(2X-3Y) is -0.4740746

> # Finding sd(2X-3Y)

> Var<- (4\*Var.X)+(9\*Var.Y)

> sd<- sqrt(Var)

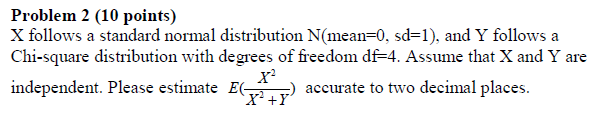
> print("The sd(2X-3Y) is")

[1] "The sd(2X-3Y) is"

> print(sd)

[1] 1.507934

* The value of sd(2X-3Y) is 1.507934



**Solution:**

> # X follows standard normal distribution

> X<- rnorm(10000, mean=0, sd=1)

> # Y follows chi-square distribution

> Y<- rchisq(10000, df=4)

> # solving the equation

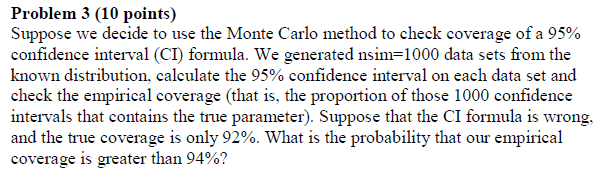
> eqn <- (X^2)/(X^2+Y)

> E.eqn<-mean(eqn)

> print(E.eqn)

[1] 0.1985588

* The estimate of E( X^2 / (X^2+Y)) is 0.1985588

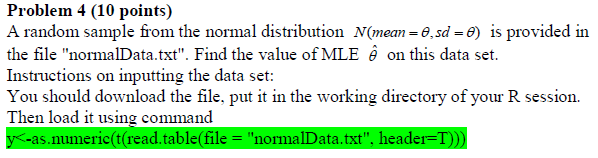


**Solution:**

> 1-pbinom(940,1000,0.92)

[1] 0.006617437

* The probability that the empirical coverage is greater than 94% is 0.006617437



**Solutions:**

> # download the file, put it in the working directory of your R session.

> getwd()

[1] "C:/Users/Neha/Documents"

> # load it using command

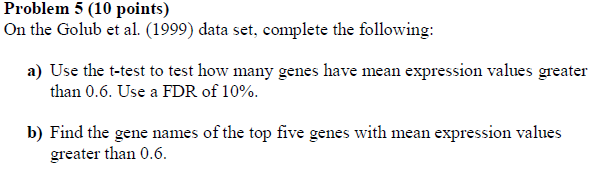
> y<-as.numeric(t(read.table(file = "normalData.txt", header=T)))

>

> c(sqrt(sum((y-mean(y))^2)/length(y)))

[1] 2.599879

* The Value of MLE for this set is 2.599879



**Solution:**

1. Use the t-test to test how many genes have mean expression values greater than 0.6. Use a FDR of 10%.

> # finding the genes with mean expression greater than 0.6

> mean <- p.values<0.05

> print(" The genes with mean expression greater than 0.6")

[1] " The genes with mean expression greater than 0.6"

> sum(p.values<0.05)

[1] 526

> # using FDR of 10%

> p.fdr<-p.adjust(p = p.values, method="fdr")

> print("the no of genes have mean expression values greater than 0.6 after 10% FDR are")

[1] "the no of genes have mean expression values greater than 0.6 after 10% FDR are"

> sum(p.fdr<0.10)

[1] 502

* The genes with mean expression greater than 0.6 = 526
* After using FDR = 502

1. Find the gene names of the top five genes with mean expression values greater than 0.6.

> genes <- order(p.fdr, decreasing = FALSE)[1:5]

> print("gene names of the top five genes with mean expression values greater than 0.6.")

[1] "gene names of the top five genes with mean expression values greater than 0.6."

> golub.gnames[genes, 2]

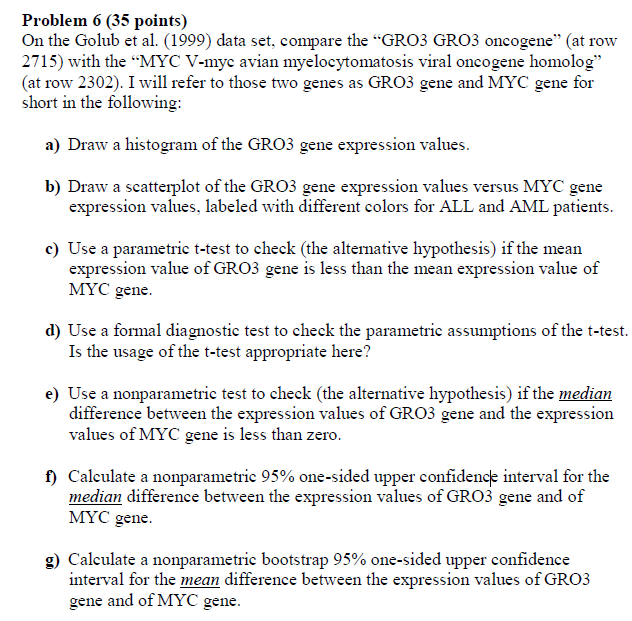
[1] "HnRNP-E2 mRNA"

[2] "Ornithine decarboxylase antizyme, ORF 1 and ORF 2"

[3] "GB DEF = Polyadenylate binding protein II"

[4] "RPS14 gene (ribosomal protein S14) extracted from Human ribosomal protein S14 gene"

[5] "GAPD Glyceraldehyde-3-phosphate dehydrogenase"



**Solution:**

**6a)** Draw a histogram of the GRO3 gene expression values.

> data(golub, package = "multtest")

> gol.fac <- factor(golub.cl,levels=0:1, labels= c("ALL","AML"))

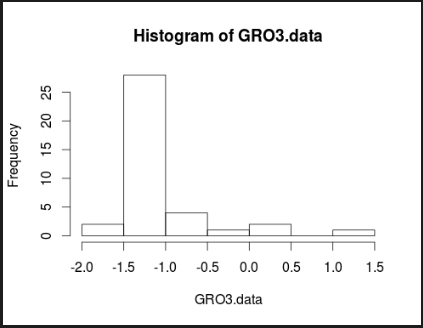
> # a) Draw a histogram of the GRO3 gene expression values.

> # row index of GRO3 gene is 2715. getting the data for GRO3

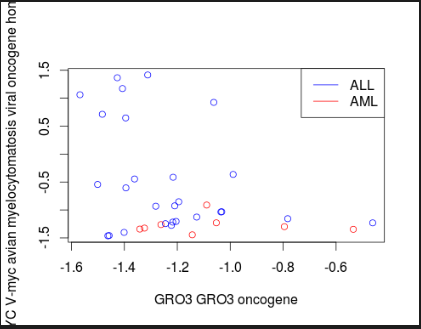
> GRO3.data = golub[2715,]

> # to draw a histogram

> hist(GRO3.data)



**6b**) b) Draw a scatterplot of the GRO3 gene expression values versus MYC gene expression values, labeled with different colors for ALL and AML patients



> data(golub, package = "multtest")

> gol.fac <- factor(golub.cl,levels=0:1, labels= c("ALL","AML"))

> # a) Draw a histogram of the GRO3 gene expression values.

> # row index of GRO3 gene is 2715. getting the data for GRO3

> GRO3.data = golub[2715,]

> # to draw a histogram

> hist(GRO3.data)

> MYC.data = golub[2302,]

> # row index of GRO3 gene is 2715. getting the data for GRO3

> GRO3.data<- golub[2715,]

> ALL.x<- golub[2715, gol.fac=="ALL"]

> AML.x<- golub[2715, gol.fac=="AML"]

> ALL.y<- golub[2302, gol.fac=="ALL"]

> AML.y<- golub[2302, gol.fac=="AML"]

> # to draw a scatterplot

> plot(ALL.x,ALL.y,xlab=golub.gnames[2715,2],ylab=golub.gnames[2302,2], col="blue")

> points(AML.x,AML.y, col="red")

> legend("topright",c("ALL","AML"),col=c("blue","red"),lty=c(1,1))

**6c)** Use a parametric t-test to check (the alternative hypothesis) if the mean expression value of GRO3 gene is less than the mean expression value of MYC gene.

H0 : Mean(GRO3)- Mean(MYC) > 0

* H0 : mean(GRO3) > Mean(MYC)

HA :  Mean(GRO3) < Mean (MYC)

The p value is 0.03718

* As p-value is small, we can reject The NULL HYPOTHESIS and accept the mean expression of GRO3 is less than mean expression of MYC

> # applying t-test

> t.test(GRO3.data, MYC.data, paired=T, alternative = "less" )

Paired t-test

data: GRO3.data and MYC.data

t = -1.8363, df = 37, p-value = 0.03718

alternative hypothesis: true difference in means is less than 0

95 percent confidence interval:

-Inf -0.02909346

sample estimates:

mean of the differences

-0.3580716

6 **d)** Use a formal diagnostic test to check the parametric assumptions of the t-test. Is the usage of the t-test appropriate here?

> shapiro.test(golub[2715,])

Shapiro-Wilk normality test

data: golub[2715, ]

W = 0.715, p-value = 2.9e-07

> shapiro.test(golub[2302,])

Shapiro-Wilk normality test

data: golub[2302, ]

W = 0.7537, p-value = 1.355e-06

* Since both test gives p-valu less than 0.05, we reject the null hypothesis. Thus we conclude that these tests do not follow normality and so t-test cannot be used here.

**6 e)** Use a nonparametric test to check (the alternative hypothesis) if the *median* difference between the expression values of GRO3 gene and the expression values of MYC gene is less than zero

Here we use Wilcox test

> wilcox.test (x= GRO3.data, y= MYC.data, paired=T, alternative="less")

Wilcoxon signed rank test with continuity

correction

data: GRO3.data and MYC.data

V = 107, p-value = 0.04208

alternative hypothesis: true location shift is less than 0

* P value = 0.04208 , we reject the null hypothesis and accept the alternate hypothesis that the difference is less than zero

6 **f)** Calculate a nonparametric 95% one-sided upper confidence interval for the *median* difference between the expression values of GRO3 gene and of MYC gene

> data.diff<-abs(golub[2715,]-golub[2302,])

> n<-length(data.diff)

> nboot<-1000

> boot.xbar<-rep(NA, nboot)

> for(i in 1:nboot) {

+ data.star <- data.diff[sample(1:n,replace=TRUE)]

+ boot.xbar[i]<-median(data.star)

+ }

> quantile(boot.xbar,c(0.95))

95%

0.795765

* 95% one sided upper CI for the median difference between expression values of bothe the genes is -Infinity , 0.795765

**6g)** Calculate a nonparametric bootstrap 95% one-sided upper confidence interval for the *mean* difference between the expression values of GRO3 gene and of MYC gene.

* Non parametric bootstrap 95% one-sisde upper CI for median difference between the expression values of the genes is –Infinity, 1.071662

> data.diff<-abs(golub[2715,]-golub[2302,])

> n<-length(data.diff)

> nboot<-1000

> t.boot = rep(NA,nboot)

> for(i in 1:nboot){

+ data.star<-data.diff[sample(1:n,replace=TRUE)]

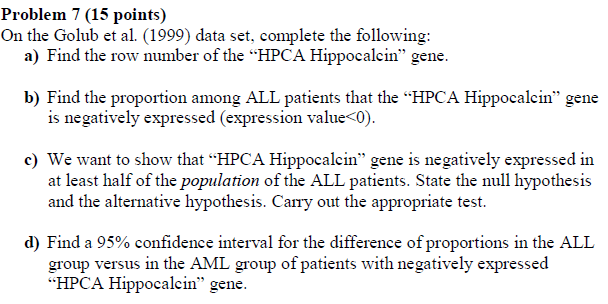
+ t.boot[i]= mean(data.star)

+ }

> quantile(t.boot,c(0.95))

95%

1.071662



Solutions:

7**a)** Find the row number of the “HPCA Hippocalcin” gene.

> # Finding the roe index through grep

>

> grep("HPCA Hippocalcin",golub.gnames[,2])

[1] 118

* The row number of HPCA Hippocalcin is 118

7**b)** Find the proportion among ALL patients that the “HPCA Hippocalcin” gene is negatively expressed (expression value<0).

> HPCA.data = golub[118, gol.fac=="ALL"]

> mean(HPCA.data<0)

[1] 0.5925926

* Proportion among ALL patients that the HPCA Hippocalcin gene is negatively expressed is 0.5925926

7**c)** We want to show that “HPCA Hippocalcin” gene is negatively expressed in at least half of the *population* of the ALL patients. State the null hypothesis and the alternative hypothesis. Carry out the appropriate test.

H0 : PALL  = ½ ; HA: PALL > ½

> p.all<-sum(golub[118,gol.fac=="ALL"]<0)

> binom.test(x=p.all,n=27,p= 0.5, alternative="greater")

Exact binomial test

data: p.all and 27

number of successes = 16, number of trials = 27,

p-value = 0.221

alternative hypothesis: true probability of success is greater than 0.5

95 percent confidence interval:

0.4170687 1.0000000

sample estimates:

probability of success

0.5925926

* As the pvalue is 0.221 , we accept Null Hypothesis

**7d)** Find a 95% confidence interval for the difference of proportions in the ALL group versus in the AML group of patients with negatively expressed “HPCA Hippocalcin” gene.

> # getting the row index for HPCA

> grep("HPCA Hippocalcin",golub.gnames[,2])

[1] 118

> # getting the data

> x.ALL <- golub[118,gol.fac=="ALL"]

> nALL <- length(x.ALL)

> x.AML<- golub[118,gol.fac=="AML"]

> nAML <- length(x.AML)

> nboot<-1000

> boot.diff <- rep(NA,nboot)

> for (i in 1:nboot) {

+ data.ALL <- x.ALL[sample(1:nALL,replace=TRUE)]

+ data.AML <- x.AML[sample(1:nAML,replace=TRUE)]

+ data.diff <- data.ALL-data.AML

+ boot.diff[i] <- mean(data.diff)

+ }

There were 50 or more warnings (use warnings() to see the first 50)

> quantile(boot.diff,c(0.025,0.975))

2.5% 97.5%

-0.4796049 0.3626952

* The 95% CI for the difference of proportions in the ALL group versus in the AML group of patients with negatively expressed “HPCA Hippocalcin” gene is -0.4796049, 0.3626952