Midterm

Yanhe Wen

3/10/2017

Question 1

(a)

```
fx1 <- function(x){2.469862*x*exp(-x^2)}
fy1 <- function(y){2*y*exp(-y^2)}
x1 <- c(1,2,3)
Ex1 <- sum(x1*fx1(x1))
cat("E(X)=",Ex1)
```

```
## E(X)= 1.092303
```

```
Ey1 <- integrate(function(y) y*fy1(y),lower = 0, upper = Inf)$value
cat("E(Y)=",Ey1)
```

```
## E(Y)= 0.8862269
```

```
Sdx1 <- sqrt(sum((x1-Ex1)^2*fx1(x1)))
cat("sd(X)=",Sdx1)
```

```
## sd(X)= 0.2925953
```

```
Sdy1 <- sqrt(integrate(function(y) (y-Ey1)^2*fy1(y) , lower = 0, upper = Inf)$value)
cat("sd(Y)=",Sdy1)
```

```
## sd(Y)= 0.4632514
```

(b)

```
E_xy <- sum(2*x1*fx1(x1)-3*Ey1)
cat("E(2X-3Y)=",E_xy)
```

```
## E(2X-3Y)= -5.791436
```

Question 2

```
X <- rnorm(1000, mean = 0, sd = 1)
Y <- rchisq(1000, df = 4)
cat("E(X^2/(X^2+Y))=",mean(X^2/(X^2+Y)))
```

```
## E(X^2/(X^2+Y))= 0.2078595
```

Question 3

```r
nsim <- 1000
cov1 <- cov2 <- rep(NA, nsim)
for(i in 1:nsim){
  data3 <- rnorm(1000, mean = 0, sd = 1)
  CI <- c(mean(data3)+qt(0.03,999)*sd(data3)/sqrt(999), mean(data3)-qt(0.03,999)*sd(data3)/sqrt(999))
  cov1[i] <- CI[1]
  cov2[i] <- CI[2]
}
cat("94% CI is: (",mean(cov1),",",mean(cov2),")")
```

```
## 94% CI is: ( -0.05901485 , 0.0601106 )
```

Question 4

```r
y <- as.numeric(t(read.table(file = "normalData.txt", header = T)))
nloglik <- function(theta) -sum(log(dnorm(y, mean = theta, sd = theta)))
cat("MLE=",optim(par = 1, nloglik)$par)
```

```
## Warning in optim(par = 1, nloglik): one-dimensional optimization by Nelder-Mead is unreliable:
## use "Brent" or optimize() directly
```

```
## MLE= 2.426563
```

Question 5

```r
library(multtest)
data("golub")
```

(a)

```r
p.value5 <- t.test(golub[201,], alternative = "greater", mu = 0.6, conf.level = 0.9)$p.value
p.values <- apply(golub[1:3025,], 1, function(x) t.test(x, alternative = "greater", mu = 0.6, conf.level
p.fdrs <- p.adjust(p.values, method = "fdr")
cat("There are",sum(p.fdrs>0.1),"genes have mean expression values greater than 0.6")
```

```
## There are 2528 genes have mean expression values greater than 0.6
```

(b)

```r
cat("Top five genes with mean expression values greater than 0.6: \n")
```

```
## Top five genes with mean expression values greater than 0.6:
```

```r
for(i in 1:5){
  cat(golub.gnames[order(p.fdrs,decreasing = FALSE)[i],2],"\n")
}
```

```
## HnRNP-E2 mRNA
## Ornithine decarboxylase antizyme, ORF 1 and ORF 2
## GB DEF = Polyadenylate binding protein II
```

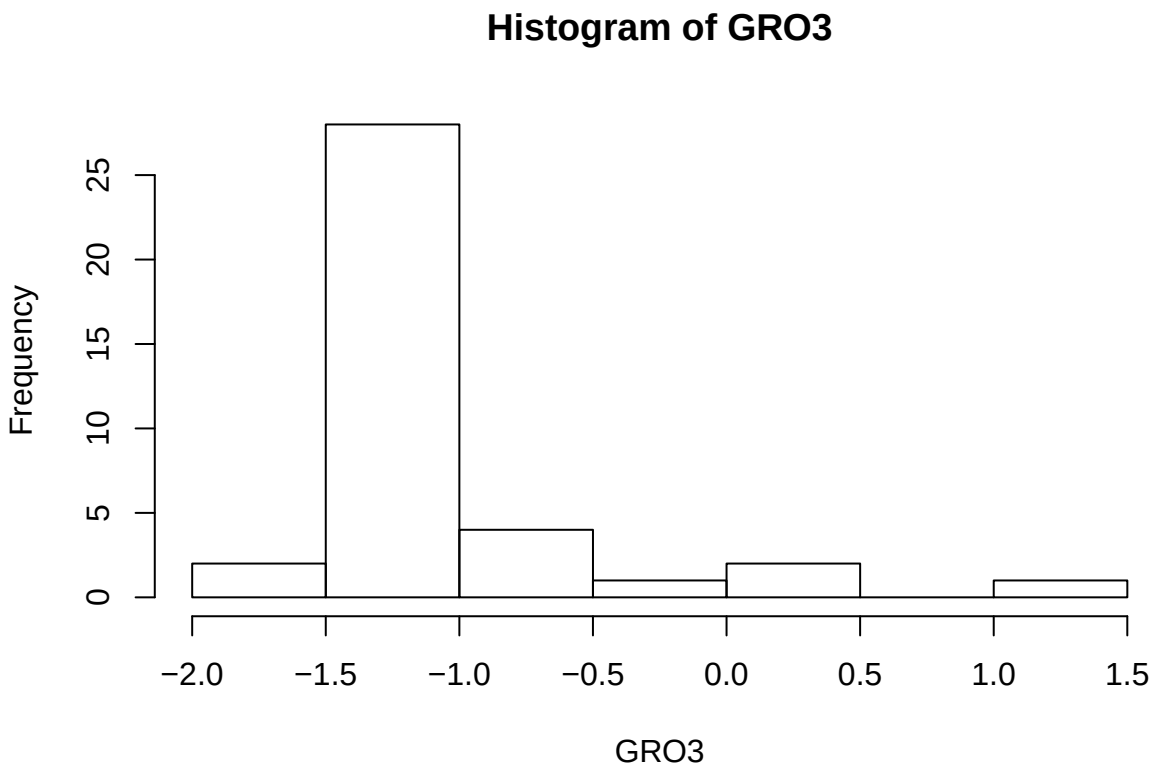
```
## RPS14 gene (ribosomal protein S14) extracted from Human ribosomal protein S14 gene
## GB DEF = HLA-B null allele mRNA
```

Question 6

```
GRO3 <- golub[2715,]
MYC <- golub[2302,]
gol.fac <- factor(golub.cl, levels=0:1, labels = c("ALL","AML"))
```

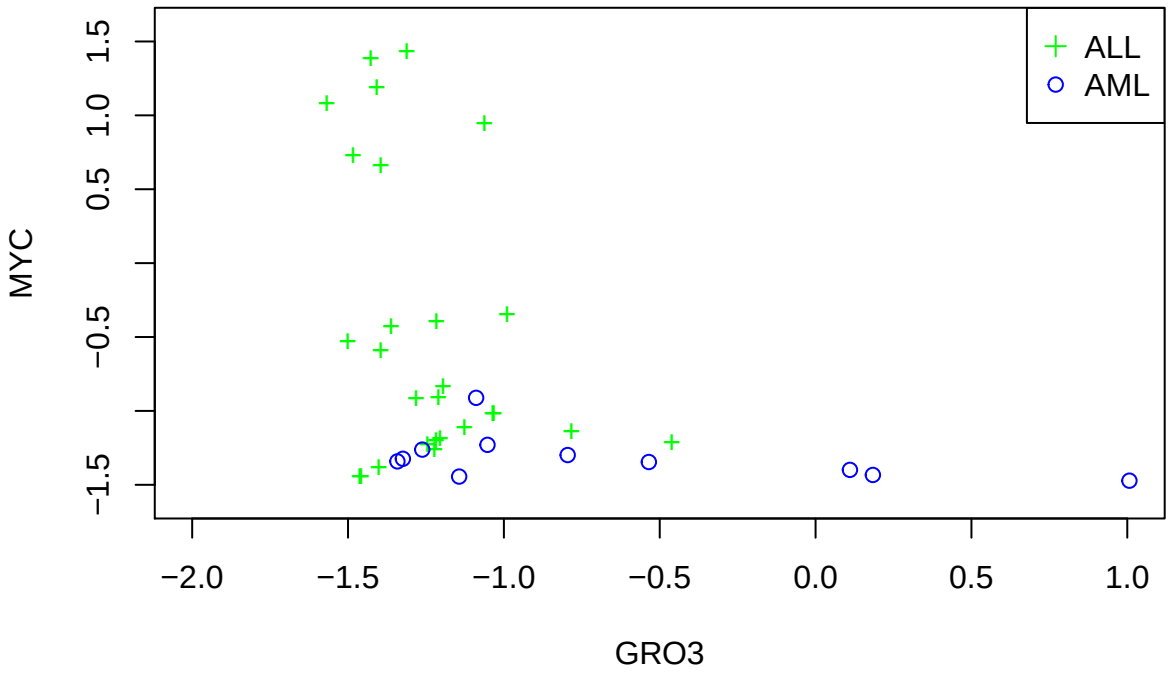
(a)

```
hist(GRO3)
```



###(b)

```
GRO3_ALL <- golub[2715,gol.fac=="ALL"]
GRO3_AML <- golub[2715,gol.fac=="AML"]
MYC_ALL <- golub[2302,gol.fac=="ALL"]
MYC_AML <- golub[2302,gol.fac=="AML"]
plot(GRO3_ALL, MYC_ALL, col = "green", pch = "+", xlab = "GRO3", ylab = "MYC", ylim = range(-1.6,1.6), 
points(GRO3_AML, MYC_AML, col = "blue")
legend("topright", c("ALL","AML"), col = c("green","blue"), pch = c(3,1))
```



###(c)

```
t.test(GRO3-MYC,alternative = "less")
```

```
##
##  One Sample t-test
##
## data:  GRO3 - MYC
## t = -1.8363, df = 37, p-value = 0.03718
## alternative hypothesis: true mean is less than 0
## 95 percent confidence interval:
##         -Inf -0.02909346
## sample estimates:
##  mean of x
## -0.3580716
```

```
cat("From GRO3 - MYC, we get p-vlaue is 0.03718, so reject null hypothesis, GRO3 < MYC")
```

```
## From GRO3 - MYC, we get p-vlaue is 0.03718, so reject null hypothesis, GRO3 < MYC
```

(d)

```
diff_GRO3_MYC <- GRO3 - MYC
t_statistic <- (mean(diff_GRO3_MYC)-0)/(sd(diff_GRO3_MYC)/sqrt(37))
pnorm(t_statistic)
```

```
## [1] 0.03499539
```

```
cat("t statistic is",t_statistic, "pnorm(t_statistic)=",pnorm(t_statistic),"
    which matches to the p-value as well")
```

```
## t statistic is -1.81197 pnorm(t_statistic)= 0.03499539
##     which matches to the p-value as well
```

(e)

```
wilcox.test(GRO3, MYC, paired = T, alternative = "less")
```

```
## Warning in wilcox.test.default(GRO3, MYC, paired = T, alternative =
## "less"): cannot compute exact p-value with zeroes
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  GRO3 and MYC
## V = 107, p-value = 0.04208
## alternative hypothesis: true location shift is less than 0
```

```
cat("p-value is less than 0.05, so we reject null hypothesis, the median difference is less than 0")
```

```
## p-value is less than 0.05, so we reject null hypothesis, the median difference is less than 0
```

(f)

```
binom.test(x=length(diff_GRO3_MYC)-sum(diff_GRO3_MYC==0), n = length(diff_GRO3_MYC), p = 0.5, alternati
```

```
##
##  Exact binomial test
##
## data:  length(diff_GRO3_MYC) - sum(diff_GRO3_MYC == 0) and length(diff_GRO3_MYC)
## number of successes = 26, number of trials = 38, p-value = 0.01678
## alternative hypothesis: true probability of success is greater than 0.5
## 95 percent confidence interval:
##  0.5391389 1.0000000
## sample estimates:
## probability of success
##              0.6842105
```

```
cat("95% CI is (0.5391389,1.0000000)")
```

```
## 95% CI is (0.5391389,1.0000000)
```

Question 7

(a)

```
HPCA_row <- grep("HPCA Hippocalcin", golub.gnames[,2])
cat("The row number of HPCA Hippocalcin is: ", HPCA_row)
```

```
## The row number of HPCA Hippocalcin is:  118
```

(b)

```
HPCA <- golub[HPCA_row,]
HPCA_ALL <- golub[HPCA_row,gol.fac=="ALL"]
```

```r
HPCA_AML <- golub[HPCA_row,gol.fac=="AML"]
cat("The proportion is:",sum(HPCA_ALL<0)/length(HPCA_ALL))
```

```
## The proportion is: 0.5925926
```

(c)

```r
cat("H0: 0.5 population of ALL patients's gene is negatively expressed; H1: >0.5")
```

```
## H0: 0.5 population of ALL patients's gene is negatively expressed; H1: >0.5
```

```r
binom.test(sum(HPCA_ALL<0), n=27, p=0.5, alternative = "greater")
```

```
##
##  Exact binomial test
##
## data:  sum(HPCA_ALL < 0) 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
```

```r
cat("p-value = 0.221, so we accept null hypothesis, more than half of population of
    genes are negtively expressed")
```

```
## p-value = 0.221, so we accept null hypothesis, more than half of population of
##     genes are negtively expressed
```

(d)

```r
prop.test(x=c(sum(HPCA_ALL<0),sum(HPCA_AML<0)),n=c(27,11), alternative = "two.sided")
```

```
## Warning in prop.test(x = c(sum(HPCA_ALL < 0), sum(HPCA_AML < 0)), n =
## c(27, : Chi-squared approximation may be incorrect
```

```
##
##  2-sample test for equality of proportions with continuity
##  correction
##
## data:  c(sum(HPCA_ALL < 0), sum(HPCA_AML < 0)) out of c(27, 11)
## X-squared = 2.5878e-32, df = 1, p-value = 1
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.3477551  0.4420312
## sample estimates:
##    prop 1    prop 2
## 0.5925926 0.5454545
```

```r
cat("95% CI is (-0.3477551,0.4420312)")
```

```
## 95% CI is (-0.3477551,0.4420312)
```