Math7340 Midterm

Chengbo Gu

Problem 1 (10 points)

X follows a distribution with pdf $f_X(x) = 2.469862(xe^{-x^2}), x = 1, 2, 3$; while Y follows a distribution with pdf $f_Y(y) = 2ye^{-y^2}, y > 0$.

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

```
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)
EX <- sum(X_range*f_X(X_range))
EX
```

```
## [1] 1.092303
```

```
# sd(X)
VarX <- sum((X_range-EX)^2*f_X(X_range))
sdX <- sqrt(VarX)
sdX
```

```
## [1] 0.2925953
```

```
f_Y <- function(y) 2*y*exp(-y^2)*(y > 0)
# E(Y)
EY <- integrate(function(y) y*f_Y(y), lower=0, upper=Inf)$value
EY
```

```
## [1] 0.8862269
```

```
# sd(Y)
VarY <- integrate(function(y) (y-EY)^2*f_Y(y), lower=0, upper=Inf)$value
sdY <- sqrt(VarY)
sdY
```

```
## [1] 0.4632514
```

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

$$E(2X - 3Y) = 2E(X) - 3E(Y)$$

```
cat("E(2X-3Y)=", 2*EX-3*EY, "\n")
```

```
## E(2X-3Y)= -0.4740746
```

$$Var(2X - 3Y) = 2^2Var(X) + 3^2Var(Y)$$
$$sd(2X - 3Y) = \sqrt{Var(2X - 3Y)}$$

```
cat("sd(2X-3Y)=", sqrt(4*VarX+9*VarY), "\n")
```

```
## sd(2X-3Y)= 1.507934
```

Problem 2 (10 points)

X follows a standard normal distribution N(mean=0, sd=1), and Y follows a Chi-square distribution with degrees of freedom df=4. Assume that X and Y are independent. Please estimate $E(\frac{X^2}{X^2+Y})$ accurate to two decimal places.

```
x <- rnorm(100000, mean=0, sd=1)
y <- rchisq(100000, df=4)
z <- x^2/(x^2+y)
res <- mean(z)
res
```

```
## [1] 0.1990918
```

```
format(round(res, 2), nsmall = 2)
```

```
## [1] "0.20"
```

Problem 3 (10 points)

Suppose we decide to use the Monte Carlo method to check coverage of a 95% confidence interval (CI) formula. We generated nsim=1000 data sets from the known distribution, calculate the 95% confidence interval on each data set and check the empirical coverage (that is, the proportion of those 1000 confidence intervals that contains the true parameter). Suppose that the CI formula is wrong, and the true coverage is only 92%. What is the probability that our empirical coverage is greater than 94%?

Let X denote the number of trails that CI contains the true parameter, then we have: X~Binom(1000, 0.92).

What we want is P(X>940).

$$P(X > 940) = 1 - P(X \leq 940) = 1 - pbinom(940, size = 1000, prob = 0.92)$$

```
cat("P(X>940)=", 1-pbinom(940, size=1000, prob=0.92), "\n")
```

```
## P(X>940)= 0.006617437
```

Problem 4 (10 points)

A random sample from the normal distribution $N(mean = \theta, sd = \theta)$ is provided in the file "normalData.txt". Find the value of MLE $\hat{\theta}$ on this data set.

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

nloglik <- function(theta) {
  -sum ( log(dnorm(obs, mean=theta, sd=theta)))
}
optim(par=c(2), nloglik)$par
```

```
## [1] 2.426563
```

Problem 5 (10 points)

On the Golub et al. (1999) data set, complete the following:

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

```
data(golub)
p.values <- apply(golub, 1, function(x) t.test(x, mu=0.6, alternative = "greater")$p.value)
p.fdr <- p.adjust(p=p.values, method = "fdr")
sum(p.fdr < 0.1)
```

```
## [1] 502
```

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

```
golub.gnames[,2][order(p.fdr)][1:5]
```

```
## [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"
```

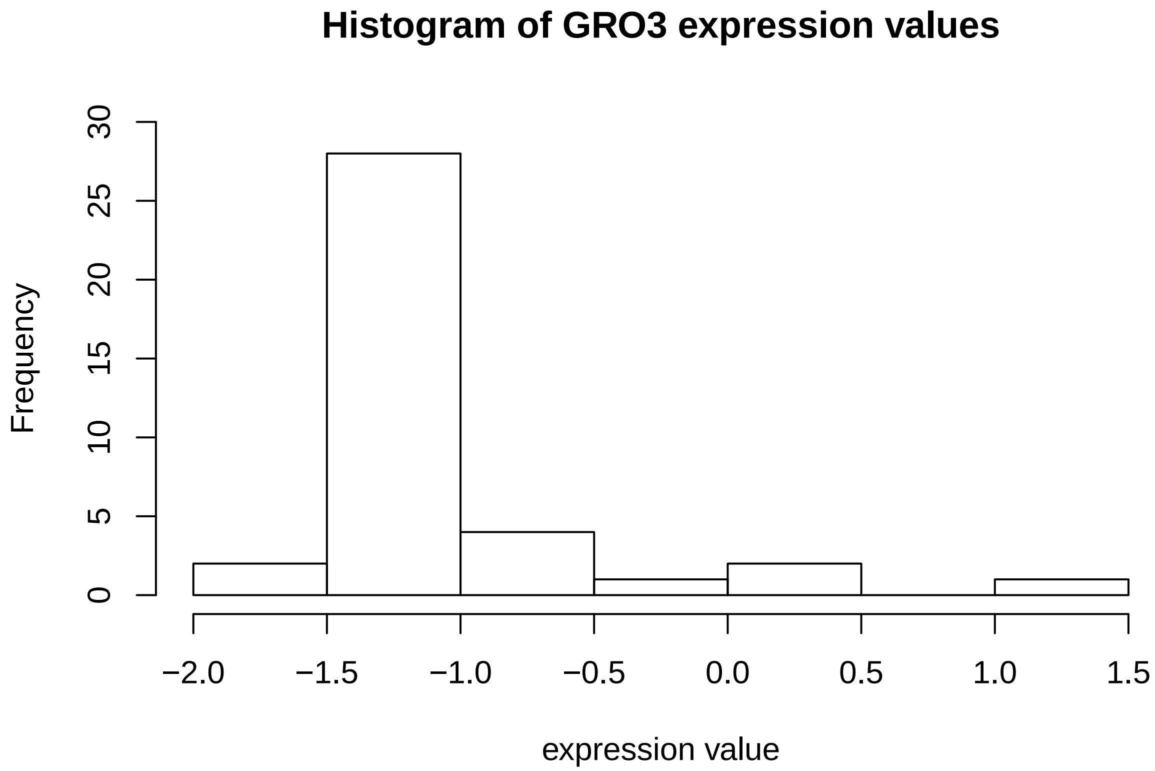
Problem 6 (35 points)

On the Golub et al. (1999) data set, compare the "RO3 GRO3 oncogene" (at row 2715) with the "MYC V-myc avian myelocytomatosis viral oncogene homolog" (at row 2302).

```
data(golub)
```

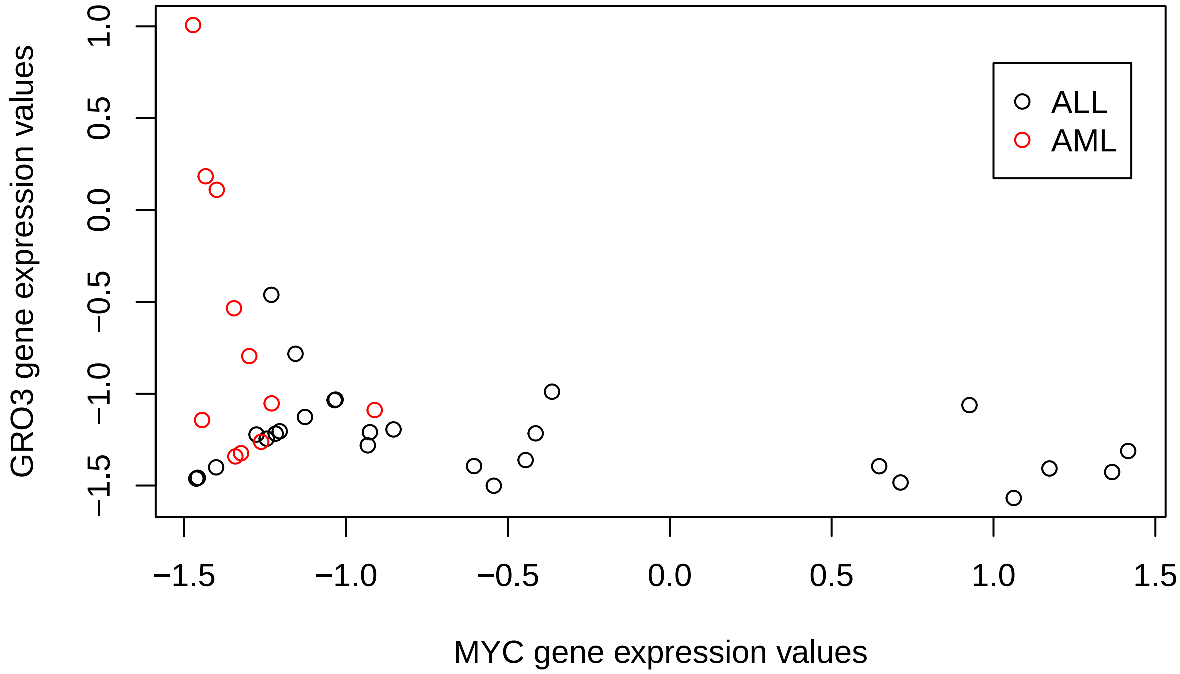
a) Draw a histogram of the GRO3 gene expression values.

```
hist(golub[2715,], main = "Histogram of GRO3 expression values",
     ylim=c(0,30), xlab="expression value")
```



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

```
gol.fac <- factor(golub.cl, levels=0:1, labels=c("ALL", "AML"))
exp2715th = golub[2715,]
exp2302th = golub[2302,]
plot(exp2715th ~ exp2302th, col=gol.fac,
     xlab = "MYC gene expression values", ylab = "GRO3 gene expression values")
legend(1,0.8 ,unique(gol.fac),col=1:length(gol.fac),pch=1)
```



c) 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.

$$H_0 : \mu_{GRO3} = \mu_{MYC}, \quad H_A : \mu_{GRO3} < \mu_{MYC}$$

```
t.test(golub[2715,]-golub[2302,], alternative = "less")
```

```
##
##  One Sample t-test
##
## data:  golub[2715, ] - golub[2302, ]
## 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
```

The p-value here is smaller than 0.05. Hence, we reject the null hypothesis and conclude that mean expression value for GRO3 gene is lower than that for MYC gene.

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

```r
shapiro.test(golub[2715,]-golub[2302,])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  golub[2715, ] - golub[2302, ]
## W = 0.90688, p-value = 0.004009
```

The p-value here is extremely small, so the normality doesn't hold. The usage of t test is not appropriate here.

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.

Let m_D denote the population median of the difference between the expression values of GRO3 gene and the expression values of MYC gene.

$$H_0: m_D = 0, \quad H_A: m_D < 0$$

```r
wilcox.test (x= golub[2715,], y= golub[2302,], paired=T, alternative="less")
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  golub[2715, ] and golub[2302, ]
## V = 107, p-value = 0.04208
## alternative hypothesis: true location shift is less than 0
```

The p-value here is 0.04208 which is less than 0.05. So we reject the null hypothesis and conclude that the median difference between the expression values of GRO3 gene and the expression values of MYC gene is less than zero.

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.

```r
wilcox.test (x= golub[2715,], y= golub[2302,], paired=T,
             alternative="less", conf.int = TRUE)
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  golub[2715, ] and golub[2302, ]
## V = 107, p-value = 0.04208
## alternative hypothesis: true location shift is less than 0
## 95 percent confidence interval:
##         -Inf -0.02023244
## sample estimates:
## (pseudo)median
##     -0.5064647
```

The nonparametric 95% one-sided upper confidence interval for the median difference is $(-\infty, -0.02023244)$.

g) 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.

```
dif <- golub[2715,] - golub[2302,]
n<-length(dif)
nboot<-1000
boot.dif <- rep(NA, nboot)
for (i in 1:nboot) {
  data.star <- dif[sample(1:n,replace=TRUE)]
  boot.dif[i]<-mean(data.star)
}
quantile(boot.dif,c(0.95))
```

```
##         95%
## -0.05714491
```

So the one-sided upper confidence interval for the mean difference between the expression values of GRO3 gene and of MYC gene is $(-\infty$,-0.0571449).

Problem 7 (15 points)

On the Golub et al. (1999) data set, complete the following:

a) Find the row number of the "HPCA Hippocalcin" gene.

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

```
## [1] 118
```

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

```
gol.fac <- factor(golub.cl, levels=0:1, labels=c("ALL", "AML"))
resALL <- golub[118, gol.fac=="ALL"] < 0
lengthALL <- length(resALL)
numALL <- sum(resALL)
numALL/lengthALL
```

```
## [1] 0.5925926
```

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.

Let p_{HPCA} denotes the proportion of patients for whom the "HPCA Hippocalcin" gene expression values is nagatively expressed.

$$H_0 : p_{ALL,HPCA} = 0.5, \quad H_A : p_{ALL,HPCA} < 0.5$$

```
binom.test(numALL, lengthALL, p=0.5, alternative = "less")
```

```
##
##  Exact binomial test
```

```
##
## data:  numALL and lengthALL
## number of successes = 16, number of trials = 27, p-value = 0.8761
## alternative hypothesis: true probability of success is less than 0.5
## 95 percent confidence interval:
##  0.0000000 0.7520656
## sample estimates:
## probability of success
##              0.5925926
```

Since the p-value here is greater than 0.05, we accept the null hypothesis that $p_{ALL,HPCA} = 0.5$. Thus, “HPCA Hippocalcin” gene is negatively expressed in at least half of the population of the ALL patients.

d) 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.

```
resAML <- golub[118, gol.fac=="AML"] < 0
lengthAML <- length(resAML)
numAML <- sum(resAML)
prop.test(x=c(numALL, numAML), n=c(lengthALL, lengthAML), alternative="two.sided")
```

```
##
##  2-sample test for equality of proportions with continuity
##  correction
##
## data:  c(numALL, numAML) out of c(lengthALL, lengthAML)
## 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
```

The 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 is $(-0.3477551, 0.4420312)$.