Northeastern University
College of Science

Module 6 Homework

1. (60 points) On the Golub et al. (1999) data, consider the "H4/j gene" gene (row 2972) and the "APS Prostate specific antigen" gene (row 2989). Setup the appropriate hypothesis for proving the following claims. Chose and carry out the appropriate tests.
(a) The mean "H4/j gene" gene expression value in the ALL group is greater than -0.9 (note that this is negative 0.9).

Answer)

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

gol.fac<-factor(golub.cl,levels=0:1, labels=c("ALL","AML"))
xALL <-golub[2972,gol.fac=="ALL"]
t.test(xALL,mu=-0.9,alternative ="greater")
```

Output)

```
One Sample t-test

data:  gol.fac
t = 2.2659, df = 26, p-value = 0.01601
alternative hypothesis: true mean is greater than -0.9
95 percent confidence interval:
 -0.844439       Inf
sample estimates:
 mean of x
-0.6753033
```

- Here we have to prove the gene expression value is greater than -0.9, then w e should set that as the alternative hypothesis. **The p-value is 0.01601**. **We reject the null hypothesis** and conclude that the mean "H4/j gene" gene exp ression value in the ALL group is greater than -0.9.

(b) The mean “H4/j gene” gene expression value in ALL group differs from the mean “H4/j gene” gene expression value in the AML group.

Answer)

```
gol.fac<-factor(golub.cl,levels=0:1, labels=c("ALL","AML"))
t.test(golub[2972,]~gol.fac)
```

Output)

```
        Welch Two Sample t-test

data:  golub[2972, ] by gol.fac
t = -1.4988, df = 29.978, p-value =
0.1444
alternative hypothesis: true difference in means between group ALL a
nd group AML is not equal to 0
95 percent confidence interval:
 -0.48627436  0.07463315
sample estimates:
mean in group ALL mean in group AML
      -0.6753033       -0.4694827
```

- Here we have to show that the "H4/j gene" expresses differently in ALL pati ents from in AML patients, using the Golub data set.we have to test H0: μX = μY versus HA: μX ≠ μY, where μX and μY denote the mean H4/j gene ex pression values in ALL patients and AML patients respectively**. p-value is 0 .1444. We don’t reject null hypothesis.**

(c) In the ALL group, the mean expression value for the “H4/j gene” gene is lower than the mean expression value for the “APS Prostate specific antigen” gene.

Answer)

```
t.test(golub[2972,gol.fac=="ALL"],golub[2989,gol.fac=="ALL"], paired
= TRUE, alternative = "less")
```

Output)

```
        Paired t-test

data:  golub[2972, gol.fac == "ALL"] and golub[2989, gol.fac == "ALL"
]
t = -1.8366, df = 26, p-value = 0.03886
alternative hypothesis: true mean difference is less than 0
95 percent confidence interval:
      -Inf -0.02175309
sample estimates:
mean difference
    -0.3050307
```

- We have to show that ALL group, the mean expression value for the “H4/j g ene” is lower than the mean expression value for the “APS Prostate specific antigen” gene. we have to test H0: $\mu X = \mu Y$ versus HA: $\mu X < \mu Y$. the “H4/j gene” clearly has lower mean expression value than APS Prostate specific an tigen gene, **p value is 0.038. so, we reject null hypothesis.**

(d) Let p_{H4j} denotes the proportion of patients for whom the "H4/j gene" expression values is greater than -0.6. We wish to show that p_{H4j} in the ALL group is less than 0.5.

Answer)

```
golub_ALL <- golub[2972,gol.fac=="ALL"]
binom.test(sum(golub_ALL > -0.6), length(golub_ALL), p=0.5,
alternative="less")
```

Output)

```
        Exact binomial test

data:  sum(golub_ALL > -0.6) and length(golub_ALL)
number of successes = 10, number of trials = 27,
p-value = 0.1239
alternative hypothesis: true probability of success is less than 0.5
95 percent confidence interval:
 0.0000000 0.5466402
sample estimates:
probability of success
             0.3703704
```

- Here we have to show that pH4j in the ALL group is less than 0.5. the proportion of pH4j patients where the value is greater than -0.6**. The p-value=0.1239. We do not reject null hypothesis.**

(e) The proportion p_{H4j} in the ALL group differs from the proportion p_{H4j} in the AML group.

Answer)

```
golub_AML <- golub[2972,gol.fac=="AML"]
prop.test(x=c(sum(golub_ALL > -0.6), sum(golub_AML > -0.6)),
n=c(length(golub_ALL),length(golub_AML)), alternative="two.sided")
```

Output)

```
        2-sample test for equality of proportions with
        continuity correction

data:  c(sum(golub_ALL > -0.6), sum(golub_AML > -0.6)) out of c(lengt
h(golub_ALL), length(golub_AML))
X-squared = 2.6901, df = 1, p-value = 0.101
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.74094690  0.02714219
sample estimates:
   prop 1    prop 2
0.3703704 0.7272727
```

- We have to show that proportion of pH4j in the ALL group differs from the proportion pH4j in the AML group. **p-value is 0.101 so we don't reject null hypothesis.**

You should state the hypothesis, show the R commands for the tests, show the output of these tests, and state your conclusion based on these outputs.

2. (10 points) Suppose that the probability to reject a biological hypothesis by the results of a certain experiment is 0.03. This experiment is repeated 3000 times.
(a) How many rejections do you expect?

Answer)

```
n<-3000
p<-0.03
expreject <- (n*p)
expreject
```

Output)

```
[1] 90
```

(b) What is the probability of less than 75 rejections?

Answer)

```
pbinom(74,3000,0.03)
```

Output)

```
[1] 0.04537989
```

3. (10 points)
For testing H_0: μ=5 versus H_A: μ>5, we considers a new α=0.1 level test which rejects when $t_{obs} = \frac{\bar{X}-5}{s/\sqrt{n}}$ falls between $t_{0.3,n-1}$ and $t_{0.4,n-1}$.
Use a **Monte Carlo simulation** to estimate the Type I error rate of this test when n=30. Do 10,000 simulation runs of data sets from the $N(\mu = 5, \sigma = 4)$. Please show the R script for the simulation, and the R outputs for running the script.

Provide your numerical estimate for the Type I error rate. Is this test valid (that is, is its Type I error rate same as the nominal α=0.1 level)?

Answer)

```
x.sim<-matrix(rnorm(10000*30, mean=5, sd=4), ncol=30)

tstat<-function(x)(mean(x)-5)/sd(x)*sqrt(length(x))

tstat.sim<-apply(x.sim,1,tstat)

power.sim<-mean(tstat.sim>qt(0.90,df=29))

power.sim+c(-1,0,1)*qnorm(0.975)*sqrt(power.sim*(1-
power.sim)/10000)
```

Output)

```
[1] 0.09509407 0.10100000 0.10690593
```

- **So, the Monte Carlo estimate of the Type I error rate is 0.101 with its 95% CI as (0.095, 0.1069). This does agree with the nominal level of α = 0.1.**

4. (20 points)

On the Golub et al. (1999) data set, do **Welch two-sample t-tests** to compare every gene's expression values in ALL group versus in AML group.

(a) Use Bonferroni and FDR adjustments both at 0.05 level. How many genes are differentially expressed according to these two criteria?

Answer)

```
data(golub, package = "multtest")

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

ALL <- golub[2972, gol.fac == "ALL"]

AML <- golub[2972, gol.fac == "AML"]

t_test <- apply(golub, 1, function(x) t.test(x~gol.fac, var.equal = F))
```

```
p.values <- sapply(t_test, function(x) x$p.value)
p.bon <-p.adjust(p=p.values, method="bonferroni")
p.fdr <-p.adjust(p=p.values, method="fdr")
sum(p.bon<0.05)
sum(p.fdr<0.05)
```

Output)

```
> sum(p.bon<0.05)
[1] 103
> sum(p.fdr<0.05)
[1] 695
```

(b) Find the gene names for the top three strongest differentially expressed genes (i.e., minimum p-values). Hint: the gene names are stored in ***golub.gnames***.

Answer)

```
data(golub, package = "multtest");
gol.fac <- factor(golub.cl,levels=0:1, labels= c("ALL","AML"))
test <- apply(golub, 1, function(x) t.test(x ~ gol.fac)$p.value)
o <- order(test,decreasing=FALSE)
golub.gnames[o[1:3],2]
```

Output)

```
[1] "Zyxin"
[2] "FAH Fumarylacetoacetate"
[3] "APLP2 Amyloid beta (A4) precursor-like protein 2"
```

Please submit your R commands together with your answers to each part of the question.