1. (50 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 -1.

(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.

(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.

(d) Let p_{low} denote the proportion of patients for whom the "H4/j gene" expression is lower than the "APS Prostate specific antigen" expression. We wish to show that p_{low} in the ALL group is greater than half. Does this test conclusion agree with the conclusion in part (c)?

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

(f) p_{H4j} in the ALL group differs from p_{H4j} in the AML group.

Please submit your R commands for the tests, the output of these tests, and stated your decision based on these outputs.

A)

$H_0 \rightarrow \mu = -1$, $H_a \rightarrow \mu > -1$

RScript:

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

Answer:

```
One Sample t-test

data:  golub[2972, gol.fac == "ALL"]
t = 3.2743, df = 26, p-value = 0.001497
alternative hypothesis: true mean is greater than -1
```

95 percent confidence interval:
 -0.844439   Inf
sample estimates:
 mean of x
-0.6753033

As we can see in t.test p value is 0.001497 which is less than 0.05 hence we will reject null hypothesis, which states that mean “H4/j gene” gene expression value in the ALL group is greater than -1.

b)

H_0 -> $\mu_X = \mu_y$, H_a -> $\mu_X \neq \mu_y$

Rscript:

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

Answer:

Welch Two Sample t-test

data: golub[2972, gol.fac == "ALL"] and golub[2972, gol.fac == "AML"]
t = -1.4988, df = 29.978, p-value = 0.1444
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.48627436 0.07463315
sample estimates:
 mean of x mean of y
-0.6753033 -0.4694827

We can see in the Welch two sample t-test that p value is 0.1444 which is greater than 0.05 hence we fail to reject null hypothesis in this case which means The mean “H4/j gene” gene expression value in ALL group does not differs from the mean “H4/j gene” gene expression value in the AML group.

C)

H_0 -> $\mu_X = \mu_y$, H_a -> $\mu_X < \mu_y$

Rscript:

```
data(golub, package = "multtest")
gol.fac<- factor(golub.cl,levels=0:1, labels=c("ALL","AML"))
t.test(golub[2972,gol.fac=="ALL"],golub[2989,gol.fac=="ALL"],paired
=T, alternative = "less")
```

Answer:

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 difference in means is less than 0
95 percent confidence interval:
-Inf -0.02175309
sample estimates:
mean of the differences
-0.3050307

We can see in the paired t-test that p value is 0.03886 which is less than 0.05 so we will reject the null hypothesis and hence can say In the ALL group, the mean expression value for the “H4/j gene” is lower than the mean expression value for the “APS Prostate specific antigen” gene.

D)

H_0 -> $P_{low} = \frac{1}{2}$, H_a -> $P_{low} > \frac{1}{2}$

Rscript:

```
data(golub, package = "multtest")
gol.fac<- factor(golub.cl,levels=0:1, labels=c("ALL","AML"))
ALL2972<- golub[2972,gol.fac=="ALL"]
ALL2989<- golub[2989,gol.fac=="ALL"]
plow <- mean(ALL2972<ALL2989)
prop.test(x=17,n=27,p=1/2, alternative="greater", correct="true")
```

Answer:

1-sample proportions test with continuity correction

data: 17 out of 27, null probability 1/2
X-squared = 1.3333, df = 1, p-value = 0.1241
alternative hypothesis: true p is greater than 0.5
95 percent confidence interval:
 0.4535203 1.0000000
sample estimates:
      p
0.6296296

Here we can see the p value is 0.1241 which is more than 0.05 so we fail to reject the null hypothesis and hence We can say that p_{low} in the ALL group is not greater than half. Now In the c part we saw that the mean gene expression of "H4/j gene" is lower than the mean expression value for the "APS Prostate specific antigen" gene which still holds true here but we are not able to say that the proportion of H4/j gene expression being less than APS Prostate specific antigen gene is greater than .5.

E)

H_0 -> $P_{h4j} = \frac{1}{2}$, H_a -> $P_{h4j} < \frac{1}{2}$

Rscript:

```
data(golub, package = "multtest")
gol.fac<- factor(golub.cl,levels=0:1, labels=c("ALL","AML"))
ALL2972<- golub[2972,gol.fac=="ALL"]
ph4j<-sum(ALL2972>-.5)
prop.test(x=ph4j,n=27,p=1/2, alternative="less", correct="true")
```

Answer:

1-sample proportions test with continuity correction

data: ph4j out of 27, null probability 1/2
X-squared = 3.7037, df = 1, p-value = 0.02715
alternative hypothesis: true p is less than 0.5
95 percent confidence interval:
 0.0000000 0.4728575
sample estimates:
      p
0.2962963

Here we can see the p value as 0.02715, which is less than 0.05, so we will reject the null hypothesis and hence we can conclude that p_{H4j} in the ALL group is less than 0.5.

F)

H_0 -> $P_1 = P_2$, H_a -> $P_1 \neq P_2$

Rscript:

```
data(golub, package = "multtest")
gol.fac<- factor(golub.cl,levels=0:1, labels=c("ALL","AML"))
ALL2972<- golub[2972,gol.fac=="ALL"]
ph4jALL<-sum(ALL2972>-.5)
AML2972<- golub[2972,gol.fac=="AML"]
ph4jAML<-sum(AML2972>-.5)
prop.test(x=c(ph4jALL,ph4jAML),n=c(27,11), alternative="two.sided",
correct="true")
```

Answer:

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

data:  c(ph4jALL, ph4jAML) out of c(27, 11)
X-squared = 0.3086, df = 1, p-value = 0.5785
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.5631762  0.2466779
sample estimates:
  prop 1    prop 2
0.2962963 0.4545455
```

Here we can see the p value is 0.5785 which is greater than 0.05 so we fail to reject null hypothesis and hence can conclude that p_{H4j} in the ALL group does not differs from p_{H4j} in the AML group.

2. (10 points) Suppose that the probability to reject a biological hypothesis by the results of a certain experiment is 0.05. Suppose that this experiment is repeated 1000 times.
(a) How many rejections do you expect?
(b) What is the probability of less than 20 rejections?

A) Expected rejections will equal to the probability times the simulations which would be .05*1000 = 50.

B)
Rscript:

pbinom(19,size=1000, p=0.05)

Answer:

2.879692e-07 is the probability of less than 20 rejections.

3. (10 points)

For testing H_0: $\mu=3$ versus H_A: $\mu>3$, we considers a new $\alpha=0.1$ level test which rejects when $t_{obs}=\frac{\bar{X}-3}{S/\sqrt{n}}$ falls between $t_{0.3,n-1}$ and $t_{0.4,n-1}$.

(a) Use a Monte Carlo simulation to estimate the Type I error rate of this test when n=20. Do 10,000 simulation runs of data sets from the N $(\mu = 3, \sigma = 4)$

(b) . Please submit 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 $\alpha=0.1$ level)?

(b) Should we use this new test in practice? Why or why not?

a)

Rscript:

```
x.sim<-matrix(rnorm(10000*20, mean=3, sd=4), ncol=20)
tstat<-function(x) (mean(x)-3)/sd(x)*sqrt(20)
tstat.sim<-apply(x.sim,1,tstat)
power.sim<-mean((tstat.sim>qt(0.3,df=19))&(tstat.sim<qt(0.4,df=19)))
power.sim
power.sim+c(-1,0,1)*qnorm(0.975)*sqrt(power.sim*(1-
power.sim)/10000)
```

Answer:

rejection rate (power) with its 95% CI - 0.09197804 0.09780000 0.10362196

b) Yes, This test is valid as we are actually getting the value almost 0.1.

B)

Answer:

I think we should not use this new 0.1 level test in practice as type 1 error rate is 0.0978 with its 95% CI as (.0919, .103) which is not within the nominal level of α = 0.05

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?

(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***.

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

A)

Rscript:

```
data(golub, package = "multtest")
gol.fac<- factor(golub.cl,levels=0:1, labels=c("ALL","AML"))
p.values<- apply(golub, 1, function(x) t.test(x ~ gol.fac)$p.value)
p.bon<-p.adjust(p=p.values, method="bonferroni")
p.fdr<-p.adjust(p=p.values, method="fdr")
sum(p.values<0.05)
sum(p.bon<0.05)
sum(p.fdr<0.05)
```

Answer:
Bonferroni adjustments give 103 genes, which are expressed, at 0.05 level.
FDR adjustments give 695 genes, which are expressed, at 0.05 level.

B)

Rscript:

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

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

5. (10 points) Read the paper "Interval estimation for a binomial proportion" by Lawrence D Brown, T Tony Cai, Anirban DasGupta (2001) Statistical Science pages 101-117. Available at link http://projecteuclid.org/download/pdf_1/euclid.ss/1009213286

(a) Program R functions to calculate the Wald CI, the Wilson CI and the Agresti–Coull CI for binomial proportion. (Formulas are in equations (1), (4) and (5) of the paper.)

(b) Run a Monte Carlo simulation to check the coverage of the Wald CI, the Wilson CI and the Agresti–Coull CI for n=40 and p=0.2 at the nominal confidence level of 95%. Do 10,000 simulation runs for calculating the empirical coverages.

Please submit your R functions in part (a). Submit your R script for the simulation in part (b). Also answer part (b) with your numerical estimates of the three coverage probabilities.

A)

Rscript:

```
Waldci<- function(x,n,alpha)
{
 k<-qnorm(1-alpha/2)
 p<-x/n
 q<- 1-p
 upper<- p+k * n^(-1/2) * (p*q)^1/2
 lower <- p-k *n^(-1/2) * (p*q)^1/2
 c(lower,upper)
}
Waldci(.2,40,0.05)
k<- qnorm(1-0.05/2)
k
Wilsonci<- function(x,n,alpha)
{
 k<- qnorm(1-alpha/2)
 pcap<- (x+((k^2)/2))/(n+k^2)
 p<-x/n
 q<- 1-p
 sd<- ((p*q)+((k^2)/4*n))^1/2
 z<- (k*(n^1/2))/(n+k^2)
 upper<- pcap+ (z*sd)
 lower<- pcap- (z*sd)
 return(lower,upper)
}
```

```
Agrestici<- function(x,n,alpha)
{
 k<- qnorm(1-alpha/2)
 pcap<- (x+((k^2)/2))/(n+k^2)
 q<- 1-pcap
 sd<-((pcap*q)^1/2) * ((n+k^2)^-1/2)
 upper<- pcap+k*sd
 lower<- pcap-k*sd
 return(lower,upper)
}
```

B)

Rscript:

```
Waldci<- function(x,n,alpha)
{
 k<-qnorm(1-alpha/2)
 p<-x/n
 q<- 1-p
 upper<- p+k * n^(-1/2) * (p*q)^1/2
 lower <- p-k *n^(-1/2) * (p*q)^1/2
 c(lower,upper)
}
Wilsonci<- function(x,n,alpha)
{
 k<- qnorm(1-alpha/2)
 pcap<- (x+((k^2)/2))/(n+(k^2))
 p<-x/n
 q<- 1-p
 sd<- ((p*q)+((k^2)/(4*n)))^(1/2)
 z<- (k*(n^(1/2)))/(n+k^2)
 upper<- pcap+ (z*sd)
 lower<- pcap- (z*sd)
 c(lower,upper)
}

Agrestici<- function(x,n,alpha)
{
 k<- qnorm(1-alpha/2)
 pcap<- (x+((k^2)/2))/(n+k^2)
 q<- 1-pcap
 sd<-((pcap*q)^1/2) * ((n+k^2)^-1/2)
 upper<- pcap+k*sd
 lower<- pcap-k*sd
 c(lower,upper)
}
```

```
n<-40
nsim<-10000
wald<-matrix(NA, nrow=nsim, ncol=2)
wilson <- matrix(NA, nrow=nsim, ncol=2)
agresti <- matrix(NA, nrow=nsim, ncol=2)
for (i in 1:nsim)
{
 x<-rbinom(40, size = n, prob=.2)
 wald[i,] <- Waldci(mean(x),n,.05)
 wilson[i,] <- Wilsonci(mean(x),n,.05)
 agresti[i,] <-Agrestici(mean(x),n,.05)
}
mean(wald[,1])
mean(wald[,2])
mean(wilson[,1])
mean(wilson[,2])
mean(agresti[,1])
mean(agresti[,2])
```

Answer:
Wald CI- 0.1750705 0.2245922
Wilson CI- 0.1049278 0.3473374
Agresti and Coull CI- 0.2241777 0.2280876