**Module 6 Home Work**

**Problem 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 appropriatehypothesis for proving the following claims. Chose and carry out the appropriatetests.

(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 plow 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 plow in the ALL group is greater than half. Does this test conclusion agree with the conclusion in part (c)?

(e) Let pH4j denotes the proportion of patients for whom the “H4/j gene” expression values is greater than -0.6. We wish to show that pH4j in the ALL group is less than 0.5.

(f) pH4j in the ALL group differs from pH4j 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.

Answer:

#Setup GOLUB for (1a-1f)

data(golub,package="multtest")

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

H4<- grep("H4/j", golub.gnames [,2])

APS<- grep("APS", golub.gnames [,2])

(a) One-sided sample t-Test

t.test(golub[H4,gol.fac=="ALL"],mu= -1, alternative = "greater")

Output: p-value = 0.001497

Our p-value is 0.001497 (p − value = 0.001497), therefore, we reject the null

hypothesis H0 and accept alternative hypothesis HA.

conclude that there is enough evidence to support the mean “H4/j” gene expression

value in the ALL group is greater than -1.

(b) Two-sided sample t-Test

length (golub[2972,gol.fac=="ALL"])

length(golub[2972,gol.fac=="AML"])

t.test(golub[2972,gol.fac=="ALL"],golub[2972,gol.fac=="AML"])

Output:

p-value = 0.1444

Our p-value is 0.1444 (p − value = 0.1444), therefore, we accept the null hypothesis H0 and will not accept alternate hypothesis HA.

We conclude the mean “H4/j gene” gene expression values in ALL group does not differ from the mean “H4/j gene” gene expression value in the AML group.

(c)

t.test(golub[2972,gol.fac=="ALL"

],golub[2989,gol.fac=="ALL"],alternative="less",paired=T)

Output: p-value = 0.03886

Our p-value is 0.03886, therefore, we reject the null hypothesis and accept alternative hypothesis HA.

We conclude that the “H4/j gene” gene is lower than the mean expression value for the “APS Prostate specific antigen” gene.

(d)

ALLH4gene<-length(golub[2972, gol.fac=="ALL"])

ALLAPS<-length(golub[2989, gol.fac=="ALL"])

AMLH4gene<-length(golub[2972,gol.fac=="AML"])

AMLAPS<-length(golub[2989, gol.fac=="AML"])

APSpALL<-golub[2989,gol.fac=="ALL"]

H4pALL<-golub[2972, gol.fac=="ALL"]

plow<- sum(H4pALL<APSpALL)

binom.test(x=plow,n=ALLH4gene,p=0.5,alternative="greater")

Output: p-value = 0.1239

Our p-value is 0.1239, therefore, we fail to reject the null hypothesis and we will not accept the alternative hypothesis HA.

We conclude that plow in the ALL group is not greater than half.

It’s not possible to compare the conclusions of test (c) and (d) because (c) ask for the mean of genes whereas (d) asks for the proportion of genes.

(e)

H4ALL<- sum(golub[2972, gol.fac=="ALL"] > -0.6)

ALLH4gene<-length(golub[2972, gol.fac=="ALL"])

binom.test(x=H4ALL, n=ALLH4gene, p=0.5, alternative="less")

Output: p-H4ALL<- sum(golub[2972, gol.fac=="ALL"] > -0.6)

ALLH4gene<-length(golub[2972, gol.fac=="ALL"])

binom.test(x=H4ALL, n=ALLH4gene, p=0.5, alternative="less")

Output: p-value = 0.1239

Our p-value is 0.1239, therefore, we reject the null hypothesis H0 and accept

alternative hypothesis HA. We conclude pH4j in the ALL group is less than 0.5.

(f)

H4ALL<-sum(golub[H4,gol.fac=="ALL"]>-0.6)

H4AML<-sum(golub[H4,gol.fac=="AML"]>-0.6)

ALLH4gene<- length(golub[H4, gol.fac=="ALL"])

AMLH4gene <- length(golub[H4, gol.fac=="AML"])

prop.test(x=c(H4ALL,H4AML), n=c(ALLH4gene,AMLH4gene),

alternative="two.sided")

Output:

p-value = 0.101

Our p-value is 0.101,herefore, we accept the null hypothesis and we reject the alternative hypothesis HA.

We conclude pH4j in the ALL group doesn’t differ from pH4j in the AML group.

**Problem 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 2000 times.

(a) How many rejections do you expect?

(b) What is the probability of less than 90 rejections?

Answer:

(a)

Let p be the probability that a hypothesis is rejected.

Let n be the total number of experiments run.

From the question we can infer the following that the expected number of rejections will be:

E(X) = n · p = 2000 · 0.05 = 100

So, we can expect 100 rejections.

(b)

We consider the Binomial distribution to calculate P(X<90) rejections

Given,

Sample size (n) =

Probability of rejection = 0.05.!Therefore, the sum of all probabilities of X, P(X) for

X = {0, . . . , 89} will give us P(X < 90).

We can compute this in R using the probability distribution function or the cumulative distribution function of the binomial distribution (dbinom and pbinom respectively) as follows:

pbinom(89,size=2000,prob=0.05)

[1] 0.1400147

**Problem 3:**

For testing H0: μ=3 versus HA: μ>3, we considers a new α=0.1 level test which

rejects when tobs= !!!

!/ ! falls between t0.3,n-1 and t0.4,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( μ= 3,σ = 4) . 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 α=0.1 level)?

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

Answer:

(a)

x.sim<-matrix(rnorm(10000\*20, mean=3, sd=4), ncol=20)

tstat<-function(x) (mean(x)-3)/sd(x)\*sqrt(length(x))

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

powersim1<-mean(tstat.sim<qt(0.3,df=19))

powersim2<- mean(tstat.sim<qt(0.4,df=19))

power.sim<- (powersim2-powersim1)

power.sim

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

Output:

Our rejection rate (TYPE I error) is 0.0995 with its 95% CI of (0.0933,!0.1050)

Type 1 error is almost same as the nominal value. Hence, the test is valid.

(b) We can use 0.1 as the level of testing but the problem here is the test does not have the power.

**Problem 4:**

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

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

b) Find the names of 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.

Answer:

(a)

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)

bon.p<-p.adjust(p=p.values,method="bonferroni")

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

sum(pt<0.05)

sum(bon.p<0.05)

sum(fdr.p<0.05)

Output:

sum(pt<0.05)

Error in pt < 0.05 :

comparison (3) is possible only for atomic and list types

sum(bon.p<0.05)

[1] 103 103 genes are differentially expressed by using Bonferroni adjustment at 0.05 level.

sum(fdr.p<0.05)

[1] 695 695 genes are differentially expressed by using Fdr adjustment at 0.05 level

(b)

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.values <- apply(golub, 1, function(x) t.test(x ~ gol.fac)$p.value)

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

orderAML<-order(fdr.p, decreasing=FALSE)

golub.gnames[orderAML[1:3],2]

Output:

[1] "Zyxin"

[2] "FAH Fumarylacetoacetate"

[3] "APLP2 Amyloid beta (A4) precursor-like protein 2"

**Problem 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.

Answer:

(a)

#Wald CT#

wald.CI <- function(X,n,confd=0.05){

z <- qnorm(1-confd/2)

p <- X/n

return (c(p,(p + c(-1,1) \* z \* sqrt((p\*(1-p))/n))))

}

#Wilson CI

wilson.CI <- function(X,n,confd=0.05){

z <- qnorm(1-confd/2)

p <- X/n

return (c(p,((1/(1+z^2/n)) \* (p + (z^2/(2\*n)) + c(-1,1)\*z\*sqrt((p\*(1-

p))/n+z^2/(4\*n^2))))))

}

#Agresti. CI

agresti.CI <- function(X,n,confd=0.05){

z <- qnorm(1-confd/2)

N <- n + z^2

P <- (X+z^2/2)/N

return (c(P,(P + c(-1,1)\*z\*sqrt((P\*(1-P))/N))))

}

(b)

n.40 <- rbinom(n=1,size=40,p=0.2)

n.40.Wald <- wald.CI(n.40,40)

n.40.Wilson <- wilson.CI(n.40,40)

n.40.AgC <- agresti.CI(n.40,40)

sim <- rbinom(n=10000,size=40,p=0.2)

confd.Wald <- NULL

confd.Wilson <- NULL

confd.AgC <- NULL

for(i in sim){

confd.Wald <- rbind(confd.Wald,wald.CI(i,40))

confd.Wilson <- rbind(confd.Wilson,wilson.CI(i,40))

confd.AgC <- rbind(confd.AgC,agresti.CI(i,40))

}

# Calculate CI Intervals

Wald.cover <- mean(0.2 > confd.Wald[,2] & 0.2 < confd.Wald[,3])

Wilson.cover <- mean(0.2 > confd.Wilson[,2] & 0.2 < confd.Wilson[,3])

AgC.cover <- mean(0.2 > confd.AgC[,2] & 0.2 < confd.AgC[,3])

print("Estimated coverage after 10,000 observations (simulations) of n=40, p=0.2")

print(paste("Wald CI",Wald.cover,sep=" = "))

print(paste("Wilson CI",Wilson.cover,sep=" = "))

print(paste("Agresti-Coull CI",AgC.cover,sep=" = "))

Output:

print("Estimated coverage after 10,000 observations (simulations) of n=40, p=0.2")

[1] "Estimated coverage after 10,000 observations (simulations) of n=40, p=0.2"

> print(paste("Wald CI",Wald.cover,sep=" = "))

[1] "Wald CI = 0.9036"

> print(paste("Wilson CI",Wilson.cover,sep=" = "))

[1] "Wilson CI = 0.9295"

> print(paste("Agresti-Coull CI",AgC.cover,sep=" = "))

[1] "Agresti-Coull CI = 0.9503"