Joshua Conte
2/20/17

Math7340: Statistics for Bioinformatics
Module 6 Homework

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 -0.9.

Answer:

1) The null hypothesis H_0: $\mu > -0.9$ and the alternative hypothesis H_A: $\mu < -0.9$.
2) I will use the Golub et al. (1999) data set, and take those H4/j gene gene expression values for the ALL patients.
3) p-value = 0.01601
4) The results are statistically significant since the p-value < α (where $\alpha = 0.05$), the data is considered to be "rare (or surprising) enough" when H_0 is true, therefore I reject null hypothesis and accept the alternative.

R Code:

```
> #a)
> #Hypothesis Testing Process
> #1) ALL patients should be greater than -0.9
> #H0: mu > -0.9 HA: mu < -0.9
>
> #2) Here we just use the Golub et al. (1999) data set, and take those
> #Gdf5 gene expression values for the ALL patients.
> rm(list=ls())  #clears environment
>
> data(golub, package = "multtest")
> gol.fac <- factor(golub.cl,levels=0:1, labels= c("ALL","AML"))
> H4j_ALL <- golub[2972,gol.fac=="ALL"]
>
> #3) summarize the data and carry out a test
> t.test(H4j_ALL, mu=-0.9, alternative = "greater")

        One Sample t-test

data:  H4j_ALL
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:
```

Math7340: Statistics for Bioinformatics
Module 6 Homework

```
 mean of x
-0.6753033
```

(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:

1) The null hypothesis H_0: $\mu \neq$ H4/j gene AML and the alternative hypothesis H_A: $\mu =$ H4/j gene AML.
2) I will use the Golub et al. (1999) data set, and take those H4/j gene gene expression values for the AML and ALL patients.
3) p-value = 0. 1444
4) The results are not statistically significant since the p-value > α (where α = 0.05), the data is not considered to be "rare (or surprising) enough" when H_0 is true, therefore there is not enough evidence to reject the null hypothesis.

R Code:

```
> #b)
> #Hypothesis Testing Process
> #1) "H4/j gene" gene expression value in ALL group differs from the mean "H4/j gene"
> #H0: mu does not equal H4/j gene AML; HA: mu equals H4/j gene AML
>
> #2) Here we just use the Golub et al. (1999) data set, and take those
> #Gdf5 gene expression values for the AML patients.
> H4j_AML <- golub[2972,gol.fac=="AML"]
>
> #3) summarize the data and carry out a test
> t.test(H4j_AML,H4j_ALL)

        Welch Two Sample t-test

data:  H4j_AML and H4j_ALL
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:
 -0.07463315  0.48627436
sample estimates:
 mean of x  mean of y
-0.4694827 -0.6753033
```

Math7340: Statistics for Bioinformatics
Module 6 Homework

(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:

1) The null hypothesis H_0: $\mu <$ APS gene ALL and the alternative hypothesis H_A: $\mu >$ APS gene ALL.
2) I will use the Golub et al. (1999) data set, and take those APS gene values for the ALL patients.
3) p-value = 0.03886
4) The results are statistically significant since the p-value < α (where α = 0.05), the data is considered to be "rare (or surprising) enough" when H0 is true, therefore I reject null hypothesis and accept the alternative.

R Code:

```
> #c)
> #Hypothesis Testing Process
> #1) “H4/j gene” gene is lower than the mean expression value for the
> #“APS Prostate specific antigen” gene
> #H0: mu < APS gene ALL; HA: mu > APS gene ALL
>
> #2) Here we just use the Golub et al. (1999) data set, and take those
> #APS gene expression values for the ALL patients.
> APS_ALL <- golub[2989,gol.fac=="ALL"]
>
> #3) summarize the data and carry out a test
> t.test(H4j_ALL,APS_ALL, paired=T, alternative = "less" )

        Paired t-test

data:  H4j_ALL and APS_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
```

Math7340: Statistics for Bioinformatics
Module 6 Homework

(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)?

Answer:

1) The null hypothesis H_0: $\mu <$ APS gene and the alternative hypothesis H_A: $\mu >$ APS gene.
2) I will use the Golub et al. (1999) data set, and take those APS gene values.
3) p-value = 0.0006145
4) The results are statistically significant since the p-value < α (where α = 0.05), the data is considered to be "rare (or surprising) enough" when H0 is true, therefore I reject null hypothesis and accept the alternative.

The test conclusion for both part (c) and (d) reject the null hypothesis.

R Code:

```
> #d)
> #Hypothesis Testing Process
> #1) “H4/j gene” gene is lower than the mean expression value for the
> #“APS Prostate specific antigen” gene
> #H0: mu < APS gene; HA: mu > APS gene
>
> #2) Here we just use the Golub et al. (1999) data set, and take
> #APS and H4j gene expression values.
> H4j <- golub[2972,]
> APS <- golub[2989,]
>
> #3) summarize the data and carry out a test
> t.test(H4j,APS, paired=T, alternative = "less" )

        Paired t-test

data:  H4j and APS
t = -3.5006, df = 37, p-value = 0.0006145
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
       -Inf -0.2366308
sample estimates:
mean of the differences
             -0.4567639
```

Math7340: Statistics for Bioinformatics
Module 6 Homework

(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.

Answer:

1) The null hypothesis H_0: $\mu > -0.6$ in the H4/j gene and the alternative hypothesis H_A: $\mu > -0.6$.
2) I will use the Golub et al. (1999) data set, and take those H4/j gene values.
3) p-value = 0.5809
4) The results are not statistically significant since the p-value > α (where α = 0.05), the data is not considered to be "rare (or surprising) enough" when H0 is true, therefore there is not enough evidence to reject the null hypothesis.

R Code:

```
> #e)
> #Hypothesis Testing Process
> #1) "H4/j gene" gene is lower than the mean expression value for the
> #"APS Prostate specific antigen" gene
> #H0: mu < APS gene; HA: mu > APS gene
>
> #2) Here we just use the Golub et al. (1999) data set, and take those
> #H4j gene expression values.
>
>
> #3) summarize the data and carry out a test
> t.test(H4j, mu = -0.6, alternative = "greater" )

        One Sample t-test

data:  H4j
t = -0.20556, df = 37, p-value = 0.5809
alternative hypothesis: true mean is greater than -0.6
95 percent confidence interval:
 -0.7447746        Inf
sample estimates:
 mean of x
-0.6157237
```

(f) The proportion pH4j in the ALL group differs from the proportion pH4j in the AML group.

Answer:

Even though the proportions are different, when the data are saved in two vectors x and y, we can just use t.test(x,y) for the Welch two-sample t-test. See 1(b) as an example.

Math7340: Statistics for Bioinformatics
Module 6 Homework

2. (10 points) Suppose that the probability to reject a biological hypothesis by the results of a certain experiment is 0.05. This experiment is repeated 2000 times.

(a) How many rejections do you expect?

Answer:

I expect 100 rejections.

Where:

N = 2000, p=0.05 so np = 100.

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

Answer:

p = 0.1649724

R code:

```
> pbinom(90,2000,0.05)
[1] 0.1649724
```

Math7340: Statistics for Bioinformatics
Module 6 Homework

3. (10 points)

For testing H_0: μ=3 versus H_A: μ>3, we considers a new α=0.1 level test which rejects when falls between and $t_{obs} = \frac{\bar{X}-3}{s/\sqrt{n}}$ 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(μ = 3, σ = 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

0.2418, this test is not valid because it does not agree with a nominal level of α=0.1

where:

Here I use N(μ = 2, σ = 4)

```
> #Problem 3
> #Calculating Power Using Monte Carlo simulation
> rm(list=ls())  #clears environment
> x.sim<-matrix(rnorm(10000*10, mean=2, sd = 4), ncol=20)
> tstat<-function(x) (mean(x)-3)/sd(x)*sqrt(length(x))
> tstat.sim<-apply(x.sim,1,tstat) #Calculate t-test statistic for each dat
a set
> power.sim<-mean(tstat.sim>qt(c(0.3,0.4),df=19)) #Calculate the rejection
rate
> #Display rejection rate (power) with its 95% CI
> power.sim+c(-1,0,1)*qnorm(0.975)*sqrt(power.sim*(1-power.sim)/10000)
[1] 0.2334079 0.2418000 0.2501921
```

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

No, because the error rate it too high.

Math7340: Statistics for Bioinformatics
Module 6 Homework

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?
Answers:

	Bonferroni	FDR
ALL	1607	2434
AML	820	2033

(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:
ALL

1. YWHAZ Tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein, zeta polypeptide
2. ZNF91 Zinc finger protein 91 (HPF7, HTF10)
3. MYL4 Myosin, light polypeptide 4, alkali; atrial, embryonic

AML

1. GB DEF = Polyadenylate binding protein II
2. AFFX-HSAC07/X00351_5_at (endogenous control)
3. FTL Ferritin, light polypeptide

R Code:

```
> #Problem 4
> rm(list=ls())  #clears environment
> data(golub, package = "multtest")
> gol.fac <- factor(golub.cl,levels=0:1, labels= c("ALL","AML"))
> p.values <- apply(golub[,gol.fac=="ALL"], 1, function(x) t.test(x)$p.val
ue)
> p.bon <-p.adjust(p=p.values, method="bonferroni")
> p.fdr <-p.adjust(p=p.values, method="fdr")
> #ALL Values
> sum(p.values<0.05)
[1] 2475
> sum(p.bon<0.05)
[1] 1607
> sum(p.fdr<0.05)
[1] 2434
> o <- order(p.values,decreasing=FALSE)
> golub.gnames[o[1:3],2]
```

Math7340: Statistics for Bioinformatics
Module 6 Homework

```
[1] "YWHAZ Tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation
protein, zeta polypeptide"
[2] "ZNF91 Zinc finger protein 91 (HPF7, HTF10)"
[3] "MYL4 Myosin, light polypeptide 4, alkali; atrial, embryonic"
>
> p.values <- apply(golub[,gol.fac=="AML"], 1, function(x) t.test(x)$p.val
ue)
> p.bon <-p.adjust(p=p.values, method="bonferroni")
> p.fdr <-p.adjust(p=p.values, method="fdr")
> #AML Values
> sum(p.values<0.05)
[1] 2129
> sum(p.bon<0.05)
[1] 820
> sum(p.fdr<0.05)
[1] 2033
> o <- order(p.values,decreasing=FALSE)
> golub.gnames[o[1:3],2]
[1] "GB DEF = Polyadenylate binding protein II"    "AFFX-HSAC07/X00351_5_a
t (endogenous control)"
[3] "FTL Ferritin, light polypeptide"
>
> #b)
> pt <- apply(golub, 1, function(x) t.test(x ~ gol.fac)$p.value)
> o <- order(pt,decreasing=FALSE)
> golub.gnames[o[1:3],2]
[1] "Zyxin"
[2] "FAH Fumarylacetoacetate"
[3] "APLP2 Amyloid beta (A4) precursor-like protein 2"
```

Math7340: Statistics for Bioinformatics
Module 6 Homework

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.)
Answers:
See code below, to get the code to work, you need to declare these variables (I put the values from part b in there):
NCL=0.95 #conf.level
p=0.2
n=40

Wald

$$CI_s = p \pm z\sqrt{p(1-p)/n}$$

where z is $1 - \frac{1}{2}\alpha$

R Code

```
#Wald
Wald.ci<-function(x,n,conf.level=NCL){
   alpha = 1-NCL
  z <- qnorm(1-alpha/2)
  x<-n*p

  p+c(-z,z)*sqrt((p*(1-p))/n)
  }
Wald.ci(x,n,NCL)
```

Wilson

$$CI = \frac{a \pm b}{c}$$

where

$$a = p + \frac{z^2}{2n}$$

$$b = z\sqrt{\frac{p(1-p) + \frac{z^2}{4n}}{n}}$$

$$a = 1 + \frac{z^2}{n}$$

Math7340: Statistics for Bioinformatics
Module 6 Homework

$$z = 1 - \frac{1}{2}\alpha$$

R Code

```
#wilson
wilson.ci<-function(x,n,conf.level=NCL){
 alpha = 1-NCL
 z <- qnorm(1-alpha/2)
 x<-n*p

 a<-p+((z^2)/(2*n))
 b<-z*sqrt(((p*(1-p))+((z^2)/(4*n)))/n)
 c<-1+((z^2)/n)

 (a+c(-b,b))/c
}
wilson.ci(x,n,NCL)
```

Agresti-Coull

$$CI = \tilde{p} \pm z\sqrt{\frac{1}{\tilde{n}}\tilde{p}(1-\tilde{p})}$$

where

$$z = 1 - \frac{1}{2}\alpha$$

$$\tilde{n} = n + z^2$$

$$\hat{p} = \frac{1}{\hat{n}}\left(x + \frac{1}{2}z^2\right)$$

R Code

```
#Agresti-Coull
AC.ci<-function(x,n,conf.level=NCL){
 alpha = 1-NCL
 z <- qnorm(1-alpha/2)
 x<-n*p

 n2=n+(z^2)
 p2=(1/n2)*(x+((z^2)/2))

 p2+c(-z,z)*sqrt((p2*(1-p2))/n2)
```

Math7340: Statistics for Bioinformatics
Module 6 Homework

```
}
AC.ci(x,n,NCL)
```

(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.

	Lower	Upper
Wald	0.07604099	0.323959
Wilson	0.1049999	0.3475731
Agresti-Coull	0.1024282	0.3501447

```
> #Problem 5
> rm(list=ls())  #clears environment
>
> #a)
> #declare variables
> NCL=0.95 #conf.level
> p=0.2
> n=40
> x=n*p
>
> #Wald
> Wald.ci<-function(x,n,conf.level=NCL){
+
+   alpha = 1-NCL
+   z <- qnorm(1-alpha/2)
+   x<-n*p
+
+   p+c(-z,z)*sqrt((p*(1-p))/n)
+
+ }
>
> Wald.ci(x,n,NCL)
[1] 0.07604099 0.32395901
>
>
> #wilson
> wilson.ci<-function(x,n,conf.level=NCL){
+
+   alpha = 1-NCL
+   z <- qnorm(1-alpha/2)
+   x<-n*p
+
+   a<-p+((z^2)/(2*n))
+   b<-z*sqrt(((p*(1-p))+((z^2)/(4*n)))/n)
+   c<-1+((z^2)/n)
+
+   (a+c(-b,b))/c
+ }
```

Math7340: Statistics for Bioinformatics
Module 6 Homework

```
>
> wilson.ci(x,n,NCL)
[1] 0.1049999 0.3475731
>
>
> #Agresti-Coull
> AC.ci<-function(x,n,conf.level=NCL){
+
+   alpha = 1-NCL
+   z <- qnorm(1-alpha/2)
+   x<-n*p
+
+   n2=n+(z^2)
+   p2=(1/n2)*(x+((z^2)/2))
+
+   p2+c(-z,z)*sqrt((p2*(1-p2))/n2)
+ }
>
> AC.ci(x,n,NCL)
[1] 0.1024282 0.3501447
>
> #b)
> #Wald Monte Carlo
> n=40
> p=0.2
>
> x.sim=rbinom(10000,size=n,prob=p)
> Wald.sim=matrix(Wald.ci(x.sim,n=n, conf.level=0.95),nrow=2)
>
> mean(Wald.sim[1,]);mean(Wald.sim[2,])
[1] 0.07604099
[1] 0.323959
>
>
> #Wilson Monte Carlo
> n=40
> p=0.2
>
> x.sim=rbinom(10000,size=n,prob=p)
> wilson.sim=matrix(wilson.ci(x.sim,n=n, conf.level=0.95),nrow=2)
>
> mean(wilson.sim[1,]);mean(wilson.sim[2,])
[1] 0.1049999
[1] 0.3475731
>
>
> #agresti-coull Monte Carlo
> n=40
> p=0.2
>
> x.sim=rbinom(10000,size=n,prob=p)
> AC.sim=matrix(AC.ci(x.sim,n=n, conf.level=0.95),nrow=2)
>
> mean(AC.sim[1,]);mean(AC.sim[2,])
[1] 0.1024282
[1] 0.3501447
```