Math7340 HW6

Chengbo Gu

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 appropriate hypothesis for proving the following claims. Chose and carry out the appropriate tests.

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

(a) The mean "H4/j gene" gene expression value in the ALL group is greater than -0.9.

$$H_0 : \mu_{ALL,H4/j} = -0.9, \quad H_A : \mu_{ALL,H4/j} > -0.9$$

```
t.test(golub[2972, gol.fac=="ALL"], mu=-0.9, alternative = "greater")
```

```
##
##  One Sample t-test
##
## data:  golub[2972, gol.fac == "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:
##  mean of x
## -0.6753033
```

The p-value is smaller than 0.05. Hence, we reject the null hypothesis and conclude that the mean "H4/j gene" gene expression value in 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.

$$H_0 : \mu_{ALL,H4/j} = \mu_{AML,H4/j}, \quad H_A : \mu_{ALL,H4/j} \neq \mu_{AML,H4/j}$$

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

```
##
##  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
```

The p-value is larger than 0.05. We accept the null hypothesis that mean "H4/j gene" genge expression values from ALL and AML groups are the same.

(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.

$$H_0 : \mu_{ALL,H4/j} = \mu_{ALL,APS}, \quad H_A : \mu_{ALL,H4/j} < \mu_{ALL,APS}$$

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

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

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

```
##
##  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
```

The p-value here is smaller than 0.05. Hence, we reject the null hypothesis and conclude that mean expression value for "H4/j" gene is lower than that for "APS Prostate specific antigen" gene in the ALL group.

(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)?

$$H_0 : p_{low} = \frac{1}{2}, \quad H_A : p_{low} > \frac{1}{2}$$

```
res <- golub[2972, gol.fac=="ALL"] < golub[2989, gol.fac=="ALL"]
binom.test(sum(res), length(res), p=1/2, alternative = "greater")
```

```
##
##  Exact binomial test
##
## data:  sum(res) and length(res)
## number of successes = 17, number of trials = 27, p-value = 0.1239
```

```
## alternative hypothesis: true probability of success is greater than 0.5
## 95 percent confidence interval:
##  0.4533598 1.0000000
## sample estimates:
## probability of success
##              0.6296296
```

Since the p-value here is greater than 0.05, we accept the null hypothesis that p_{low} is exactly $\frac{1}{2}$.

(e) 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.

$$H_0 : p_{ALL,H4J} = 0.5, \quad H_A : p_{ALL,H4J} < 0.5$$

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

```
##
##  Exact binomial test
##
## data:  sum(res) and length(res)
## 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
```

The p-value here is greater than 0.05. Thus we accept the null hypothesis that p_{H4j} equals to 0.5.

(f) The proportion p_{H4j} in the ALL group differs from the proportion pH4j in the AML group.

$$H_0 : p_{ALL,H4J} = p_{AML,H4J}, \quad H_A : p_{ALL,H4J} \neq p_{AML,H4J}$$

```
resALL <- golub[2972, gol.fac=="ALL"] < -0.6
obsALL <- sum(resALL); nALL <- length(resALL)
resAML <- golub[2972, gol.fac=="AML"] < -0.6
obsAML <- sum(resAML); nAML <- length(resAML)
prop.test( x=c(obsALL, obsAML), n=c( nALL, nAML ), alternative="two.sided")
```

```
##
##  2-sample test for equality of proportions with continuity
##  correction
##
## data:  c(obsALL, obsAML) out of c(nALL, nAML)
## X-squared = 2.6901, df = 1, p-value = 0.101
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.02714219  0.74094690
## sample estimates:
##    prop 1    prop 2
## 0.6296296 0.2727273
```

The p-value is greater than 0.05. So we accept the null hypothesis that p_{H4j} proportions are the same in ALL and AML groups.

Problem 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.

The number of rejections $X \backsim Binom(size = 2000, \quad prob = 0.05)$.

(a) How many rejections do you expect?

$$E(X) = size * prob = 2000 * 0.05 = 100$$

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

$$P(X < 90) = P(X \leq 89) = pbinom(89, size = 2000, prob = 0.05) = 0.1400$$

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

```
## [1] 0.1400147
```

Problem 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 $N(\mu = 3, \sigma = 4)$. 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?

```
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)
power.sim <- mean(tstat.sim > qt(0.3, df=19) & tstat.sim < qt(0.4, df=19))
power.sim+c(-1,0,1)*qnorm(0.975)*sqrt(power.sim*(1-power.sim)/10000)
```

```
## [1] 0.09908914 0.10510000 0.11111086
```

So the Monte Carlo estimate of the Type I error rate is 0.1051 with its 95% CI as (0.0990891, 0.1111109). This does agree with the nominal level of $\alpha = 0.1$. This test is valid.

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

No, we shouldn't. Because this new test violates the original intention of hypothesis test. The decision rule is to reject the null hypothesis H_0 if the observed value t_{obs} is in the critical region, and to accept or "fail to reject" the hypothesis otherwise. For single-tail test of level $\alpha = 0.1$, one can say $(-\infty, t_{0.1})$ or $(t_{0.9}, \infty)$ are the critical regions but not $(t_{0.3}, t_{0.4})$.

Problem 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?

```
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)
```

```
## [1] 1078
```

```
sum(p.bon<0.05)
```

```
## [1] 103
```

```
sum(p.fdr<0.05)
```

```
## [1] 695
```

103 genes are differentially expressed according to Bonferroni.

695 genes are differentially expressed according to FDR.

(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.

```
golub.gnames[,2][order(p.values)][1:3]
```

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

```
golub.gnames[,2][order(p.bon)][1:3]
```

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

```
golub.gnames[,2][order(p.fdr)][1:3]
```

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

The top three strongest differentially expressed genes are Zyxin, FAH Fumarylacetoacetate and 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.

(a) Program R functions to calculate the Wald CI, the Wilson CI and the Agresti–Coull CI for binomial proportion.

```
# Wald.CI
Wald.CI <- function(x, n, conf.level) {
  p <- x/n
  z <- qnorm(1-(1-conf.level)/2)
  p + c(-1,1)*z*sqrt(p*(1-p)/n)
}

# Wilson.CI
Wilson.CI <- function(x, n, conf.level) {
  p <- x/n
  z <- qnorm(1-(1-conf.level)/2)
  (x+z^2/2)/(n+z^2) + c(-1, 1)*(z*sqrt(n)/(n+z^2))*sqrt(p*(1-p)+z^2/(4*n))
}

# AC.CI
AC.CI <- function(x, n, conf.level) {
  z <- qnorm(1-(1-conf.level)/2)
  x.s <- x + z^2/2
  n.s <- n + z^2
  p.s <- x.s/n.s
  q.s <- 1-p.s
  p.s + c(-1, 1)*z*sqrt(p.s*q.s/n.s)
}
```

(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.

```
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,] < p & Wald.sim[2,] > p)
```

```
## [1] 0.9076
```

```
Wilson.sim <- matrix(Wilson.CI(x.sim, n=n, conf.level=0.95), nrow=2)
mean(Wilson.sim[1,] < p & Wilson.sim[2,] > p)
```

```
## [1] 0.9312
```

```
AC.sim <- matrix(AC.CI(x.sim, n=n, conf.level=0.95), nrow=2)
mean(AC.sim[1,] < p & AC.sim[2,] > p)
```

```
## [1] 0.9518
```

So the coverages are 0.9076, 0.9312 and 0.9518 for Wald CI, Wilson CI and Agresti-Coull CI respectively.