Sardana_Module6HW

Prachi Sardana

2023-02-21

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).

Set the null hypothesis mu = -0.9 Setting the alternative hypothesis > -0.9 p value is 0.01 which is close to 0 , hence we reject the null hypothesis that gene expression value of H4/J gene is greater than -0.9

```
library(multtest); data(golub)
```

```
## Loading required package: BiocGenerics
```

```
##
## Attaching package: 'BiocGenerics'
```

```
## The following objects are masked from 'package:stats':
##
##     IQR, mad, sd, var, xtabs
```

```
## The following objects are masked from 'package:base':
##
##     anyDuplicated, aperm, append, as.data.frame, basename, cbind,
##     colnames, dirname, do.call, duplicated, eval, evalq, Filter, Find,
##     get, grep, grepl, intersect, is.unsorted, lapply, Map, mapply,
##     match, mget, order, paste, pmax, pmax.int, pmin, pmin.int,
##     Position, rank, rbind, Reduce, rownames, sapply, setdiff, sort,
##     table, tapply, union, unique, unsplit, which.max, which.min
```

```
## Loading required package: Biobase
```

```
## Welcome to Bioconductor
##
##     Vignettes contain introductory material; view with
##     'browseVignettes()'. To cite Bioconductor, see
##     'citation("Biobase")', and for packages 'citation("pkgname")'.
```

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

```
##
##  One Sample t-test
##
## data:  x
## 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
```

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.

alternative hypothesis set to value not equal to 0 here the p value is 0.14 which is greater thann 0.05,hence accepting null hypothesis.

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

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.

p value = 0.04

The mean expression value for the "H4/j gene" gene is significantly lower than the mean expression value for the "APS Prostate specific antigen" gene in the ALL group,because the p valued of the test is less than the significance level of 0.05, which means we reject the null hypothesis.

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

# H4/j gene expression
x <- golub[2972,gol.fac =="ALL"]

# APS Prostate specific antigen
y <- golub[2989,gol.fac == "ALL"]

# t- test statistics when H4gene is lower than the mean expression value for APS.
t.test(x,y,alternative ="less",mu = 0)
```

```
##
##  Welch Two Sample t-test
##
## data:  x and y
## t = -1.7275, df = 45.763, p-value = 0.04541
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##          -Inf -0.008596676
## sample estimates:
##  mean of x  mean of y
## -0.6753033 -0.3702726
```

d) 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.

We'll accept the null hypothesis as the p value is close to 0.1.

```
# H4/j gene expression
X <-golub[2972,gol.fac=="ALL"]

# proportion of pH4j patients greater than -0.6(Binomial test)
binom.test(sum(X > -0.6), length(X), p=0.5, alternative = "less")
```

```
##
##  Exact binomial test
##
## data:  sum(X > -0.6) and length(X)
## 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
```

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

p value is 0.101 hence we accept the null hypothesis that the proportion of pH4J differs from the AML group.

```
# All group H4 gene
z_h4_all <- golub[2972,gol.fac=="ALL"]
# AML group H4 gene
z_h4_aml <- golub[2972,gol.fac=="AML"]

prop.test(x=c(sum(z_h4_all > -0.6), sum(z_h4_aml > -0.6)),n=c(length(z_h4_all),length(z_h4_aml)), alter
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(sum(z_h4_all > -0.6), sum(z_h4_aml > -0.6)) out of c(length(z_h4_all), length(z_h4_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
```

2. a) # The probability to reject a biological hypothesis by the results of a certain experiment is 0.03 # The experiment is repeated 3000 times # according to binomial expression

```
n = 3000
p = 0.03
E_x = n*p
E_x
```

```
## [1] 90
```

b) Probability less than 75 rejections

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

```
## [1] 0.04537989
```

3.) Test The output test is invalid as the the alpha significance =0.1 and test is not close to 0.1 to be valid. Here the numerical estimate is 0.05.

```
# Number of simulations = 10000
number_sim <- 10000

# sample size
n <- 30

# population mean
population_mean <- 5

# Standard deviation
standard_dev <- 4

alpha <- 0.1

# Calculating critical t value

tvalue_lower <- qt(0.3, n-1)
tvalue_higher <- qt(0.4, n-1)

# vector initialization for test results
testresult <- numeric(number_sim)
# initializing simulation
for (i in 1:number_sim) {
  x <- rnorm(n, mean=population_mean, sd=standard_dev)  # random sample generation
  x_mean <- mean(x)  # sample mean, standard deviation calculation
  s <- sd(x)
```

```
  #  t statistic value calculation
  t_stat <- (x_mean - population_mean)/(s/sqrt(n))
  # Condition if t statistic falls between the critical values for hypothesis rejection
  if (t_stat > tvalue_lower  & t_stat < tvalue_higher) {
    testresult[i] <- 0 # no HO rejection
  } else {
    testresult[i] <- 1 # HO rejection
  }
}

# Calculating Type I error rate
type1error<- mean(testresult)
type1error
```

```
## [1] 0.8963
```

```
cat("Type I error rate is", type1error)
```

```
## Type I error rate is 0.8963
```

4. To perform Welch two-sample t-tests to compare every gene's expression values in ALL group versus in AML group.

a) To Use Bonferroni and FDR adjustments both at 0.05 level and count the differentially expressed genes

b) To find the gene names for the top three strongest differentially expressed genes

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

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

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

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

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

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

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

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

```r
# Top three strongest differentially expressed genes

exp_value <- order(p.values, decreasing = FALSE)
three <- exp_value[1:3]
cat("\nThe top three strongest differentially expressed genes are:\n")
```

```
##
## The top three strongest differentially expressed genes are:
```

```r
for (i in three) {
   cat(golub.gnames[i, 2],"\n")
}
```

```
## Zyxin
## FAH Fumarylacetoacetate
## APLP2 Amyloid beta (A4) precursor-like protein 2
```