Gulishana Adilijiang (adiliji2)

STAT 530 Bioinformatics: Homework 3

Problem1:

For a fixed Pj that is less than threshold and is small enough to be included in the new vector x, the two FDR-adjusted p-values Pj* and Pj^+ will have two possible relationships based on the distribution of the p-values in the original data set.

(1) Pj^+ will be equal to Pj*, if the p-values in original data set is skewed and closer to 0, which means most of them are significant;

(2) most of Pj^+ will be larger than Pj* (only several smallest ones may keep the same), if the p-values in original data set is uniformly distributed, which means most of them are not significant.

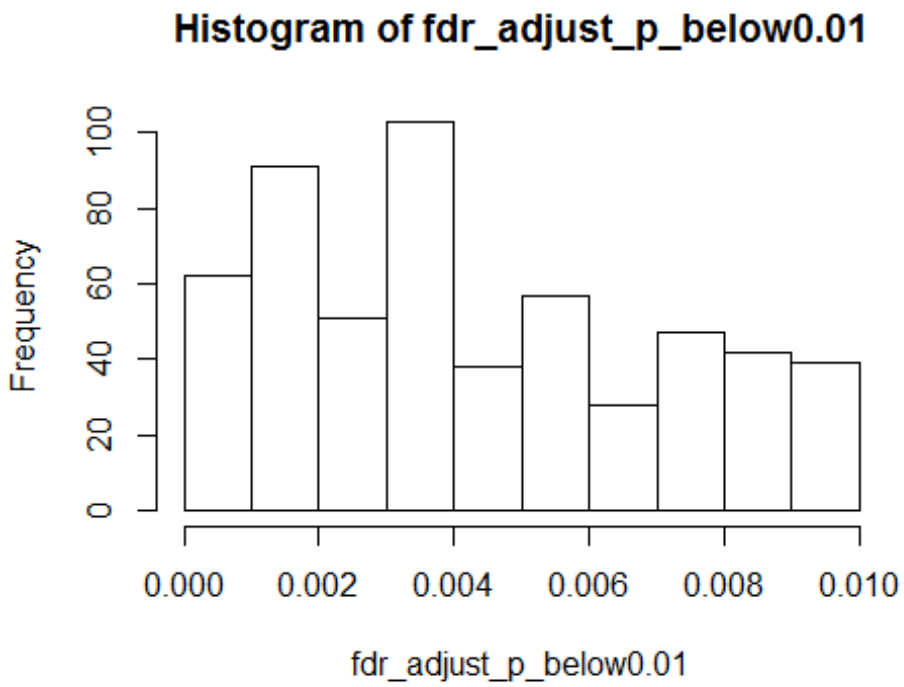
Below is an R simulation of the problem in these two cases.

R Notebook

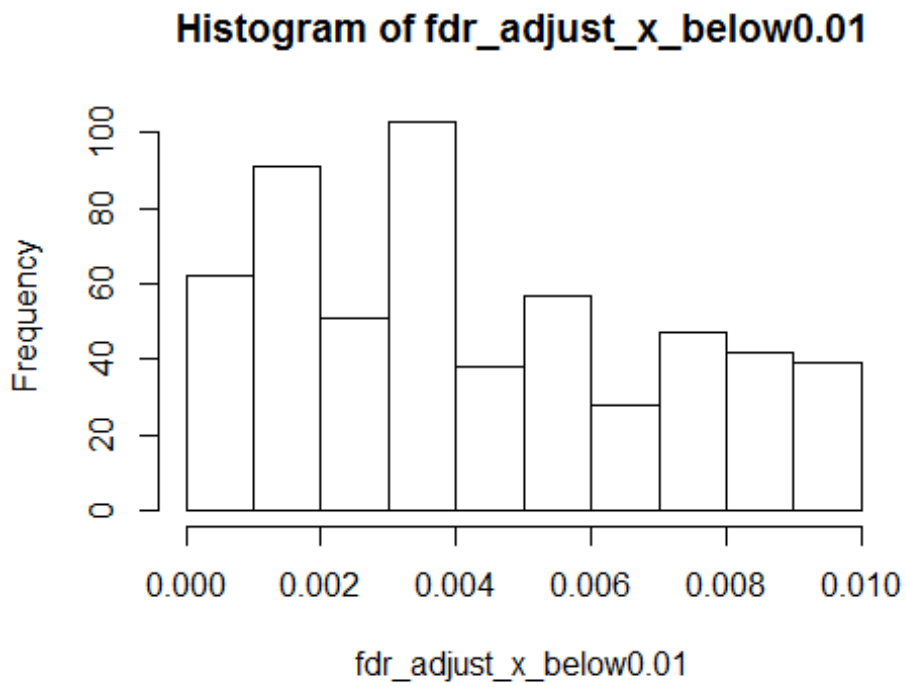
```
# Problem 1

# (1) compare two different normal distributions
p <- NULL
for (i in 1:1000) {
  sample_A = rnorm(n = 5,mean = 10,sd = 3)
  sample_B = rnorm(n = 5,mean = 30,sd = 9)
  t_test_results <- t.test(sample_A,sample_B,paired = FALSE, alternative = "t
wo.sided")
  p[i] <- t_test_results$p.value
}

# first rearrange p-values from the smallest to the largest
p = p[order(p)]

# FDR adjust p-values in p vecoter
m = length(p)
fdr_adjust_p = p.adjust(p,method = "fdr",n=m)
# list the first 30 values of p
head(fdr_adjust_p,n=30)
```

```
##  [1] 8.501882e-07 2.193563e-05 1.002929e-04 1.681134e-04 1.870403e-04
##  [6] 1.870403e-04 1.870403e-04 1.870403e-04 2.550008e-04 2.550008e-04
## [11] 3.463441e-04 3.463441e-04 3.463441e-04 3.463441e-04 3.463441e-04
## [16] 3.463441e-04 3.463441e-04 3.463441e-04 3.532150e-04 3.556171e-04
```

```
## [21] 3.903921e-04 3.903921e-04 3.903921e-04 4.321543e-04 4.321543e-04
## [26] 4.653809e-04 4.653809e-04 4.653809e-04 4.653809e-04 4.653809e-04

# then only store the smallest p-values that are less than 0.01 into a new ve
ctor x
x = p[which(p<0.01)]
# FDR adjust p-values in x using same total m
fdr_adjust_x = p.adjust(x,method = "fdr",n=m)
# list the first 30 values of FDR-adjusted x
head(fdr_adjust_x,n=30)

##  [1] 8.501882e-07 2.193563e-05 1.002929e-04 1.681134e-04 1.870403e-04
##  [6] 1.870403e-04 1.870403e-04 1.870403e-04 2.550008e-04 2.550008e-04
## [11] 3.463441e-04 3.463441e-04 3.463441e-04 3.463441e-04 3.463441e-04
## [16] 3.463441e-04 3.463441e-04 3.463441e-04 3.532150e-04 3.556171e-04
## [21] 3.903921e-04 3.903921e-04 3.903921e-04 4.321543e-04 4.321543e-04
## [26] 4.653809e-04 4.653809e-04 4.653809e-04 4.653809e-04 4.653809e-04

# list the last 30 values of FDR-adjusted x
N=length(fdr_adjust_x)
fdr_adjust_x[(N-29):N]

##  [1] 0.01320310 0.01322972 0.01326845 0.01326845 0.01332900 0.01332900
##  [7] 0.01338861 0.01360196 0.01368924 0.01369752 0.01377926 0.01383631
## [13] 0.01384312 0.01389477 0.01389477 0.01389477 0.01389477 0.01391102
## [19] 0.01391102 0.01391102 0.01394220 0.01408961 0.01412602 0.01415018
## [25] 0.01417414 0.01417414 0.01417414 0.01417414 0.01431435 0.01434787

# compare those to the FDR adjusted p-values of the same orginal values from
ordered p
fdr_adjust_p[(N-29):N]

##  [1] 0.01320310 0.01322972 0.01326845 0.01326845 0.01332900 0.01332900
##  [7] 0.01338861 0.01360196 0.01368924 0.01369752 0.01377926 0.01383631
## [13] 0.01384312 0.01389477 0.01389477 0.01389477 0.01389477 0.01391102
## [19] 0.01391102 0.01391102 0.01394220 0.01408961 0.01412602 0.01415018
## [25] 0.01417414 0.01417414 0.01417414 0.01417414 0.01431435 0.01434787

# if we only look at the p-value distribution histogram of FDR-adjusted p wit
hin the range (0, 0.01)
fdr_adjust_p_below0.01 = fdr_adjust_p[which(fdr_adjust_p<0.01)]
length(fdr_adjust_p_below0.01)

## [1] 558

hist(fdr_adjust_p_below0.01)
```




```
# if we only look at the p-value distribution histogram of FDR-adjusted x wit
hin the range (0, 0.01)
# so as to compare with previous histogram: fdr_adjust_p_below0.01
fdr_adjust_x_below0.01 = fdr_adjust_x[which(fdr_adjust_x<0.01)]
length(fdr_adjust_x_below0.01)

## [1] 558

hist(fdr_adjust_x_below0.01)
```




```r
# (2) compare two same normal distributions
p <- NULL
for (i in 1:1000) {
  sample_A = rnorm(n = 5,mean = 10,sd = 3)
  sample_B = rnorm(n = 5,mean = 10,sd = 3)
  t_test_results <- t.test(sample_A,sample_B,paired = FALSE, alternative = "t
wo.sided")
  p[i] <- t_test_results$p.value
}

# first rearrange p-values from the smallest to the largest
p = p[order(p)]

# FDR adjust p-values in p vecoter
m = length(p)
fdr_adjust_p = p.adjust(p,method = "fdr",n=m)
# list the first 30 values of FDR-adjusted p
head(fdr_adjust_p,n=30)
```

```
##  [1] 0.7871731 0.8816098 0.8816098 0.8816098 0.8816098 0.8816098 0.8816098
##  [8] 0.8816098 0.8816098 0.8816098 0.8816098 0.8816098 0.8816098 0.8816098
## [15] 0.8816098 0.8816098 0.8816098 0.8816098 0.8816098 0.8816098 0.8816098
## [22] 0.8816098 0.8816098 0.8816098 0.8816098 0.8816098 0.8816098 0.8816098
## [29] 0.8816098 0.8816098
```

```r
# then only store the smallest p-values that are less than 0.1 into a new vec
tor x
```

```r
x = p[which(p<0.1)]
# FDR adjust p-values in x using same total m
fdr_adjust_x = p.adjust(x,method = "fdr",n=m)
# list the first 30 values of FDR-adjusted x
head(fdr_adjust_x,n=30)
```

```
##  [1] 0.7871731 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000
##  [8] 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000
## [15] 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000
## [22] 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000
## [29] 1.0000000 1.0000000
```

```r
# list the last 30 values of FDR-adjusted x
N=length(fdr_adjust_x)
fdr_adjust_x[(N-29):N]
```

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

```r
# compare those to the FDR adjusted p-values of the same orginal values from
ordered p
fdr_adjust_p[(N-29):N]
```

```
##  [1] 0.8816098 0.8816098 0.8816098 0.8816098 0.8816098 0.8816098 0.8816098
##  [8] 0.8816098 0.8816098 0.8816098 0.8816098 0.8816098 0.8816098 0.8816098
## [15] 0.8816098 0.8816098 0.8816098 0.8816098 0.8816098 0.8816098 0.8816098
## [22] 0.8816098 0.8816098 0.8816098 0.8816098 0.8816098 0.8816098 0.8816098
## [29] 0.8816098 0.8816098
```

Problem2:

Here I used the definition of FDR-adjusted p-value from the suggested Youtube video.

FDR adjusted p-value of a fixed Pj = the smaller of two options below:

(1) previous FDR adjusted p-value of Pj' (the smallest p-value that is larger than Pj)

(2) Pj * (total # p-values / p-value rank of Pj) = Pj * (total # p-values / # p-values <= Pj)

Suppose data set A has m_A total p-values. And for a fixed Pj_A, there are n_A p-values smaller than Pj_A. Another data set B has m_B total p-values. If we put the fixed Pj_A from A into B, there are n_B p-values smaller than Pj_A in B.

Then if only apply FDR only to A, the adjusted p-value of fixed Pj_A is the smaller of two options below:

(1) previous FDR adjusted p-value of Pj'_A (the smallest p-value that is larger than Pj_A in A)

(2) $Pj_A * m_A/(n_A+1)$

If combine p-values from both A and B and then apply FDR to Pj_A, then the adjusted p-value of fixed Pj_A in "A+B" is the smaller of two options below:

(1) previous FDR adjusted p-value of Pj'_{A+B} (the smallest p-value that is larger than Pj_A in "A+B")

(2) $Pj_A * (m_A+m_B)/(n_{AB}+1) = Pj_A * (m_A+m_B)/ (n_A+n_B+1)$

Therefore, if we want to discover more significant p-values from A in FDR adjusted "A+B", then it needs the new adjustment of the fixed Pj_A from A be smaller, which will make more p-values from A be smaller than previous significance level α.

That is, to better achieve both conditions below:

(1) $Pj'_{A+B} < Pj'_A$

(2) $Pj_A * (m_A+m_B)/ (n_A+n_B+1) < Pj_A * m_A/(n_A+1)$

Consider if we make it an extreme case that for all Pj_A the #(2) condition is valid, then for all Pj'_A the #(1) condition is valid, which is just a previous FDR adjustment of the smallest p-value that is larger than Pj_A. Therefore, here we only consider the #(2) condition, which ends in a solution that:

$$n_B/m_B > (n_A+1)/ m_A$$

This means if before FDR adjustment, the **original B data set has a higher proportion of smaller (significant) p-values than that in the original A data set at the same significance level α**, then combining A and B will discover more significant p-values from A among FDR adjusted p-values of "A+B".