Quiz 5

Monikrishna Roy

2021-11-10

Question 1

For each row, calculate the t-SCORE between the two groups. Store this information in a data frame called t-OBSERVED.

```r
# code from previous Quiz
library(GEOquery)
dataset <- "GSE19804"
gsets <- getGEO(dataset, GSEMatrix = T, getGPL = T)
gset <- gsets[[1]]
expr <- exprs(gset)

pdata <- pData(gset)
control <- rownames(pdata[grep("Lung Normal", pdata$title), ])
cancer <- rownames(pdata[grep("Lung Cancer", pdata$title), ])

# Function to calculate t-score
cal_t_score <- function(x, cancer, control) {
  t.test(x[cancer], x[control])$statistic
}

TScore <- apply(expr, MARGIN = 1, FUN = cal_t_score, cancer, control)
t_OBSERVED = data.frame(TScore)
```

Question 2

Permute the samples of the original matrix and recalculate the t-SCORE 100 times for each row. Store the new scores in a matrix called t_NULL_DISTRIBUTION that will have 54675 rows (total number of genes in the given matrix) and 100 columns.

```r
library(tidyverse)
t_NULL_DISTRIBUTION <- lapply(c(1:100), function(i) {
  message(i)
  permuteExpr <- expr
  colnames(permuteExpr) <- sample(colnames(expr))
  apply(permuteExpr, MARGIN = 1, FUN = cal_t_score, cancer, control)
}) %>% do.call(what = cbind) %>% as.data.frame()
```

Question 3

Count how many times an observed value, from vector t-OBSERVED, is more extreme than the values in the respective row in the permuted matrix (t_NULL_DISTRIBUTION). Divide this count by the number of columns (100) and note that this value will be the empirical p-value. Save the p-values in the vector pT.

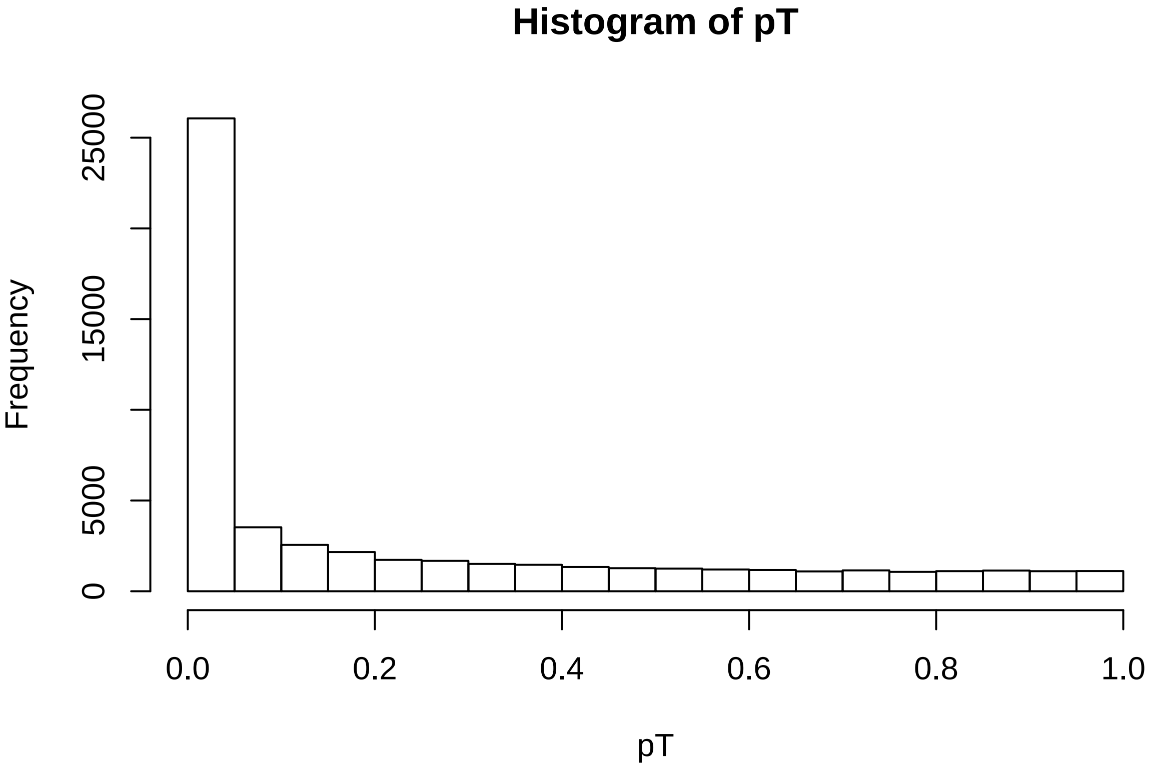
Hint: More extreme in two-tailed test means the number of values in the null that has absolute value higher that the absolute observed value.

```
pT <- lapply(rownames(t_OBSERVED), function(r) {
  sum(abs(t_NULL_DISTRIBUTION[r, ]) > abs(t_OBSERVED[r, ])) / ncol(t_NULL_DISTRIBUTION)
}) %>% unlist()
```

Question 4

Plot the histogram of the p-values

```
hist(pT)
```


Histogram of pT
Frequency
pT