Quiz 6

Monikrishna Roy

2021-11-16

Question 1

For each row in the expression matrix, calculate the e-SCORE (Euclidean distance between the means of the two groups) for each row. Store it in a data frame called e-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 Euclidean distance between the means of the two groups
cal_e_score <- function(x, cancer, control) {
  dist(rbind(mean(x[cancer]), mean(x[control])), method = "euclidean")
}

eSCORE <- apply(expr, MARGIN = 1, FUN = cal_e_score, cancer, control)
e_OBSERVED = data.frame(eSCORE)
```

Question 2

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

```r
library(tidyverse)
e_NULL_DISTRIBUTION <- lapply(c(1:100), function(i) {
  message(i)
  samples <- sample(ncol(expr), replace = F)
  controls <- samples[1:length(control)]
  cancers <- samples[(length(control) + 1):length(samples)]
  apply(
    expr,
    MARGIN = 1,
    FUN = cal_e_score,
    cancer = cancers,
    control = controls
  )
}) %>% do.call(what = cbind) %>% as.data.frame()
```

Question 3

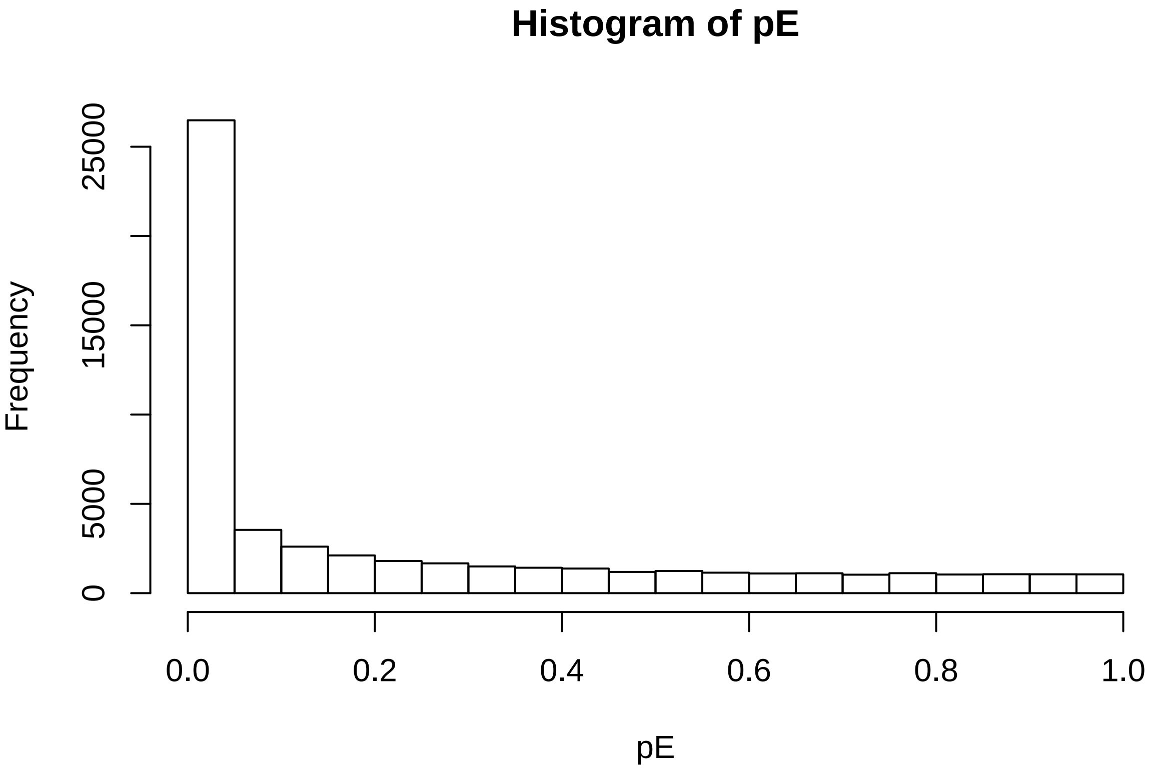
Count how many times an observed value, from vector e-OBSERVED, is more extreme than the values in the respective row in the permuted matrix. Divide this count by the number of columns (100) and note that this value will be the empirical p-value. Save the p-values in the vector pE.

```
pE <- lapply(rownames(e_OBSERVED), function(r) {
  mean(abs(e_NULL_DISTRIBUTION[r, ]) > abs(e_OBSERVED[r, ]))
}) %>% unlist()
```

Question 4

Plot the histogram of the p-values (pE)

```
hist(pE)
```


Histogram of pE
Frequency
pE