---

title: "m254-hw1 Name: Jiahao Tian"

output: html\_document

---

```{r setup, include=FALSE}

knitr::opts\_chunk$set(echo = TRUE)

```

```{r}

library(ISLR)

dim(NCI60$data)

head(NCI60$labs)

table(NCI60$labs)

```

\*\*Problem 1\*\*

```{r}

NSCLC <- NCI60$data[NCI60$labs=='NSCLC',]

RENAL <- NCI60$data[NCI60$labs=='RENAL',]

p\_values <- mapply(function(i)

t.test(NSCLC[,i], RENAL[,i], var.equal = T,

paired = F, alternative = "two.sided")$p.value,

1:ncol(NSCLC))

sort\_p <- sort(p\_values)

plot(sort\_p, xlab = "Rank", ylab = "p value",

cex=0.1, col='blue', main = "Sorted p value")

plot(ecdf(sort\_p), main="Empirical Dist. of Ordered p-values", col="blue")

lines(x=seq(0,1,0.01), y=seq(0,1,0.01), col="red", lwd=1.5)

text(0.8, 0.6, "Theoretical cdf", col="red")

text(0.25, 0.6, "Empirical cdf", col="blue")

plot(sort\_p, cex=0.5, xlim = c(1, 25), ylim = c(0, 0.0008),

main = "FDR and FWER", xlab = 'index i', ylab = 'p-value', pch=8)

x <- c(1:6830); fdr <- c(x/6830\*0.05); bonforroni <- rep(0.05 / 6830, times=length(x));

holm <- c(0.05/(length(x)-x+1))

sum(sort\_p <= bonforroni[1])

lines(x, fdr, lwd=1.5)

lines(x, bonforroni, lwd=1.5)

lines(x, holm, lwd=0.5)

text(20, 0.00001, "Holm's")

text(23, 0.00001, "Bonforroni")

text(18, 0.0002, "FDR")

sum(p\_values<=0.05)

```

\*\*Problem 2\*\*

```{r}

## Descriptive study

table(NCI60$labs)

## Exploratory study

plot(round(apply(NCI60$data, 2, mean)[1:20], 4), xlab="Cell type",

ylab="Gene expression",

main="Exploratory of first 10 genes", pch=3)

## Inferential study

plot(NCI60$data[,1], NCI60$data[,2], xlab="Gene 1", ylab="Gene 2", cex=0.4,

pch=1, main = "Inferential study of gene 1 and 2")

## Predictive study

library(e1071)

dat <- data.frame(NCI60$data[,1:2], y=factor(NCI60$labs))

mod <- svm(y~., data=dat, kernel="linear", cost=10, scale = F)

plot(mod, dat)

```

\*\*Problem 3\*\*

library(tidyverse)

```{r}

rm(list=ls())

babies <- read.table("babies.txt", header=TRUE)

bwt.nonsmoke <- subset(babies, smoke==0)$bwt

bwt.smoke <- subset(babies, smoke==1)$bwt

#1. We will generate the following statistics based on a sample size of 10 and observe the following diﬀerence:

permutation\_test <- function(N=10, method='median', B=1000, randomseed=1) {

n <- N

set.seed(randomseed)

nonsmokers <- sample(bwt.nonsmoke , n)

smokers <- sample(bwt.smoke , n)

if(method=='mean') {

diff <- mean(smokers) - mean(nonsmokers)

} else if(method=='median') diff <- median(smokers) - median(nonsmokers)

else return("Choose between 'mean' and 'median' !")

if(B > 100000) return("Iteration cannot be greater than 1000000 !")

diff\_permute <- c()

for (i in 1:B) {

dat <- c(smokers, nonsmokers)

shuffle <- sample(dat)

smokers\_star <- shuffle[1:n]

nonsmokers\_star <- shuffle[(n+1):(2\*n)]

if(method=='mean') {

diff\_star <- mean(smokers\_star) - mean(nonsmokers\_star)

} else if(method=='median') {

diff\_star <- median(smokers\_star) - median(nonsmokers\_star)

}

diff\_permute <- c(diff\_permute, diff\_star)

}

p <- (sum(diff\_permute <= -abs(diff)) + sum(diff\_permute >= abs(diff))) / B

return(list(test\_statstic=diff, permutation=diff\_permute, p\_value=p))

}

result\_mean <- permutation\_test(N=10, method = 'mean', B=1000, randomseed=1)

result\_mean$p\_value

hist(result\_mean$permutation, breaks = 20, xlab='null statistic',

main = 'Hist. of null dist. (N=10, B=1000, mean)')

abline(v=result\_mean$test\_statstic, lwd=5)

text(-10, 80, paste("Test stat. with p value:", result\_mean$p\_value), cex=1.5)

#2. Repeat the above exercise, but instead of the diﬀerences in mean, consider the diﬀerences in median.

result\_med <- permutation\_test(N=10, method = 'median', B=1000, randomseed=1)

result\_med$p\_value

hist(result\_med$permutation, breaks = 20, xlab='null statistic',

main = 'Hist. of null dist. (N=10, B=1000, median)')

abline(v=result\_med$test\_statstic, lwd=5)

text(-10, 80, paste("Test stat. with p value:", result\_med$p\_value), cex=1.5)

```

\*\*Problem 4\*\*

```{r}

#1. Please display the Q-Q plots and argue whether the Gaussian assumption is reasonable for this data set.

qqnorm(bwt.nonsmoke, main = "QQ plot for nonsmokers")

qqline(bwt.nonsmoke,col=2)

qqnorm(bwt.smoke, main = "QQ plot for smokers")

qqline(bwt.smoke,col=2)

#2. Perform the t test using the following R codes. Compare the resulting p-value with what you

# obtain from the permutation test in Problem 3. Do you reach the same conclusion?

n <- 10

nonsmokers <- sample(bwt.nonsmoke , n)

smokers <- sample(bwt.smoke , n)

result <- t.test(nonsmokers, smokers, alternative = 'two.sided')

result$p.value

```

\*\*Problem 7\*\*

```{r}

dist\_sample\_mean <- function(sample\_size=10, B=1000) {

n <- sample\_size

x <- matrix(rnorm(n\*B), ncol = B)

hist(apply(x, 2, mean), breaks=100, freq=T, xlab='Sample mean',

main = paste('Hist. of sample mean (sample distribution) [ N=', sample\_size, ']'))

}

dist\_sample\_mean(10, 1000000)

dist\_sample\_mean(100, 1000000)

dist\_sample\_mean(1000, 1000000)

```