Question 2

Comparing two sub-populations

Consider the TitanicSurvival Data Set. The interest lies in comparing the *age* of the two sub-populations P_1: female passengers and P_2: male passengers on titanic.

```
Titanic.Female = Titanic[Titanic$sex == "female",]
Titanic.Female = Titanic.Female$age
Titanic.Male = Titanic[Titanic$sex == "male",]
Titanic.Male = Titanic.Male$age
pop = list(pop1 = Titanic.Female, pop2 = Titanic.Male)
```

We will now consider 3 discrepancy measures:

$$D_1(P_1, P_2) = \frac{\bar{Y_1} - \bar{Y_2}}{\tilde{\sigma}\sqrt{1/n_1 + 1/n2}}$$

$$D_2(P_1, P_2) = \frac{SD(P_1)}{SD(P_2)} - 1$$

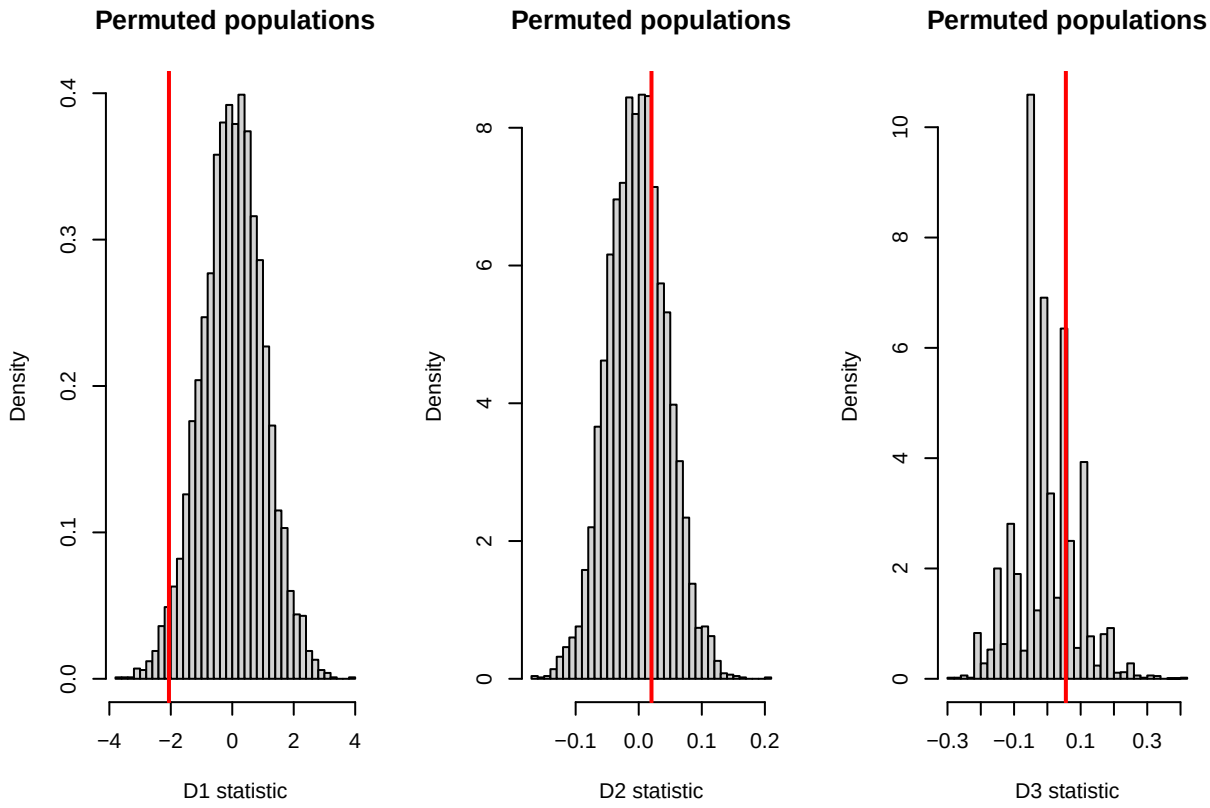
$$D_3(P_1, P_2) = \frac{IQR(P_1)}{IQR(P_2)} - 1$$

We will write the necessary functions - mixRandomly, D1Fn, D2Fn, D3Fn (hidden due to space constraint).

Let us now generate the historgrams of the three discrepency measures based on 5000 shuffles on the two subpopulations P_1: female passengers on Titanic P_2: male passengers on Titanic. We will also superimpose the observed discrepency measure on these histograms.

```
par(mfrow = c(1,3))

# Plot for D1
D1Vals <- sapply(1:5000, FUN = function(...){D1Fn(mixRandomly(pop))})

hist(D1Vals, breaks=40, probability = TRUE,
     main = "Permuted populations", xlab="D1 statistic",
     col="lightgrey")
abline(v=D1Fn(pop), col = "red", lwd=2)

#Plot for D2
D2Vals <- sapply(1:5000, FUN = function(...){D2Fn(mixRandomly(pop))})
hist(D2Vals, breaks=40, probability = TRUE,
     main = "Permuted populations", xlab="D2 statistic",
     col="lightgrey")
abline(v=D2Fn(pop), col = "red", lwd=2)

#Plot for D3
D3Vals <- sapply(1:5000, FUN = function(...){D3Fn(mixRandomly(pop))})
```

```
hist(D3Vals, breaks=40, probability = TRUE,
     main = "Permuted populations", xlab="D3 statistic",
     col="lightgrey")
abline(v=D3Fn(pop), col = "red", lwd=2)
```



We will now use all the three discrepancy measures D_1, D_2, D_3 to perform a multiple test to compare the two subpopulations P_1 and P_2. We will use $M = M^* = 300$ for multiple testing.

```
discrepancies <- list(D1Fn , D2Fn , D3Fn)
### The following takes a long time (about 20 minutes)
### for B_outer = B_inner = 1,000 say
### So for illustration much smaller values than would be sensible are
### used here
set.seed(341)
SLstar=calculateSLmulti(pop, discrepancies, B_outer = 300, B_inner=300)
SLstar
```

```
## [1] 0.6266667
```

Conclusion

• Since the p-value/significance level is large (≈ 0.63), there is no evidence against the hypothesis that the Male and Female passengers were randomly drawn from the same population based on the discrepency measures D_1, D_2, D_3 of age of passengers