Assignment 3 Question 4

Fenglin Chen

14/11/2021

## Summary

The significance level is a measure of how unusual the difference between two population attributes & are compared to the difference between attributes of randomly mixed sub-populations.

There are 4 main steps:

1. State the null hypothesis H0: and are drawn from the same population.
2. Formulate a discrepancy measure or test statistic that quantifies how inconsistent the null hypothesis is with the observations.
3. Apply the discrepancy measure to the observed populations.
4. Apply the discrepancy measure to randomly shuffled sub-populations and calculate the proportion (p-value) of random discrepancies higher than the original.

Finally, either conclude that there is no evidence or there is sufficient evidence against H0 at some significance level.

## Mathematical Formula

The observed discrepancy (baseline) is calculated on the original populations:

Random sub-populations are generated by shuffling and times:

Finally, calculate the p-value as the proportion of shuffled discrepancy values greater than the observed discrepancy:

## General Code

# Generate randomly shuffled sub-populations  
mixRandomly <- function(pop) {  
 pop1 <- pop$pop1  
 n\_pop1 <- nrow(pop1)  
   
 pop2 <- pop$pop2  
 n\_pop2 <- nrow(pop2)  
   
 mix <- rbind(pop1,pop2)  
 select4pop1 <- sample(1:(n\_pop1 + n\_pop2), n\_pop1,   
 replace = FALSE)  
   
 new\_pop1 <- mix[select4pop1,]   
 new\_pop2 <- mix[-select4pop1,]  
 list(pop1=new\_pop1, pop2=new\_pop2)  
}

# Define the discrepancy measure  
getDiscrepancy <- function(variate, attr) {  
 function(pop) {attr(pop$pop1[, variate]) -   
 attr(pop$pop2[,variate])}  
}

## Code Example

In this example, we calculate the significance level that New York City’s Gen Ed class size is different from its Collaborative Team Teaching (CTT) class size.

boroughs <- read.csv("Borough\_Summary.csv", header=TRUE)  
pop <- list(pop1 = subset(boroughs, boroughs$PROGRAM.TYPE == "GEN ED"),   
 pop2 = subset(boroughs, boroughs$PROGRAM.TYPE == "CTT"))  
set.seed(341)  
  
discFn <- getDiscrepancy("AVERAGE.CLASS.SIZE", mean)  
disc.random <- sapply(1:1000, FUN = function(...){   
 discFn(mixRandomly(pop))   
})  
  
disc.prop <- mean( abs(disc.random) >= discFn(pop) )  
disc.prop

## [1] 0.29

Since the significance level/p-value is very high (0.29), we conclude that there is insufficient evidence against the two populations coming from the same distribution.

