A3Q3

Fenglin Chen

15/11/2021

knitr::opts\_chunk$set(fig.width=10, fig.height=5)

### Question 3

#### Description of the Context

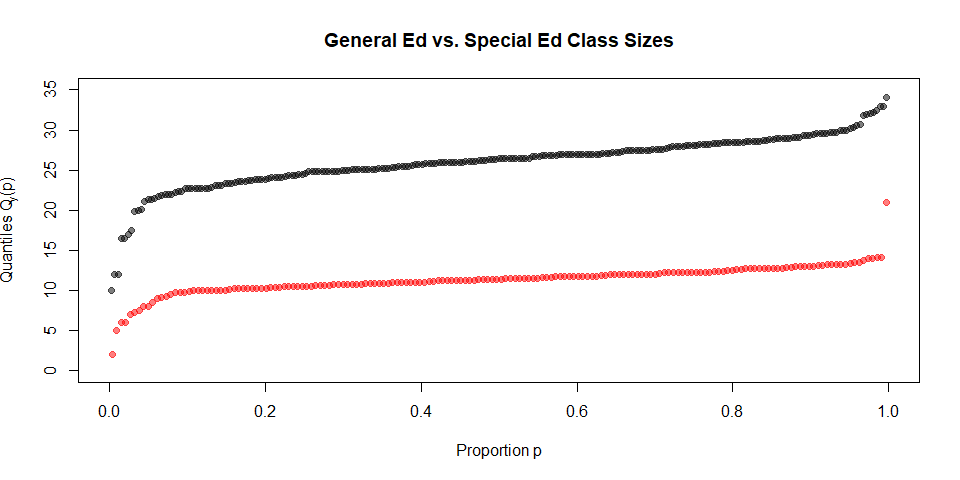
This dataset contains class sizes for five different boroughs in New York City: Brooklyn, Manhattan, Queens, Staten Island, and Bronx. The data is collected from 2010 to 2011, and is aggregated by location, grade, and program type.

The two sub-populations of interest are the average class sizes of General Education (GEN ED) and Self-Contained Special Education (SPEC ED).

#### Compare the Sub-Populations

boroughs <- read.csv("Borough\_Summary.csv", header=TRUE)  
pop <- list(pop1 = subset(boroughs, boroughs$PROGRAM.TYPE == "GEN ED"),   
 pop2 = subset(boroughs, boroughs$PROGRAM.TYPE == "SPEC ED"))

qvals <- sort(pop[[1]]$AVERAGE.CLASS.SIZE)  
pvals <- ppoints(length(qvals))  
plot(pvals, qvals, pch = 19, col = adjustcolor("black", alpha = 0.5),   
 xlim = c(0,1), ylim = c(0, 35),   
 xlab = "Proportion p", ylab = bquote("Quantiles Q"["y"] \* "(p)"),   
 main = "General Ed vs. Special Ed Class Sizes")  
qvals <- sort(pop[[2]]$AVERAGE.CLASS.SIZE)  
pvals <- ppoints(length(qvals))  
points(pvals, qvals, pch = 19, col = adjustcolor("red", alpha = 0.5))

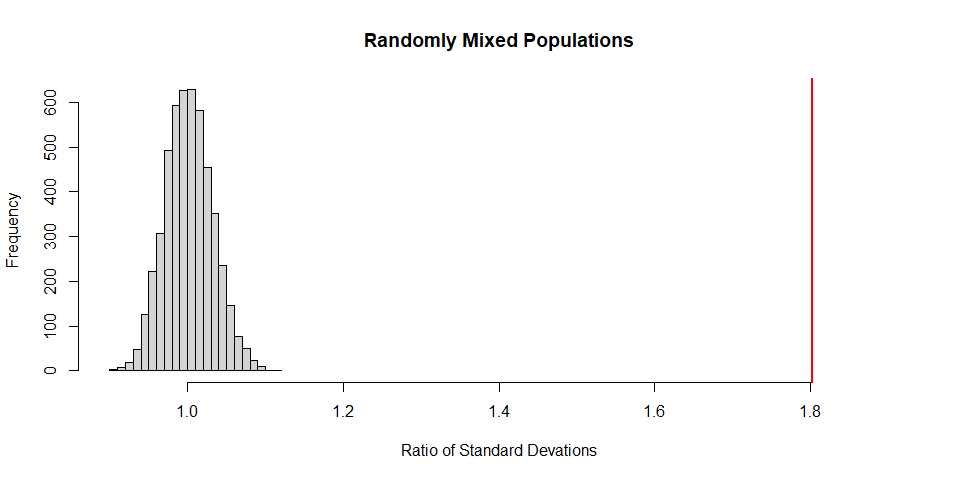
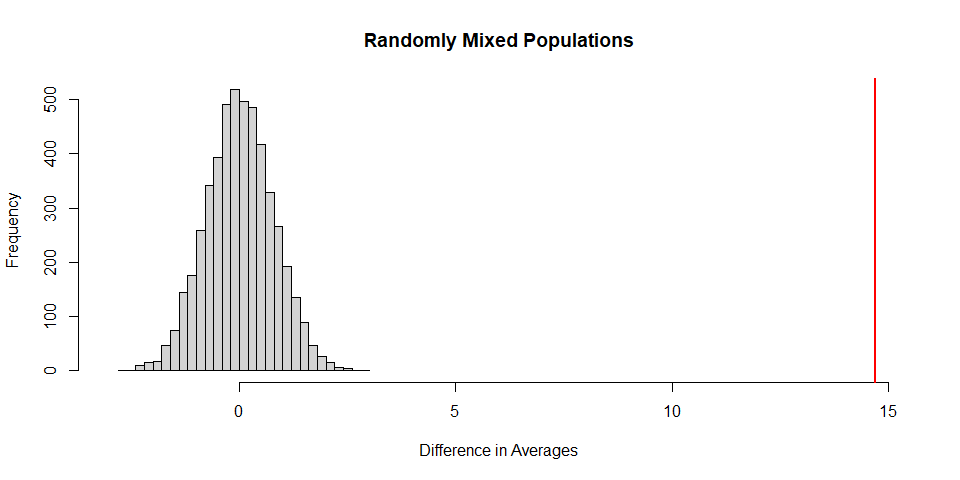


The quantile plot shows a clear separation between the sub-populations.

The functions getAveDiffsFn(), getSDRatioFn(), and mixRandomly() from class are used to compare the sub-populations.

diffAveLengths <- getAveDiffsFn("AVERAGE.CLASS.SIZE")  
ratioSDLengths <- getSDRatioFn("AVERAGE.CLASS.SIZE")  
round(c(diffAveLengths(pop), ratioSDLengths(pop)), 3)

## [1] 14.704 1.803



#### Description of the Comparison

The populations have a difference in class size of 14.704 and a ratio of standard deviations of 1.803.

The graphs show that the population attributes are very far from the distribution of randomly mixed attributes. Assuming a normal distribution, the p-value for both difference in mean and ratio of standard deviations are close to 0.

#### Conclusion

Both the numerical and graphical summaries show that it is almost impossible for the two sub-populations to be picked from a random mix, as there is sufficient evidence that “H0: pop1 and pop2 are from the same distribution” can be rejected.