A5Q2

Description

The data-set 'cars.csv' contains all information of a wide variety of used cars. It includes data on used car transmission types and price listed in US dollars. We will take first 1000 rows as the population $\mathcal{P}$. The interest is in comparing the two sub-populations of transmissions: $\mathcal{P}_1$: automatic and $\mathcal{P}_2$: mechanical. We will consider two discrepancy measures: $D_1(\mathcal{P}_1, \mathcal{P}_2) = \frac{median(\mathcal{P}_1) - median(\mathcal{P}_2)}{IQR(\mathcal{P})}$ and $D_2(\mathcal{P}_1, \mathcal{P}_2) = \frac{IQR(\mathcal{P}_1)}{IQR(\mathcal{P}_2)} - 1$. We will create histograms of two discrepancy measures based on 5000 shuffles of two sub-populations. We will find a P-value and give a conclusion.

Code

Get sub-populations and define discrepancy measure functions:

```
car <- read.csv('cars.csv',header=T)[1:1000,]
car.auto = car$price_usd[car$transmission=="automatic"]
car.mec = car$price_usd[car$transmission=="mechanical"]
car.pop = list(pop1 = car.auto , pop2 = car.mec)

D1 <- function(pop) {
    pop1 <- pop$pop1
    m1 <- median(pop1)
    pop2 <- pop$pop2
    m2 <- median(pop2)
    IQR.P <- IQR(c(pop1,pop1))
    test <- (m1 - m2) /  IQR.P
    test
}

D2 <- function(pop) {
    pop1 <- pop$pop1
    IQR1 <- IQR(pop1)
    pop2 <- pop$pop2
    IQR2 <- IQR(pop2)
    test <- IQR1/IQR2 - 1
    test
  }
```

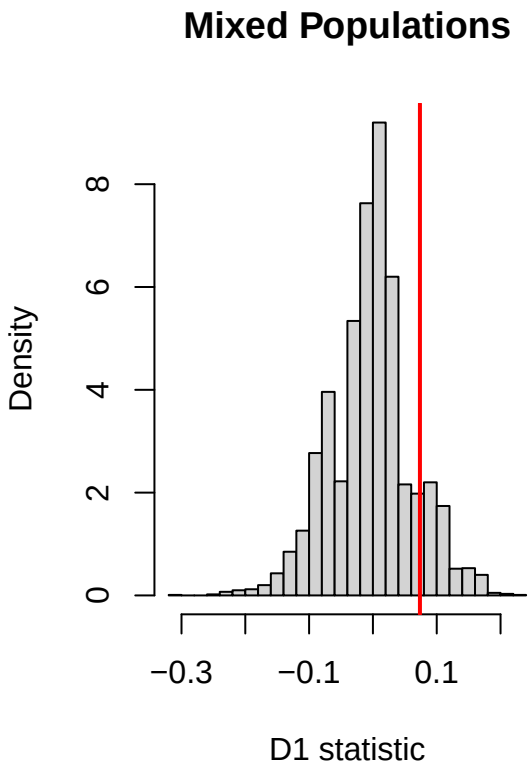
Draw histograms:

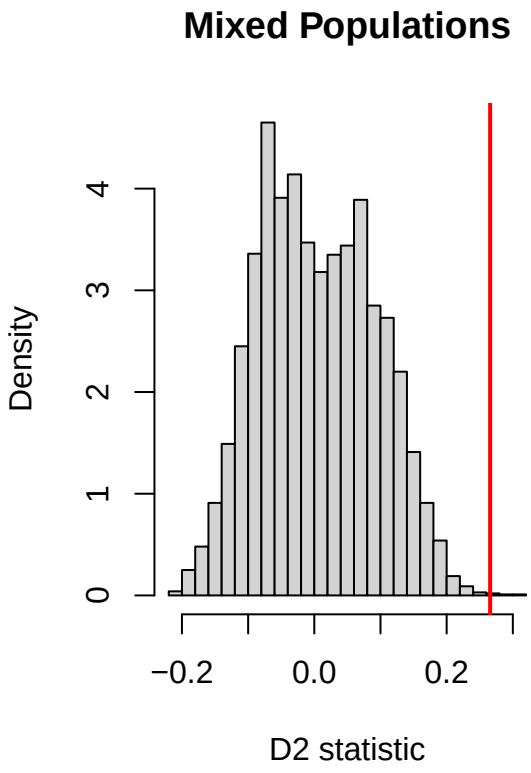
```
par(mfrow = c(1,2))

D1stat <- sapply(1:5000, FUN = function(...){D1(mixRandomly(car.pop))})
hist(D1stat, breaks=25, probability = TRUE,
     main = "Mixed Populations", xlab="D1 statistic")
```

```r
abline(v=D1(car.pop), col = "red", lwd=2)

D2stat <- sapply(1:5000, FUN = function(...){D2(mixRandomly(car.pop))})
hist(D2stat, breaks=25, probability = TRUE,
     main = "Mixed Populations", xlab="D2 statistic")
abline(v=D2(car.pop), col = "red", lwd=2)
```





Calculate P-values of two discrepancy measures:

```r
d1ob <- D1(car.pop)
d2ob <- D2(car.pop)
c(mean(abs(D1stat)>= abs(d1ob)), mean(abs(D2stat)>= abs(d2ob)))
```

```
## [1] 0.2836 0.0006
```

Conclusion:

We can see that from the first graph, the true value is close to the central tendency. From the second graph, the true value is far from the center and it is located at the right tail. Also, we calculated the P-values of D1 and D2. The D1 P-value is large and it is greater than 0.1, so there is no evidence against the null hypothesis. The D2 P-value is really small as it is smaller than 0.001, so it has a very strong evidence against the null hypothesis. Therefore, two sub-populations may be a random mix of the population when we consider measures of location, but it may not be a random mix of the population when we consider measures of spreads.