A4Q3

Summary

Sometimes, we divide the population into two sub-populations to check if there's a certain relation between two attributes of sub-populations. The test compares the difference of two attributes $a(\mathcal{P}_1)$ and $a(\mathcal{P}_2)$ with a randomly mixed sub-population:

1. We usually make an assumption before we compare the difference. We called this assumption **the null hypothesis** H_0. For H_0, we say that $\mathcal{P}_1$ and $\mathcal{P}_2$ are drawn randomly from the same population. However, by defining the null hyppthesis, we don't say two attribute are equal to each other.
2. We then calculate the **discrepancy measure** $D(\mathcal{P}_1, \mathcal{P}_2)$. The measure gives us a general idea that the data is consistent with the null hypothesis or not. There are two ways of measurement – location and spread:

$$\text{Location: } D(\mathcal{P}_1, \mathcal{P}_2) = |\bar{y_1} - \bar{y_2}|$$
$$\text{Spread: } D(\mathcal{P}_1, \mathcal{P}_2) = |\frac{SD(P_1)}{SD(P_2)} - 1|$$

3. We calculate the observed p-value. p-value is the probability that a randomly mixed sub-population has the discrepancy greater than or equal to the observed discrepancy we calculated above. If the p-value is really small, then we can say H_0 is true, otherwise we are against it.
4. By shuffling two sub-populations M times, we record every discrepancy measure and calculate the p-value:

$$p\text{-value} = Pr(D \geq d_{obs} | H_0 \text{ is true}) \approx \frac{1}{M} \sum_{i=1}^{M} I(D(\mathcal{P}_{1,i}^*, \mathcal{P}_{2,i}^*) \geq d_{obs})$$

Example

```r
cars <- read.csv('cars.csv',header=T, nrows = 100)
pop <- list(pop1 = cars[cars[, "transmission"] == 'mechanical', ],
            pop2 = cars[cars[, "transmission"] == 'automatic', ])
```

```r
diffAvePrice <- getAveDiffsFn("price_usd")
ratioSDPrice <- getSDRatioFn("price_usd")
round(c(diffAvePrice(pop), ratioSDPrice(pop)), 3)
```

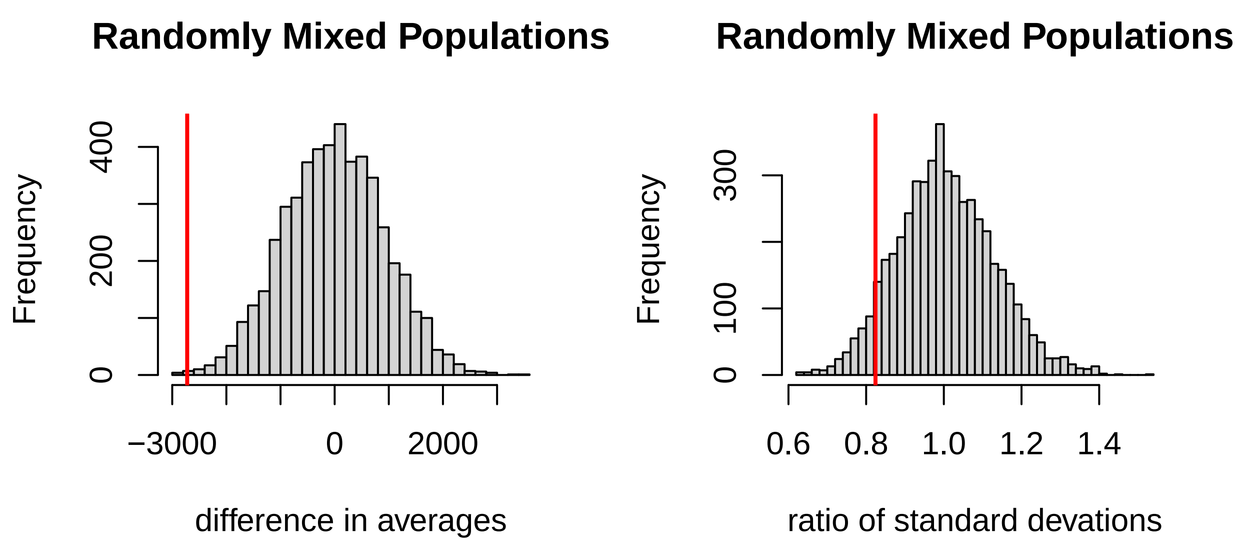
```
## [1] -2723.885     0.824
```

```r
set.seed(341)

diffPrice <- sapply(1:5000, FUN = function(...) {
  tmixedPop = mixRandomly(pop)
  c(diffAvePrice(tmixedPop), ratioSDPrice(tmixedPop))
})
```

```r
par(mfrow = c(1, 2))
hist(diffPrice[1, ], breaks = "FD", main = "Randomly Mixed Populations",
     xlab = "difference in averages", col = "lightgrey")
abline(v = diffAvePrice(pop), col = "red", lwd = 2)

hist(diffPrice[2, ], breaks = "FD", main = "Randomly Mixed Populations",
     xlab = "ratio of standard devations",
    col = "lightgrey")
abline(v = ratioSDPrice(pop), col = "red", lwd = 2)
```




```r
sum(abs(diffPrice) >= abs(diffAvePrice(pop))) / length(diffPrice)
```

```
## [1] 0.0015
```

Conclusion: Suppose that the pair ($\mathcal{P}_{mechanical}$,$\mathcal{P}_{automatic}$) is random draw: the probability of at least as large as the observed value is 0.0015. Therefore, there is a strong evidence against the null hypothesis that the pair ($\mathcal{P}_{mechanical}$,$\mathcal{P}_{automatic}$) was randomly drawn. We can see in the graph that the data seems to be symmetric around 0, and there are only a few data that have the absolute value greater the real value. It matches the p-value we calculated above.