Permutation Test

Boyie Chen

12/16/2019

The following example is from **Chapter 3 Introduction to Hypothesis Testing: Permutation Tests Section 3.3**

- Story：調查在某間 Bar 中，老主顧對 hot wings 和 beer 的消費量，同時紀錄性別
- Want to know： Now we only focus on the consumption of hot wings. Do males consume more hot wings than females?

$$H_0 : \text{Gender does not affect the consumption of hot wings}$$
$$H_a : \text{Men do consume more hot wings}$$

```r
Beerwings = read.csv("https://sites.google.com/site/chiharahesterberg/data2/Beerwings.csv")

tapply(Beerwings$Hotwings, Beerwings$Gender, mean)
```

```
##         F         M
##  9.333333 14.533333
```

```r
#equivalent to the following approach:
mean(subset(Beerwings$Hotwings, subset = Beerwings$Gender == 'F'))
```

```
## [1] 9.333333
```

```r
mean(subset(Beerwings$Hotwings, subset = Beerwings$Gender == 'M'))
```

```
## [1] 14.53333
```

```r
observed = 14.5333 - 9.3333 #store observed mean differences
```

Men consume 14.5333 hot wings on average; Women consume 9.3333 hot wings on average.

Are we able to conclude that there are difference in the two average consumptions between genders?

Permutation under Null Hypothesis

If the null hypothesis is TURE, then we are allowed to see all observations as in one group.

```r
#Get hotwings variable
hotwings = Beerwings$Hotwings

#Equivalent Approach
hotwings = subset(Beerwings, select = Hotwings, drop = TRUE)
#`drop = TRUE`` to convert hotwings to a vector (without this, hotwings will be a
#30x1 data frame
```

Then we can do the permutation, and see what will happen if we randomly assign them into different groups (Male & Female).

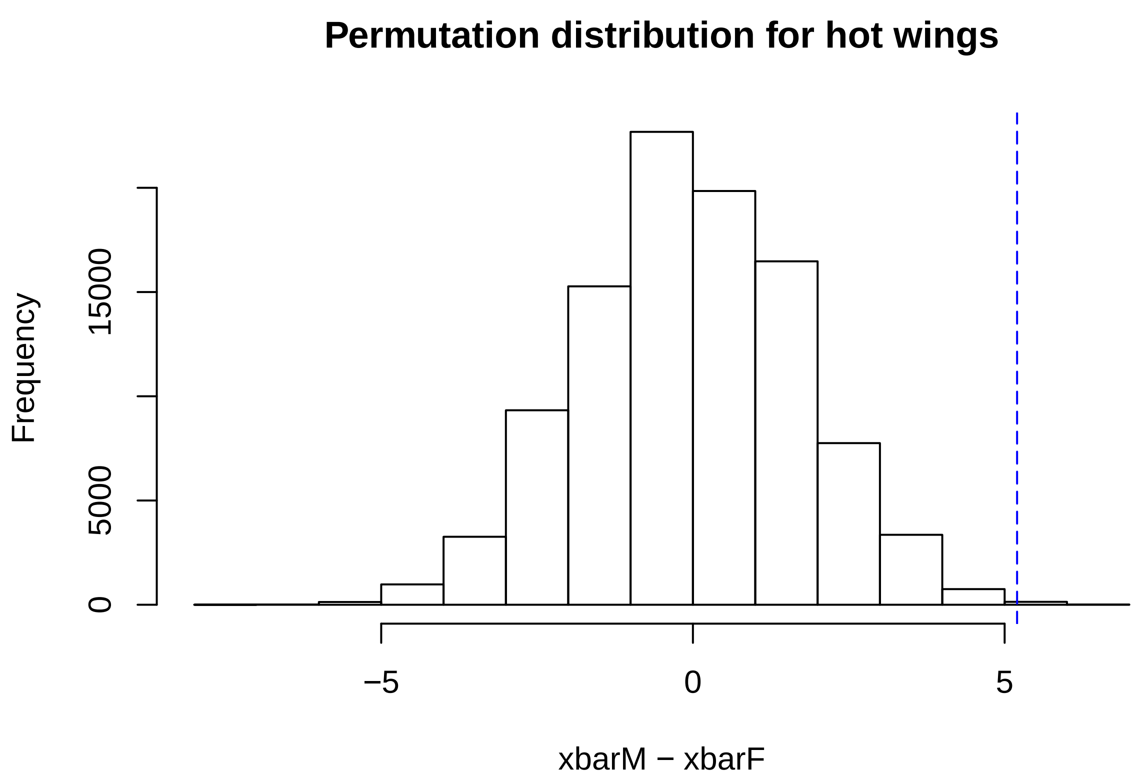
What we concern is that whether the mean differences are different among two groups. Thus our BS statisitc is the difference in sample means.

```r
#set.seed(0)
N = 10^5-1  #set number of times to repeat this process
result = numeric(N) # space to save the random differences
for(i in 1:N){
  index = sample(30, size=15, replace = FALSE) # sample of numbers from 1:30
  result[i] = mean(hotwings[index]) - mean(hotwings[-index])
}
```

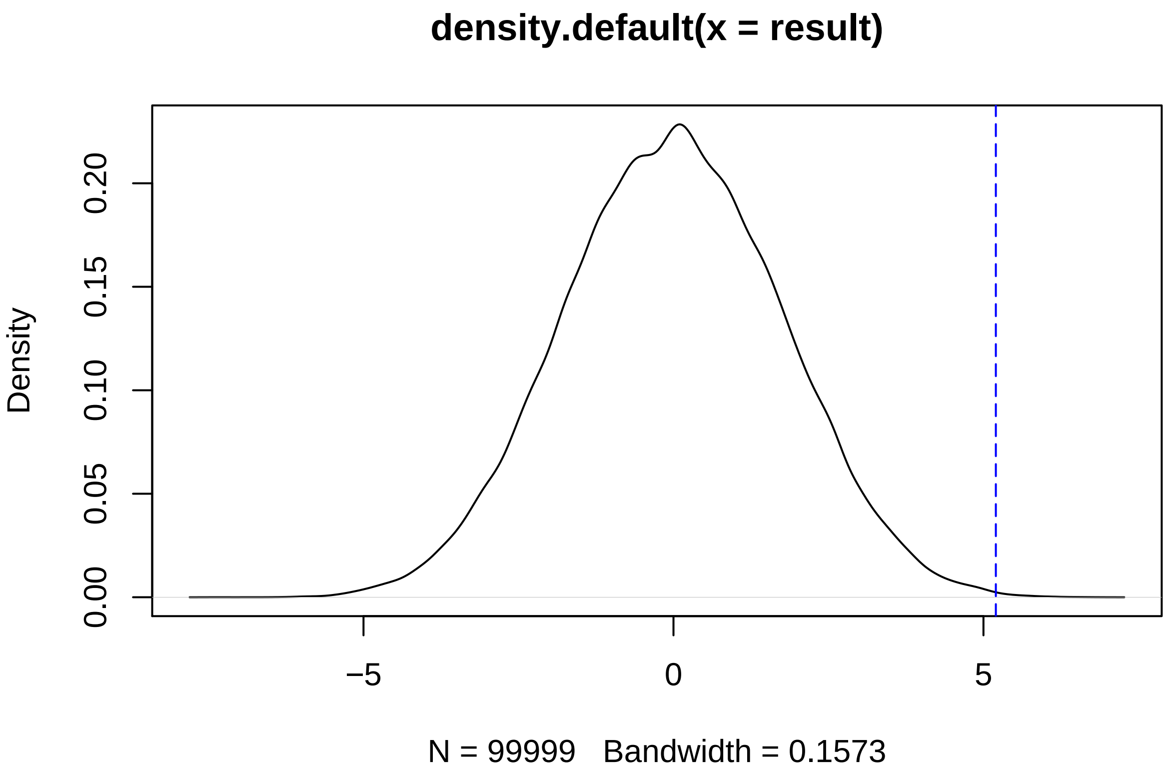
Note that `[-i]` means that we "skip" the index i

Distribution of BS Statistic under null hypothesis

```r
hist(result, xlab = "xbarM - xbarF", main = "Permutation distribution for hot wings")
abline(v = observed, col = "blue", lty=5)
```



```
#Alternative View
plot(density(result))
abline(v = observed, col = "blue", lty=5)
```




```
#Compute P-value
(sum(result >= observed)+1)/(N+ 1)  #P-value
```

```
## [1] 0.00078
```

Why we add 1?

Somtimes the p-value (prob. of extreme events) is very small, our resamples may not catch them. That would happen when the times of resampling is small.

For example:

```
set.seed(1234)
N = 100
result = numeric(N)
for(i in 1:N){
  index = sample(30, size=15, replace = FALSE) # sample of numbers from 1:30
  result[i] = mean(hotwings[index]) - mean(hotwings[-index])
}
sum(result >= observed)/(N)  #Actual relative frequency over permutations
```

```
## [1] 0
```

```r
(sum(result >= observed)+1)/(N+ 1)  #Adjusted P-value
```

```
## [1] 0.00990099
```

In this case, we do not observe any extreme events compared to the RS we observed. But it is doubtful to conclude that the p-value is 0.

Thus a conservative way to calculate p-value is add 1 to the numerator and to the denominator.