Permutation Test

Zhihan Fang

Import Happiness data and subset

```
#check data
summary(happiness.data)

#data clean and subset, either
happiness <- subset(happiness.data,happiness.data$AGE>0 & happiness.data$AGE<120)

#we want to compare the happiness level of two countries, Russia and China
two.country.happiness <-subset(happiness,happiness$COUNTRY=='Russia'|happiness$COUNTRY=='China')

#happiness value of Russia
russia.happiness <- happiness[happiness$COUNTRY=='Russia',6]
#happiness value of China
china.happiness <- happiness[happiness$COUNTRY=='China',6]
```

Plot distribution of happiness

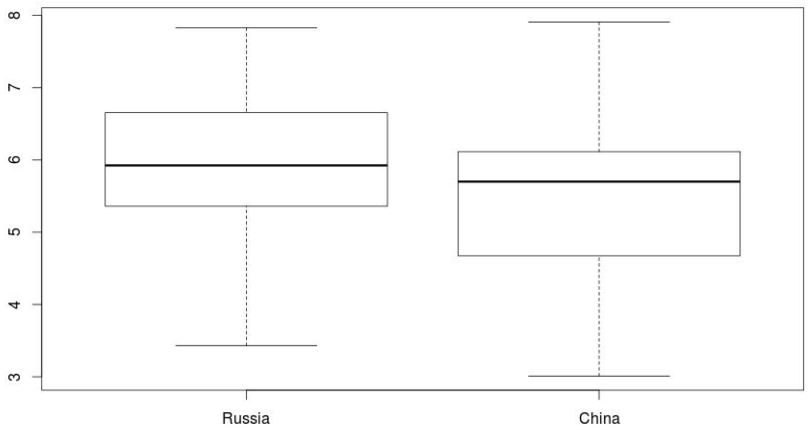
```
#happiness value of Russia
russia.happiness <- happiness
[happiness$COUNTRY=='Russia',6]

#happiness value of China
china.happiness <- happiness
[happiness$COUNTRY=='China',6]

#plot happiness distribution
boxplot(russia.happiness,china.happiness,names = c
('Russia','China'))

#mean differences
D <- mean(russia.happiness) - mean(china.
happiness)
```

Here, we got D = 0.4712674



Hypothesis and null Hypothesis

Are Russian happier than Chinese?

From the boxplot, it looks like the hypothesis is true, how much confidence do we on the hypothesis?

Hypothesis: Russian are happier than Chinese, in general.

In other words, the mean value D represents the real trends, the data is not randomly generated.

Are Russian happier than Chinese?

Or how much possibility are those data randomly generated?

Null hypothesis: Russian are not happier than Chinese.

In other words, the mean values D can not represents the reality if the nationality is randomly assigned to a group of people.

We need to disprove null hypothesis to defend our hypothesis.

Mix data

#number of people from russia
l_russia <- length(russia.happiness)
#number of people from china
l_china <- length(china.happiness)
l <- l_russia + l_china

There are 72 Russian and 64 Chinese people in the data set. If we mix them together, the total number of people is 136.

```
> l_russia <- length(russia.happiness)
> l_china <- length(china.happiness)
> l <- l_russia + l_china
> l_russia
[1] 72
> l_china
[1] 64
> l
[1] 136
```

Randomly selected nationality

#set null country

```
null_country <- rep("Russia",l)
null_country[sample(l,l_china)] <- 'China'
null <- data.frame(null_country,two.country.happiness[,6])
```

The above code randomly select 72 people as Russia and 64 people from China.That means now the nationality is randomly assigned to a group of people

.

```
> null_country <- rep("Russia",l)
> null_country[sample(l,l_china)] <- 'China'
> null <- data.frame(null_country,two.country.happiness[,6])
> null
   null_country two.country.happiness...6.
1         China                       6.48
2        Russia                       5.60
3         China                       6.55
4        Russia                       5.69
5         China                       4.44
6        Russia                       6.10
7        Russia                       5.82
8         China                       4.39
9        Russia                       4.16
10       Russia                       4.22
11        China                       7.22
12       Russia                       6.10
13        China                       7.08
14       Russia                       6.10
15        China                       5.36
16        China                       5.28
17       Russia                       7.24
18       Russia                       5.38
19        China                       5.31
```

difference of random data

```
#get random generated happiness of each country
russia_null <-null[null$null_country=='Russia',2]
china_null <-null[null$null_country=='China',2]
#the difference of mean value
D_null <- mean(russia_null) - mean(china_null)
```

What is the difference of happiness if the nationality is randomly assigned to a group of people?

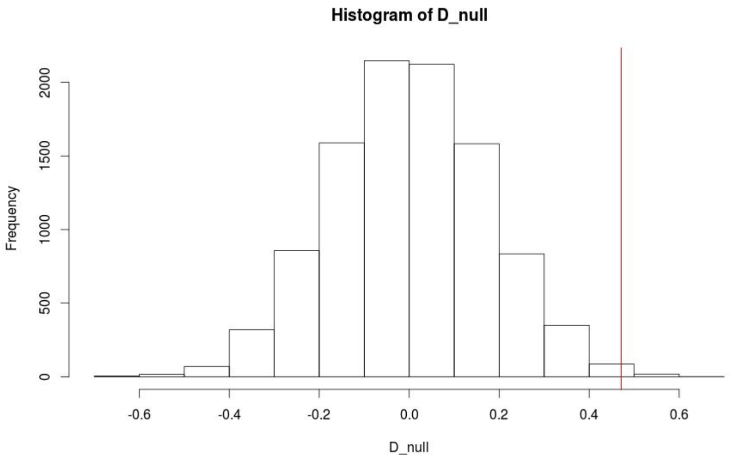
```
> russia_null <-null[null$null_country=='Russia',2]
> china_null <-null[null$null_country=='China',2]
> D_null <- mean(russia_null) - mean(china_null)
> D_null
[1] -0.2429687
```

Permutation Test

```
#Do the same test 10000
for(i in 1:10000){
   null_country <- rep("Russia",l)
   null_country[sample(l,l_china)] <- 'China'
   null <- data.frame(null_country,two.country.happiness[,6])
   russia_null <-null[null$null_country=='Russia',2]
   china_null <-null[null$null_country=='China',2]
   D_null[i] <- mean(russia_null) - mean(china_null) }

#plot the distribution of  the mean value of random cases
hist(D_null)
#add a line where mean value = D
abline(v=D,col='red')
```

We do the same thing 10000 times, every time we randomly assign nationality to the same group of people? Then check how much possibility we can get the the mean value or higher compared with original data.



P value

What's the possibility to get a mean value that is no less than D based on random data?

The possibility is 0.0028 = 0.28%. Therefore p =0.28%

p is much smaller than 5%.

So we can reject the null hypothesis and successfully defend the original hypothesis.

Hypothesis: Russian are happier than Chinese, in general.

```
> p <- length(D_null[D_null>=D])/length(D_null)
> p
[1] 0.0028
```