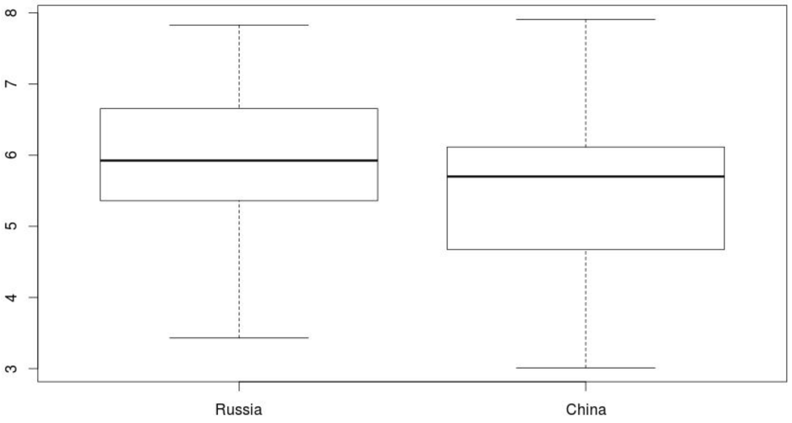
Z-TEST

Zhihan Fang

Permutation Test

Last week, we rejected our null hypothesis and supported our alternative hypothesis by permutation test.

Permutation test: estimate possibility that getting same difference from random data by trials.

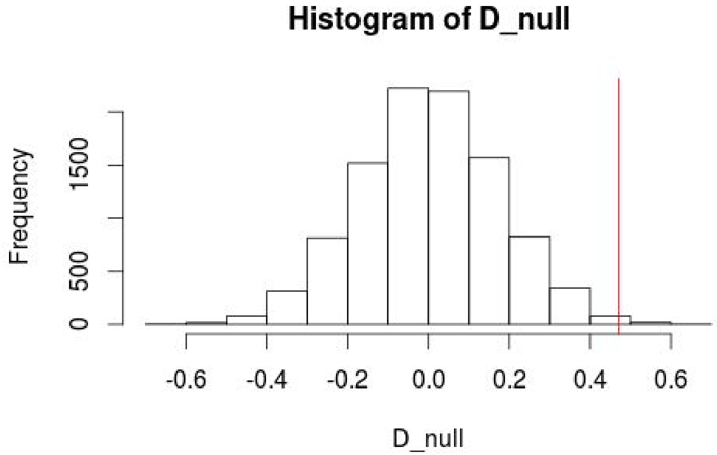


Permutation Test

number of trial : n = 10000, p = 0.37%

number of trial : n = 100000, p = 0.35%

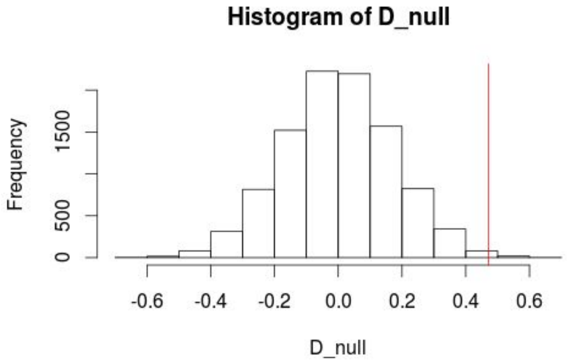
When n goes to infinity?



Permutation Test

Weakness of Permutation Test:

1. based on trials, it is time consuming (run 10000), even unsolvable sometimes.
2. It can not make sure you can get the accurate p value when the number trials is small.



Z Test

According to central limit theorem, the distribution of D_null is a normal distribution when n goes infinity.

We want to use inference and math to calculate the normal distribution formula and p value.

Null Hypothesis:

There is no happiness difference between China and Russia.

Alternative Hypothesis:

Russian are happier than Chinese.

population difference mean

Null Hypothesis:

There is no happiness difference between China and Russia.

$$\mu_{\bar{x}_R} = \mu_{\bar{x}_C}$$

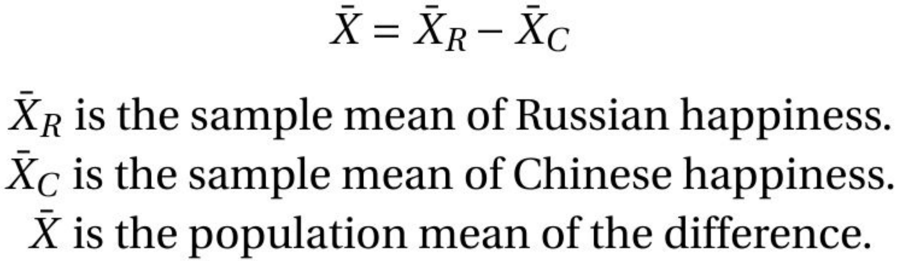
$$\Rightarrow \mu = \mu_{\bar{x}_R} - \mu_{\bar{x}_C} = 0$$

$\mu_{\bar{x}_R}$ is the population mean of Russian happiness.

$\mu_{\bar{x}_C}$ is the population mean of Chinese happiness.

$\mu_{\bar{x}}$ is the population mean of the difference.

sample mean difference



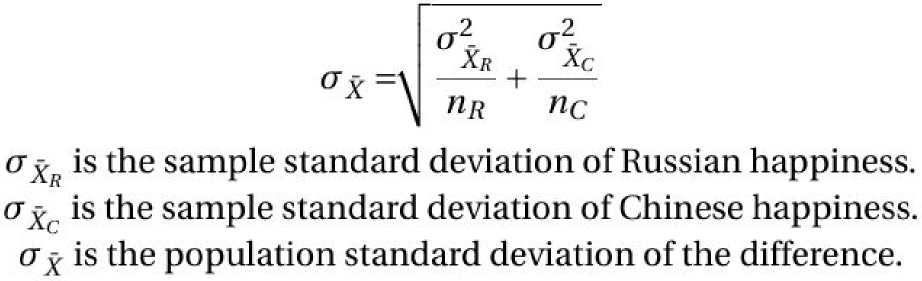
$$\bar{X} = \bar{X}_R - \bar{X}_C$$

$\bar{X}_R$ is the sample mean of Russian happiness.
$\bar{X}_C$ is the sample mean of Chinese happiness.
$\bar{X}$ is the population mean of the difference.

The mean difference distribution is the normal distribution.

population standard deviation

We can use sample standard deviation to estimate the population standard deviation.



$$\sigma_{\bar{X}} = \sqrt{\frac{\sigma^2_{\bar{X}_R}}{n_R} + \frac{\sigma^2_{\bar{X}_C}}{n_C}}$$

$\sigma_{\bar{X}_R}$ is the sample standard deviation of Russian happiness.

$\sigma_{\bar{X}_C}$ is the sample standard deviation of Chinese happiness.

$\sigma_{\bar{X}}$ is the population standard deviation of the difference.

normal distribution curve

The buell curve can be determined when the population mean and standard deviation are both known. z score is how many standard deviation to the center of curve.

$$z = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} = \frac{\bar{X}_R - \bar{X}_C}{\sqrt{\frac{\sigma^2_{\bar{X}_R}}{n_R} + \frac{\sigma^2_{\bar{X}_C}}{n_C}}}$$

Code Import your data

```
#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]
```

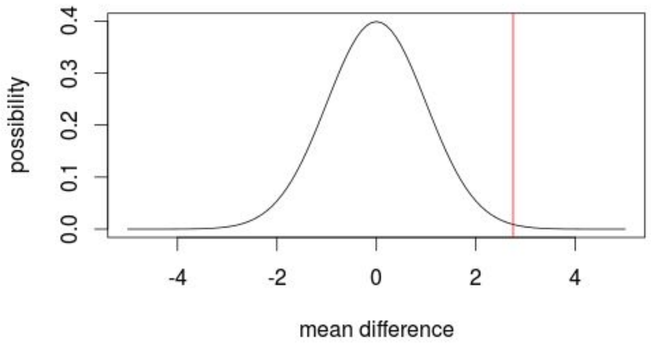
sample standard deviation

```
# standard deviation of two samples
sd_china <- sd(china.happiness)
sd_russia <- sd(russia.happiness)

#length of china and russia
l_china <- length(china.happiness)
l_russia <- length(russia.happiness)

#standard deviation of difference population
sd_russia_china <- sqrt(sd_china^2/l_china+sd_russia^2/l_russia)

#z score
zeta <- (mean(russia.happiness)-mean(china.happiness))/sd_russia_china
zeta

zeta = 2.750132
```

p value and plot

```
#plot red line
plot(x=seq(from = -5, to= 5, by=0.1),y=dnorm(seq(from = -5, to= 5,
by=0.1),mean=0),type='l',xlab = 'mean difference',
ylab='possibility')
abline(v=zeta, col='red')

#get p
p = 1-pnorm(zeta)
p
```

p = 0.298%


possibility
mean difference