

Sample means of Non-normal Distribution Data Follow Normal Distribution

មធ្យមរបស់ សំណាកទិន្នន័យ Non-normal Distribution គោរព Normal Distribution

ដោយ ជឹម សុជាតិ មន្ត្រីវិទ្យាស្ថានសិក្សាចិន
រាជបណ្ឌិត្យសភាកម្ពុជា

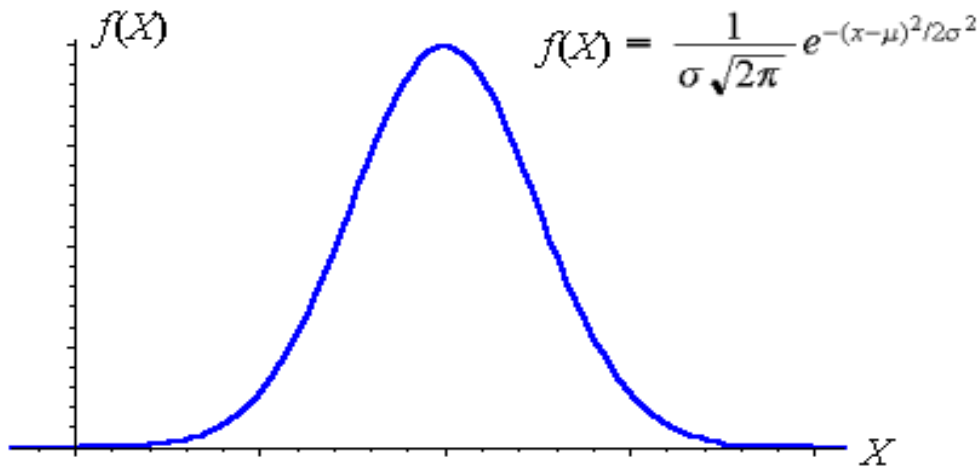
អ្នកសិក្សាស្រាវជ្រាវពីស្ថិតិវិទ្យា តែងតែដឹងច្បាស់ពី ការសន្មត ដែលទិន្នន័យចុះស្រាវជ្រាវជាទិន្នន័យដែលគោរពតាម Normal Distribution។ កាលណាគេចង់ដឹងពីមធ្យម mean របស់ Population ១ អ្នកសិក្សាស្រាវជ្រាវបានយកគម្រូសំណាក (sample) ដែលជាចំនួនគម្រូសំណាក ៣០ ឬត្រឹមតែ ១០០ ទិន្នន័យ ដោយអ្នកសិក្សាស្រាវជ្រាវមិនមានលទ្ធភាពរាប់ចំនួនទាំងអស់របស់ Population ដែលមានរាប់ពាន់ ឬរាប់ម៉ឺនទិន្នន័យ។ ក្នុងការគណនា (Estimate) មធ្យមរបស់ Population យើងយកមធ្យម របស់សំណាក ដែលបានមកពី ចំនួនបូកសរុបរបស់ទិន្នន័យចែកនឹងចំនួនទិន្នន័យក្នុងសំណាក។ អ្នកសិក្សាស្រាវជ្រាវតែងគិតថា ទិន្នន័យគោរពតាម Normal Distribution ប៉ុន្តែក្នុងអត្ថបទនេះ យើងសិក្សាថា៖ មធ្យម (means) របស់ សំណាកទិន្នន័យ Non-normal Distribution គោរព Normal Distribution ។ វិធីសាស្ត្រដែលប្រើក្នុងការសិក្សា គឺប្រើ កម្មវិធីស្ថិតិ R ដោយប្រើក្នុង replicate ដើម្បីបង្កើត samples n=30, 20 ទៅលើ Distribution ៥ មាន Normal Distribution, Chi-square Distribution, Poisson Distribution, Uniform Distribution និង F Distribution ។ បន្ទាប់មក យើងសរសេរក្នុងគណនា មធ្យមរបស់ sample ទាំង ៣០, ២០ គូស Histogram តេស្ត សមត្ថិកម្ម ទិន្នន័យមធ្យមសំណាកទាំង ៣០, ២០។

យើងសិក្សា Distribution ដូចខាងក្រោម៖

Normal Distribution

Normal Distribution ជា Distribution ដែលមានរាងស៊ីមេទ្រី ដែលគេបានប្រើប្រើក្នុងស្ថិតិ។ អនុគមន៍ Normal Distribution មានដូចខាងក្រោម៖

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



ការសិក្សារបស់យើង បានសរសេរកូដក្នុងកម្មវិធី R `normal_data<-rnorm(30, mean=10, sd=6)` និងយើងបានទិន្នន័យចំនួន ៣០ ដូចខាងក្រោម៖

```
> normal_data
[1]  7.14744548  5.06694536  5.59318319 -8.20316774 16.08530681  5.47071910
[7]  8.83434776 11.66135015  0.56062666  3.67374808  7.28122449  2.17312991
[13]  9.69682770 20.65372651 -4.14367406  8.99839810 19.20653741  6.24343143
[19] 18.02224754 10.44281091  2.72553977  2.89194541 19.59883333  0.02527424
[25]  6.80956709 10.92490367 16.86132532  4.40893221  7.28612081 17.24575552
```

```
> normal_sample30_means<-replicate(20, mean(rnorm(30,mean=10, sd=6)))
> normal_sample30_means
[1]  8.400973  9.047888 11.533097  9.771983  8.552117  8.786912  8.115315
[8]  9.086434  8.789313  9.367800 10.178594  9.954003 13.286603 10.333165
[15]  9.272985 10.426682 10.590380  9.359398  9.401357  8.090990
```

បន្ទាប់មកយើង តេស្តសម្មតិកម្ម H_0 : ទិន្នន័យគោរពតាម Normal Distribution លើទិន្នន័យទាំង ៣០ ដែលយើងបាន បង្កើត (Generate) ដែលជាទិន្នន័យគោរពតាម Normal Distribution ដោយយើងសរសេរកូដ កំណត់ Distribution ឲ្យទិន្នន័យ។

```
> shapiro.test(normal_data)

      Shapiro-Wilk normality test

data:  normal_data
W = 0.96549, p-value = 0.4241

      Shapiro-Wilk normality test

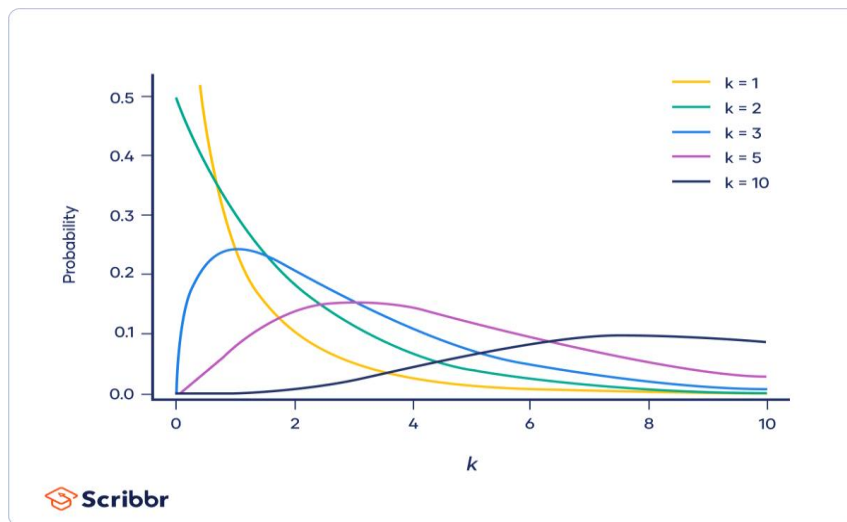
data:  normal_sample30_means
W = 0.92484, p-value = 0.1228
```

លទ្ធផល របស់ Shapiro Test $p_value=0.4241$ ដូចនេះតេស្តសម្មតិកម្ម H_0 : ទិន្នន័យគោរពតាម Normal Distribution មិនត្រូវបានបដិសេធ ដូចនេះទិន្នន័យទាំងនេះគោរពតាម Normal Distribution ($p_value<0.05$ reject null hypothesis)។ បន្ទាប់មកទៀត យើងសរសេរកូដ `normal_sample30_means<-replicate(20, mean(rnorm(30,mean=10, sd=6)))` ដើម្បីបង្កើតមធ្យម Mean Samples ចំនួន ២០ និងយើងតេស្តសម្មតិកម្ម H_0 : ទិន្នន័យគោរពតាម Normal Distribution ដោយ $p_value=0.1228$ យើងឃើញថា មធ្យម Mean Samples ចំនួន ២០ គោរពតាម Normal Distribution ប៉ុន្តែភាព Normal ថយចុះ។

Chi-square Distribution

Chi-square Distribution ត្រូវបានគេប្រើដើម្បីសិក្សាពីទិន្នន័យជាក់ស្តែង និងទិន្នន័យ
តាមទ្រឹស្តី។ Chi-square Distribution មានអនុគមន៍ដូចខាងក្រោម៖

$$f(x; k) = \frac{x^{\frac{k}{2}-1} e^{-\frac{x}{2}}}{2^{\frac{k}{2}} \Gamma(\frac{k}{2})}$$



ការសិក្សារបស់យើងបានសរសេរក្នុងកម្មវិធី R `chisquare_data<-rchisq(60, df=20)`
ដើម្បីបង្កើតទិន្នន័យ គោរពតាម Chi-square Distribution និងយើងតេស្ត Normality របស់
ទិន្នន័យទាំងនេះ លទ្ធផលរបស់ Shapiro Test `p_value=0.09306` ដូចនេះតេស្ត
សម្មតិកម្ម H_0 : ទិន្នន័យគោរពតាម Normal Distribution អាចត្រូវបានបដិសេធ ដូចនេះ
ទិន្នន័យទាំងនេះ មិនគោរពតាម Normal Distribution។ បន្ទាប់មកទៀត យើងសរសេរកូដ
`chisquare_sample20_means<-replicate(20, mean(rchisq(60, df=30)))` ដើម្បីបង្កើត មធ្យម
Mean Samples ចំនួន ២០ និងយើងតេស្តសម្មតិកម្ម H_0 : ទិន្នន័យគោរពតាម Normal
Distribution ដោយ `p_value=0.5657` យើងឃើញថា មធ្យម Mean Samples ចំនួន ២០ គោរព
តាម Normal Distribution។

```
> chisquare_data<-rchisq(60, df=20)
> hist(chisquare_data)
> shapiro.test(chisquare_data)
```

Shapiro-Wilk normality test

```
data:  chisquare_data
W = 0.96603, p-value = 0.09306
```

Shapiro-Wilk normality test

```
data:  chisquare_sample20_means
W = 0.96108, p-value = 0.5657
```

Uniform Distribution

Uniform Distribution ជា Distribution ដែលកើតតាមតម្លៃ Min ទាបបំផុត និង Max ខ្ពស់បំផុត។ ការសិក្សារបស់យើងបានសរសេរក្នុងក្នុងកម្មវិធី R `uniform_data<-runif(20, min=30, max=80)` ដើម្បីបង្កើតទិន្នន័យ គោរពតាម Uniform Distribution និងយើងតេស្ត Normality របស់ទិន្នន័យទាំងនេះ លទ្ធផល របស់ Shapiro Test $p_value=0.03511$ ដូចនេះតេស្តសម្មតិកម្ម H_0 : ទិន្នន័យគោរពតាម Normal Distribution ត្រូវបានបដិសេធ ដូចនេះទិន្នន័យទាំងនេះមិនគោរពតាម Normal Distribution។ បន្ទាប់មកទៀត យើងសរសេរកូដ `uniform_sample20_means<-replicate(20, mean(runif(min=60, max=80)))` ដើម្បីបង្កើតមធ្យម Mean Samples ចំនួន ២០ និងយើងតេស្តសម្មតិកម្ម H_0 : ទិន្នន័យគោរពតាម Normal Distribution ដោយ $p_value=0.692$ យើងឃើញថា មធ្យម Mean Samples ចំនួន ២០ គោរពតាម Normal Distribution។

```

> uniform_data<-runif(20, min=30, max=80)
> hist(uniform_data)
> shapiro.test(uniform_data)

      Shapiro-Wilk normality test

data:  uniform_data
W = 0.89626, p-value = 0.03511

> shapiro.test(uniform_sample20_means)

      Shapiro-Wilk normality test

data:  uniform_sample20_means
W = 0.96706, p-value = 0.692

```

Poisson Distribution

Poisson Distribution មានអនុគមន៍៖

$$f(x; k) = \Pr(x = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

ការសិក្សារបស់យើងបានសរសេរក្នុងក្នុងកម្មវិធី R `poisson_data<-rpois(20, lamda=6)` ដើម្បីបង្កើតទិន្នន័យ គោរពតាម Uniform Distribution និងយើងតេស្ត Normality របស់ទិន្នន័យទាំងនេះ លទ្ធផល របស់ Shapiro Test `p_value=0.2063` ដូចនេះតេស្តសម្មតិកម្ម H_0 : ទិន្នន័យគោរពតាម Normal Distribution អាចមិនត្រូវបានបដិសេធ ដូចនេះទិន្នន័យទាំងនេះគោរពតាម Normal Distribution។ បន្ទាប់មកទៀត យើងសរសេរកូដ `poisson_sample20_means<-replicate(20, mean(rpois(20, lamda=6)))` ដើម្បីបង្កើត មធ្យម Mean Samples ចំនួន ២០ និងយើងតេស្តសម្មតិកម្ម H_0 : ទិន្នន័យគោរពតាម Normal Distribution ដោយ `p_value=0.8231` យើងឃើញថា មធ្យម Mean Samples ចំនួន ២០ គោរពតាម Normal Distribution។

```

> poisson_data<-rpois(20,lambda=6)
> poisson_data
[1] 11 7 8 11 6 5 3 8 1 5 8 3 5 11 2 3 8 9 2 4
> hist(poisson_data)
> shapiro.test(poisson_data)

      Shapiro-Wilk normality test

data:  poisson_data
W = 0.93656, p-value = 0.2063

> poisson_sample20_means<-replicate(20,mean(rpois(20,lambda=6)))
> hist(poisson_sample20_means)
> shapiro.test( poisson_sample20_means)

      Shapiro-Wilk normality test

data:  poisson_sample20_means
W = 0.97333, p-value = 0.8231

```

F Distribution

ការសិក្សារបស់យើងបានសរសេរក្នុងកម្មវិធី R $F_data \leftarrow rf(20, df1=10, df2=16)$ ដើម្បីបង្កើតទិន្នន័យ គោរពតាម F Distribution និងយើងតេស្ត Normality របស់ទិន្នន័យទាំងនេះ លទ្ធផល របស់ Shapiro Test $p_value=0.286$ ដូចនេះតេស្តសម្មតិកម្ម H_0 : ទិន្នន័យគោរពតាម Normal Distribution មិនត្រូវបានបដិសេធ ដូចនេះទិន្នន័យទាំងនេះគោរពតាម Normal Distribution។ បន្ទាប់មកទៀត យើងសរសេរកូដ $F_sample20_means \leftarrow replicate(20, mean(rf(20, df1=10, df2=16)))$ ដើម្បីបង្កើត មធ្យម Mean Samples ចំនួន ២០ និងយើងតេស្តសម្មតិកម្ម H_0 : ទិន្នន័យគោរពតាម Normal Distribution ដោយ $p_value=0.1378$ យើងឃើញថា មធ្យម Mean Samples ចំនួន ២០ គោរពតាម Normal Distribution តែភាព Normal ថយចុះ។

```
> F_data<-rf(20, df1=10, df2=16)
> hist(F_data)
> shapiro.test(F_data)
```

Shapiro-Wilk normality test

```
data: F_data
W = 0.94408, p-value = 0.286
```

Shapiro-Wilk normality test

```
data: F_sample20_means
W = 0.92744, p-value = 0.1378
```

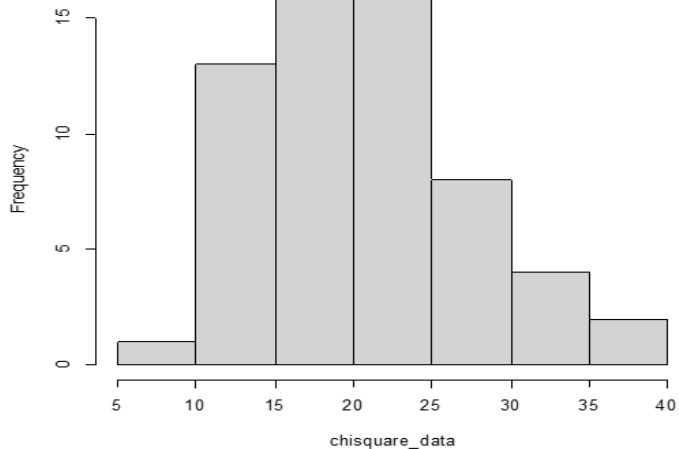
ជាសន្និដ្ឋាន៖ នៅក្នុងការសិក្សារបស់យើង មធ្យម means របស់ សំណាកទិន្នន័យ Non-normal Distribution គោរព Normal Distribution មានតម្លៃ P value ធំ លើកលែងតែទិន្នន័យ F Distribution។ មធ្យម (means) របស់ទិន្នន័យ ពី Normal Distribution មានការធ្លាក់ចុះ ភាព Normal ដោយសារការចាប់គម្រូសំណាក (Sampling) ធ្វើឱ្យខូចទម្រង់ Distribution។ បញ្ហាសម្រាប់ការចាប់គម្រូសំណាក (Sampling) ដោយកូដកម្មវិធីស្ថិតិ R មិនចាប់ ទិន្នន័យក្នុង Population រួម។

ឯកសារយោង៖

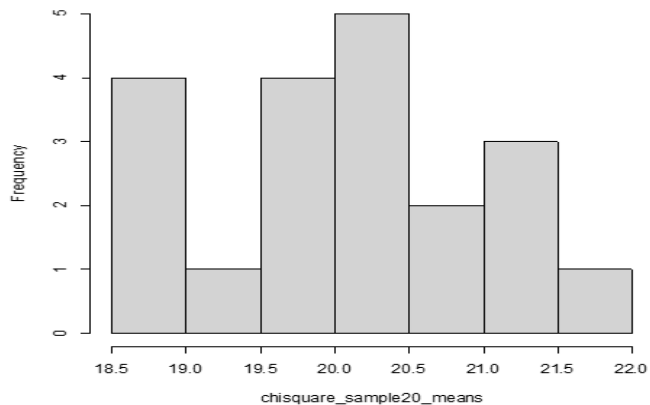
១. R version 4.5.0 (2025-04-11 ucrt)
២. Brad Duthie (២០២២) Creating simulated data sets in R
៣. <https://www.scribbr.com/statistics/chi-square-distributions/>
៤. <https://www.geeksforgeeks.org/poisson-distribution/>
៥. <https://alevelmaths.co.uk/statistics/the-normal-distribution/>

ឧបសម្ព័ន្ធ៖ Histogram របស់ទិន្នន័យវិភាគក្នុងការសិក្សា

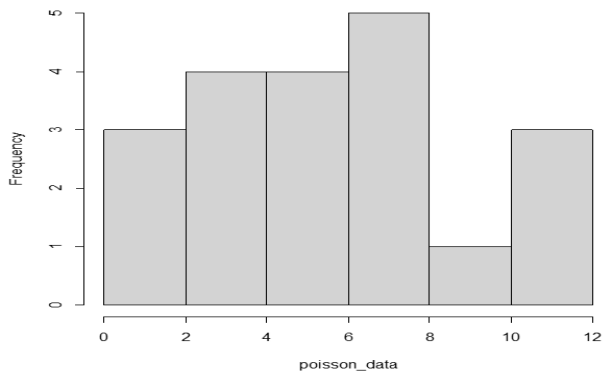
Histogram of chisquare_data



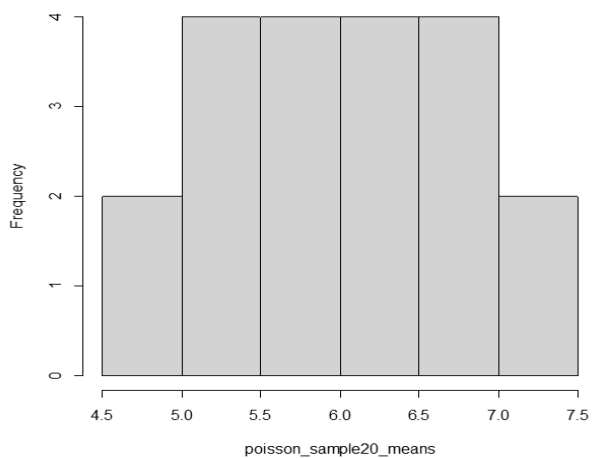
Histogram of chisquare_sample20_means



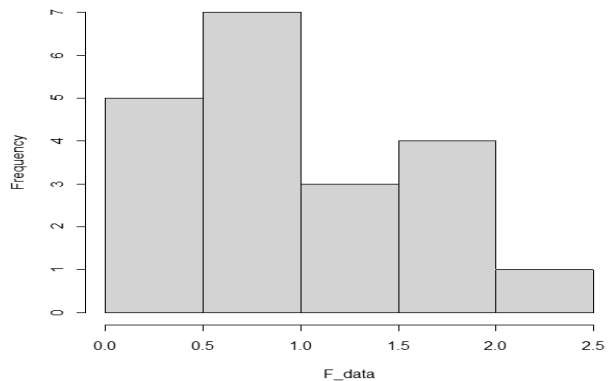
Histogram of poisson_data



Histogram of poisson_sample20_means



Histogram of F_data



Histogram of F_sample20_means

