

I. S-plus Environment

S-plus ត្រូវបានដំណើរការនៅក្នុង window មួយប្រភេទ អោយឈ្មោះថា
Command Window S-plus ដែលមានទំរង់៖

Working data will be in/Project1

>.....

S-plus មាន Directory ពិសេសមួយ អោយឈ្មោះថា "Project1"។

"Project1" ផ្ទុកនូវ objects ជាច្រើនដូចជា៖ Data file , Graphics, Function,.....។

រាល់ objects ទាំងអស់ដែលបានបង្កើតនៅក្នុង S-plus ត្រូវបានរក្សាដោយ
ស្វ័យប្រវត្តិនៅក្នុង Directory នេះ។

1. Command ls() :

ប្រើដើម្បី list នូវ objects ទាំងអស់ដែលមាននៅក្នុង Directory "Project1"

Objects ដែលមាននៅក្នុង Directory នេះត្រូវបានបែងចែកជាពីរ៖

- Objects ដែលបានបង្កើតឡើងដោយ S-plus
- Objects ដែលបានបង្កើតឡើងដោយអ្នកប្រើប្រាស់

2. Copy object:

បើ Mat ជា object ដែលមានស្រាប់នៅក្នុង Directory នោះការចំលង object នេះ

ត្រូវបានសរសេរ៖

>Mat1<- Mat

Mat1 មាន Content ដូច Mat

3. Remove Object:

rm (object1, object2,....., objectn) ប្រើសំរាប់លុប object នីមួយៗពី

Directory ។

ចំណាំ៖ គេអាចលុប objects ទាំងអស់ព្រមពេលជាមួយគ្នាដោយប្រើ Command :

remove(ls())

ដូចគ្នាដែរ គេអាចហៅ objects ដែលលុបហើយមកវិញ តាមរយៈ ៖

Click Edit -----> Restore Data Object-----> Select Objects ដែលត្រូវការ---> Ok

4. ការបង្កើត Editor:

>options(editor="notepad")

> fix(ឈ្មោះ) ហៅ editor "notepad" មកសរសេរ Program

Ex: > ?matrix សួរទៅលើការប្រើប្រាស់ Command matrix ។

ចំណាំ៖ ដើម្បី Run function , គេសរសេរ៖

>ឈ្មោះ() Enter

II.Manipulation of Data:

១. ការបង្កើត vector:

Vector ជាបណ្តុំនៃធាតុអាចជា Numeric or Non-Numeric ដែលមានទំរង់ជា array មួយវិមាត្រ ($n \times 1$) or ($1 \times n$) ហើយដែលជាទូទៅនៅក្នុងតារាងទិន្នន័យតំណាងដោយជួរឈរមួយ។ គេអាចបង្កើត vector តាមច្រើនរូបភាព៖

- Vector ចំនួនគត់ ធម្មតា, a : b ; c()
- Vector string c("a", "b"), letters; LETTERS
- Vector ដែលជាលេខគូ ឬ សេស; seq()

ទំរង់ទី១៖ seq(from, to, by)

ប្រើសំរាប់បង្កើតតំលៃលេខគត់ ឬ ធម្មតា ដែលមានគំលាតពីមួយទៅមួយស្មើគ្នា កំណត់ដោយ argument " by"

ទំរង់ទី២៖ seq(from, to, length)

ប្រើសំរាប់បង្កើតតំលៃលេខ ធម្មតាដែលមានគំលាតពីមួយទៅមួយស្មើគ្នា ហើយចំនួនតំលៃលេខដែលត្រូវបង្កើតនេះត្រូវបានកំណត់ដោយ argument " length " ។

២. ការអានធាតុនៃ vector:

_____ជាទូទៅគេអាចអានធាតុនៃ vector តាមទំរង់ Syntax ដូចខាងក្រោម៖

X[]

X ជា object vector

ចន្លោះនៅក្នុងដង្កៀប ត្រូវបញ្ជាក់ទីតាំងធាតុនៃ vector
 Ex: បើ X ជា object vector ដែលមាន ទំរង់ ៖

> X<- c(23, 45 ,15 ,60, 75,56,80)

1. ចូរអានយក ៤ ធាតុដំបូងនៃ Vector X?

ចំណាំ៖ length() ជាcommand ដែលប្រើសំរាប់រាប់ធាតុនៃ vector

>length(X)

2. ចូរអានយកធាតុទាំងឡាយណាដែលស្ថិតនៅលើទីតាំង គូនៃ vector X ?

3. ចូរអានយកធាតុ X ទាំងឡាយណាដែលតូចជាង 60 ?

4. ចូរអានយកធាតុ Xទាំងឡាយណាដែលស្ថិតនៅចន្លោះ [45, 75]?

X<=75 ; A={ 23,45,15,60,75,56}

X>=45 ; B={45, 60, 75, 56, 80}

A *.B ={45, 60, 75, 56}

5. បើ ធាតុ X មួយត្រូវបានជ្រើសរើសដោយចៃដន្យ ចូរកំណត់ ប្រូបាប៊ីលីតេ ដើម្បី
 អោយ X ជាធាតុដែលធំជាង 60?

$\Omega=\{23,45,15,60,75,56,80\}$ សំណុំនៃធាតុករណីអាច

$A=\{ 75,80\}$ សំណុំនៃធាតុ ករណីស្រប

$$P(A)=\frac{Card(A)}{Card(\Omega)} = 2/7 =0.2857=28.57\%$$

ចំណាំ៖ ជាទូទៅនៅក្នុងជីវភាពរស់នៅប្រចាំថ្ងៃ គេតែងជួបប្រទះនូវទិន្នន័យដែល

បាត់បង់ (Missing data)។ទិន្នន័យដែលបាត់បង់ (Missing data) នេះ នៅក្នុង
 S-plus តាងដោយតំលៃ NA (Non-Available)។ តំលៃនេះបង្កបញ្ហា ដ៏ស្មុកស្មាញ
 ក្នុងការវាយតំលៃជាប្រូបាប៊ីលីតេ។ដូច្នេះដើម្បីដំណើរការជាមួយតំលៃ missing
 នេះ គេប្រើ Command ៖

is.na ()

សំរាប់ទាញតំលៃ missing (NA) ចេញពី vector។

Ex: គេមាន vector Y មួយដែលមានទំរង់ដូចខាងក្រោម៖

`>Y<-c(45,60,35,NA,55,50,NA,75)`

1. ចូរអានយកតំលៃ NA ពី vector នេះ?
2. តើនៅក្នុង vector Y មាន NA ចំនួនប៉ុន្មាន?
3. ចូរអានយកធាតុ vector Y ទាំងឡាយណាដែលគ្មាន NA?
4. បើធាតុ vector Y មួយត្រូវបានជ្រើសរើសដោយចៃដន្យ កំណត់ប្រូបាប៊ីលីតេ ដើម្បីអោយវាជាធាតុដែលតូចជាង 55?

$A=\{45,60,35,55,50,75\}$ $B=\{45,35,NA,50,NA\}$; $A * B= E =\{45,35,50\}$;

$$P(E)=\frac{Card(E)}{Card(A)}=3/6=1/2=0.5$$

៣. Data Frame:

DataFrame គឺជាតារាងទិន្នន័យដែលមានទំរង់ជា array ពីរវិមាត្រ (mxn) ដែលយើងអាចអានវាជា ទំរង់ Frame តែនៅក្នុង object Explorer នៃ S-plus តែប៉ុណ្ណោះ។

Ex: គេមានតារាងទិន្នន័យ "market.survey" នៅក្នុង DataSet មួយនៃ object explorer នៃ S-plus។

`>market.survey<-market.survey` ចំលង តារាងទិន្នន័យ "market.survey" ពី DataSet មួយ នៃ object explorer នៃ S-plus មក Directory "Project1"។

នៅក្នុង object explorer មាន DataSets ជាច្រើនដែលផ្ទុកទៅដោយតារាងទិន្នន័យ ក្នុងនោះមាន DataSet តែមួយប៉ុណ្ណោះដែលមានទំនាក់ទំនងដោយផ្ទាល់ជាមួយ Directory "Project1"។

ការអាន DataFrame:

ជាទូទៅបើ "DATA" ជា Object DataFrame នោះ គេអាន Object នេះទៅតាមទំរង់ Syntax ដូច ខាងក្រោម៖

DATA[,]

- ខាងមុខក្បៀសតំណាងអោយទីតាំង ជួរដេកនៃតារាងទិន្នន័យ DATA
- ខាងក្រោយក្បៀសតំណាងអោយទីតាំង ជួរឈរនៃតារាងទិន្នន័យ DATA

Ex: សន្មតថាគេមានតារាងទិន្នន័យ "market.survey" នៅក្នុង Directory "Project1" ។

ក. ចូរអានយក 10 ជួរដេកដំបូងនៃតារាងទិន្នន័យនេះ?

>DATA<-market.survey ចំលង content នៃ market.survey ដោយអោយឈ្មោះថា DATA

ខ. ចូរអានយក 10 ជួរដេកដំបូង ដោយផ្ដាច់យកតែជួរឈរទី 2, 4 និង 5?

គ. ចូររាប់ចំនួន NA នៅក្នុងជួរឈរ "Income"?

ចំណាំ៖ levels() ជា command ដែលប្រើសំរាប់កំណត់ចំនួនព្រឹត្តិការណ៍ ឯកធាតុ

(Categories)នៅក្នុងជួរឈរនីមួយៗនៃតារាងទិន្នន័យដែលមាន data type

ជាប្រភេទ "Non-Numeric"។

ឃ. ចូរកំណត់ចំនួនព្រឹត្តិការណ៍ ឯកធាតុដែលមាននៅក្នុង ជួរឈរ Income?

$\Omega_{Income} = \{ "<7.5", "7.5 - 15", "15 - 25", "25 - 35", "35 - 45", "45 - 75", ">75" \}$

ង. ចូរកំណត់ប្រូបាប៊ីលីតេ នៃអ្នកដែលមាន Income "15 - 25"?

$P("15 - 25") = ?$

ចំណាំ៖ dim() ជា Command ដែលប្រើសំរាប់កំណត់វិមាត្រនៃតារាងទិន្នន័យ

>dim(DATA)

1000 10

>dim(DATA)[1] កំណត់ចំនួនជួរដេក

1000

>dim(DATA)[2] កំណត់ចំនួនជួរឈរ

10

ចម្លើយ ទី ង៖

រៀបចំ ១៖ គណនាដោយប្រើទំរង់ vector

>_M<-length(DATA[,2][DATA[,2]=="15-25" & !is.na(DATA[,2])])

> N<-length(DATA[,2][!is.na(DATA[,2])])

> M/N

[1] 0.2356688

> round(M/N , digits=4) បង្កត់យក ៤ ខ្ទង់ក្រោយក្បៀស

[1] 0.2357

រៀបចំ 2៖ គណនាដោយប្រើទំរង់ DataFrame

```

> M1<-dim( DATA[ DATA[,2]== "15-25" & !is.na( DATA[,2] ) , ])[1]
> N1<- dim( DATA[!is.na( DATA[,2] ) , ])[1]
> M1/N1
[1] 0.2356688

```

ចំណាំ៖ summary() ជា Command ដែលប្រើប្រាស់សំរាប់សង្ខេប ព័ត៌មាននៅក្នុងជួរឈរនីមួយៗ នៃតារាងទិន្នន័យ។

. បើទិន្នន័យជាប្រភេទ “Non-Numeric” នោះ វាសង្ខេបទៅតាម ព្រឹត្តិការណ៍ឯកធាតុនីមួយៗ (Category)ដែលមាននៅក្នុងជួរឈរនោះ។

. បើទិន្នន័យជាប្រភេទ “Numeric” នោះ វាសង្ខេបតាមរយៈ ការគណនា តំលៃ៖
Min , Max, Mean, Range, Standard deviation, Variance ,

Ex: ដោយអនុវត្តទៅលើជួរឈរទី២ នៃតារាងទិន្នន័យ DATA (market.survey) , ចូរ៖

១. រាប់ចំនួនទៅតាមព្រឹត្តិការណ៍ឯកធាតុនីមួយៗ (Category) នៅក្នុងជួរឈរនេះ?

```

> summary(DATA[,2])
<7.5   7.5-15   15-25   25-35   35-45   45-75   >75   NA's
   96    114    185    171    107    80    32    215

```

```

> summary( DATA[,2][!is.na( DATA[,2] ) ] )
<7.5  7.5-15  15-25  25-35  35-45  45-75  >75
   96   114   185   171   107   80   32

```

២. គណនាប្រូបាប៊ីលីតេ នៃព្រឹត្តិការណ៍ឯកធាតុនីមួយៗ (Category) ដែលមាននៅក្នុងជួរឈរនេះ?

P("<7.5") , P("7.5-15") , P("15-25") , P("25-35") , P("35-45") , P("45-75") , P(">75")?

ចម្លើយ៖

```

> summary( DATA[,2][ !is.na( DATA[,2] ) ] )/length( DATA[,2][ !is.na( DATA[,2] ) ] )
<7.5      7.5-15      15-25      25-35      35-45      45-75      >75
0.122293 0.1452229  0.2356688 0.2178344  0.1363057 0.1019108 0.04076433

```

ឬ

```
> summary( DATA[ , 2][!is.na( DATA[ , 2] ) ] ) /dim( DATA[ !is.na( DATA[,2] ) , ] )[1]
<7.5      7.5-15      15-25      25-35      35-45      45-75      >75
0.122293  0.1452229  0.2356688  0.2178344  0.1363057  0.1019108  0.04076433
```

ក្នុងករណីដែលយើងចង់ផ្តល់ឈ្មោះអោយធាតុនីមួយៗនៃ vector, គេអាចសរសេរបានដូចខាងក្រោម៖

```
> Proba<-summary(DATA[,2][!is.na(DATA[,2])]/length(DATA[,2][!is.na(DATA[,2])])
> names(Proba)<-c("P(<7.5)","P(7.5-15)","P(15-25)","P(25-35)","P(35-45)",
                  "P(45-75)","P(>75)")
```

```
> Proba
```

```
 P(<7.5)  P(7.5-15)  P(15-25)  P(25-35)  P(35-45)  P(45-75)  P(>75)
0.122293 0.1452229 0.2356688 0.2178344 0.1363057 0.1019108 0.04076433
```

៣. ចូរគណនាប្រូបាប៊ីលីតេនៃអ្នកដែលមាន Income “ 25-35” ហើយដែលមាន Education ត្រឹម “BA” ?

$$P(\text{“Income=25-35”} * \text{“ Education= BA”}) = 3.32 \%$$

បើ A ជាព្រឹត្តិការណ៍នៃអ្នកដែលមាន “Income=25-35”

B ជាព្រឹត្តិការណ៍នៃអ្នកដែលមាន “ Education= BA”

$$\text{នោះគេអាចសរសេរ ៖ } P(A*B) = \frac{Card(A \cap B)}{Card(\Omega)}$$

ចម្លើយ៖

```
> M<-dim( DATA[( DATA[ ,2]=="25-35" & !is.na( DATA[,2])) & DATA[,5]=="BA" & !is.na(DATA[,5]) , ])[1]
> N<-dim(DATA[!is.na(DATA[,2]) & !is.na(DATA[,5]),,])[1]
> M/N
[1] 0.03324808
```

៤. ចូរគណនាប្រូបាប៊ីលីតេនៃអ្នកដែលមាន Income “ 25-35” ក្នុងចំណោមអ្នកដែលមាន Education ត្រឹម “BA”?

$$P(\text{“Income=25-35”} \setminus \text{“Education=BA”}) = 20.63\%$$

បើ B តំណាងអោយព្រឹត្តិការណ៍ “Income=25-35”

A តំណាងអោយ ព្រឹត្តិការណ៍ “Education=BA”

នោះ គេអាចគណនាដូចខាងក្រោម៖

$$P(B \setminus A) = \frac{P(A * B)}{P(A)} = \frac{\frac{Card(A \cap B)}{Card(\Omega)}}{\frac{Card(A)}{Card(\Omega)}} = \frac{Card(A \cap B)}{Card(A)}$$

របៀបទី១៖

```
>M<-dim( DATA[( DATA[,2]=="25-35"&lis.na(DATA[,2]) ) & ( DATA[,5]=="BA"&lis.na(DATA[,5]) ) , ])[1]
```

```
> N<-dim(DATA[ DATA[,5]=="BA" & lis.na(DATA[,5]) & lis.na(DATA[,2]) , ])[1]
```

```
> M/N
```

```
[1] 0.2063492
```

របៀបទី២៖

```
>DATA1<-DATA[ lis.na( DATA[,2] ) & lis.na( DATA[,5] ) , ]
```

```
> M<-dim( DATA1[ ( DATA1[, 2]=="25-35" ) & ( DATA1[, 5]=="BA" ) , ])[1]
```

```
> N<-dim( DATA1[DATA1[,5]=="BA" , ])[1]
```

```
> M/N
```

```
[1] 0.2063492
```

៤. ចូរគណនាប្រូបាប៊ីលីតេនៃអ្នកដែលមាន Income “ 25-35” ក្នុងចំណោមអ្នកដែលមាន Education ត្រឹម “BA” និងមានអាយុ (age) នៅចន្លោះ “25 – 34” ?

បើ B តំណាងអោយព្រឹត្តិការណ៍ “Income=25-35”

A តំណាងអោយព្រឹត្តិការណ៍ “Education=BA”

C តំណាងអោយព្រឹត្តិការណ៍ “Age = 25-34”

នោះ គេអាចគណនាដូចខាងក្រោម៖

$$P(B \setminus (A \cap C)) = \frac{P(B * (A * C))}{P(A * C)} = \frac{\frac{Card(A \cap B \cap C)}{Card(\Omega)}}{\frac{Card(A \cap C)}{Card(\Omega)}} = \frac{Card(A \cap B \cap C)}{Card(A \cap C)}$$

= 38.70%

របៀបទី១៖

```
> M<-dim( DATA[ ( DATA[,2]=="25-35" & lis.na( DATA[,2] ) ) & ( DATA[,5]=="BA" & lis.na(DATA[,5]) ) &
```



```
( DATA[,4]=="25-34" & !is.na( DATA[,4] ) ) , ] ) [1]
> N<-dim( DATA[ ( DATA[,5]=="BA" & !is.na( DATA[,5] ) ) & ( DATA[,4]=="25-34" & !is.na(DATA[,4])) &
!is.na(DATA[,2]), ] ) [1]
> M/N
[1] 0.3870968
```

របៀបទី២៖

```
>DATA1<-DATA[ !is.na( DATA[,2] ) & !is.na(DATA[,5]) & !is.na( DATA[,4] ) , ]
>M<-dim( DATA1[ ( DATA1[, 2]=="25-35" ) & ( DATA1[,5]=="BA" ) & ( DATA1[, 4]=="25-34" ) , ] ) [1]
> N<-dim( DATA1[ ( DATA1[,5]=="BA" ) & ( DATA1[,4]=="25-34" ) , ] ) [1]
>M/N
[1] 0.3870968
```

Loop នៅក្នុង S-plus:

ជាទូទៅ loop នៅក្នុង S-plus មានទំរង់៖

```
for( i in 1: n ) { Expression}
```

Ex: ចូរគណនា កន្សោមពិជគណិតដូចខាងក្រោម៖

2ⁱ កាលណា i ទទួលតំលៃ 1, 2, 3, 4, 5?

ចម្លើយ៖

```
> Z<-1 ប្រកាស Z ជា object vector ដើម្បីយក Z ទៅដាក់ index
> for( i in 1:5 ) { Z[ i ]<- 2 ^ i }
> Z
```

```
[1] 2 4 8 16 32
```

ចំណាំ៖ ការប្រើប្រាស់ command នីមួយៗរបស់ S-plus:

- Command matrix() : មានទំរង់ទូទៅដូចខាងក្រោម៖

```
matrix(X, nrow, ncol, byrow= T ឬ F)
```

X ជា vector

nrow ប្រើសំរាប់កំណត់ចំនួនជួរដេកនៃ matrix

ncol ប្រើសំរាប់កំណត់ចំនួនជួរឈរ នៃ matrix

byrow=T ប្រើសំរាប់រៀបចំធាតុនៃ vector តាមលក្ខណៈ ជួរដេក

byrow=F ប្រើសំរាប់រៀបធាតុនៃ vector តាមលក្ខណៈ ជួរឈរ

Ex: គេមាន vector X ដូចខាងក្រោម៖

```
>X<-c(20,30,40,34,12,50,35,45)
```

```
>matrix(X,nrow=2,ncol=4,byrow=T)
```

```
  [,1] [,2] [,3] [,4]  
[1,]  20  30  40  34  
[2,]  12  50  35  45
```

```
> matrix(X,nrow=2,ncol=4,byrow=F)
```

```
  [,1] [,2] [,3] [,4]  
[1,]  20  40  12  35  
[2,]  30  34  50  45
```

- Command cbind() & rbind() : ប្រើសំរាប់ភ្ជាប់ vectors ជាច្រើនដែលមាន ប្រវែងស្មើគ្នា បង្កើតជា array matrix។

Ex: គេមាន vectors ពីរដូចខាងក្រោម៖

```
> Y1<- c(23,25,34,45,40)
```

```
> Y2<- c(45,43,36,38,25)
```

```
> cbind(Y1,Y2)
```

```
  Y1 Y2  
[1,] 23 45  
[2,] 25 43  
[3,] 34 36  
[4,] 45 38  
[5,] 40 25
```

```
> rbind(Y1,Y2)
```

```
  [,1] [,2] [,3] [,4] [,5]  
Y1  23  25  34  45  40  
Y2  45  43  36  38  25
```

- Command `dimnames()` :

ប្រើសំរាប់ផ្តល់ឈ្មោះអោយជួរដេក និង ជួរឈរនីមួយៗនៃ Matrix ។

```
dimnames( object matrix)<- list( A, B)
```

Object matrix ជា object ដែលត្រូវផ្តល់ឈ្មោះទៅលើជួរដេក និង ជួរឈរ

A & B ជា vector name ដែលត្រូវដាក់ឈ្មោះអោយជួរដេក និង ជួរឈររៀងគ្នានៃ matrix

Ex:

```
>X<-c(20,30,40,34,12,50,35,45)
```

```
> Tab<-matrix(X,nrow=2,ncol=4,byrow=F)
```

```
> dimnames(Tab)<-list(c("A","B") , c("E", "F", "G", "H" ))
```

```
> Tab
```

	E	F	G	H
A	20	40	12	35
B	30	34	50	45

Ex:

គេមានតារាងព័ត៌មានដូចខាងក្រោម៖

Items BY Sectors of Industry

	Item1	Item2	Item3	Total
Sector1	1000	1200	3000	?
Sector2	500	2000	1500	?
Sector3	2500	1250	2080	?
Sector4	900	800	1900	?
Total	?	?	?	?

1. ចូរ Input តារាងព័ត៌មាននេះចូលក្នុង S-plus?
2. ចូរប្រើ loop ដើម្បីបូក Items សរុបតាម Sector of Industry នីមួយៗនៅលើជួរដេក និង ចំនួនសរុបនៃ Item នីមួយៗនៅលើជួរឈរ?

ចម្លើយ៖

1.

```
>X<-c(1000,1200,3000,500,2000,1500,2500,1250,2080,900,800,1900)
>Tab<-matrix(X, nrow=4, ncol=3, byrow=T)
>A<-c("Sector1","Sector2", "Sector3", "Sector4")
>B<-c("Item1", "Item2", "Item3")
>dimnames(Tab)<-list( A , B)
```

2.

```
>Row<-1
>for( i in 1:dim( Tab )[ 1 ] ) {Row[ i ]<- sum( Tab[ i , ] ) }
>Tab<-cbind(Tab,Total=Row)
>Col<-1
>for( j in 1:dim( Tab )[2] ) { Col[ j ] <- sum(Tab[ , j]) }
>Tab<-rbind(Tab, Total= Col)
> Tab
```

	Item1	Item2	Item3	Total
Sector1	1000	1200	3000	5200
Sector2	500	2000	1500	4000
Sector3	2500	1250	2080	5830
Sector4	900	800	1900	3600
Total	4900	5250	8480	18630

Ex: ដោយអនុវត្តទៅលើ អថេរ "Income" និង "Age" នៅក្នុងតារាងទិន្នន័យ DATA (market.survey)
 ចូរ បង្កើតតារាងរាប់ចំនួនដែលមានជាលក្ខណៈខ្វែងគ្នារវាងអថេរ "Income" និង "Age" ដូចខាងក្រោម៖

Class of Income BY Age categories

	18-24	25-34	35-44	45-54	55-64	65+	Total
<7.5							
7.5-15							
15-25							
25-35							
35-45							
45-75							
>75							
Total							

ចម្លើយ៖

```
function(DATA, R, C)
{
  #This function is used for counting the observed frequency between two
  #non-numerical variables.

  DATA <- DATA[!is.na(DATA[, R]) & !is.na(DATA[, C]), ]
  SubFun <- function(DATA, Inc, R, C)
  {
    Result <- summary(DATA[, C][DATA[, R] == Inc])
    return(Result)
  }
  Income <- levels(DATA[, R])
  Age <- levels(DATA[, C])
  Tab <- matrix(1:(length(Income) * length(Age)), length(Income), length(Age))
  for(i in 1:length(Income)) {
    Tab[i, ] <- SubFun(DATA, Income[i], R, C)
  }
  dimnames(Tab) <- list(Income, Age)
  Row <- 1
  for(i in 1:dim(Tab)[1]) {
    Row[i] <- sum(Tab[i, ])
  }
  Tab <- cbind(Tab, Total = Row)
  Col <- 1
  for(j in 1:dim(Tab)[2]) {
    Col[j] <- sum(Tab[, j])
  }
  Tab <- rbind(Tab, Total = Col)
  return(Tab)
}
```

Ex: ចូរសរសេរ Program S-plus ទូទៅមួយសំរាប់ការវាយតម្លៃជាភាគរយ (ប្រូបាប៊ីលីតេ) ជាលក្ខណៈ ខ្សែងគ្នារវាងអថេរពីរដែលជាប្រភេទ Non-Numeric។ សន្មតថាគេមានអថេរពីរ A & B ៖

- បើអថេរ A មាន ៤ ថ្នាក់ (Category): A1,A2,A3 & A4
- បើអថេរ B មាន ៥ ថ្នាក់ (Category) : B1,B2,B3,B4 & B5

នោះ គេទទួលបានតារាងព័ត៌មានដូចខាងក្រោម៖

Probability of total (Percent of Total)

Percent of Total = Count / Total of valid observations

	A1	A2	A3	A4	Total
B1	P(A1*B1)	P(A2*B1)	P(A3*B1)	P(A4*B1)	P(B1)
B2	P(A1*B2)	P(A2*B2)	P(A3*B2)	P(A4*B2)	P(B2)
B3	P(A1*B3)	P(A2*B3)	P(A3*B3)	P(A4*B3)	P(B3)
B4	P(A1*B4)	P(A2*B4)	P(A3*B4)	P(A4*B4)	P(B4)
B5	P(A1*B5)	P(A2*B5)	P(A3*B5)	P(A4*B5)	P(B5)
Total	P(A1)	P(A2)	P(A3)	P(A4)	1

```
function(DATA, R, C)
{
  #This Function is used to calculate the probability of multiplication of events
  #within the total of observations. The two requirement variables are Non-numeric.

  DATA <- DATA[!is.na(DATA[, R]) & !is.na(DATA[, C]), ]
  SubFun <- function(DATA, Inc, R, C)
    { Result <- summary(DATA[, C][DATA[, R] == Inc])
      return(Result)
    }
  Income <- levels(DATA[, R])
  Age <- levels(DATA[, C])
  Tab <- matrix(1, length(Income), length(Age))
  for(i in 1:length(Income)) {
    Tab[i, ] <- SubFun(DATA, Income[i], R, C)
  }
  dimnames(Tab) <- list(Income, Age)
  Row <- 1
  for(i in 1:dim(Tab)[1]) {
    Row[i] <- sum(Tab[i, ])
  }
  Tab <- cbind(Tab, Total = Row)
  Col <- 1
  for(j in 1:dim(Tab)[2]) {
    Col[j] <- sum(Tab[, j])
  }
  Tab <- rbind(Tab, Total = Col)/dim(DATA)[1]
  Tab<-round(Tab,digits=4)
  return(Tab)
}
```

Ex: ចូរសរសេរ Program S-plus ទូទៅមួយសំរាប់ការវាយតម្លៃជាប្រូបាប៊ីលីតេមានលក្ខខណ្ឌ ធៀបនឹងជួរដេក(Conditional probability within row)រវាងអថេរពីរដែលជាប្រភេទ Non-Numeric។

- បើអថេរទីមួយមាន ៤ ថ្នាក់ (Category) A : A1,A2,A3 & A4
- បើអថេរទីពីរមាន ៥ ថ្នាក់ (Category) B : B1,B2,B3,B4 & B5

នោះ គេទទួលបានតារាងព័ត៌មានដូចខាងក្រោម៖

Conditional Probability within row(Percent within row)

Percent = Count / Total in Row

	A1	A2	A3	A4	Total
B1	P(A1 \ B1)	P(A2 \ B1)	P(A3 \ B1)	P(A4 \ B1)	1
B2	P(A1 \ B2)	P(A2 \ B2)	P(A3 \ B2)	P(A4 \ B2)	1
B3	P(A1 \ B3)	P(A2 \ B3)	P(A3 \ B3)	P(A4 \ B3)	1
B4	P(A1 \ B4)	P(A2 \ B4)	P(A3 \ B4)	P(A4 \ B4)	1
B5	P(A1 \ B5)	P(A2 \ B5)	P(A3 \ B5)	P(A4 \ B5)	1
Total	P(A1)	P(A2)	P(A3)	P(A4)	1

```
function(DATA, R, C)
{
  #This function is used to calculate the conditional probability within row
  #(Percent within row). The two requirement variables are non-numeric.

  DATA <- DATA[!is.na(DATA[, R]) & !is.na(DATA[, C]), ]
  SubFun <- function(DATA, Inc, R, C)
  {
    Result <- summary(DATA[, C][DATA[, R] == Inc])
    return(Result)
  }
  Income <- levels(DATA[, R])
  Age <- levels(DATA[, C])
  Tab <- matrix(1:(length(Income) * length(Age)), length(Income), length(Age))
  for(i in 1:length(Income)) {
    Tab[i, ] <- SubFun(DATA, Income[i], R, C)
  }
  dimnames(Tab) <- list(Income, Age)
  Col <- 1
  for(j in 1:dim(Tab)[2]) {
    Col[j] <- sum(Tab[, j])
  }
  Tab <- rbind(Tab, Total = Col)
  Row <- 1
  for(i in 1:dim(Tab)[1]) {
    Tab[i, ] <- Tab[i, ]/sum(Tab[i, ])
    Row[i] <- sum(Tab[i, ])
  }
  Tab <- round(Tab, digits = 4)
  Tab <- cbind(Tab, Total = Row)
  return(Tab)
}
```

Ex: ចូរសរសេរ Program S-plus ទូទៅមួយសំរាប់ការវាយតម្លៃជាប្រូបាប៊ីលីតេមានលក្ខខណ្ឌ រៀប
 នឹងជួរឈរ (Conditional probability within column)រវាងអថេរពីរដែលជាប្រភេទ Non-Numeric។

- បើអថេរទីមួយមាន ៤ ថ្នាក់ (Category) A : A1,A2,A3 & A4
- បើអថេរទីពីរមាន ៥ ថ្នាក់ (Category) B : B1,B2,B3,B4 & B5

នោះ គេទទួលបានតារាងព័ត៌មានដូចខាងក្រោម៖

Conditional Probability within column(Percent within column)

Percent within column = Count / Total in Column

	A1	A2	A3	A4	Total
B1	$P(B1 \setminus A1)$	$P(B1 \setminus A2)$	$P(B1 \setminus A3)$	$P(B1 \setminus A4)$	$P(B1)$
B2	$P(B2 \setminus A1)$	$P(B2 \setminus A2)$	$P(B2 \setminus A3)$	$P(B2 \setminus A4)$	$P(B2)$
B3	$P(B3 \setminus A1)$	$P(B3 \setminus A2)$	$P(B3 \setminus A3)$	$P(B3 \setminus A4)$	$P(B3)$
B4	$P(B4 \setminus A1)$	$P(B4 \setminus A2)$	$P(B4 \setminus A3)$	$P(B4 \setminus A4)$	$P(B4)$
B5	$P(B5 \setminus A1)$	$P(B5 \setminus A2)$	$P(B5 \setminus A3)$	$P(B5 \setminus A4)$	$P(B5)$
Total	1	1	1	1	1

```
function(DATA, R, C)
{ DATA <- DATA[!is.na(DATA[, R]) & !is.na(DATA[, C]), ]
  SubFun <- function(DATA, Inc, R, C)
  { Result <- summary(DATA[, C][DATA[, R] == Inc])
    return(Result)
  }
  Income <- levels(DATA[, R])
  Age <- levels(DATA[, C])
  Tab <- matrix(0, length(Income), length(Age))
  for(i in 1:length(Income)) {
    Tab[i, ] <- SubFun(DATA, Income[i], R, C)
  }
  dimnames(Tab) <- list(Income, Age)
  Row <- 1
  for(i in 1:dim(Tab)[1]) {
    Row[i] <- sum(Tab[i, ])
  }
  Tab <- cbind(Tab, Total = Row)
  Col <- 1
  for(j in 1:dim(Tab)[2]) {
    Tab[, j] <- Tab[, j]/sum(Tab[, j])
    Col[j] <- sum(Tab[, j]) }
  Tab <- rbind(Tab, Total = Col)
  Tab <- round(Tab, digits = 4)
  return(Tab)
}
```


ចំណាំ៖

គេអាចប្រើ command “ table() ” សំរាប់ការរាប់ចំនួនខ្ទែងគ្នារវាងអថេរពីរដែលជាប្រភេទ Non-Numeric ៖

table(Variable1,Variable2)

ដែល ៖ Variable1 មានតួនាទីជា ជួរដកនៃ តារាងពត៌មាន

Variable2 មានតួនាទីជាជួរឈរនៃតារាងពត៌មាន

```
function(DATA, R, C)
{
  #This function is used for counting the observed frequency between two
  #non-numerical variables.
  DATA <- DATA[!is.na(DATA[, R]) & !is.na(DATA[, C]), ]
  Tab <- table(DATA[, R], DATA[, C])
  Row <- 1
  for(i in 1:dim(Tab)[1]) {
    Row[i] <- sum(Tab[i, ])
  }
  Tab <- cbind(Tab, Total = Row)
  Col <- 1
  for(j in 1:dim(Tab)[2]) {
    Col[j] <- sum(Tab[, j])
  }
  Tab <- rbind(Tab, Total = Col)
  return(Tab)
}
```

Crosstabs Analysis:

គេមាន Command " crosstabs() " ដែលប្រើសំរាប់ការបង្កើតតារាងពត៌មាន វាយតំលៃជាលក្ខណៈ ខ្វែងគ្នារវាងអថេរពីរដែលជាប្រភេទ Non-numeric ។ ក្នុងនោះ ថត (cell) នីមួយៗនៃតារាង crosstabs មានបួនតំលៃ៖

- Observed Frequency (ចំនួន) fo_{ij}
- Percent within row (Conditional probability under row)
- Percent within column (Conditional probability under column)
- Percent of Total (Probability of total)

Model of crosstabs

+-----+	
N	
N/RowTotal	
N/ColTotal	
N/Total	
+-----+	

crosstabs(~ Variable1+Variable2, na.action = na.exclude)

ដែល៖

~ (tild) មានន័យថាកំណត់ Model នៃ តារាង crosstabs

Variable1 មានតួនាទីជាជួរដេកនៃតារាងពត៌មាន crosstabs

Variable2 មានតួនាទីជាជួរឈរនៃតារាងពត៌មាន crosstabs

na.action=na.exclude ប្រើសំរាប់លុបគូទិន្នន័យណាដែលមាន NA

សំគាល់៖

សន្មតថាគេមានអថេរពីរ A & B ដូចខាងក្រោម៖

A មាន ៣ ព្រឹត្តិការណ៍ឯកធាតុ៖ A1,A2 & A3

B មាន ៤ ព្រឹត្តិការណ៍ឯកធាតុ៖ B1,B2,B3 & B4

Obs. A B

1 A1 B2

2 A2 B3 ដោយធ្វើ

3 A1 B1 ----->

4 A3 B2 Crosstabs Analysis

. . .

. . .

. . .

N A3 B1

	A1	A2	A3	Total
B1	f_{O11} $P(A1 B1)=f_{O11}/f_{O1.}$ $P(B1 A1)=f_{O11}/f_{O.1}$ $P(A1*B1)=f_{O11}/N$	f_{O12}	f_{O13}	$f_{O1.}$ $P(B1)=f_{O1.}/N$
B2	f_{O21}	f_{O22}	f_{O23}	$f_{O2.}$ $P(B2)=f_{O2.}/N$
B3	f_{O31}	f_{O32}	f_{O33}	$f_{O3.}$ $P(B3)=f_{O3.}/N$
B4	f_{O41}	f_{O42}	f_{O43}	$f_{O4.}$ $P(B4)=f_{O4.}/N$
Total	$f_{O.1}$ $P(A1)=f_{O.1}/N$	$f_{O.2}$ $P(A2)=f_{O.2}/N$	$f_{O.3}$ $P(A3)=f_{O.3}/N$	N

Ex: គេបានធ្វើការអង្កេតនៅក្នុងហាងលក់ទំនិញដ៏ធំមួយ ទៅលើអតិថិជនចំនួន ២០០ នាក់។ ក្នុងចំណោមអតិថិជនទាំងនេះ គេរកឃើញថា មាន ១២០ នាក់ បានឃើញការផ្សាយពាណិជ្ជកម្មរបស់ក្រុមហ៊ុន។ គេដឹងទៀតថាក្នុងចំណោមអតិថិជនទាំង ២០០នាក់នេះ មាន ៨០នាក់បានទិញសម្ភារៈមួយចំនួន។ ព័ត៌មានបន្ថែមទៀតបានបង្ហាញអោយឃើញថា ក្នុងចំណោមអ្នកបានទិញសម្ភារៈមាន ៦០ នាក់បានឃើញការផ្សាយពាណិជ្ជកម្ម។

ចូរប្រៀបព័ត៌មានទាំងនេះជាតារាង crosstabs?

វិភាគ៖ នៅឧទាហរណ៍នេះមានអថេរពីរ ដែលបង្កជាតារាង crosstabs៖

អថេរទី១៖ ការផ្សាយពាណិជ្ជកម្ម -----> បានឃើញ ឬ មិនបានឃើញ

អថេរទី២៖ ការទិញសម្ភារៈ -----> បានទិញ ឬ មិនបានទិញ

	បានឃើញ (A1)	មិនបានឃើញ (A2)	សរុប
បានទិញ (B1)	60 $P(A1 B1)=60/80=75\%$ $P(B1 A1)=60/120=50\%$ $P(A1*B1)=60/200=30\%$	20 $P(A2 B1)=20/80$ $P(B1 A2)=20/80$ $P(A2*B1)=20/200$	80 $P(B1)=80/200$
មិនបានទិញ (B2)	60 $P(A1 B2)=60/120$ $P(B2 A1)=60/120$ $P(A1*B2)=60/200$	60 $P(A2 B2)=60/120$ $P(B2 A2)=60/80$ $P(A2*B2)=60/200$	120 $P(B2)=120/200$
សរុប	120 $P(A1)=120/200$	80 $P(A2)=80/200$	200

Random Variable (អថេរចៃដន្យ)៖

អថេរចៃដន្យ៖ អថេរចៃដន្យដាច់ និង អថេរចៃដន្យជាប់។

ក. អថេរចៃដន្យដាច់៖ (*Discrete Random Variable*)

ជាអថេរចៃដន្យដែលអាចទទួលតំលៃដាច់ដោយឡែកពីគ្នានៅក្នុងសំនុំរាប់អស់ ឬ រាប់មិនអស់។

ឧទាហរណ៍៖

- អថេរចៃដន្យជាប់ X ដែលសំគាល់បរិមាណសមាជិកនៅក្នុងគ្រួសារ។

$X: 2 \ 3 \ 4 \ 5 \ 6 \ 7 \dots\dots\dots$

- អថេរចៃដន្យជាប់ X ដែលសំគាល់ចំនួនអវត្តមានរបស់និស្សិតនៅក្នុងការសិក្សា មុខវិជ្ជាស្ថិតិ។

$X: 0 \ 1 \ 2 \ 3 \ 4 \dots\dots\dots$

ខ. អថេរចៃដន្យជាប់៖ (Continuous Random Variable)

ជាអថេរចៃដន្យដែលអាចទទួលបានតំលៃទាំងអស់នៅក្នុង ចន្លោះ , ផ្ទៃក្រឡា ឬ លំហរជាក់លាក់មួយ។ អថេរចៃដន្យជាប់ជាចំនួនរាប់មិនអស់។

Expectation, Variance and Standard deviation

1. Expectation:

$$E(X) = \sum_{i=1}^n x_i p_i = \bar{X} = \frac{1}{n} \sum_{i=1}^n x_i$$

mean(X, na.rm=T)

ដែល ៖

X ជាអថេរចៃដន្យដែលជាប្រភេទ Numeric

na.rm=T ប្រើសំរាប់លុបតំលៃ NA ចេញពី ប្រមាណវិធី

2. Variance:

$$V(x) = \sum_{i=1}^n (x_i - E(X))^2 p_i \quad s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{X})^2$$

var(X, na.method= "available")

ដែល៖

X ជាអថេរចៃដន្យដែលជាប្រភេទ Numeric

na.method="available" ប្រើសំរាប់លុបតំលៃ NA ចេញពី ប្រមាណវិធី

3. Standard deviation (Std.dev)

$$\sigma(X) = \sqrt{V(X)} = S = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{X})^2}$$

stdev(X, na.rm=T)

Ex: ចូរសរសេរProgram S-plus ទូទៅមួយសំរាប់ ការបង្កើតតារាងប្រៀបធៀបតំលៃមធ្យម
(Compare means) ដូចខាងក្រោម៖

បើអថេរ X ជាប្រភេទ Numeric ហើយ

បើ អថេរ A ជាប្រភេទ Non-Numeric ដែលមាន ៥ ថ្នាក់ (Category): A1, A2, A3 & A4

Compare Means Table

	N	Min	Max	Mean	Variance	Std.deviation
A1						
A2						
A3						
A4						
Total						

```
function(DATA, X, A)
{
  #This function is used to construct "Compare means" table.
  #The first requirement variable is Numeric & the second one
  #is Non-Numeric
  DATA <- DATA[!is.na(DATA[, X]) & !is.na(DATA[, A]), ]
  SubFun <- function(DATA, X, A, Class)
  {
    I <- DATA[, X][DATA[, A] == Class]
    Result <- c(min(I), max(I), mean(I), var(I), stdev(I))
    return(Result)
  }
  Class <- levels(DATA[, A])
  Colname <- c("Min", "Max", "Mean", "Variance", "Std.dev")
  Tab <- matrix(0, length(Class), length(Colname))
  for(i in 1:length(Class)) {
    Tab[i, ] <- SubFun(DATA, X, A, Class[i])
  }
  dimnames(Tab) <- list(Class, Colname)
  I <- DATA[, X]
  Total <- c(min(I), max(I), mean(I), var(I), stdev(I))
  Tab <- rbind(Tab, Total)
  return(Tab)
}
```

Normal Distribution:

សន្មតថា X ជាអថេរចៃដន្យជាប់ដែលអាចទទួលនូវគ្រប់តំលៃនៅក្នុងចន្លោះ $]-\infty, +\infty[$ ។
 X ជាអថេរនៃរបាយទិន្នន័យណ័រម៉ាល់ (Normal Distribution) កាលណា X មានអនុគមន៍
របាយដង់ស៊ីតេ (Density Distribution) $f(x)$, កំណត់ ៖

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, -\infty < x < +\infty$$

ដែល $E(X) = \mu$ និង $V(X) = \sigma^2$ ។ គេសរសេរ ៖

$$X \sim N(\mu, \sigma^2)$$

S-plus:

នៅក្នុង S-plus គេមាន *command* "pnorm" ដែលប្រើសំរាប់ការគណនាអាំងតេក្រាល ពី
 $-\infty$ ទៅដល់តំលៃ x ណាមួយ៖

$$\text{pnorm}(x, \text{mean}, \text{Std.deviation})$$

$$P(X < x) = \int_{-\infty}^x \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx$$

Ex: ការអង្កេតទៅលើនិស្សិតមួយចំនួនបានអោយដឹងថា ការចំណាយជាមធ្យមប្រចាំសប្តាហ៍
(X) របស់និស្សិតនីមួយៗគឺ 55 \$ និងមានគំលាតស្តង់ដា 20 \$ ។ ចូរគណនាប្រូបាប៊ីលីតេដូច
ខាងក្រោម៖

- 1- $P(X < 45 \$)$?
- 2- $P(45\$ < X < 65 \$)$?
- 3- $P(X > 65 \$)$?

ចម្លើយ៖

ដោយការចំណាយជាមធ្យមប្រចាំសប្តាហ៍ (X) របស់និស្សិតនីមួយៗមានរបាយទិន្នន័យ
ណ័រម៉ាល់ $X \sim N(55, 20^2)$ នោះអនុគមន៍របាយដង់ស៊ីតេ $f(x)$ សរសេរ៖

$$f(x) = \frac{1}{20\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-55}{20}\right)^2}$$

$$1- P(X < 45 \$) = \int_{-\infty}^{45} \frac{1}{20\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-55}{20}\right)^2} dx = 30.85\%$$

$$>\text{pnorm}(45, 55, 20)$$

0.3085375

$$2- P(45 \$ < X < 65 \$) = \int_{45}^{65} \frac{1}{20\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-55}{20}\right)^2} dx = \int_{-\infty}^{65} \frac{1}{20\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-55}{20}\right)^2} dx - \int_{-\infty}^{45} \frac{1}{20\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-55}{20}\right)^2} dx \\ = 38.29\%$$

>pnorm(65, 55, 20) – pnorm(45, 55, 20)

0.3829249

$$3- P(X > 65 \$) = \int_{65}^{+\infty} \frac{1}{20\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-55}{20}\right)^2} dx = 1 - \int_{-\infty}^{65} \frac{1}{20\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-55}{20}\right)^2} dx = 30.85\%$$

>1 – pnorm(65,55,20)

0.3085375

Ex: គេដឹងថារបាយពិន្ទុនៃការប្រលងមុខវិជ្ជា Statistics Analysis របស់និស្សិតមានរបាយ

ទិន្នន័យ ណ័រម៉ាល់ $X \sim N(75, 10^2)$ ។

ចូរកំណត់ប្រូបាប៊ីលីតេនៃអ្នកដែលមានពិន្ទុលើសពី 85 ដោយប្រើរគោលការណ៍នៃរបាយ

ទិន្នន័យណ័រម៉ាល់បង្រួមកណ្តាល Z ៖

$$P(X > 85) ?$$

ចម្លើយ៖

1- គោលការណ៍របាយទិន្នន័យណ័រម៉ាល់

$$P(X > 85) = \int_{85}^{+\infty} \frac{1}{10\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-75}{10}\right)^2} dx = 1 - \int_{-\infty}^{85} \frac{1}{10\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-75}{10}\right)^2} dx = 15.86\%$$

>1 – pnorm(85, 75, 10)

0.1586553

2- គោលការណ៍របាយទិន្នន័យណ័រម៉ាល់បង្រួមកណ្តាល

$$P(X > 85) = P\left(\frac{X - 75}{10} > \frac{85 - 75}{10}\right) = P(Z > 1) = 1 - P(Z < 1) \\ = 1 - \int_{-\infty}^1 \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz = 15.86\%$$

>1- pnorm(1, 0, 1)

0.1586553

Ex: គេដឹងថាពិន្ទុនៃការប្រឡងរបស់និស្សិតមានរបាយទិន្នន័យណ័រម៉ាល់ $N(75, 20^2)$ ។ គេដឹងទៀតថាភាគរយនិស្សិតដែលរៀនពូកែមាន 15% នៃនិស្សិតទាំងអស់។

តើក្នុងចំណោមនិស្សិតដែលរៀនពូកែ និស្សិតដែលមានពិន្ទុទាបជាងគេបំផុតស្មើប៉ុន្មាន?

ចម្លើយ៖

បើ X ជាអថេរចៃដន្យដែលសំគាល់ពិន្ទុរបស់និស្សិត នោះភាគរយនិស្សិតដែលរៀនពូកែ

កំណត់៖ $P(X \geq x) = 0.15 \rightarrow x = ?$

ក្នុងករណីនេះ $P(X < x) = 0.85 \rightarrow x = \text{qnorm}(0.85, 75, 20) = 95.72$ ពិន្ទុ

Binomial Distribution (ច្បាប់ទ្វេធានៈ)

ច្បាប់ទ្វេធានៈសិក្សាអំពីការកំណត់ឧកាសនៃការកើតឡើងនៃព្រឹត្តិការណ៍ជោគជ័យ x នៅក្នុងការសាកល្បង ការពិសោធន៍ ឬ ការអង្កេត n ដងមិនអាស្រ័យគ្នា, កំណត់៖

$$P(X = x) = C_n^x p^x (1 - p)^{n-x}, \quad x : 0, 1, 2, \dots, n$$

n ជាចំនួនដងនៃការអង្កេត

x ជាចំនួនព្រឹត្តិការណ៍ជោគជ័យក្នុងការអង្កេត n ដងមិនអាស្រ័យគ្នា

p ជាប្រូបាប៊ីលីតេនៃព្រឹត្តិការណ៍ជោគជ័យក្នុងការអង្កេតម្តងៗ

S-plus: `dbinom` : density of Binomial Distribution

`dbinom(x, n, p)`

Ex: និស្សិតម្នាក់ត្រូវឆ្លងកាត់ការប្រឡងមួយដែលមាន 120 សំណួរ ។ ក្នុងសំណួរនីមួយៗមាន 4 ចម្លើយ, ក្នុងនោះចម្លើយដែលត្រឹមត្រូវមានតែមួយប៉ុណ្ណោះ ។ ដោយនិស្សិតរូបនោះពុំមានឧកាសគ្រប់គ្រាន់សំរាប់ការសិក្សា គាត់សំរេចចិត្តចូលរួមការប្រឡងដោយជ្រើសរើស យកចម្លើយក្នុងសំណួរនីមួយៗ ដោយចៃដន្យ។ និស្សិតនេះអាចប្រឡងជាប់ដ៏រាបណាគាត់ឆ្លើយត្រូវយ៉ាងតិច 50% នៃសំណួរ ទាំងអស់ ។ តើនិស្សិតរូបនេះអាចប្រឡងជាប់ដែរឬទេ?

ចម្លើយ៖

បើ X ជាចំនួនសំនួរដែលអាចឆ្លើយត្រូវ, នោះ ឧកាសដើម្បីអោយគាត់ប្រលងជាប់ កំណត់៖

$$P(X \geq 60) = P(X=60) + P(X=61) + \dots + P(X=120)$$

$$\begin{aligned} &= \sum_{x=60}^{120} P(X=x) \\ &= \sum_{x=60}^{120} C_{120}^x p^x (1-p)^{120-x} \\ &= \sum_{x=60}^{120} C_{120}^x (1/4)^x (1-1/4)^{120-x} \\ &= 0 \end{aligned}$$

S-plus:

```
function()
{
  n <- 120
  x <- 60: n
  p <- 1/4
  Proba <- sum(dbinom(x, n, p))
  Proba <- round(Proba, digits = 4)
  return(Proba)
}
```

សន្និដ្ឋាន៖ ដោយប្រូបាប៊ីលីតេ $P(X \geq 60) = 0$ មានន័យថានិស្សិតរូបនេះគ្មានឧកាសទាល់តែ

សោះ ដើម្បីប្រលងជាប់។

យើងដឹងថានិស្សិតរូបនេះគ្មានឧកាសទាល់តែសោះ ដើម្បីប្រលងជាប់ ក៏ប៉ុន្តែគាត់អាចឆ្លើយត្រូវខ្លះដែរ។ កំណត់រកចំនួនសំនួរជាមធ្យមដែលគាត់អាចឆ្លើយត្រូវ?

$$E(X) = n \cdot p = 120 \cdot 1/4 = 30 \text{ សំនួរ}$$

ក្នុងករណីនេះនិស្សិតរូបនេះអាចឆ្លើយត្រូវជាមធ្យម 30 សំនួរ ។

ក្នុងករណីដែលយើងរំលឹករបស់គាត់អាចឡើងចុះ តើចំនួនសំនួរជាមធ្យមដែលគាត់អាចឆ្លើយត្រូវនឹងមានការប្រែប្រួលក្នុងគំលាតជាមធ្យមប៉ុន្មានដែរ?

$$V(X)=n \cdot p \cdot (1-p) \rightarrow \sigma(X)=\sqrt{n \cdot p \cdot (1-p)}$$

$$\sigma(X)=\sqrt{n \cdot p \cdot (1-p)}=\sqrt{120 \cdot 1/4 \cdot (1-1/4)} \approx 5 \text{ សំនួរ}$$

មានន័យថាបើយើងអ៊ុនឡែង និស្សិតរូបនេះអាចគូសត្រូវជាមធ្យម 35 សំនួរ , ក៏ប៉ុន្តែបើយើងអ៊ុន
 និស្សិតរូបនេះធ្លាក់ចុះ នោះគាត់អាចឆ្លើយត្រូវជាមធ្យម តែ 25 សំនួរតែប៉ុណ្ណោះ។ដូច្នេះ គេអាច
 និយាយថា និស្សិតរូបនេះអាចឆ្លើយត្រូវជាមធ្យម ប្រែប្រួលពី 25 ទៅ 35 សំនួរ។