Statistiques en Scilab

1. Statistiques descriptives univariées

a. Série statistique associée à un échantillon

i. Population, individus, échantillon et variable

La population est l'ensemble des éléments dont on étudie les données

Les éléments de la population sont les individus

Un échantillon est une liste d'individus, issus de la population, sur lesquels on observe une caractéristique appelée variable

Les valeurs prises par la variable *X* sont les modalités de *X* notées $x_1, x_2, \dots, x_N, \dots$

La liste $x_1, x_2, \dots, x_N, \dots$ est une série statistique simple :

Exemple :

Modalités x_i	1	1	2	3	2	2	4	3	2	1	2	5	4

Même série ordonnée

Modalités x_i	1	1	1	2	2	2	2	2	3	3	4	4	5

Si certaines valeurs de $x_1, x_2, \dots, x_N$ sont égales, on les regrouper en $y_1, y_2, \dots, y_p$ où $p \leq N$, en indiquant le nombre n_i de fois où y_i apparaît : on parle alors de série dépouillée

Exemple :

Modalités y_i	1	2	3	4	5
Effectifs n_i	3	5	2	2	1

b. Classes

i. Définitions

Les classes correspondent aux regroupements des modalités par intervalles

Le réel $c_{i+1} - c_i$ est l'amplitude de la classe $]c_i, c_{i+1}]$

ii. Création de classes en Scilab

La création de classes en Scilab passe par `s=linspace(inf, sup, n+1)`

Où :

- `inf` est l'extrémité inférieure de la 1ère classe
- `sup` est l'extrémité supérieure de la dernière classe
- `n` est le nombre de classes

Pour obtenir le nombre `b` de valeurs dans chaque classe et le `a` de la classe à laquelle appartient chaque valeur : `[a,b]=dsearch(x,s)`

iii. Exemple en Scilab

Soit la série statistique :

```
x  =

7.   3.   8.   7.   3.   10.   8.   3.   6.   4.   2.
5.   4.   3.   5.   11.   9.   12.   9.   9.
```

On se propose de regrouper cette série en 4 classes dont les extrémités sont 1 et 15 :

```
--> s=linspace(1,15,6)

 s  =

   1.   3.8   6.6   9.4   12.2   15.
```

Le nombre de valeurs par classe est `b` et la classe à laquelle appartient chaque valeur est `a` :

```
--> [a,b]=dsearch(x,s)

 b  =

   5.   5.   7.   3.   0.

 a  =

3.   1.   3.   3.   1.   4.   3.   1.   2.   2.   1.
2.   2.   1.   2.   4.   3.   4.   3.   3.
```

c. Effectifs

i. Définitions

L'effectif n_i de la modalité y_i ou de la classe $]c_i, c_{i+1}]$ est le nombre d'individus de cette modalité ou de cette classe

L'effectif cumulé d'une modalité est la somme des effectifs des modalités qui lui sont inférieures ou égales

ii. Calcul des effectifs cumulés en Scilab : `cumsum(m( :,c))` lorsque les effectifs se trouvent dans la *c*-ème colonne du tableau

iii. Calcul de l'effectif d'une modalité en Scilab

La fonction `tabul(x,'i')` ordonne la série X dans l'ordre croissant et donne l'effectif de chaque modalité de la série

```
x  =

7.   3.   8.   7.   3.   10.   8.   3.   6.   4.   2.
5.   4.   3.   5.   11.   9.   12.   9.   9.

--> m=tabul(x,'i')

 m  =

   2.    1.

   3.    4.

   4.    2.

   5.    2.

   6.    1.

   7.    2.

   8.    2.

   9.    3.

   10.   1.

   11.   1.

   12.   1.
```

Calcul des effectifs cumulés :

```
->  cumsum(m(:,2))
 ans  =
   1.
   5.
   7.
   9.
   10.
   12.
   14.
   17.
   18.
   19.
   20.
```

d. Fréquences

La fréquence f_i de x_i ou de $]c_i, c_{i+1}]$ est $f_i = \frac{n_i}{N}$

e. Fréquences cumulées

La fréquence cumulée d'une modalité est la somme des fréquences des modalités qui lui sont inférieures ou égales

f. Graphiques

i. Diagramme en bâton : `bar(x,n)`

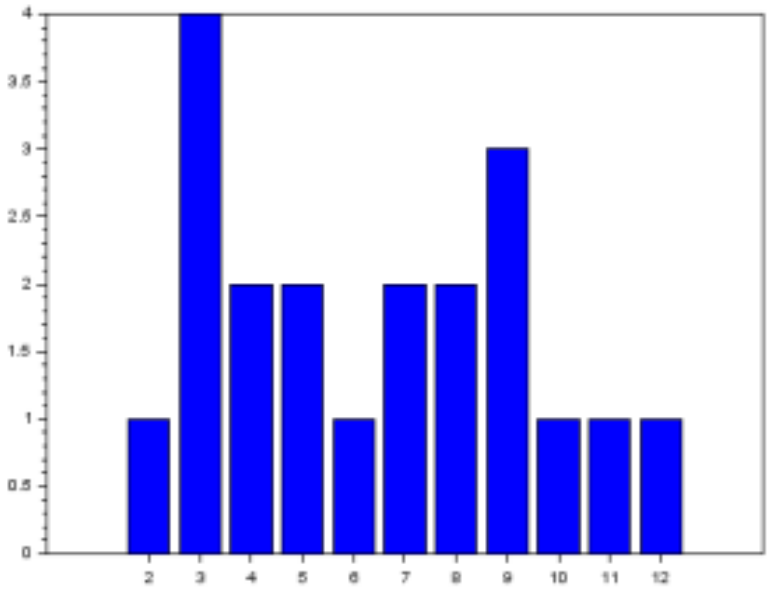
où :

`x`= série ordonnée étudiée

`n` = série des effectifs correspondants

Exemple :

```
--> m=tabul(x,'i')
 m  =
   2.    1.
   3.    4.
   4.    2.
   5.    2.
   6.    1.
   7.    2.
   8.    2.
   9.    3.
   10.   1.
   11.   1.
   12.   1.

--> bar(m(:,1),m(:,2))
```



ii. Histogramme

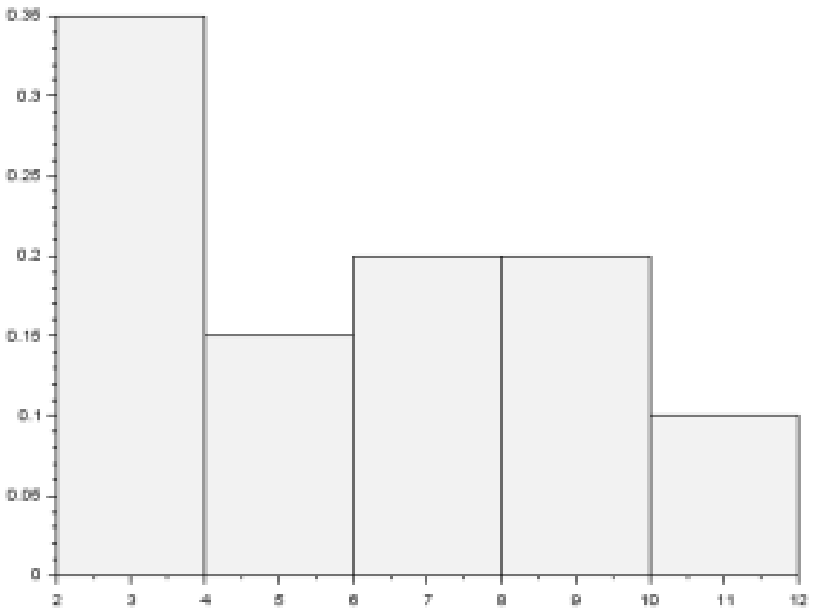
1. Principe

On représente la série statistique groupée par classes en plaçant les c_i sur un axe horizontal et en traçant, à la verticale, un rectangle de base $[c_i, c_{i+1}]$ dont l'aire est proportionnelle à n_i.

2. Création en Scilab : `histplot(c,x)` où `c` est le nombre de classes

3. Exemple

```
->  x=[7.   3.   8.   7.   3.   10.   8.   3.   6.   4.   2.
5.   4.   3.   5.   11.   9.   12.   9.   9.]
 x  =
7.   3.   8.   7.   3.   10.   8.   3.   6.   4.   2.   5.   4.
3.   5.   11.   9.   12.   9.   9.

--> histplot(5,x)
 ans  =
   0.35   0.15   0.2   0.2   0.1
```



iii. Diagramme en secteurs ou diagramme circulaire ou camembert

1. Définition

   Chaque modalité (ou chaque classe) est représentée par un secteur circulaire dont l'angle au centre est proportionnel à l'effectif ou à la modalité

2. Création en Scilab : `pie(n,'x1','x2',…,'xp')` où `n` est un vecteur de taille p et `'x1','x2',…,'xp'` sont des légendes

3. Exemple

```
--> x=[7.   3.   8.   7.   3.   10.   8.   3.   6.   4.   2.
5.   4.   3.   5.   11.   9.   12.   9.   9.]

--> m=tabul(x,'i')

 m   =

   2.    1.

   3.    4.

   4.    2.

   5.    2.

   6.    1.

   7.    2.

   8.    2.

   9.    3.

   10.   1.

   11.   1.

   12.   1.

pie(m(:,2),['2','3','4','5','6','7','8','9','10',
'11','12'])
```

g. Indicateurs de position

i. Moyenne

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Si la série est groupée par modalités (y_i, n_i) :

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{N} n_i\, y_i$$

En Scilab :

```
--> mean(x)
 ans  =
   6.4
```

ii. Médiane

La médiane d'une série statistique ordonnée est le réel M_e qui partage la série en 2 séries d'effectifs égaux. C'est la valeur en laquelle la fréquence cumulée est égale à $\frac{1}{2}$.

En Scilab :

```
--> median(x)
 ans  =
   6.5
```

iii. Mode

Le mode d'une série statistique est la valeur de la variable qui a le plus grand effectif. Il peut y avoir plusieurs modes

iv. Quantiles

Le 1er quartile d'une série statistique est le réel q_1 correspondant à 25% des fréquences cumulées

Le 3ème quartile d'une série statistique est le réel q_3 correspondant à 75% des fréquences cumulées

En Scilab :

```
--> quart(x)
 ans  =
   3.5
```

```
   6.5
   9.
```

Le k-ème décile d'une série statistique est le réel correspondant à $10k\%$ des fréquences cumulées

h. Indicateurs de dispersion

i. Etendue

L'étendue est la différence entre la plus grande modalité et la plus petite modalité

```
--> max(x)-min(x)
 ans  =
   10.
```

ii. Variance empirique et écart type empirique

$$V(X) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{X})^2$$

Si la série est groupée par modalités (y_i, n_i) :

$$V(X) = \frac{1}{N}\sum_{i=1}^{N} n_i.(x_i - \bar{X})^2$$

L'écart type est $\sigma = \sqrt{V(X)}$

En Scilab :

```
--> variance(x)
 ans  =
   9.0947368
--> stdev(x)
 ans  =
   3.0157481
```

La variance empirique, estimateur sans biais de la population toute entière, est :

$$\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{X})^2$$

iii. Ecart inter-quantile : $q_3 - q_1$

2. Statistiques descriptives bivariées

a. Série statistique à 2 variables et nuage de points associé

On appelle série statistique double la donnée de la liste (x_i, y_i), où $i \in [1, n]$ et $j \in [1, n]$, des valeurs ou modalité prises par 2 variables *X* et *Y*, chaque couple (x_i, y_i) étant associé à un seul individu de l'échantillon.

On appelle nuage de points d'une série statistique double, l'ensemble des points $M_i(x_i, y_i)$,

b. Point moyen du nuage

On appelle point moyen du nuage le point de coordonnées $(\overline{X}, \overline{Y})$

Les droites de régression linéaire passent par le même point moyen.

c. Covariance empirique

$$\mathrm{cov}(X,Y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{X})(y_i - \overline{Y})$$

En Scilab : `corr(X,Y,1)`
Exemple :

```
-->
x=1:20,y=[10,12,15,13,16,14,17,11,20,18,19,21,22,25,26,27,23
,28,30,29]
 x  =
1.   2.   3.   4.   5.   6.   7.   8.   9.   10.   11.   12.
13.   14.   15.   16.   17.   18.   19.   20.

 y  =

10.   12.   15.   13.   16.   14.   17.   11.   20.   18.
19.   21.   22.   25.   26.   27.   23.   28.   30.   29.

--> corr(x,y,1)
 ans  =

   33.4
```

d. Coefficient de corrélation empirique

$$\mathrm{r}(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sigma_x \sigma_y}$$

```
--> r=corr(x,y,1)/(stdev(x)*stdev(y))
 r  =

   0.896673
```

e. Droites de régression

i. Droite de régression de *Y* en *X*

L'unique droite minimisant la somme S est : $y = \frac{\text{cov}(X,Y)}{V(X)}(x - \bar{X}) + \bar{Y}$, où :

$$S = \sum_{i=1}^{n} (y_i - a\,x_i - b)^2$$

X est alors la variable explicative et *Y* la variable expliquée

ii. Droite de régression de *X* en *Y*

L'unique droite minimisant la somme S est : $x = \frac{\text{cov}(X,Y)}{V(Y)}(y - \bar{Y}) + \bar{X}$, où :

$$S = \sum_{i=1}^{n} (x_i - a\,y_i - b)^2$$

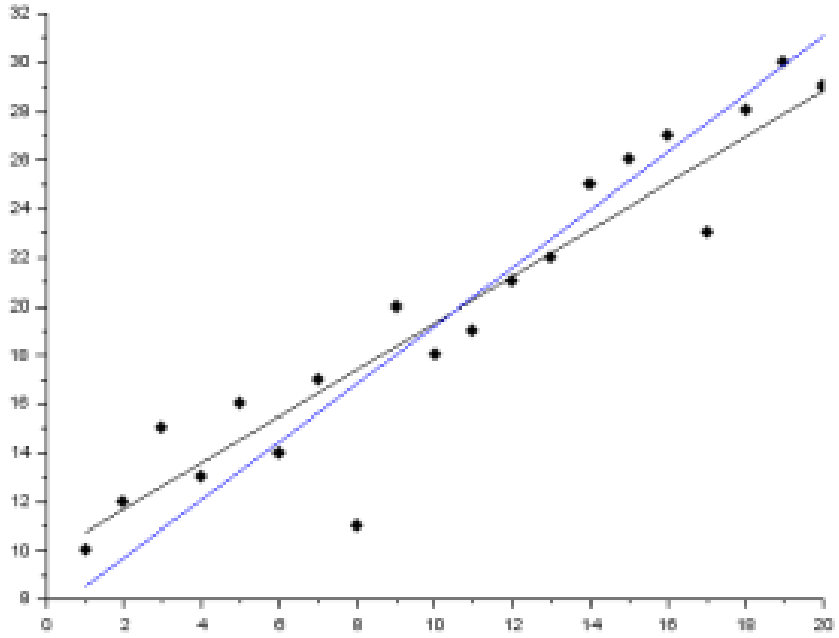
Or :

$$x = \frac{\text{cov}(X,Y)}{V(Y)}(y - \bar{Y}) + \bar{X} \Leftrightarrow y = \frac{V(Y)}{\text{cov}(X,Y)}(x - \bar{X}) + \bar{Y}$$

Y est alors la variable explicative et *X* la variable expliquée

En Scilab :

```
-->
x=1:20,y=[10,12,15,13,16,14,17,11,20,18,19,21,22,25,26,
27,23,28,30,29]

 x  =

1.   2.   3.   4.   5.   6.   7.   8.   9.   10.   11.
12.   13.   14.   15.   16.   17.   18.   19.   20.

 y  =

10.   12.   15.   13.   16.   14.   17.   11.   20.
18.   19.   21.   22.   25.   26.   27.   23.   28.
30.   29.

--> plot2d(x,y,-4),plot2d(x,corr(x,y,1)/variance(x)*(x-
mean(x))+mean(y),1),plot2d(x,variance(y)/corr(x,y,1)*(x
-mean(x))+mean(y),2)
```


32
30
28
26
24
22
20
18
16
14
12
10
8
0
2
4
6
8
10
12
14
16
18
20