Statistique descriptive

DUT Informatique, semestre 3

Version 1.0

1er septembre 2011

Ph. Roux

2009-2011

Table des matières

Avertissement

Pour bien utiliser ce polycopié, il faut le lire au fur et à mesure de l'avancement du cours magistral, et prendre le temps de refaire les exercices types qui y sont proposés.

- Les définitions et théorèmes sont numérotés suivant le même ordre que dans le cours magistral. Ils apparaissent dans un cadre grisé et sont en général suivis de leur démonstration, signalée par une barre dans la marge et un □ à la fin.
- La table des matières et l'index (à la fin du document) permettent de retrouver une notion précise dans ce polycopié.
- Les méthodes et techniques qui seront approfondies en TD sont signalées par un cadre (sans couleurs)
- Des exercices types corrigés, et surtout rédigés comme vous devriez le faire en DS, sont signalés par le symbole : ✎
- Les erreurs et les confusions les plus fréquentes sont signalées dans des cadres rouges avec le symbole :
-

1 Représentation des données statistiques

La statistique est la partie des mathématiques qui s'intéresse à la détermination des caractéristiques d'un ensemble de données (généralement très important). Ici nous ne nous intéresserons pas aux problèmes posés par la collecte d'un grand nombre de données, mais seulement aux deux aspects suivants :

- le traitement des données collectées qui est appelé *la statistique descriptive*
- l'interprétation des données qui est appelée *l'inférence statistique*

Il existe deux types de données statistiques :

- les caractères qualitatifs (qui ne peuvent pas être mesurées) qui constitue une nomenclature, par exemple le sexe (Masculin/Féminin), les couleurs ...
- les caractères quantitatifs (qui peuvent être mesurées) et que l'on peut représenter par un nombre réel, par exemple taille, le poids, ...

Dans ce cours nous nous intéresserons aux caractères quantitatifs, qu'on va représenter par des variables statistiques.

Définition 1.1 (série statistique) *Soit P un ensemble, appelé « population », alors un variable statistique est une application :*

$$\begin{array}{rccc} X: & P & \longrightarrow & \mathbb{R} \\ & i & \longmapsto & X(i) \end{array}$$

$X(P)$ est appelé univers image de X et on dira que :

- *X est discrète si $X(P)$ est un ensemble discret (fini ou dénombrable)*
- *X est continue si $X(P)$ est un intervalle de $\mathbb{R}$*

$N = \mathrm{Card}(P)$ sera appelé l'effectif total

Dans la pratique on s'intéresse surtout à l'univers image $X(P)$ [1] et on identifie souvent une variable statistique avec la liste des valeurs $X(i)$ prises par la variable. Dans ce cas on parle en général de *série statistique*.

✎ **1.1** Exemples de séries statistiques : on considère la population P composée des étudiants inscrits au semestre 1 du DUT INFORMATIQUE à Lannion (on admettra qu'on a un effectif total $N = 100$) et nous définissons deux variables statistiques :

- $X =$« nombre d'absence injustifiées (en maths) au semestre 1 »
- $Y =$« taille (exprimée en cm) de chaque individu »

ce sont bien des caractères quantitatifs qui peuvent donc être décrits par une série statistique, qui correspondent aux listes de nombres :
$X =$(2, 4, 2, 1, 0, 0, 3, 1, 3, 0, 2, 3, 0, 3, 1, 0, 2, 0, 2, 1, 0, 3, 1, 2, 2, 3, 1, 3, 2, 2, 0, 2, 0, 3, 2, 1, 2, 2, 3, 2, 0, 2, 0, 2, 0, 2, 1, 2, 0, 0, 3, 1, 0, 2, 0, 2, 1, 2, 2, 2, 1, 2, 0, 2, 1, 0, 0, 4, 1, 0, 3, 0, 2, 0, 0, 3, 1, 3, 0, 0, 1, 2, 1, 2, 6, 2, 2, 1, 1, 2, 3, 1, 0, 1, 1, 1, 3, 1, 3, 2)
et
$Y =$(188, 169, 181, 180, 159, 180, 164, 184, 177, 159, 187, 174, 170, 173, 175, 150, 171, 166, 173, 182, 167, 173, 174, 175, 166, 173, 180, 188, 165, 173, 178, 182, 175,

1. pour des raisons de confidentialité on ne peut pas en général identifier chaque individu de la population P

186, 162, 193, 173, 184, 184, 184, 182, 187, 193, 161, 193, 169, 163, 179, 179, 184, 182, 182, 172, 175, 193, 170, 176, 166, 177, 163, 191, 171, 189, 183, 178, 197, 166, 187, 180, 172, 181, 167, 177, 177, 186, 174, 168, 182, 183, 182, 170, 186, 193, 167, 184, 159, 169, 192, 172, 167, 178, 176, 177, 175, 172, 175, 182, 176, 179, 188)
c'est ce qu'on appelle des données brutes.

Évidement il n'est pas très parlant de représenter une série statistiques sous forme de données brutes c'est pourquoi on va utiliser diverses représentations permettant de faire ressortir la répartition des données.

Définition 1.2 (modalités)
Soit X une variable statistique alors on appelle modalités *de X :*

- *l'ensemble des valeurs différentes $\{x_1; x_2; \ldots; x_n\}$ prisent par X si X est discrète*
- *un ensemble d'intervalles $\{[x_1; x_2[; [x_2; x_3[; \ldots; [x_{n-1}; x_n]\}$ formant une partition de $X(P)$ si X est continue*

un tableau statistique est un tableau regroupant les caractéristiques d'une série statistiques par modalités :

série discrète

modalités	x_1	x_2	x_3	$\ldots$
effectif n_i				
fréquence f_i				

série continue

modalités	$[x_1; x_2[$	$[x_2, x_3[$	$\ldots$
effectif n_i			
fréquence f_i			

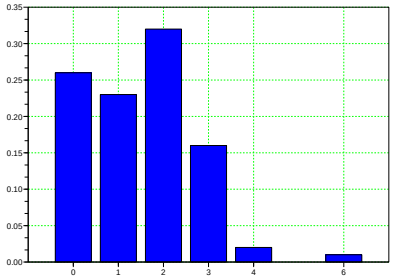
Si N est l'effectif total alors $f_i = \frac{n_i}{N}$ $\forall i = 1, \ldots, n$.

On représentera graphiquement les données d'une série discrète (resp. continue) en dessinant des colonnes de hauteurs (resp. surface) proportionnelle à l'effectif de chaque modalité, c'est ce qu'on appelle un diagramme en bâtons (resp. un histogramme).

✎ **1.2 tableau statistique d'une série statistique** Pour faire le tableau statistique de la série X, dont les modalités sont $\{0; 1; 2; 3; 4; 6\}$, il faudra à partir des données brutes construire le tableau :

$X_i =$	0	1	2	3	4	6	
n_i	26	23	32	16	2	1	$\to \sum n_i = N = 100$
f_i	0.26	0.23	0.32	0.16	0.02	0.01	$\to \sum f_i = 1$

On pourra aussi représenter graphiquement la série statistique à l'aide d'un diagramme en bâtons :

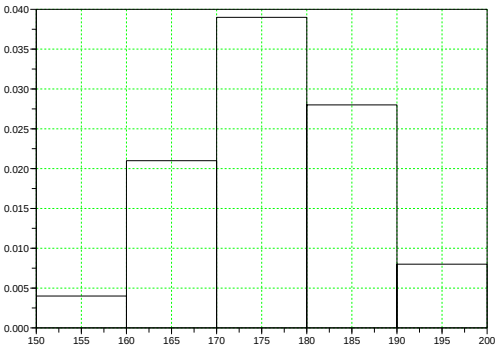


ou chaque colonne à une hauteur proportionnelle à l'effectif de la modalité.

Pour la série Y, même si les tailles ne prennent que des valeurs entières, il n'est pas opportun de la représenter comme une série discrète puisque dans ce cas on se retrouverait avec 34 modalités différentes ! C'est pour cette raison qu'on va considérer cette série comme une série continue prenant ses valeurs dans l'intervalle $[150; 200]$. Si on découpe cet intervalle en modalités de même taille on obtient :

$Y_i =$	$[150, 160[$	$[160, 170[$	$[170, 180[$	$[180, 190[$	$[190, 200[$
n_i	4	21	39	28	8
f_i	0.04	0.21	0.39	0.28	0.08

ce qui donne :



Proposition 1.3 (effectifs corrigés) *Soit X une variable statistique continue alors ayant des modalités $[x_1; x_2[$ $[x_2; x_3[$ $\ldots [x_{n-1}; x_n]$ d'amplitude différentes alors pour chaque modalité on appelle « effectif corrigé » ou « hauteur corrigée » la hauteur qu'il faut donner dans l'histogramme à la colonne pour que sa surface soit proportionnelle à son effectif réel.*

Dans un histogramme normalisé on calcule les hauteurs corrigées par la formule :

$$h_i = \frac{f_i}{|x_{i+1} - x_i|} = \frac{n_i}{N|x_{i+1} - x_i|}$$

de telle sorte que la surface totale des colonnes soit égale à 1.

Preuve : chaque modalité $[x_i, x_{i+1}[$ à pour largeur $|x_{i+1} - x_i|$ et pour hauteur h_i sa surface est donc :

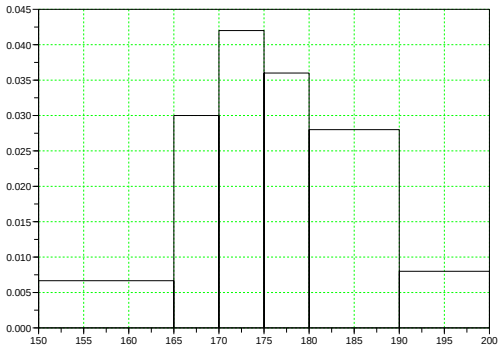
$$h_i \times |x_{i+1} - x_i| = f_i = \frac{n_i}{N}$$

est bien proportionnel à l'effectif réel, de plus il est évident que la somme des aires est alors $\sum_i f_i = 1$. □

✎ **1.3** Reprenons le cas de la variable Y que l'on étudie avec les modalités : $[150, 165[$, $[165, 170[$, $[170, 175[$, $[175, 180[$, $[180, 190[$, $[190, 200[$
Après avoir calculé les effectifs de chaque modalité, il faut cette fois calculer les hauteurs corrigées :

$Y_i =$	$[150, 165[$	$[165, 170[$	$[170, 175[$	$[175, 180[$	$[180, 190[$	$[190, 200[$
f_i	0.1	0.15	0.21	0.18	0.28	0.08
$\|x_{i+1} - x_i\|$	15	5	5	5	10	10
h_i	0.0066667	0.03	0.042	0.036	0.028	0.008

pour pouvoir dessiner l'histogramme



Un autre représentation permettant d'identifier la nature de la répartition des données statistiques est donnée par la fonction de répartition, qui est étroitement associée à la notion de cumul des fréquences.

Définition 1.4 (Fonction de répartition) *Soit X une série statistique alors la fonction de répartition de X*

$$\begin{array}{rccc} F: & \mathbb{R} & \longrightarrow & [0;1] \\ & x & \longmapsto & \displaystyle\sum_{\{i|x_i \leq x\}} f_i \end{array}$$

On retiendra que la fonction de répartition vérifie les conditions suivantes :

Proposition 1.5 *Soit F la fonction de répartition d'une série statistique X alors*
- *F est croissante*
- $\lim_{t\to-\infty} F(t) = 0$
- $\lim_{t\to+\infty} F(t) = 1$
- *si X est une variable discrète alors F est constante par morceau.*
- *si X est une variable continue alors F on la représentera par une fonction continue et affine par morceau.*

Preuve :
- si $s \leq t$ alors
$$F(s) = \sum_{\{i|x_i \leq s\}} f_i \leq \sum_{\{i|x_i \leq t\}} f_i = F(t)$$
- si $s < x_1$ (la première modalité) alors
$$F(s) = \sum_{\{i|x_i \leq s\}} f_i = 0 \Longrightarrow \lim_{s\to-\infty} F(s) = 0$$
- inversement si $t > x_n$ (dernière modalité) alors
$$F(t) = \sum_{\{i|x_i \leq t\}} f_i = f_1 + \cdots + f_n = 1 \Longrightarrow \lim_{t\to+\infty} F(t) = 1$$

Si X est une variable discrète, l'effectif d'une modalité $]-\infty, x]$ ne va varier que lorsque x prend pour valeur une des modalités x_i. Il est donc normal que le graphe de F soit constant par morceau. Par contre la représentation du graphe de F par une fonction affine par morceau dans le cas où X est une variable continue, est une approximation qui se justifiera d'un point de vue pratique (voir les exemples). □

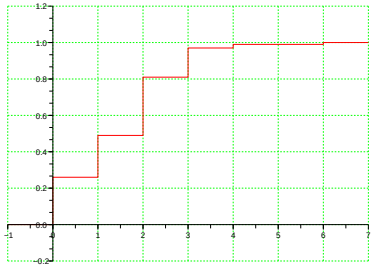
✎ **1.4** Pour construire la fonction de répartition de la variable X il faut donc faire le cumul des fréquences. Pour cela repartons du tableau statistique et ajoutons y une ligne pour calculer le cumul des fréquences :

$X_i =$	0	1	2	3	4	6
f_i	0.26	0.23	0.32	0.16	0.02	0.01
cumul	0.26	0.49	0.81	0.97	0.99	1

Cette dernière ligne donne la valeur de la fonction de répartition sur l'intervalle qui démarre à la modalité correspondante :

I	$]-\infty,0[$	$[0,1[$	$[1,2[$	$[2,3[$	$[3,4[$	$[4,6[$	$[6,+\infty[$
$F(t)$	0	0.26	0.49	0.81	0.97	0.99	1

ce qui donne :



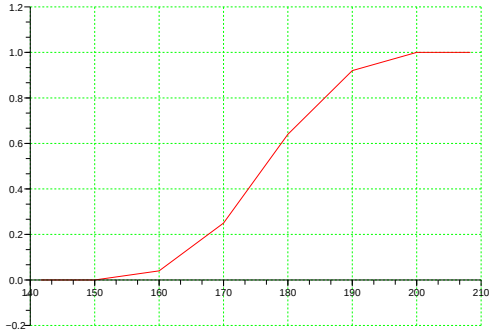
de même pour Y on repart du tableau statistique :

$Y_i =$	$[150,160[$	$[160,170[$	$[170,180[$	$[180,190[$	$[190,200[$
f_i	0.04	0.21	0.39	0.28	0.08
cumul	0.04	0.25	0.64	0.92	1

permet de calculer les valeurs de la fonction de répartition :

t	$]-\infty,150[$	160	170	180	190	200	$[200,+\infty[$
$F(t)$	0	0.04	0.25	0.64	0.92	1	1

par contre on doit représenter la fonction de répartition comme une fonction continue affine par morceau (*i.e.* son graphe est une suite de segments de droites) entre chaque borne des modalités :



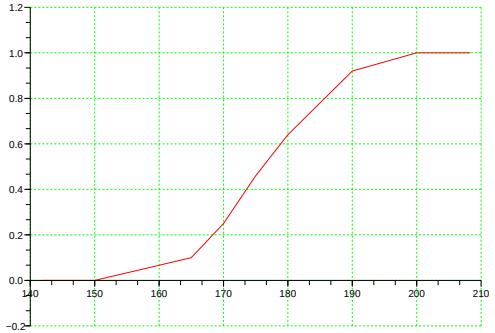
ce procédé permet d'obtenir des résultats numériquement cohérents quand on change les modalités lors de l'étude d'une série statistique. Reprenons le cas de Y avec des modalités de taille variable, ici on aura pas de problème d'effectifs corrigés, le tableau des fréquences devient :

$Y_i =$	$[150,165[$	$[165,170[$	$[170,175[$	$[175,180[$	$[180,190[$	$[190,200[$
f_i	0.1	0.15	0.21	0.18	0.28	0.08
cumul	0.1	0.25	0.46	0.64	0.92	1

ce qui donne cette fois pour la fonction de répartition :

t	$]-\infty,150[$	165	170	175	180	190	200	$[200,+\infty[$
$F(t)$	0	0.1	0.25	0.46	0.64	0.92	1	1

et pour le graphe :



2 Paramètres statistiques

L'analyse des données statistiques doit permettre d'isoler certaines valeurs remarque définissant la manière dont son réparties les données statistiques. Le plus connu est la moyenne empirique, mais il y a aussi la variance, la corrélation ….

Dans le cas d'une variable continue, il n'est pas possible de calculer la valeur exacte des paramètres statistiques, puisqu'on ne connait plus exactement les valeurs valeurs mais seulement leur nombre dans un intervalle donné ! Pour chaque indicateur statisque on aura donc deux définitions différentes : une dans le cas des séries discrètes et l'autre dans le cas des séries continues regroupées par intervalles.

La moyenne permet de cerner facilement la valeur centrale de la série statistique.

Définition 2.1 (Moyenne empirique)
Soit X une série statistique alors la moyenne de X, notée $\overline{X}$, est la valeur :

$$\overline{X} = \underbrace{\frac{1}{N}\sum_{k\in P} X(k)}_{\text{données brutes}} = \underbrace{\frac{1}{N}\sum_{i=1}^{n} n_i x_i}_{\text{effectifs par modalités}} = \underbrace{\sum_{i=1}^{n} f_i x_i}_{\text{fréquences par modalités}}$$

Dans le cas d'une série continue regroupée par intervalle on a la définition :

Définition 2.2 (Moyenne approchée)
Soit X une variable continue regroupée suivant les modalités $[x_1; x_2[$ $[x_2; x_3[$ $\ldots[x_{n-1}; x_n]$. Si on pose $xc_i = \frac{x_{i+1}+x_i}{2}$ le milieu de chaque modalité alors la moyenne approchée est définie par

$$\text{moyenne approchée } = \sum_{i=1}^{n} f_i \times xc_i$$

✎ **2.1** Pour une variable discrète X on peut calculer la moyenne directement à partir du tableau statistique :

$X_i =$	0	1	2	3	4	6
n_i	26	23	32	16	2	1
f_i	0.26	0.23	0.32	0.16	0.02	0.01

en faisant

$$\overline{X} = 0.26 \times 0 + 0.23 \times 1 + 0.32 \times 2 + 0.16 \times 3 + 0.02 \times 4 + 0.01 \times 6 = 1.49$$

de même pour la variable Y on trouvera $\overline{Y} = 176.71$. Mais dans le cas d'une variable continue comme Y, si on a pas accès aux données brutes on peut quand même calculer une valeur approchée de la moyenne à partir du tableau des fréquences :

$Y_i =$	$[150, 160[$	$[160, 170[$	$[170, 180[$	$[180, 190[$	$[190, 200[$
f_i	0.04	0.21	0.39	0.28	0.08

$$\begin{aligned}\text{moyenne approchée} &= 0.04 \times 155 + 0.21 \times 165 + 0.39 \times 175 + 0.28 \times 185 + 0.08 \times 195 \\ &= 176.5 \approx 176.71 = \overline{Y}\end{aligned}$$

la valeur est proche de la moyenne calculée avec les données brutes. Si on change les modalités on va trouver un résultat proche :

$Y_i =$	$[150, 165[$	$[165, 170[$	$[170, 175[$	$[175, 180[$	$[180, 190[$	$[190, 200[$
f_i	0.1	0.15	0.21	0.18	0.28	0.08

cette fois on trouve :

$$\begin{aligned}&\text{moyenne approchée} \\ &= 0.1 \times 157.5 + 0.15 \times 167.5 + 0.21 \times 172.5 + 0.18 \times 177.5 + 0.28 \times 185 + 0.08 \times 195 \\ &= 176.45 \approx 176.71 = \overline{Y}\end{aligned}$$

On retiendra aussi que si on manipule plusieurs séries statistiques sur la même population alors la moyenne est linéaire :

Proposition 2.3 (linéarité de la moyenne)
Soit X, Y deux séries statistiques (sur une même population) et $a, b \in \mathbb{R}$ alors

$$\overline{aX + bY} = a\overline{X} + b\overline{Y}$$

en particulier $\overline{aX + b} = a\overline{X} + b$.

La moyenne n'est pas suffisante pour comprendre la manière dont sont réparties les valeurs d'une série statistique X. On a besoin d'avoir une indication sur la dispersion des données. Le meilleur de ces indicateurs est l'écart-type qui est définit comme suit.

Définition 2.4 (variance et écart-type) *Soit X une série statistique de moyenne $\mu = \overline{X}$ alors on appelle variance de X la valeur*

$$\mathbb{V}\text{ar}(X) = \overline{(X - \mu)^2}$$

et l'écart-type de X est $\sigma_X = \sqrt{\mathbb{V}\text{ar}(X)}$.

On remarquera qu'on a toujours $\mathbb{V}\text{ar}(X) \geq 0$ puisque c'est la moyenne de nombres positifs et l'écart-type est donc toujours bien défini !

L'écart-type σ donne une estimation de l'intervalle $[\overline{X} - \sigma; \overline{X} + \sigma]$ centré sur la moyenne où sont concentrées les 2/3 des données. il permet donc de comprendre à quel point les données sont concentrées (ou pas) autour de la moyenne.

Pour une variable continue si on a pas accès aux données brutes on pourra aussi calculer une variance et un écart-type approché (comme pour la moyenne approché).

Dans la pratique on calculera la variance d'une série via le théorème suivant.

Théorème 2.5 (Koening)
Soit X une variable statistique alors

$$\mathbb{V}\text{ar}(X) = \overline{X^2} - \overline{X}^2$$

Preuve : On a que $(X-\mu)^2 = X^2 - 2\mu X + \mu^2$ donc en utilisant la linéarité de la moyenne on obtient

$$\begin{aligned}\mathbb{V}\text{ar}(X) &= \overline{(X-\mu)^2}\\ &= \overline{X^2 - 2\mu X + \mu^2}\\ &= \overline{X^2} - 2\mu\overline{X} + \mu^2\\ &= \overline{X^2} - 2\overline{X}\times\overline{X} + \overline{X}^2 \quad \text{car } \mu = \overline{X}\\ &= \overline{X^2} - \overline{X}^2\end{aligned}$$

□

Dans la démonstration de la formule de Koening si on remplace μ par $\overline{X}$ il faut bien faire attention que :

$$\overline{\overline{X}^2} = \overline{X}^2 \neq \overline{X^2} \quad et \quad \overline{\overline{X}X} = \overline{X}\,\overline{X} = \overline{X}^2 \neq \overline{X^2}$$

2.2 Calculons la variance de X à partir du tableau statistique :

$X_i =$	0	1	2	3	4	6
n_i	26	23	32	16	2	1
f_i	0.26	0.23	0.32	0.16	0.02	0.01

on a déjà obtenu que $\overline{X} = 1.49$ de même on obtient

$$\begin{aligned}\overline{X^2} &= 0.26\times 0 + 0.23\times 1 + 0.32\times 4 + 0.16\times 9 + 0.02\times 16 + 0.01\times 36\\ &= 3.63\end{aligned}$$

donc en utilisant le théorème de Koening on en tire pour la variance $\mathbb{V}\text{ar}(X) = 3.63 - 1.49^2 = 3.63 - 2.2201 = 1.4099$. On aurait trouvé le même résultat en calculant :

$$\begin{aligned}\mathbb{V}\text{ar}(X) &= 0.26\times(0-1.49) + 0.23\times(1-1.49) + 0.32\times(2-1.49)\\ &\quad +0.16\times(3-1.49) + 0.02\times(4-1.49) + 0.01\times(6-1.49)\\ &= 1.4099\end{aligned}$$

dans tous les cas on trouve l'écart-type $\sigma_X = \sqrt{1.4099} = 1.1873921$.

Pour les cas où on a pas directement accès aux données brutes (en particulier pour les variables statistiques continues) il existe d'autres indicateurs de centralité et de dispersion qui peuvent être calculés à partir de l'histogramme ou du graphe de la fonction de répartition .

Définition 2.6 (Mode et médiane)
Soit X une série statistique alors

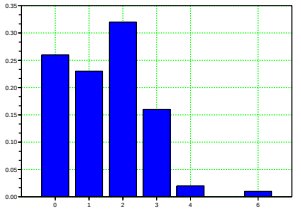
- *Le mode est la modalité la plus représentée (effectif maximum)*
- *la médiane m est la valeur qui permet de couper la population en 2 part égales : $\{i|X(i) \geq m\}$ et $\{i|X(i) \leq m\}$*
  - *si l'effectif total $N = 2p+1$ est impair alors la médiane est la $p+1^{ième}$ valeur*
  - *si l'effectif total $N = 2p$ est pair alors la médiane est la moyenne des $p^{ième}$ et $p+1^{ième}$ valeurs*

On peut aussi caractériser la médiane à partir de la fonction de répartition, ce qui permet d'en calculer une valeur approchée si on a pas accès aux donnée brutes.

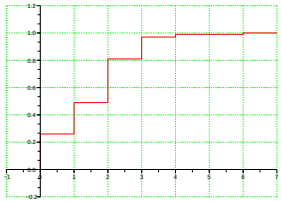
Proposition 2.7 (médiane et fonction de répartition) *Soit X une série statistique alors et F la fonction de répartition associée comme étant la solution de $F(x) = 0.5$ que l'on pourra déterminer graphiquement pour une variable statistique continue. Pour une variable statistique discrète la médiane il y a 2 cas :*

- *la première valeur telle que $F(x) > 0.5$ (si $F(x) = 0.5$ n'a pas de solution)*
- *la moyenne des valeurs telle que $F(x) = 0.5$ (si $F(x) = 0.5$ a plusieurs solutions)*

2.3 On reprend la variable X (nombre d'absences injustifiées) la moyenne est $\overline{X} = 1.49$ le mode et la médiane se trouvent facilement sur l'histogramme et le graphe de la fonction de répartition



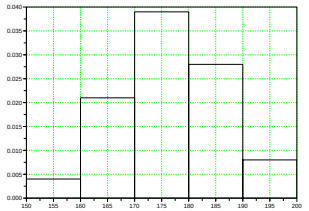
mode 2

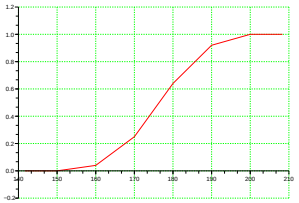


mediane 2

on peut remarquer qu'ici il n'y a pas de solution à l'équation $F(x) = 0.5$ et que la première valeur telle que $F(x) > 0.5$ est bien $x = 2$ d'après le tableau :

Pour la série continue Y on a pour moyenne brute $\overline{Y} = 176.71$.



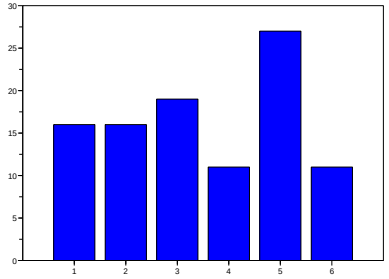


mode $175 \in [170, 180[$ mediane 177

alors que graphiquement on trouve que la médiane vaut ≈ 176.4.
La médiane est en général un bon indicateur de centralité mais sa détermination se fait graphiquement (dont de manière imprécise), au contraire le mode est plus facile à déterminer mais peut être très éloigné de la moyenne pour des séries fortement dispersées. Par exemple pour une série Z de résultats de $N = 100$ lancés d'un dé à 6 faces on obtient :

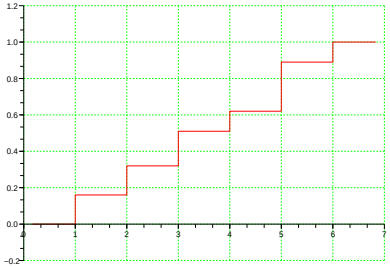
$Z_i =$	1	2	3	4	5	6
n_i	16	16	19	11	27	11
f_i	0.16	0.16	0.19	0.11	0.27	0.11

effectif total $N = 100$



le mode sera donc 5 assez éloigné de la moyenne empirique $\overline{Z} = 3.5$. Par contre le calcul de la médiane donne 3 d'après la fonction de répartition :

I	$]-\infty, 1[$	$[1, 2[$	$[2, 3[$	$[3, 4[$	$[4, 5[$	$[5, 6[$	$[6, +\infty[$
$F(t)$	0	0.16	0.32	0.51	0.62	0.89	1



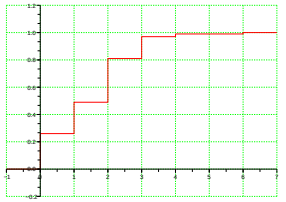
Définition 2.8 (quartiles)
Soit X une série statistique alors on définie les quartiles Q_1, Q_3 à partir de la fonction de répartition de la manière suivante :

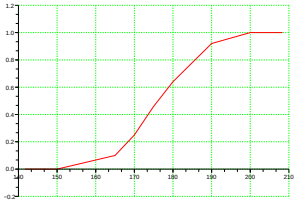
- *Q_i est la solution de $F(Q_1) = 0.25$ et $F(Q_3) = 0.75$ ou à défaut (pour une variable discrète) :*
  - *$Q_i = x$ la première valeur telle que $F(x) > 0.25 \times i$ (si $F(x) = 0.25 \times i$ n'a pas de solution)*
  - *la moyenne des valeurs telle que $F(x) = 0.25 \times i$ (si $F(x) = 0.25 \times i$ a plusieurs solutions)*

l'inter-quartile est la valeur $Q_3 - Q_1$

La définition des quartiles est très similaire à celle de la médiane (d'ailleurs la médiane peut être vue comme le quartile Q_2). L'inter-quartile est un indicateur de dispersion, il donne la largeur d'un intervalle $[Q_1; Q_3]$ contenant plus de 50% des données. De ce fait en général l'inter-quartile est $\approx 1.3333\sigma_X$.

2.4 on peut reprendre le calcul des inter-quartile pour les séries X et Y :





- écart-type $\sigma_X = 1.1873921$
- inter-quartille $q_3 - q_1 = 2 - 0 = 2$

- écart-type $\sigma_Y = 9.1129523$
- inter-quartille $q_3 - q_1 = 12.5$

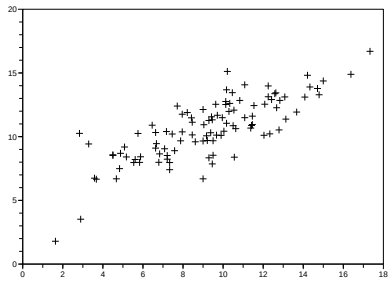
Pour une variable statistique continue la médianes et les quartilles sont calculées graphiquement à partir du graphe de la fonction de répartition (ou plutôt de son approximation affine).

3 Corrélation

En statistique il est souvent important de rechercher s'il existe un lien entre deux variables X et Y, lien qui, dans l'idéal, pourrait s'exprimer par une équation comme $Y = aX + b$. Prenons un exemple simple : on considère la population des étudiants inscrit au S1 du DUT INFORMATIQUE de Lannion en 2008 ($N = 103$ étudiants) et les deux variables aléatoires associées à cette population

$M =\ll$ note de maths au S1 $\gg$et $U =\ll$ moyenne générale au S1 $\gg$

On voudrait quantifier le lien entre ces deux variables statistiques. Si on place les points de coordonnées (M_i, U_i) sur un graphe on obtient le "nuage" suivant :



ces points semblent grossièrement alignés sur une droite, cela signifie qu'on peut quantifier le lien entre M et U par une équation de la forme $U = aM + b$. Le but est donc de trouver les coefficients a et b de telle sorte que la droite $y = ax + b$ passe au plus près du maximum de points comme sur la figure FIG.1. C'est ce qu'on appelle faire une *régression linéaire*.

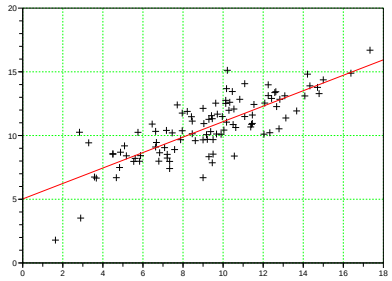


FIGURE 1 – droite de régression

Cette méthode empirique n'est pas satisfaisante, pour obtenir une méthode plus consistante nous allons devoir définir une nouvelle quantité statistique : la

corrélation.

Définition 3.1 (corrélation) *Soient X et Y deux séries statistiques d'effectif N alors la corrélation de X et Y est la valeur*

$$\sigma_{X,Y} = \frac{1}{N}\sum_{i=1}^{N}(X_i - \overline{X})(Y_i - \overline{Y})$$

Cette définition est une généralisation de la définition de la variance , d'ailleurs on peut remarquer que $\sigma_{X,X} = \mathbb{V}\mathrm{ar}(X) = \sigma_X^2$. Et comme pour la variance on a une formule de Koening pour simplifier le calcul de la covariance :

Théorème 3.2 (Koening) *Soient X et Y deux séries statistiques de moyennes $\overline{X}$ et $\overline{Y}$ alors*

$$\sigma_{X,Y} = \overline{XY} - \overline{X}\,\overline{Y}$$

Preuve : La démonstration de cette formule est très similaire à c'autre formule de Koening exprimant la variance d'une série statistique en fonction de la moyenne des carrés :

$$\begin{aligned}
\sigma_{X,Y} &= \overline{(X - \overline{X})(Y - \overline{Y})} \\
&= \overline{XY - \overline{X}Y - X\overline{Y} + \overline{X} \times \overline{Y}} \\
&= \overline{XY} - \overline{X} \times \overline{Y} - \overline{X} \times \overline{Y} + \overline{X} \times \overline{Y} \\
&= \overline{XY} - \overline{X} \times \overline{Y}
\end{aligned}$$

□

On a aussi l'inégalité suivante qui va prendre tout son sens plus loin.

Proposition 3.3 (corrélation) *Soient X et Y deux séries statistiques d'écart-type σ_X et σ_Y alors*

$$|\sigma_{X,Y}| \leq \sigma_X \sigma_Y$$

Preuve : On pose $S = (X - \overline{X})$ et $T = (Y - \overline{Y})$ il est alors évident que

$$\overline{S^2} = \overline{(X - \overline{X})^2} = \mathbb{V}\mathrm{ar}(X), \quad \overline{T^2} = \overline{(Y - \overline{Y})^2} = \mathbb{V}\mathrm{ar}(Y), \quad \overline{ST} = \sigma_{X,Y}$$

Maintenant on décide de calculer :

$$\overline{(S + \lambda T)^2} = \overline{S^2 + 2\lambda ST + \lambda^2 T^2} = \overline{S^2} + 2\lambda\overline{ST} + \lambda^2\overline{T^2} = \sigma_X^2 - 2\lambda\sigma_{X,Y} + \lambda^2\sigma_Y^2$$

c'est donc un trinôme du second degré par rapport à la variable λ, mais ce polynôme est toujours positif car $\overline{(S + \lambda T)^2} \geq 0$, donc son discriminant est négatif ou nul :

$$\sigma_X^2 - 2\lambda\sigma_{X,Y} + \lambda^2\sigma_Y^2 \geq 0 \Longrightarrow \sigma_{X,Y}^2 - \sigma_X^2\sigma_Y^2 \leq 0 \Longrightarrow |\sigma_{X,Y}| \leq \sigma_X\sigma_Y$$

□

Maintenant pour calculer l'équation de la "meilleure" droite passant parmi les points du nuage, on va pouvoir utiliser une méthode découverte par Gauss au début du $XIX^{\text{ième}}$ siècle : la méthode des moindres carrés. Cette méthode consiste à minimiser l'erreur globale qu'on comment en écrivant que $Y = aX + b$:

Théorème 3.4 (méthode des moindres carrés) *Soient X et Y deux séries statistiques alors la méthode des moindres carrés consiste à chercher la "meilleure" droite, d'équation $y = ax + b$, passant par le nuage de points (X, Y) comme étant la droite qui minimise la somme des carré des écarts entre les points $(X(i), Y(i))$ et $(X(i), aX(i) + b)$ c'est à dire on cherche a et b qui minimisent la fonction*

$$\phi(a,b) = \sum_{i=1}^{N} (Y(i) - aX(i) - b)^2$$

En utilisant nos connaissances en analyse, on peut maintenant trouver le minimum de la fonction $\phi(a, b)$.

Théorème 3.5 (ajustement linéaire) *Soient X et Y deux séries statistiques alors la droite d'ajustement linéaire de Y en X a pour équation $y = ax + b$ avec :*

$$a = \frac{\sigma_{X,Y}}{\mathbb{V}\mathrm{ar}(X)} \quad et \quad b = \overline{Y} - a\overline{X}$$

Enfin on peut quantifier la qualité de l'approximation $Y = aX + b$ en utilisant le coefficient de corrélation.

Théorème 3.6 (coefficient de corrélation) *Soient X et Y deux séries statistiques, on suppose que la droite d'ajustement linéaire de Y en X à pour équation $y = ax + b$ et on définit sont coefficient de corrélation par :*

$$\rho = \frac{\sigma_{X,Y}^2}{\mathbb{V}\mathrm{ar}(X)\mathbb{V}\mathrm{ar}(Y)} \in [0, 1]$$

alors le nuage de points (X, Y) est d'autant plus proche de la droite d'ajustement linéaire de Y en X que ρ est proche de 1. En particulier tous les points du nuage sont sur la droite $y = ax + b$ si et seulement si $\rho = 1$.

La corrélation sera dite forte (resp. faible) si $\rho \geq \frac{\sqrt{3}}{2} \approx 0.866$ (resp. $\rho < \frac{\sqrt{3}}{2} \approx 0.866$)

Preuve : le but est de trouver 2 équations qui permettent de trouver les a et b qui minimise la fonction $\phi(a, b)$. Pour cela on va minimiser ϕ d'abord par rapport à b et ensuite par rapport à a :

- recherche du minimum par rapport à b

$$\begin{aligned}
\phi(a,b) &= \sum_{i=1}^{N} (Y(i) - (aX(i) + b))^2 \\
&= \sum_{i=1}^{N} ((Y(i) - aX(i)) - b)^2 \\
&= \sum_{i=1}^{N} (Y(i) - aX(i))^2 - 2b(Y(i) - aX(i)) + b^2 \\
&= N\left(b^2 - 2b\overline{Y - aX} + \overline{(Y - aX)^2}\right)
\end{aligned}$$

ce trinôme atteint son minimum quand $b = \overline{Y - aX} = \overline{Y} - a\overline{X}$, on peut remarquer que cette équation signifie que la "meilleur droite passe par le point "moyen" $(\overline{X}, \overline{Y})$ puisque $\overline{Y} = a\overline{X} + b$.

- recherche du minimum par rapport à a

$$\begin{aligned}
\phi(a,b) &= \sum_{i=1}^{N} (Y(i) - (aX(i) + b))^2 \\
&= \sum_{i=1}^{N} ((Y(i) - aX(i)) - (\overline{Y} - a\overline{X}))^2 \\
&= \sum_{i=1}^{N} ((Y(i) - \overline{Y}) - a(X(i) - \overline{X}))^2 \\
&= \sum_{i=1}^{N} (Y(i) - \overline{Y})^2 - 2a(Y(i) - \overline{Y})(X(i) - \overline{X}) + a^2(X(i) - \overline{X})^2 \\
&= N\left(a^2\overline{(X - \overline{X})^2} - 2a\overline{(Y - \overline{Y})(X - \overline{X})} + \overline{(Y - \overline{Y})^2}\right) \\
&= N\left(a^2\mathbb{V}\mathrm{ar}(X) - 2a\sigma_{X,Y} + \mathbb{V}\mathrm{ar}(Y)\right)
\end{aligned}$$

ce trinôme atteint son minimum quand $a = \frac{\sigma_{X,Y}}{\mathbb{V}\mathrm{ar}(X)}$

- Pour finir calculons la somme des écarts pour la "meilleure" droite :

$$\begin{aligned}
\phi(a,b) &= N\left(\frac{\sigma_{X,Y}^2}{\mathbb{V}\mathrm{ar}(X)^2}\mathbb{V}\mathrm{ar}(X) - 2\frac{\sigma_{X,Y}^2}{\mathbb{V}\mathrm{ar}(X)} - \mathbb{V}\mathrm{ar}(Y)\right) \\
&= N\left(\mathbb{V}\mathrm{ar}(Y) - \frac{\sigma_{X,Y}^2}{\mathbb{V}\mathrm{ar}(X)}\right) = N\mathbb{V}\mathrm{ar}(Y)\left(1 - \frac{\sigma_{X,Y}^2}{\mathbb{V}\mathrm{ar}(X)\mathbb{V}\mathrm{ar}(Y)}\right) \\
&= N\mathbb{V}\mathrm{ar}(Y)(1 - \rho)
\end{aligned}$$

$\phi(a, b)$ étant toujours positive (par définition) on doit avoir que $\rho \leq 1$, ce qui est bien vérifié d'après la proposition 3.3. Maintenant on en déduit facilement que $\rho = 1 \iff \phi(a, b) = 0$ ce qui signifie bien que tous les écarts sont nuls si $\rho = 1$.

□

✎ **3.1** Reprenons l'exemple des notes du premier semestre :

$M =$« note de maths au S1 » et $U =$« moyenne générale au S1 »

et appliquons la méthode des moindres carrés. On calcule d'abord les différents paramètres :

- $\overline{M} = 9.2500971$ $\mathbb{V}\mathrm{ar}(M) = 9.8811854$ et $\sigma_M = 3.1434353$
- $\overline{U} = 10.636117$ $\mathbb{V}\mathrm{ar}(U) = 5.7622801$ et $\sigma_U = 2.400475$
- $\overline{MU} = 104.37805$

On en déduit l'ajustement de U par rapport à M

- $a = \frac{\sigma_{M,U}}{\mathbb{V}\mathrm{ar}(U)} = \frac{5.9929373}{3.1434353^2} = 0.6064998$
- $b = \overline{U} - a\overline{M} = 10.636117 - 0.6064998 \times 9.2500971 = 5.0259342$
- $\rho = \frac{\sigma_{M,U}^2}{\mathbb{V}\mathrm{ar}(M)\mathbb{V}\mathrm{ar}(U)} = \frac{5.9929373^2}{9.8811854 \times 5.7622801} = 0.6307773$ (corrélation faible)

ce qui donne :

$$U = 0.6064998M + 5.0259342$$

qui correspond bien à ce qu'on obtient graphiquement (voir droite rouge sur les figures FIG.1 et FIG.2).

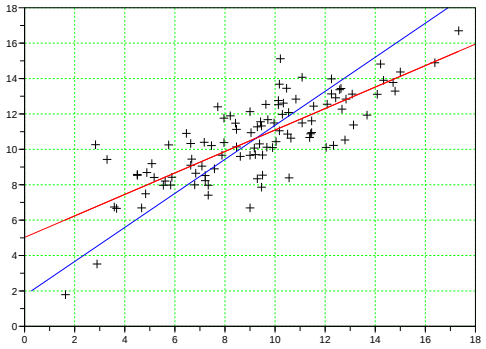


FIGURE 2 – comparaison des ajustements de U par rapport à M (en rouge) et de M par rapport à n (en bleu) avec une corrélation faible $\rho = 0.6307773$

On aurait pu de la même manière chercher à faire l'ajustement de M par rapport à U. Il suffit pour cela d'échanger les rôles de U et M. On peut remarquer que ce changement ne modifie pas la valeur de la corrélation (car $\sigma_{X,Y} = \sigma_{Y,X}$) ni du coefficient de corrélation $\rho = 0.6307773$ (corrélation faible), on peut donc reprendre ces valeurs pour calculer les nouveaux coefficients de la droite d'ajustement de M par rapport à U :

- $a = \frac{\sigma_{U,M}}{\mathrm{Var}(U)} = \frac{5.9929373}{2.400475^2}$
- $b = \overline{M} - a\overline{U} = 9.2500971 - 1.0400288 \times 10.636117 = -1.8117705$

on trouverait donc :

$$M = 1.0400288U - 1.8117705$$

on remarquera que l'ajustement de X par rapport à Y n'est pas le même que celui trouvé en ajustant Y par rapport à X, on peut s'en convaincre facilement sur la figure FIG.2 ! Ceci est lié au fait que la corrélation soit faible. Les deux méthodes donnent la même équation si et seulement si $\rho = 1$.

4 Calculs statistiques avec Scilab

À chaque TP vous trouverez un fichier `*.sce` contenant des séries statistiques sous forme de d'une matrice à une ligne ou 1 colonne :

```
X=[11;3;1;10;3;1;13;2;1;10;2;4;4;9;4;10; ...]
```

Il vous suffira de charger cette matrice dans l'environnement scilab pour pouvoir commencer l'étude de la série statistique.

4.1 séries statistiques discrètes

- **L'effectif total** N de la variable statistique X (le nombre de valeurs valeurs) s'obtient en récupérant la taille de la matrice :

```
-->X
 X  =
    11.
    3.
    1.
    10.
    3.
    1.
[More (y or n ) ?]
```

```
-->size(X)
 ans  =

    2000.    1.

-->N=length(X)
 N  =

    2000.
```

- Scilab permet de calculer facilement **les effectifs** n_i **et les fréquences** f_i des modalités de la variable X avec la commande `tabul` (il y a aussi `nfreq` mais moins pratique) :

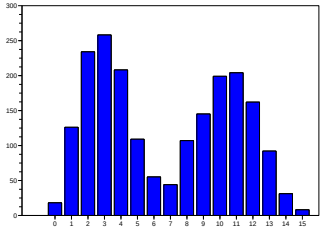
```
ordre croissant
-->m=tabul(X,'i')
 m  =

    0.     18.
    1.     126.
    2.     234.
    3.     258.
    4.     208.
    5.     109.
    6.     55.
    7.     44.
    8.     107.
    9.     145.
    10.    199.
    11.    204.
    12.    162.
    13.    92.
    14.    31.
    15.    8.
```

```
valeurs
-->x=m(:,1)
 x  =

    0.
    1.
    2.
    3.
    4.
    5.
    6.
    7.
    8.
    9.
    10.
    11.
    12.
    13.
    14.
    15.
```

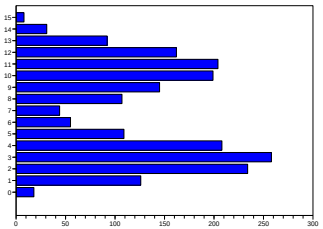
```
effectifs
-->n=m(:,2)
 n  =

    18.
    126.
    234.
    258.
    208.
    109.
    55.
    44.
    107.
    145.
    199.
    204.
    162.
    92.
    31.
    8.
```

```
fréquences
-->f=n/N
 f  =

    0.009
    0.063
    0.117
    0.129
    0.104
    0.0545
    0.0275
    0.022
    0.0535
    0.0725
    0.0995
    0.102
    0.081
    0.046
    0.0155
    0.004
```

- On peut ensuite tracer différents types d'**histogrammes** de la variable X avec les commandes `bar`, `barh` pour les diagrammes en batôn :

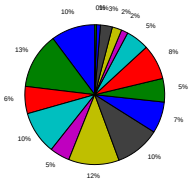


`bar(x,f)`



`barh(x,f)`

et `pie` pour les « camenberts » :



`pie(f)`

`pie(f,bool2s((x==3)|(x==10)),string(x))`

- Scilab permet de calculer facilement les principaux **estimateurs de centralité** avec `mean` et `mediane` ou par calcul :

```
-->median(X)//médiane
 ans  =

    6.
```

```
-->mean(X)//moyenne
 ans  =

    6.736
```

```
-->sum(X)/N
 ans  =

    6.736
```

- On peut facilement calculer les **estimateurs de dispersion** définis en cours :

```
-->//calcul de la variance
-->sum((X-mean(X)).^2)/N
 ans  =

    16.377304
```

```
-->//par la formule de Koening
-->mean(X.^2)-mean(X)^2
 ans  =
    16.377304

-->sqrt(ans)//écart-type
 ans  =
    4.0468882
```


Attention scilab possède aussi une fonction `variance` et une fonction écart-type (`stdev` mais qui utilisent **une définition légèrement différente** de celle vue en cours.

```
-->//estimateurs non-biaisés
-->variance(X)
 ans  =

    16.385497

-->stdev(X)//écart-type
 ans  =

    4.0479003
```

```
-->// vérification
-->sum((X-mean(X)).^2)/(N-1)
 ans  =

    16.385497

-->sqrt(ans)
 ans  =

    4.0479003
```

En fait ce sont les **estimateurs non-biaisés** qui sont calculés par Scilab.

- On peut enfin calculer **les quartiles et les déciles** :

```
-->//quartiles
-->Q=quart(X)
 Q  =

    3.
    6.
    10.

-->//inter-quartile
-->iqr(X)
 ans  =

    7.

-->//vérification
-->Q(3)-Q(1)
 ans  =

    7.
```

```
-->p=perctl(X,90)//9ième décile
 p  =

    12.    822.

-->p=perctl(X,[10:10:100])
 p  =

    2.     1466.
    3.     1817.
    3.     298.
    4.     483.
    6.     371.
    9.     1397.
    10.    968.
    11.    1152.
    12.    822.
    15.    229.
```

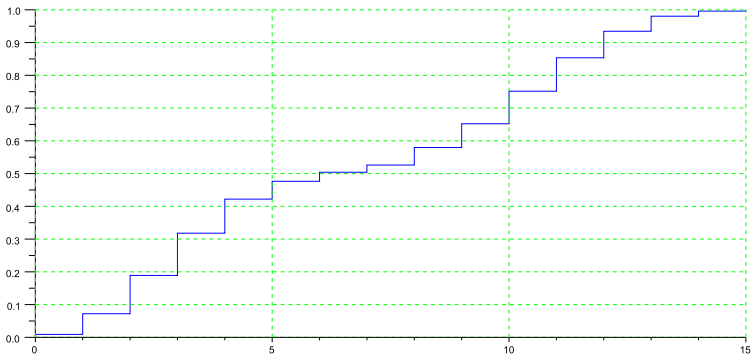
la deuxième colonne du résultat renvoyé par `perctl` correspond à la position du décile (la valeur dans la première colonne) dans la série statistique (matrice X).

- Le calcul des **effectifs cumulés** peut se faire à partir de la matrice contenant les effectifs avec la fonction `cumsum` :

```
cumul des effectifs
-->F=cumsum(n)/N
 F  =

    0.009
    0.072
    0.189
    0.318
    0.422
    0.4765
    0.504
    0.526
    0.5795
    0.652
    0.7515
    0.8535
    0.9345
    0.9805
    0.996
    1.
```

```
cummul fréquence
-->F=cumsum(f)
 F  =

    0.009
    0.072
    0.189
    0.318
    0.422
    0.4765
    0.504
    0.526
    0.5795
    0.652
    0.7515
    0.8535
    0.9345
    0.9805
    0.996
    1.
```

- on peut ensuite tracer le **graphe de la fonction de répartition d'une série statistique discrète** avec la commande `plot2d2`



`plot2d2(x,F,2),xgrid(3)`

Attention, cette commande ne dessine le graphe de la fonction de répartition F que sur l'intervalle $[\min_i(x_i), \max_i(x_i)]$ ce graphe doit être prolongé par :
- 0 pour $x < \min_i(x_i)$
- 1 pour $x > \max_i(x_i)$

4.2 séries statistiques regroupées par intervalles

Le cas des séries statistiques regroupées par intervalles est plus compliqué :

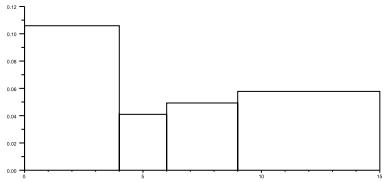
- On peut reprendre l'étude de la série X en groupant les valeurs en 4 intervalles : $[0,4]$ $]4,7]$ $]7,11]$ et $]11,15]$. Pour cela il faut définir une matrice I contenant les bornes des intervalles

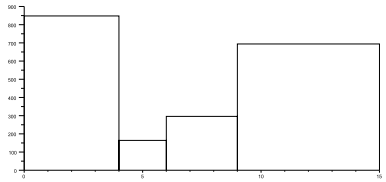
```
intervalles
-->I=[0,4,6,9,15]
 I  =

    0.    4.    6.    9.    15.
```

```
les centres des intervalles
-->xc=[2 5 7.5 12]
 xc  =

    2.    5.    7.5    12.
```

attention il y a une entrée de plus dans la matrice I que le nombre d'intervalles !

- on peut faire l'histogramme en regroupant les valeurs par classes avec `histplot`. Dans l'histogramme obtenu **la surface de chaque colonne est proportionnelle à l'effectif par défaut** et l'échelle de l'axe des ordonnées est telle que la surface des colonnes vaut 1. Si on veut avoir les effectifs réels en ordonnée (au lieu des effectifs corrigés) il faut rajouter l'option `normalization=%f` :





`histplot([0,4,7,11,15],X)` `histplot([0,4,7,11,15],X,normalization=%f)`

- si on veut récupérer les valeurs exactes des effectifs et des fréquences il faut utiliser `dsearch` :

```
-->[ind,n]=dsearch(X,I);n
 n  =

    848.    164.    296.    694.
```

```
-->f=n/N
 f  =

    0.4235764    0.0819181    0.1478521    0.3466533
```

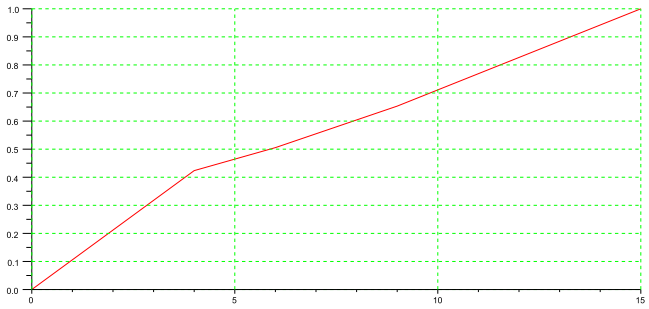
on peut vérifier que les effectifs corrigés sont bien obtenus en divisant les effectifs réels par la longueur de l'intervalle correspondant :

```
longueur des intervalles
-->l=[4 2 3 6]
 l  =

    4.    2.    3.    6.
```

```
fréquences corrigées
-->h=f./l
 h  =

    0.1058941    0.0409590    0.0492840    0.0577756
```

- on peut ensuite calculer les fréquences cumulées et tracer le **graphe de la fonction de répartition d'une série statistique continue** avec les commandes `cumsum` et `plot2d` :

```
-->F=cumsum(f)
 F  =

  0.4235764   0.5054945   0.6533467    1.

-->plot2d(I,[0 F],5),xgrid(3)
```



ce qui permet de déterminer graphiquement médiane et quartiles.

La matrice I définissant les bord intervalles possède une case de plus que la matrice F (pour k intervalles il y a $k+1$ bornes). Il faut donc rajoputer une valeur dans F : c'est le 0 dans `[0 F]` qui la valeur de départ (donc à gauche) des fréquences cumulées.

- on peut aussi calculer les valeurs approchées des différents paramètres statistiques en remplaçant les modalités discrètes x par les centres des intervalles xc. On peut comparer les valeurs trouvées aux valeurs exactes :

```
moyenne approchée
-->M=sum(xc.*f)
 M  =

    6.5254745
```

```
variance approchée
-->sum(f.*(xc-M).^2)
 ans  =

    19.395205
```

```
écart-type approché
-->sqrt(ans)
 ans  =

    4.4039988
```


Index