QUELQUES GRAPHIQUES DE PLUS !

Montrer et voir en statistique

Gérard CHAUVAT
IUT - GEII de Tours

Bernard Parzysz, dans son article paru dans cette même revue [4], montre bien les avantages et les inconvénients du résumé graphique en statistique, en insistant sur la "bonne adéquation" de celui-ci aux données.

Un certain nombre de règles, rappelées notamment par J.F. Pichard [5], permettent de réaliser et de décoder correctement ces graphiques statistiques selon le type du caractère statistique étudié. Cependant, il arrive parfois que ces règles ne soient pas clairement définies pour certains types de graphique ou que des graphiques pourtant réalisés selon les règles de l'art ne permettent pas de conclure franchement.

Je vais illustrer ces difficultés à travers trois exemples classiques mettant en jeu à chaque fois un "petit graphique" de plus…

Le polygone des densités

Le premier exemple peut être trouvé dans tout ouvrage d'initiation à la statistique qui cherche à alerter le débutant sur les difficultés issues du choix de classes d'amplitudes inégales pour résumer un caractère statistique quantitatif continu ou traité comme tel. Un caractère est considéré comme continu lorsqu'il est susceptible de prendre toutes les valeurs d'un intervalle réel (non vide, non réduit à un point !).

La représentation graphique d'un tel caractère est l'histogramme. Il n'est pas inutile d'insister sur cette terminologie normée (NF X06-003 définition 2.9, p.4) puisqu'il n'est pas rare de trouver des logiciels, d'origine anglophone, qui désignent par ce terme ce qu'il vaut mieux appeler diagramme à bandes et qu'il convient de

réserver à la représentation des caractères qualitatifs.

La norme française dit : « Après avoir fait choix d'une unité sur un axe, on porte sur cet axe les limites des classes dans lesquelles on a réparti les observations et on construit une série de rectangles ayant pour base chaque intervalle de classe et ayant une aire proportionnelle à l'effectif ou à la fréquence de la classe ». Tout semble dit, et clairement, sauf la raison pour laquelle l'aire des rectangles doit être *proportionnelle* à l'effectif ou à la fréquence de la classe.

Dans la pratique, les valeurs observées d'un caractère quantitatif continu sont comptabilisées dans un certain nombre de classes : intervalles réels qui doivent former une partition de **R** ou d'un intervalle I de **R** dans lequel elles sont susceptibles d'apparaître, de sorte que le tableau de distribution du caractère observé est de la forme :

Numéro	Classe	Effectif	Fréquence
1	C_1	n_1	f_1
…			
i	C_i	n_i	f_i
…			
r	C_r	n_r	f_r
total		N	1

Avec $\bigcup_{i=1}^{i=r} C_i = I \subseteq IR$, $\forall i \neq j \;\; C_i \bigcap C_j = \varnothing$,

et : $\sum_{i=1}^{i=r} n_i = N$, $\sum_{i=1}^{i=r} f_i = 1$.

Le choix d'une unité sur le premier axe de l'histogramme permet de placer les limites des classes, chaque limite appartenant à une classe et une seule. Malheureusement la norme ne dit pas combien de classes il faut retenir, ni avec quelles bornes, ce qui n'est pas sans poser quelques problèmes dans la pratique : des choix inadéquats pouvant conduire à *montrer* des propriétés fausses, notamment en ce qui concerne le mode du caractère. S'il n'existe pas de règle universelle en la matière, il est pour le moins conseillé, lorsque c'est raisonnable, de choisir des classes d'amplitudes égales ; dans le cas contraire les risques d'erreurs de représentation sont plus grands comme on va le voir.

Que cherche-t-on à *montrer* ? Avec le regroupement en classes on a perdu les informations concernant les valeurs observées ; on connaît seulement le nombre (ou la proportion) de valeurs observées appartenant à chaque classe. C'est cette répartition du nombre de valeurs dans chaque classe qu'on voudrait montrer de façon à voir, par exemple, quelle(s) classe(s) contien(nen)t le plus de valeurs observées. Il est donc tentant de représenter le tableau de distribution (C_i ; n_i) par une réunion de segments d'ordonnée n_i au-dessus de chaque intervalle C_i. Ceci génère deux problèmes :

i) deux classes de même effectif mais telles que l'une a une amplitude double de l'autre seraient représentées par des traits situés à la même hauteur,

ii) on serait tenté, par une lecture directe abscisse-ordonnée, d'associer à chaque valeur de la classe l'effectif n_i de l'ensemble.

On remédie à ces problèmes en représentant le tableau de distribution (C_i ; n_i) par une réunion de rectangles de base C_i dont

l'aire est proportionnelle à n_i ou f_i . Cette façon de faire repose en fait sur deux hypothèses :

- l'équi-répartition (ou répartition uniforme) des valeurs observées à l'intérieur d'une même classe,
- les classes non bornées, donc d'amplitude infinie, ont un effectif nul,

et implique que la somme des aires de tous les rectangles qui constituent l'histogramme des effectifs (resp. des fréquences) est proportionnelle à N (resp. 100% ou 1).

Que représente alors la courbe qui délimite les hauteurs de ces rectangles ? Si on note a_i l'amplitude d'une classe bornée, la hauteur du rectangle correspondant doit être proportionnelle à $d_i = n_i / a_i$ (*densité d'effectifs*) ou $d_i = f_i / a_i$ (*densité de fréquences*), de sorte que la limite supérieure des rectangles représente la fonction affine par intervalle (en escalier) définie pour tout réel t par : $\varphi(t)=\sum_{i=1}^{i=r} d_i I_{Ci}(t)$, où I_{Ci} est la fonction indicatrice de la classe C_i . Ainsi, pour tout x_i appartenant à C_i , $\varphi(x_i)=d_i$ et non pas n_i ou f_i .[1]

Exemple

Une étude portant sur la durée de vie de 100 appareils électriques du même type a permis d'établir le tableau suivant. Déterminer la classe modale.

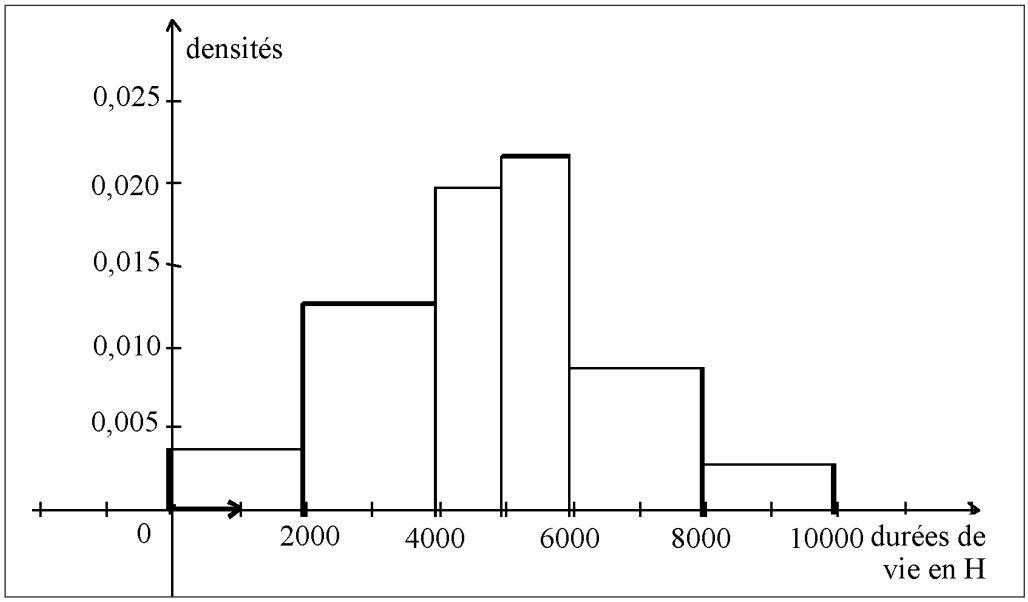
Durée de vie (en h)	Nb d'appareils
[0 ; 2000[	8
[2000 ; 4000[	26
[4000 ; 5000[	20
[5000 ; 6000[	22
[6000 ; 8000[	18
[8000 ; 10000[	6

Le mode est la valeur du caractère la plus fréquemment observée. En étendant trop vite cette définition aux classes, on peut être tenté de répondre que la classe modale de l'exemple est [2000 ; 4000[. Pour tenir compte de l'inégalité des amplitudes, on définira la classe modale (ou les classes modales) comme celle(s) de densité maximale.

Durée de vie (en h)	ni	ai	di
[0 ; 2000[	8	2000	0,004
[2000 ; 4000[	26	2000	0,013
[4000 ; 5000[	20	1000	0,020
[5000 ; 6000[	**22**	**1000**	**0,022**
[6000 ; 8000[	18	2000	0,009
[8000 ; 10000[	6	2000	0,003

La classe modale est donc l'intervalle [5000 ; 6000[ce qui se *voit* bien sur l'histogramme des densités d'effectifs ci-après. Par contre, il n'est pas forcément facile de décider, à l'aide de ce graphique, lequel des rectangles

1 A ce propos et afin d'éviter cette confusion de lecture, il me paraît utile d'insister sur la nécessité de porter les densités en ordonnées, même dans le cas où les amplitudes des classes sont égales, contrairement à l'usage qui prend en compte le fait que les densités sont alors proportionnelles aux effectifs…



associés respectivement à [4000 ; 5000[et [6000 ; 8000[a la plus grande aire…

Finalement, en respectant les règles énoncées dans les normes françaises, on construit un graphique qui *montre* correctement la répartition des valeurs d'un caractère quantitatif continu. La connaissance de ces règles est également nécessaire pour *voir* à bon escient ; en particulier pour ne pas se méprendre sur le sens de la courbe qui limite supérieurement l'histogramme. La notion de densité (d'effectifs ou de fréquences) apparaît ici comme fondamentale d'autant plus qu'on sait qu'en statistique mathématique les caractères quantitatifs continus sont modélisés par des aléas numériques définis par leur fonction densité de probabilité pour laquelle l'histogramme du caractère constitue une *bonne estimation* (pour en savoir plus, voir [2]).

Cependant, s'il paraît judicieux que l'aire totale de l'histogramme soit proportionnelle à N ou 100%, l'hypothèse de l'équi-répartition des valeurs dans chaque classe, bien que conduisant à des figures simples (rectangles) et à une fonction simple (constante par intervalle), peut être contestée. C'est ici que les manuels de statistique vous proposent… un petit graphique de plus !

C'est en effet souvent ainsi qu'apparaît le polygone des effectifs : juste un petit graphique supplémentaire, simple à réaliser, pour mieux suggérer l'évolution des effectifs ; l'évolution discontinue, par paliers, de l'histogramme paraissant choquante. En fait cet objet n'est pas toujours facile à manipuler…

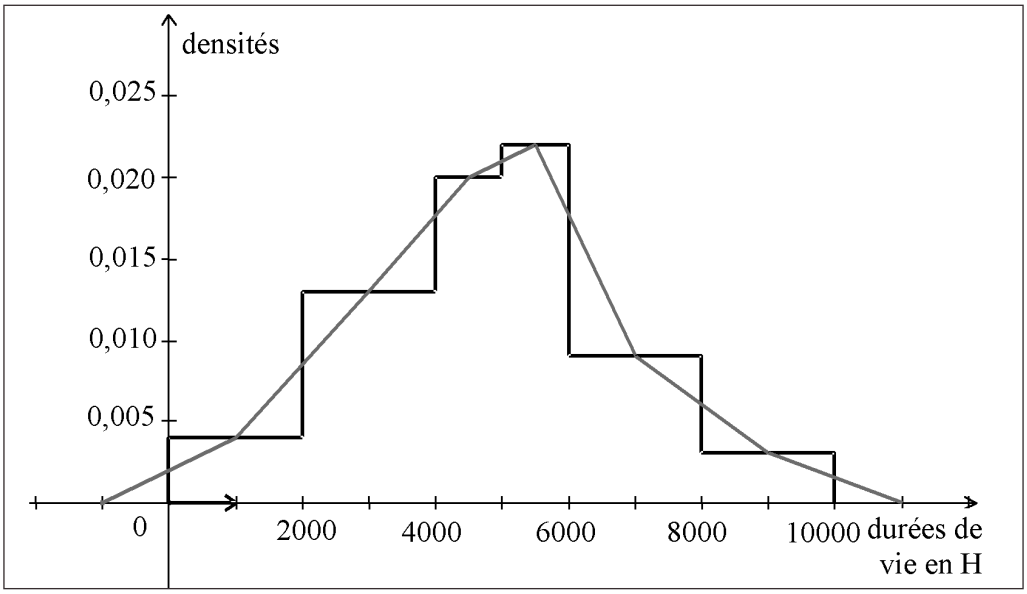
Les normes françaises se gardent bien de définir le polygone des effectifs ; on trou-

vera seulement la définition (2.12 p.4) du polygone des effectifs (ou des fréquences) **cumulé(e)s**. Ce dernier permet d'estimer la fonction de répartition théorique et s'avère plus utile et plus facile à interpréter que le polygone des effectifs (voir [6]).

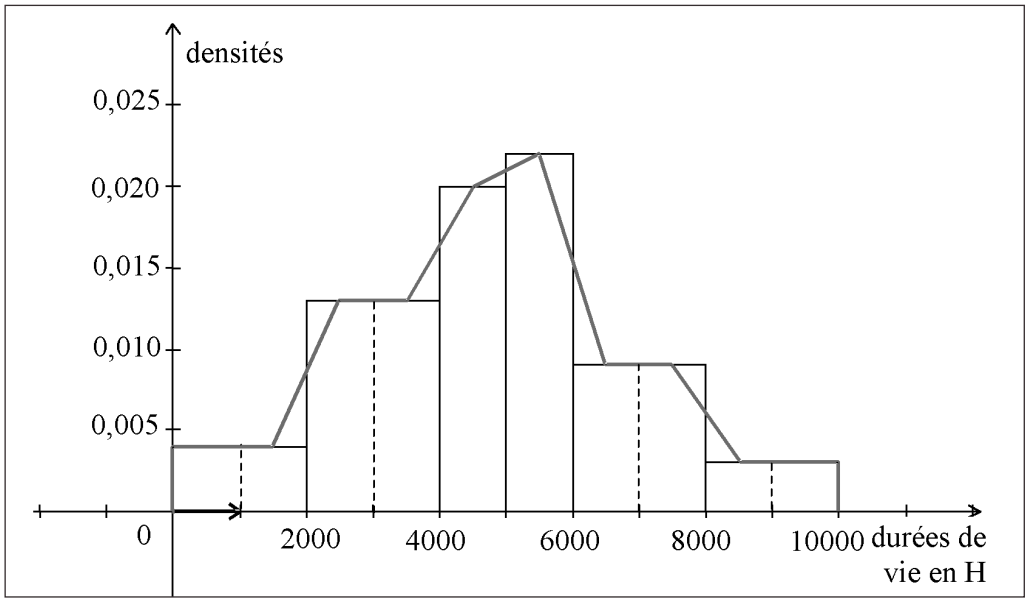
Les manuels se rallient donc à l'usage suivant : le polygone des effectifs est la ligne polygonale qui relie les points de coordonnées $(c_i\ ;\ n_i)$ où c_i est le centre de la classe C_i , *dans les cas où les classes sont bornées et d'amplitudes égales*. En fait, pour satisfaire la **propriété (P) : "l'aire totale située sous le polygone des effectifs (au-dessus de l'axe) est proportionnelle à N"**, on ajoute des classes fictives C_0 et C_{r+1}, avant la première et à la suite de la dernière classe observée, de même amplitude que les autres classes et d'effectifs nuls.

Cet ajout peut poser un problème d'interprétation : dans l'exemple ci-dessus, que signifierait une classe susceptible de contenir des durées de vie négatives ? De plus la définition précédente ne s'étend pas aux classes d'amplitudes inégales sans perdre la propriété (P). Ainsi dans la figure suivante, tracée selon ces principes, la ligne polygonale n'a pas beaucoup de sens.

Si on veut donner du sens à cette ligne, il faut au moins respecter la propriété (P) et se souvenir qu'elle représente l'évolution des densités ; il vaudrait mieux d'ailleurs l'appeler *polygone des densités d'effectifs* (ou *des fréquences*). Dans le cas d'amplitudes inégales, une solution simple consiste à déterminer, lorsque c'est possible, le *PGCD* [2] *a* des amplitudes et à subdiviser



2 au sens suivant : a est la plus grande valeur réelle telle que, pour toute amplitude ai d'une classe, il existe un entier positif k_i tel que $a_i = k_i\ a$.



toutes les classes en sous-classes d'amplitude a pour se ramener au cas des amplitudes égales. Pour l'exemple précédent, on obtient la figure ci-dessus.

B. Parzysz, dans l'article cité plus haut, note bien que l'interprétation du polygone des effectifs (dans le cas d'égalité des amplitudes des classes) repose sur l'hypothèse « que la distribution est *affine* du centre d'une classe à celui de la suivante ». Mais pourquoi faire cette hypothèse de distribution affine d'un centre à un autre plutôt que l'hypothèse d'une distribution affine à l'intérieur de chaque classe ?

J'y vois deux raisons qui méritent d'être examinées : la première est la simplicité et la crédibilité, la seconde résulte de l'existence et de l'unicité de la solution.

La première hypothèse évite tout calcul puisqu'il suffit de joindre par des segments les points $(c_i\ ;\ n_i)$ et c'est toujours possible de manière unique. Elle conduit, en général, à supposer l'existence de deux variations affines distinctes à l'intérieur d'une même classe, ce qui n'est pas choquant et correspond à la prise en compte, pour la distribution des densités à l'intérieur d'une classe, du niveau des effectifs ou fréquences dans les classes immédiatement précédente et suivante.

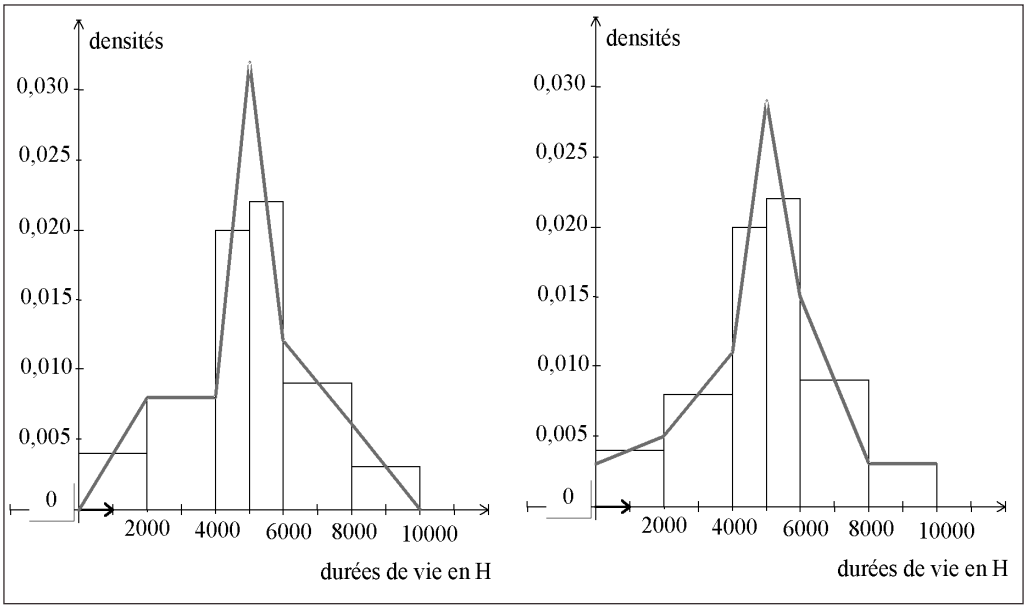
Sans être infaisables, les calculs sous-tendus par la seconde hypothèse sont beaucoup plus compliqués et surtout, si on veut respecter la propriété (P) et la positivité des densités, ne conduisent pas forcément à une solution unique et donc à une interprétation universelle. Ainsi, dans l'exemple choisi plus haut il n'y a pas de solution avec des densi-

tés toutes positives. Par contre, une petite modification des données (voir tableau ci-après) conduit à une infinité de solutions dont les deux représentées dans les figures ci-dessous.

Durée de vie (en h)	ni	ai	di
[0 ; 2000[	8	2000	0,004
[2000 ; 4000[	16	2000	0,008
[4000 ; 5000[	20	1000	0,020
[5000 ; 6000[	22	1000	0,022
[6000 ; 8000[	18	2000	0,009
[8000 ; 10000[	6	2000	0,003

En tout état de causes, le recours au “polygone des effectifs” est délicat ! Il me semble qu’il ne devrait se faire que sous les conditions suivantes :

- souligner qu’il représente des densités d’effectifs ou de fréquences et donc l’appeler *polygone des densités* ;
- rappeler que l’aire située sous le polygone et au-dessus de l’axe représente l’effectif total ou l’unité (100%) ;
- le réserver aux répartitions en classes d’amplitudes égales ou, pour le moins, montrer que le principe de construction ne s’étend pas sans précaution aux classes d’amplitudes inégales ;
- donner au moins un exemple d’utilisation et d’interprétation en faisant calculer une approximation de l’effectif (ou de la proportion) des observations situées entre deux valeurs qui ne délimitent pas une classe (voir [4] p. 100).



Sinon, il vaut mieux s'abstenir de le tracer. Voilà un petit graphique de moins !

En fait, les statisticiens ne s'en tiennent pas là ! Face aux inconvénients inhérents aux histogrammes et polygones des effectifs, ils ont cherché d'autres méthodes d'estimation de la fonction de densité théorique lorsqu'on ne dispose d'aucun renseignement a priori sur sa forme. On trouvera dans [2] la description des trois méthodes les plus courantes : la méthode du noyau, les fonctions splines et les fonctions orthogonales.

La plus simple de ces méthodes est sans doute la **méthode du noyau**. Supposant donné un échantillon aléatoire de n valeurs x_i d'un caractère quantitatif, elle consiste à estimer la fonction de densité par la fonction définie, pour tout réel x, par :

$$f_n(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

où $h > 0$ (dépendant de n) est la largeur de la fenêtre ou constante de lissage et K une fonction, appelée noyau, jouissant de "bonnes" propriétés pour assurer la convergence de f_n vers la fonction densité. En particulier si K est une densité (fonction positive continue par morceaux dont l'intégrale sur $\mathbf{R}$ est égale à l'unité), f_n aussi.

Historiquement, la première forme d'estimateur à noyau fut introduite en 1956 par M. Rosenblatt [7] avec K égale à la fonction indicatrice de l'intervalle $[-1/2 ; 1/2 [$.

Ce choix conduit à estimer $f(x)$ par $f_n(x)=\frac{n_x}{nh}$ où n_x est le nombre de valeurs observées situées à l'intérieur de l'intervalle $[x - h/2 ; x + h/2 [$.

La représentation graphique de la fonction f_n fournit un petit graphique de plus ! On la détermine en calculant $f_n(t_i)$ pour un assez grand nombre de valeurs $t_1 , \ldots , t_k$ "bien choisies" et en interpolant la courbe entre les points de coordonnées $(t_i, f_n(t_i))$.

L'exemple ci-contre (représentation graphique de f_{50} avec $h = 0{,}05$) est traité par Girard [3, pp. 57-59] à partir de données extraites de l'ouvrage de Gilbert Saporta : *Probabilités, Analyse des Données et Statistique*, Technip, 1990.

La courbe ainsi obtenue est sans doute plus crédible que le polygone des densités d'effectifs mais il faut noter qu'elle dépend tout de même fortement du noyau K et du facteur de lissage h choisis.

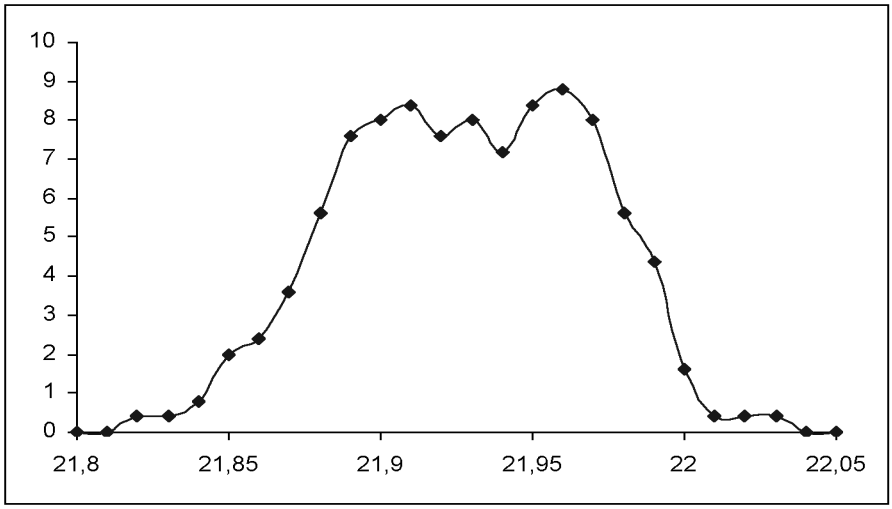
Le noyau de Rosenblatt conduit à des fonctions f_n discontinues, ne serait-ce qu'en $x_1 - h/2$ et $x_n + h/2$ (l'apparente continuité de f_{50} dans l'exemple ci-dessus résulte de l'interpolation). D'autres auteurs ont proposé des noyaux plus lisses : $K(x)=\frac{15}{16}\left(1-x^2\right)^2.I_{[-1;1]}(x)$ (où $I_{[-1\,;\,1]}$ est la fonction indicatrice de l'intervalle $[-1 ; 1]$) ou $K(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}$ (noyau gaussien).

La question se pose alors de déterminer le meilleur noyau et les meilleures valeurs de h pour estimer une densité f donnée ; pour en savoir un peu plus, consulter [2] et l'article de Berlinet Alain et Devroye Luc (1989) : Estimation d'une densité : le point sur la méthode du noyau *in Statistique et Analyse des Données*, Vol. 14 n°1, pp. 1-32 (revue de l'Association pour la Statistique et ses Utilisations).

Données :

21.86	21.92	21.91	21.97	22.01
21.84	21.90	21.91	21.98	21.96
21.88	21.91	21.92	21.95	21.95
21.90	21.89	21.91	21.89	21.95
21.92	21.91	21.93	21.98	21.97
21.87	21.87	21.96	21.96	21.96
21.90	21.89	21.91	21.98	21.95
21.87	21.90	21.97	21.95	21.94
21.90	21.89	21.97	21.97	21.97
21.93	21.92	21.97	21.94	21.95

Représentation graphique de f_{50} avec $h = 0{,}05$



Effectifs ou fréquences ?

Le deuxième exemple est emprunté à Chauvat-Réau [1]. On dispose de la répartition des étudiants dans un certain nombre d'U.F.R.[3] de l'Université de Tours pour les années scolaires 92-93 et 93-94[4] :

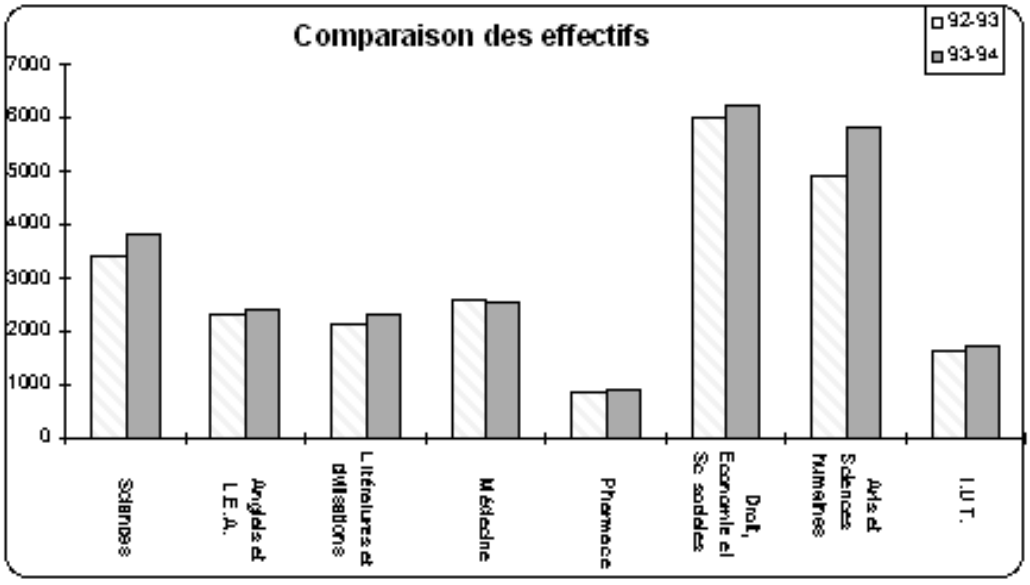
3 Unité de Formation et de Recherche

4 L'effectif correspond au nombre d'inscrits en début d'année scolaire.

Répartition des étudiants de l'université de Tours *(comparaison des effectifs)*		
U.F.R.	**92-93**	**93-94**
Sciences	3 416	3 796
Anglais et L.E.A.	2 339	2 420
Littératures et civilis.	2 137	2 348
Médecine	2 581	2 551
Pharmacie	865	917
Droit, Eco. et Sc. soc.	6 009	6 252
Arts et Sc. humaines	4 917	5 850
I.U.T.	1 658	1 741
Total	**23 922**	**25 875**

On souhaite représenter l'évolution des effectifs d'une année scolaire à l'autre. Le caractère étudié étant qualitatif, on utilise un diagramme à bandes, l'ordre des modalités n'ayant pas d'importance, la largeur des bandes non plus, la hauteur ou la largeur de celles-ci devant être proportionnelle aux effectifs des modalités. Pour faciliter la comparaison et *montrer* l'évolution U.F.R. par U.F.R., plutôt que juxtaposer les graphiques par année, on juxtapose les bandes correspondant à chaque année pour chaque U.F.R. sur un seul graphique, en prenant soin de différencier les bandes selon l'année et en fournissant une légende.

Ce graphique permet de *voir* que toutes les U.F.R. sauf Médecine ont accru leur nombre d'inscrits. Pour mieux *voir* que Médecine a perdu en effectif, il suffirait d'agrandir l'échelle verticale. On constate aussi que l'U.F.R. Arts et



Sciences humaines semble avoir progressé le plus en effectif d'une année sur l'autre.

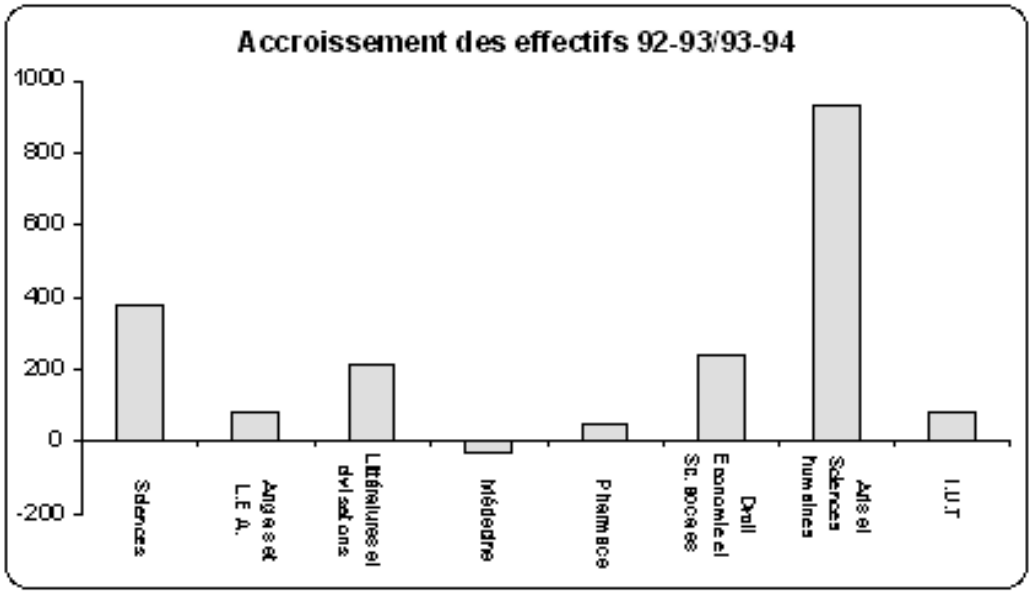
Si on s'intéresse particulièrement à l'accroissement des effectifs, on peut en fait les calculer à partir du tableau initial et les représenter par un diagramme à bandes classique (un petit graphique de plus ?).

Ici, plus de doute, on *voit* bien que Arts et Sciences humaines l'emporte et que seule Médecine a diminué ses inscriptions. Mais on a perdu les informations concernant les effectifs de chaque année…

Un observateur malicieux objecte alors qu'il préfèrerait une visualisation de l'évolution des fréquences (proportions d'inscrits dans chaque U.F.R.). On tente de lui rétorquer que dans un diagramme à bandes des fréquences les hauteurs des bandes sont pro-

Répartition des étudiants
de l'université de Tours
(évolution des effectifs)

U.F.R.	**92-93**	**93-94**	**Δ**
Sciences	3 416	3 796	380
Anglais et L.E.A.	2 339	2 420	81
Littératures et civilis.	2 137	2 348	211
Médecine	2 581	2 551	–30
Pharmacie	865	917	52
Droit, Eco. et Sc. soc.	6 009	6 252	243
Arts et Sc. humaines	4 917	5 850	933
I.U.T.	1 658	1 741	83
Total	**23922**	**25875**	**1953**

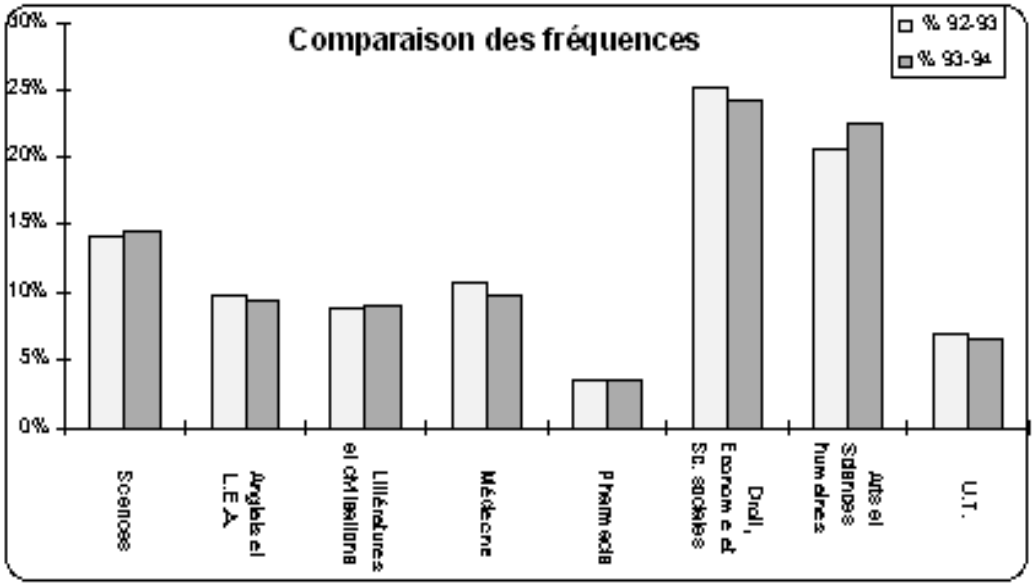


portionnelles aux fréquences, et comme les fréquences sont proportionnelles aux effectifs, on ne voit pas ce que cela va changer ! Il insiste…

Surprise ! Quatre U.F.R. qui gagnent en effectifs perdent en fréquences. On imagine les polémiques : « Oui Monsieur, l'U.F.R. de Droit, Économie et Sciences sociales est dynamique, le nombre d'inscrits s'accroît cette année !… » ; « Non Monsieur ! En réalité, il y avait 25,12% des étudiants tourangeaux inscrits dans votre U.F.R. en 92-93, mais il n'y en a plus que 24,16% en

Répartition des étudiants
de l'université de Tours
(comparaison des fréquences)

U.F.R.	**92-93**	**93-94**	**%92-93**	**%93-94**
Sciences	3 416	3 796	14,28	14,67
Anglais et L.E.A.	2 339	2 420	9,78	9,35
Littératures et civilis.	2 137	2 348	8,93	9,07
Médecine	2 581	2 551	10,79	9,86
Pharmacie	865	917	3,62	3,54
Droit, Eco. et Sc. soc.	6 009	6 252	25,12	24,16
Arts et Sc. humaines	4 917	5 850	20,55	22,61
I.U.T.	1 658	1 741	6,93	6,73
Total	**23 922**	**25 875**	**100**	**100**



93-94 !… », « Voyons Messieurs ! Inutile de vous disputer, vous voyez bien qu'on peut faire dire ce que l'on veut aux chiffres, ne me parlez pas de statistique ! »

Évidemment on peut débattre longtemps pour savoir laquelle des deux informations est la plus pertinente : l'évolution des effectifs ou l'évolution des fréquences. Puisque les conclusions peuvent être opposées, le statisticien ne devrait pas avoir à privilégier l'une ou l'autre. Est-il possible de *montrer* les deux points de vues, effectifs et fréquences, en un seul graphique en laissant l'utilisateur *voir* l'un ou l'autre ou les deux ?

Les résultats effectifs et fréquences diffèrent parce que l'on compare deux années pour lesquelles l'effectif total a varié : augmentation de 1953 inscriptions, ici. Il semble donc judicieux de prendre en compte cette variation globale. Si aucune tendance particulière ne s'était manifestée, l'accroissement d'effectifs aurait été réparti *équitablement* sur chaque U.F.R..

Mais que signifie équitablement ? Non pas que chacune des huit Unités aurait dû augmenter son effectif du huitième de 1953 puisque ce serait mettre sur le même plan Droit et Pharmacie. Il faut prendre en compte la taille de l'U.F.R. à l'époque de référence (92-93).

Le pourcentage de variation est la notion qui permet de relativiser une variation à la taille initiale de la quantité qui varie. Ainsi, avec des notations *évidentes*, le pourcentage de variation des effectifs est donné par :

$$PV_{2/1}(n) = (n_2 - n_1) \times 100/n_1 ,$$

celui des fréquences par :

$$PV_{2/1}(f) = (f_2 - f_1) \times 100/f_1 .$$

En remarquant que $f_1 = n_1/N_1$ et $f_2 = n_2/N_2$, on vérifie facilement les équivalences suivantes :

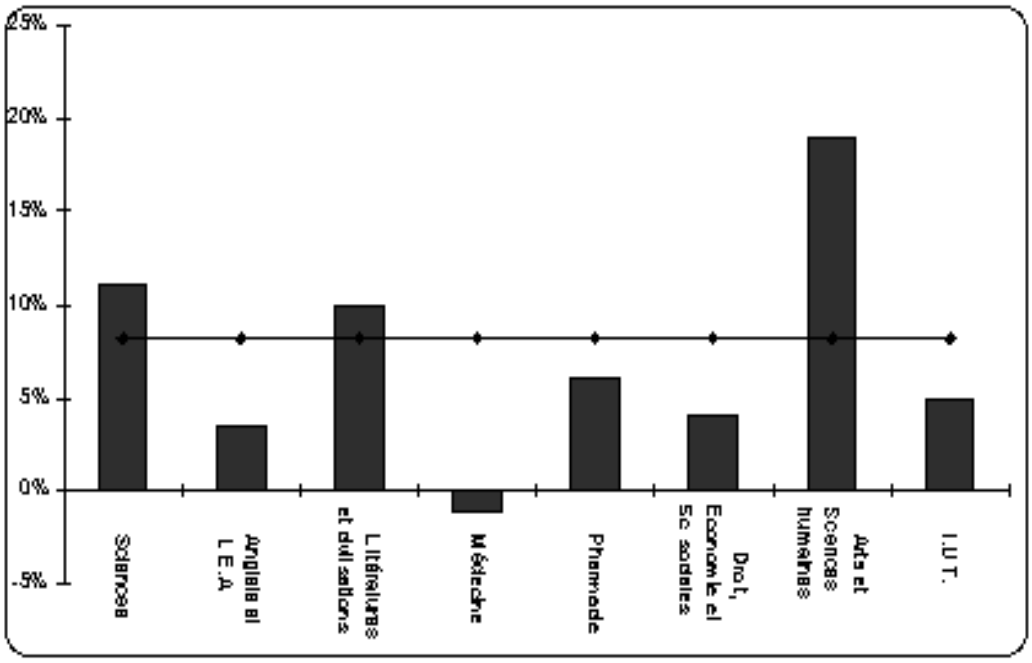
$$n_2 < n_1 \Leftrightarrow PV_{2/1}(n) < 0$$
$$\text{et } n_2 < n_1 \Leftrightarrow PV_{2/1}(f) < PV_{2/1}(N)$$

$$f_2 < f_1 \Leftrightarrow PV_{2/1}(f) < 0$$
$$\text{et } f_2 < f_1 \Leftrightarrow PV_{2/1}(n) < PV_{2/1}(N)$$

A partir du tableau initial, on calcule les pourcentages de variation des effectifs pour chaque U.F.R. et pour leur ensemble. On représente ensuite ces pourcentages de variation par un diagramme à bandes auquel on ajoute une droite d'ordonnée égale au pourcentage de variation de l'effectif total (cf. figure de la page suivante).

Répartition des étudiants de l'université de Tours
(pourcentages de variation des effectifs)

U.F.R.	92-93	93-94	%dN/N
Sciences	3 416	3 796	11,12%
Angl. et L.E.A.	2 339	2 420	3,46%
Litt. et civ.	2 137	2 348	9,87%
Médecine	2 581	2 551	-1,16%
Pharmacie	865	917	6,01%
Dr. Eco. & Sc. s.	6 009	6 252	4,04%
Arts et Sc. h.	4 917	5 850	18,97%
I.U.T.	1 658	1 741	5,01%
Total	**23 922**	**25 875**	**8,16%**



La lecture graphique est alors *évidente* : tous les rectangles situés sous l'axe des 0% correspondent à des U.F.R. qui ont perdu des inscrits ; tous les rectangles qui dépassent la ligne représentant le pourcentage de variation totale (ici 8,16%) représentent des U.F.R. qui ont augmenté aussi bien en effectifs qu'en fréquences ; les autres ont seulement gagné en effectifs mais perdu en fréquences.

Bien sûr, on aurait pu procéder de même avec les pourcentages de variation des fréquences à condition d'ajouter une droite dont l'ordonnée serait égale au pourcentage de variation *rétrograde* de l'effectif total :

$$PV_{1/2}(\mathrm{N}) = (\mathrm{N}_1 - \mathrm{N}_2) \times 100/\mathrm{N}_2 .$$

Les écarts standardisés

Le dernier exemple est emprunté à l'ouvrage de R. Tomassone et *alii.* sur la régression [8, pp. 22-24]. On dispose, dans le tableau ci-contre, de 16 valeurs de 5 différents couples de variables statistiques : (X,Y_1), (X,Y_2), (X,Y_3), (X,Y_4) et (U,V).

L'étude classique concernant un couple de variables statistiques consiste à rechercher une éventuelle relation fonctionnelle entre la seconde variable, dite *expliquée*, et la première variable, dite *explicative* (ou *contrôlée*). Le plus souvent on recherche une relation de type affine. Les coefficients de cette régression affine et le coefficient de corrélation linéaire, indice de sa validité, sont don-

i	X(i)	Y1(i)	Y2(i)	Y3(i)	Y4(i)	U(i)	V(i)
1	7,000	5,535	0,116	7,399	3,864	13,715	5,654
2	8,000	9,942	3,770	8,546	4,942	13,715	7,072
3	9,000	4,249	7,426	8,468	7,504	13,715	8,491
4	10,000	8,656	8,781	9,616	8,581	13,715	9,909
5	12,000	10,737	12,678	10,685	12,221	13,715	9,909
6	13,000	15,144	12,889	10,607	8,842	13,715	9,909
7	14,000	13,939	14,253	10,529	9,919	13,715	11,327
8	14,000	9,450	16,545	11,754	15,860	13,715	11,327
9	15,000	7,124	15,620	11,676	13,966	13,715	12,746
10	17,000	13,693	17,206	12,745	19,092	13,715	12,746
11	18,000	18,100	16,281	13,893	17,198	13,715	12,746
12	19,000	11,285	17,647	12,590	12,334	13,715	14,164
13	19,000	21,385	14,211	15,040	19,761	13,715	15,582
14	20,000	15,692	15,577	13,737	16,382	13,715	15,582
15	21,000	18,977	14,652	14,884	18,945	13,715	17,001
16	23,000	17,690	13,947	29,431	12,187	33,282	27,435

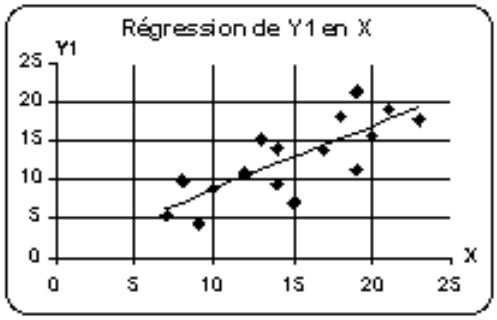
	X(i)	Y1(i)	Y2(i)	Y3(i)	Y4(i)	U(i)	V(i)
moyenne	**14,938**	***12,600***	***12,600***	***12,600***	***12,600***	**14,938**	***12,600***
variance	**22,434**	***23,775***	***23,775***	***23,775***	***23,775***	**22,434**	***23,775***
covariance		***18,142***	***18,142***	***18,142***	***18,142***		***18,142***

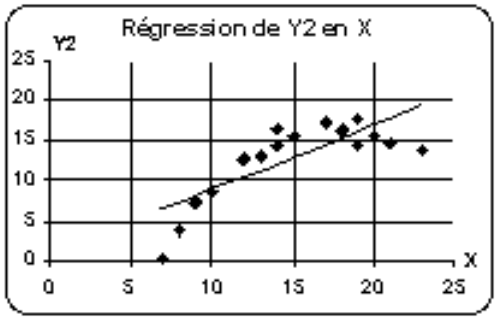
nés par des calculs bien connus et disponibles sur toute calculatrice et logiciel statistiques.

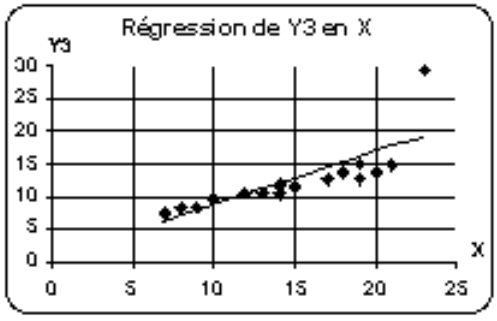
Or, ici, et c'est tout l'intérêt des valeurs fictives choisies, tous les paramètres intervenant dans ces calculs sont égaux, au moins au millième près, pour chacun des couples étudiés. Ainsi, tout tableur fournira les résultats indiqués en gras dans le tableau inférieur.

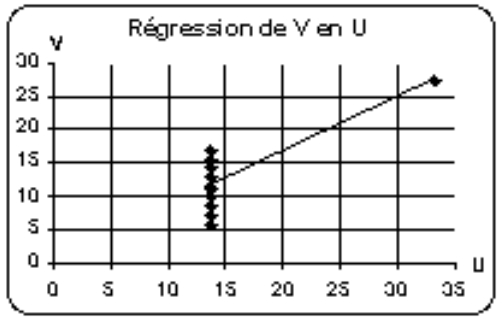
Il en résulte que les coefficients de la régression (ajustement affine selon les moindres carrés) valent a = 0,809 et b = 0,520 dans les cinq cas, avec un coefficient de corrélation linéaire toujours égal à 0,786. Si on s'en tient à ces calculs aveugles, la conclusion sera donc toujours la même. Il est clair que la moindre des prudences conseille de jeter un coup d'œil aux graphiques représentant le nuage de points défini par chaque couple et la droite de régression avant de conclure !

Ce type de graphique (cf. page suivante) est déjà suffisant pour mettre en doute l'exis-









tence d'une relation affine entre les variables (voir X,Y_2) ou pour mettre en évidence des points *suspects* (voir X,Y_3) ou un contrôle de la variable explicative (plan d'expérience) douteux (voir U,V). Ces situations sont sans doute extrêmes, voire caricaturales, mais il n'en reste pas moins qu'il faut toujours procéder à cet examen graphique primaire.

Cependant le cas de Y_1 et Y_4 reste délicat même si le graphique de (X,Y_4) *montre* que le nuage de points semble s'éloigner davantage de la droite de régression pour les valeurs élevées de X.

Le coefficient de corrélation, sans être bon (i.e. supérieur à 0,866), n'est pas franchement mauvais. Doit-on valider, au seul vu du nuage de points, l'hypothèse d'une relation affine ?

Cette hypothèse, en fait, s'exprime de la façon suivante : on suppose qu'il existe deux réels a et b, et n variables aléatoires U_i , tels que

i) pour tout $i = 1, \ldots, n$, $Y_i = a.X_i + b + U_i$ (relation affine perturbée par un aléa)

ii) pour tout $i = 1, \ldots, n$, $E(U_i) = 0$ (moyenne nulle) et $V(U_i) = s^2$ (variance constante)

iii) pour tout $i = 1, \ldots, n$, pour tout $j = 1, \ldots, n$, $i \neq j$, $Cov(U_i , U_j) = 0$ (covariances nulles)

On voit que la validité de l'hypothèse résulte de l'analyse des écarts (ou résidus) entre la réalité observée (Y_i) et le modèle affine ($a.X_i + b$). Ces écarts observés confirment-ils les hypothèses ii) et iii) ?

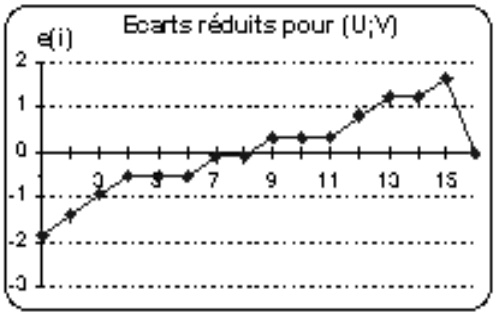
Dans la pratique, afin de disposer d'un critère indépendant de l'ordre de grandeur des valeurs observées, on considère les écarts *réduits* (ou *standardisés*) obtenus en divisant les écarts $y_i - (ax_i + b)$ par leur écart type. Si on ajoute l'hypothèse :

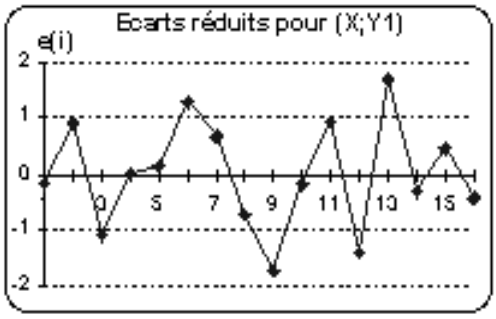
iv) pour tout $i = 1, \ldots, n$, U_i suit une loi normale de moyenne nulle et de variance s^2

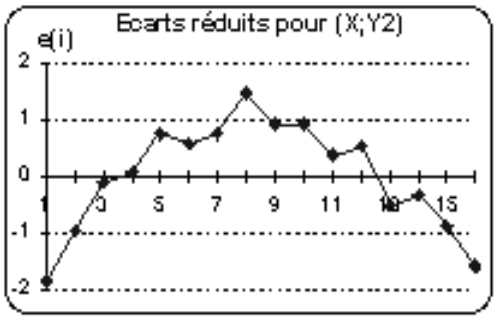
on peut démontrer que les écarts réduits suivent une loi de Student à $n - 2$ degrés de liberté. Les écarts réduits $e(i)$, calculés sur l'échantillon observé, confirmeront donc l'hypothèse d'une relation affine s'ils se répartissent aléatoirement entre environ -2 et 2 sans tendance particulière.

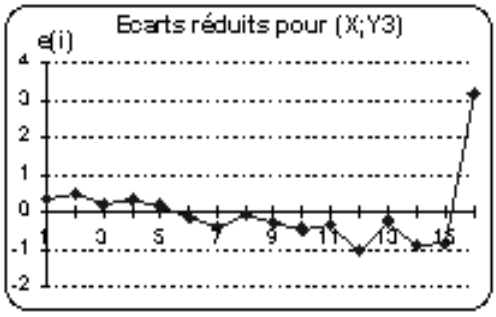
Ce que l'on vérifie avec un petit graphique de plus !

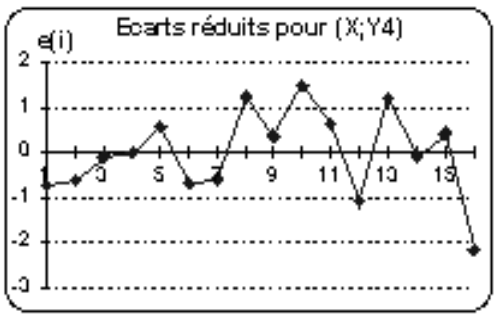
La lecture de ces graphiques est immédiate. Seul le premier confirme les hypothèses sur la composante résiduelle aléatoire. On remarque, en particulier que les écarts réduits pour $(X ; Y_4)$ semblent s'amplifier











lorsque i augmente. On peut penser alors à un modèle multiplicatif du type :

$$Y_i = (aX_i + b)\times U_i.$$

Dans le cas de séries moins caricaturales que celles examinées ici, le graphique des écarts réduits s'avère encore plus indispensable si on veut juger sérieusement la qualité de la régression au-delà de *l'impression* fournie par le nuage de points.

Conclusion

Les trois exemples traités dans cet article avaient pour ambition, non seulement de montrer l'intérêt du graphique en Statistique, mais aussi combien son maniement peut être délicat même lorsque les règles de tracé sont clairement définies. La nécessité de connaître ces règles pour *montrer* et pour *voir* est évidente. Mais il est tout aussi indispensable de connaître (et de ne pas perdre) le sens de ce qui est montrable et visible ; l'interprétation correcte des graphiques ne peut se faire qu'à ce prix ! Du coup, contrairement aux idées reçues sur le dessin censé faciliter la compréhension, le recours au graphique n'est pas si simple et demande un travail spécifique. Un travail où les mathématiques ne sont jamais loin ; ici par exemple : densité de probabilité, pourcentage de variation, loi des écarts réduits…

Mais qui s'en plaindra ?

Bibliographie

[1] CHAUVAT Gérard, REAU Jean-Philippe (1995) : *Statistique descriptive*, coll. « les Fondamentaux », Hachette Supérieur.

[2] DELECROIX Michel (1983) : *Histogrammes et estimation de la densité* ; Que sais-je ? n°2055, PUF.

[3] GIRARD Jean Claude (1996) : De l'histogramme à la fonction densité de probabilité in *Enseigner les statistiques du CM à la seconde. Pourquoi ? Comment ?*, IREM de Lyon.

[4] PARZYSZ Bernard (1999) : Heurs et malheurs du su et du perçu en statistique. Des données à leurs représentations graphiques in *Repères-IREM* n°35, pp. 91-112.

[5] PICHARD Jean-François (1992) : Représentations graphiques en statistiques in *Bulletin Inter-IREM « Des chiffres et des lettres »*, IREM de Rouen, pp. 75-101.

[6] POMBOURCQ Pascale (2000) : La fonction de répartition. Pour quoi faire ? in *Repères-IREM* n°38, pp. 91-106.

[7] ROSENBLATT M. (1956) : Remarks on some nonparametric estimates of a density function. *Annals of Mathematics Statistics*, vol. 27, pp. 832-837.

[8] TOMASSONE Richard, AUDRAIN Sylvie, LESQUOY-de TURKHEIM Elisabeth, MILLIER Claude (1983) : *La régression. Nouveaux regards sur une ancienne méthode statistique*. Masson (2ième éd. 1992).