Le tableau statistique se présente en général sous la forme :

x_i	n_i	f_i

Exemple 1

Distribution du nombre de suicides de femmes par an et par État pour 8 États allemands et pendant 14 ans (d'après Von Bortkiewicz, 1898, cité par Kendall) (tableau 6.1) :

TABLEAU 6.1

Nombre de suicides x_i	0	1	2	3	4	5	6	7	8	9	$\geqslant 10$
Effectif n_i	9	19	17	20	15	11	8	2	3	5	3
Total $n = 112$											

6.1.2 Variables continues ou assimilées

On regroupe les valeurs en k classes d'extrémités $e_0, e_1, ..., e_k$ et l'on note pour chaque classe $[e_{i-1}, e_i[$ l'effectif n_i et la fréquence f_i ainsi que les fréquences cumulées $F_i = \sum_{j=1}^{i} f_j$, ou proportion des individus pour lesquels $X < e_i$.

Le tableau statistique se présente en général comme suit :

e_{i-1}	n_i	f_i	F_{i-1}
e_i			F_i

Par convention la borne supérieure d'une classe est toujours exclue de cette classe.

Exemple 2

TABLEAU 6.2

DISTRIBUTION DES REVENUS IMPOSABLES DES FRANÇAIS, EN 1970

Tranche des revenus en francs ln $(R - 2\,500)$	% du nombre total de contribuables	% cumulés
2 500		
	0.67	
5 000 — 3.39		0.67
	30.18	
10 000 — 3.87		30.85
	27.50	
15 000 — 4.10		58.35
	17.09	
20 000 — 4.24		75.44
	14.45	
30 000 — 4.44		39.89
	7.01	
50 000 — 4.68		96.90
	1.66	
70 000 — 4.83		98.56
	0.81	
100 000 — 4.99		99.37
	0.51	
200 000 — 5.30		99.88
	0.10	
400 000 — 5.60		99.98
	0.02	
		100
$n = 10\,503\,244$		

(Statistiques et Études financières, juillet-août 1972.)

Les centres de classes c_i valent $c_i = \dfrac{e_{i-1} + e_i}{2}$.

Les amplitudes de classes h_i valent $h_i = e_i - e_{i-1}$.

On notera que dans l'exemple 2 (tableau 6.2) les classes ne sont pas d'égale amplitude.

6.2 REPRÉSENTATIONS GRAPHIQUES

6.2.1 Variables discrètes : le diagramme en bâtons

On porte en ordonnée f_i en fonction de x_i. Le fait de porter f_i ou n_i n'est qu'une question d'échelle, mais a son intérêt lorsqu'on veut comparer plusieurs distributions d'effectif total différent.

Exemple 1 (fig. 6.1)

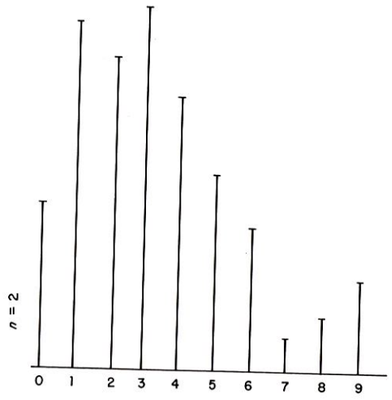

FIG. 6.1

6.2.2 Variable continue répartie en classes : l'histogramme et les estimations de densité

L'histogramme est analogue à la courbe de densité : ici le rectangle construit sur chaque classe a une surface égale à la fréquence de la classe. Dans le cas de classes d'égale amplitude on reporte directement en ordonnée f_i (à l'échelle près) sinon f_i/a_i.

Exemple 2 (fig. 6.2)

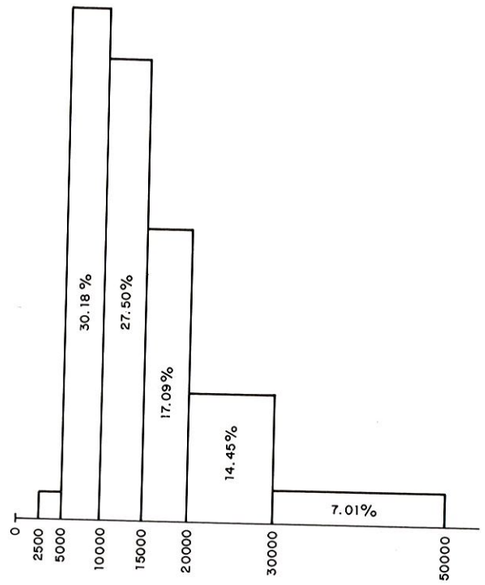

FIG. 6.2

La détermination du nombre de classes d'un histogramme est délicate et on ne dispose pas de règles absolues. Un trop faible nombre de classes fait perdre de l'information et aboutit à gommer les différences pouvant exister entre des groupes de l'ensemble étudié. En revanche un trop grand nombre de classes aboutit à des graphiques incohérents : certaines classes deviennent vides ou presque, car n est fini.

On peut d'ailleurs critiquer le fait de représenter par une fonction en escalier la distribution d'une variable continue : l'histogramme est une approximation assez pauvre d'une fonction de densité et il serait plus logique de chercher une fonction plus régulière.

La théorie de l'estimation de densité permet de proposer des solutions à ce problème. Ce qui suit n'est qu'un exposé rudimentaire de la méthode du noyau.

Considérons tout d'abord le cas d'histogrammes à classes d'égales amplitudes h.

L'histogramme aboutit à estimer la densité de probabilité de X au point x par n_i/nh si x appartient à la $i^{\text{ème}}$ classe de l'histogramme. La densité est donc la même quelle que soit la position de x entre les extrêmités de cette classe.

Une première amélioration consiste à utiliser la méthode de la «fenêtre mobile» : on construit autour de x une classe de longueur h : $I_x = [x - h/2;\ x + h/2[$ et on compte le nombre d'observations n_x appartenant à cette classe. On estime alors $f(x)$ par n_x/nh et on peut construire point par point une courbe de densité estimée $\hat{f}(x)$.

On peut remarquer alors que $\hat{f}(x)$ vérifie la formule générale :

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

où K est la fonction indicatrice de l'intervalle $[-1/2,\ 1/2[$:

$K(u) = 0$ si $u \geqslant 1/2$ ou $u < 1/2$; $K(u) = 1$ si $1/2 \leqslant u < 1/2$.

$K\left(\frac{x - x_i}{h}\right)$ vaut donc 1 si $x_i \in I_x$.

$\hat{f}(x)$ est donc une moyenne arithmétique de fonctions donnant à chaque observation x_i un poids $1/h$ si elle appartient à l'intervalle I_x, 0 sinon.

Cette méthode donne cependant une estimation $\hat{f}$ peu régulière. Pour obtenir une fonction suffisamment «lisse», il est alors possible de généraliser la formule précédente en prenant pour K (le noyau) une autre expression, en général une densité symétrique.

En pratique on utilise fréquemment un noyau gaussien $K(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}u^2\right)$ ou parabolique $K(u) = \frac{3}{4\sqrt{5}}\left(1 - \frac{u^2}{5}\right)$ pour $|u| < \sqrt{5}$.

Ce dernier, appelé noyau d'Epanechnikov, a des propriétés mathématiques intéressantes. La constante h appelée constante de lissage joue un rôle important analogue à celui de la largeur des classes de l'histogramme : si h est faible $\hat{f}$ sera très peu régulière, si h est grand $\hat{f}$ très lisse.

Bien que l'on sache que h doit être proportionnel à $n^{-1/5}$ sa valeur «optimale» se détermine en fait empiriquement.

Il n'est pas nécessaire que K soit une densité et des noyaux pouvant prendre des valeurs négatives sont même conseillés par certains auteurs.

L'exemple suivant illustre l'efficacité de l'estimation de densité par rapport à un histogramme classique. Le noyau utilisé est celui proposé par M. Lejeune :

$$K(u) = \frac{105}{64}(1-u^2)^2(1-3u^2) \quad \text{pour} \quad |u| \leqslant 1$$

avec une constante h égale à 30% de l'étendue de l'échantillon. L'estimation de densité montre clairement une distribution bimodale (fig. 6.3) alors qu'un histogramme en 6 classes d'amplitude 0.03 (fig. 6.4) ne le permet pas. Les données représentaient les hauteurs de 50 pièces usinées (tableau 6.3).

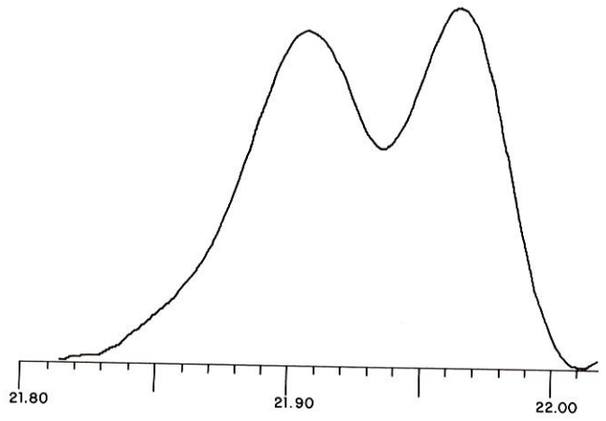

FIG. 6.3

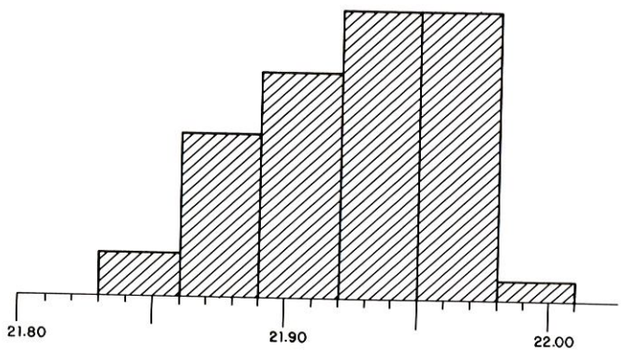

FIG. 6.4

TABLEAU 6.3

(1) 21.86	(18) 21.9	(35) 21.98
(2) 21.84	(19) 21.89	(36) 21.96
(3) 21.88	(20) 21.92	(37) 21.98
(4) 21.9	(21) 21.91	(38) 21.95
(5) 21.92	(22) 21.91	(39) 21.97
(6) 21.87	(23) 21.92	(40) 21.94
(7) 21.9	(24) 21.91	(41) 22.01
(8) 21.87	(25) 21.93	(42) 21.96
(9) 21.9	(26) 21.96	(43) 21.95
(10) 21.93	(27) 21.91	(44) 21.95
(11) 21.92	(28) 21.97	(45) 21.97
(12) 21.9	(29) 21.97	(46) 21.96
(13) 21.91	(30) 21.97	(47) 21.95
(14) 21.89	(31) 21.97	(48) 21.94
(15) 21.91	(32) 21.98	(49) 21.97
(16) 21.87	(33) 21.95	(50) 21.95
(17) 21.89	(34) 21.89	

6.2.3 Le diagramme *stem and leaf*

Il s'agit d'une sorte d'histogramme horizontal, proposé par J. W. Tukey, réalisé à l'aide des valeurs numériques de la série étudiée. Chaque donnée y est représentée par sa tige (*stem*) qui correspond aux chiffres principaux et sa feuille (*leaf*) qui correspond aux chiffres suivants. Ainsi à une variable X prenant les 24 valeurs suivantes :

8	13	27	32	25	16	32	27	8
28	79	25	35	25	38	29	80	50
38	30	20	20	49	9			

correspond le diagramme où la première colonne est le chiffre des dizaines :

0	8	8	9						
1	3	6							
2	0	0	5	5	5	7	7	8	9
3	0	2	2	5	8	8			
4	9								
5	0								
6									
7	9								
8	0								

On a donc à la fois l'allure de la distribution et les valeurs numériques des observations.