HISTOGRAMME (C5, F3, K15)
(17 / 12 / 2019, © Monfort, Dicostat2005, 2005-2019)

La notion élémentaire, mais fondamentale, d'histogramme, est utilisée en **Statistique descriptive** pour représenter ou visualiser une **variable statistique** à valeurs souvent réelles, et souvent « scalaires » (ie à 1 dimension).

L'histogramme d'une **variable statistique** est l'analogue de (ie s'interprète comme) la **densité de probabilité** d'une **va** ou d'une **statistique**, dont il constitue un estimateur naturel (cf **statistique naturelle**).

(i) Soit $(\Omega, \mathcal{T}, P)$ un **espace probabilisé**, $X = (X_1, ..., X_N) : \Omega \mapsto \mathbf{R}^N$ un N-**échantillon aléatoire** et $(I_k)_{k=1,...,K}$ une **partition** finie de la droite numérique **R**, constituée de K intervalles disjoints non vides I_k , avec :

(1) $I_1 =]-\infty, a_1]$, $I_k =]a_{k-1}, a_k]$, $\forall\ k \in N_{K-1}^* \setminus \{1\}$, et $I_K =]a_{K-1}, +\infty[$.

On note N_k , la **fréquence absolue** de la classe n° k, $\forall\ k \in N_K^*$, ie le nombre de coordonnées X_n de X tq $X_n \in I_k$:

(2) $N_k = \text{Card } A_k$, avec $A_k = \{n \in N_N^* : X_n \in I_k\}$.

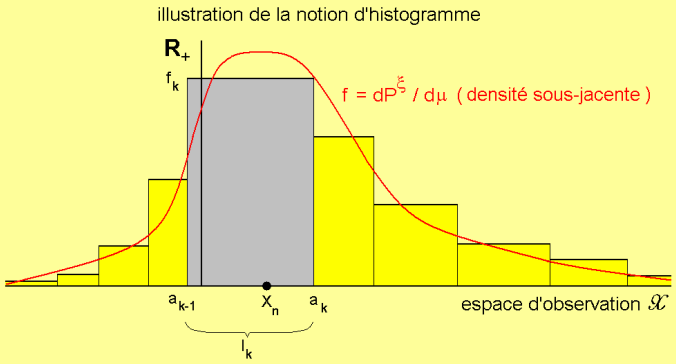
Chaque **fréquence relative** f_k se déduit de N_k selon :

(3) $f_k = N_k / N$, avec $N = \Sigma_k N_k$.

On appelle alors **histogramme** de X le graphe de la **fonction étagée** $h : \mathbf{R} \mapsto \mathbf{R}_+$ définie par :

(4) $h = \Sigma_k f_k \mathbf{1}(I_k)$,

où **1**(A) désigne la **fonction indicatrice** de A (cf graphique ci-dessous).



Comme pour une densité de probabilité, on a $\int h\, d\lambda = 1$.

(ii) Les principales questions statistiques liées à la notion d'histogramme sont les suivantes :

(a) choix du nombre K de classes, ou de leurs longueurs, ou de leurs bornes a_k elles-mêmes ;

(b) **estimation de la densité** (ou **estimation de la fr**) de la **loi de probabilité** qui génère les **observations** : **ajustement** d'une loi (ie de sa densité) sur les observations X_n et **tests d'adéquation** de cette loi à ces observations ;

(c) évaluation de la **perte d'information** résultant du groupement des X_n en classes I_k (cf **groupement de classes**, **correction de groupement**), par référence à la **densité empirique** ou à la **fr empirique**.

(iii) Un **histogramme** h est un estimateur naturel (**estimateur par le noyau** le plus simple) de la densité de probabilité théorique (cf **statistique naturelle**). Mais cet estimateur est en général moins efficace que ceux obtenus, par d'autres procédés, à partir de l'ensemble des observations X_n .

(iv) D'un point de vue terminologique :

(a) l'**ensemble** des coordonnées de X tq $X_n \in I_k$ (ou parfois même I_k lui-même) est appelé une **classe** de l'histogramme ;

(b) les extrémités de I_k sont appelées **limites de classe**, ou **extrémités de classe** ;

(c) l'intervalle ainsi défini entre les extrémités est dit **intervalle de classe** et la longueur de I_k est dite **longueur de classe** ;

(d) la fréquence N_k (resp f_k) est appelée **fréquence absolue** (resp **fréquence relative**) de la classe k.

(v) La notion d'histogramme s'étend directement aux échantillons de variables statistiques vectorielles. Ainsi, la représentation graphique, dans $\mathbf{R}^2$ x $\mathbf{R}_+$, d'un histogramme relatif à deux variables s'appelle un **stéréogramme**.

On peut aussi l'étendre à une **variable qualitative** (ie non numérique) : les observations de ces variables figurent dans un **tableau statistique** à une dimensions mettant en parallèle les modalités de la variable et les **fréquences relatives** associées (classement dans un **tableau de contingence** à une dimension).

(v) Le concept d'histogramme est donc un concept statistique central, très « profond », puisqu'il permet d'analyser d'emblée plusieurs notions : **classification**, **échantillonnage**, **estimation** et **tests** relatifs à la densité, approches paramétriques ou non paramétriques, **forme** et **forme légale**, **régression** (cas multidimensionnel), **mélange de lois**, **lacunes** résultant des procédés d'observation, etc.

C’est en raison de son importance conceptuelle que sa représentation graphique symbolique (ci-dessous) sert d’image de référence dans la **page d’accueil** de ce site.