Statistique relative

I_ Généralités

Soit X : Ð à þ une application avec Ð un ensemble de cardinaux n dit population.

A. Effectifs

On suppose que X(Ð) = { x_1 ; … ; x_n } avec $x_1 < x_2 < \ldots < x_n$ et on pose n_i = | $X^{-1}(x_i)$| et $f_i = n_i / n$ pour i = 1, … k.
On a ainsi Œ f_i = 1 car Œ n_i = n.

L'effectif (respectivement la fréquence) en x, est n_i (respectivement f_i).
L'effectif cumulé (respectivement la fréquence cumulée) en x_i est Œ n_j (respectivement Œ f_j).

La donnée (x_i ; n_i) est dite série statistique discrète.

B. Classes

On suppose que X(Ð) = Š]a_i ; a_{i+1}] _ þ
Et on pose $_{ni}$ = | X^{-1}( ]a_i ; a_{i+1}]) | et $f_i = n_i / n$.

Le nombre n_i (respectivement f_i) est dit l'effectif de la classe]a_i ; a_{i+1}].
Le cardinal X^{-1}( ]a_i ; a_{i+1}]) est l'effectif cumulé en a_i. Le rapport entre l'effectif cumulé et l'effectif total est dit fréquence cumulée.

La donnée d( ]a_i ; a_{i+1}], n) est dite une série statistique groupée ou continue.

On appelle $a_{i+1} - a_i$, l'amplitude de la classe]a_i ; a_{i+1}]. La plus petite amplitude de la série statistique est dite l'amplitude élémentaire. ($a_{i+1} - a_i$) / 2 est dite le centre de la classe.

C. Application

On peut mettre à profit la connaissance des fréquences cumulées d'une série statistique x pour tester graphiquement l'hypothèse selon laquelle cette série suit une loi normale.

Si X est une variable aléatoire : XX N (m ; â), en changeant l'axe de graduation de la fonction des probabilités cumulées de la loi N (m ; â), on transforme la courbe de la fonction de répartition en droite D : y = (x – m) /â

On appelle la droite D, la droite d'Henri.

Pour tester si la distribution fournie par une série statistique continue suit "approximativement" une loi normale, on trace le *nuage de points (x_i ; t_i)* où t_i est défini par Ë (t_i) = P (X < t_i) et Ë (t_i) = F_i, la fréquence cumulée. On appelle t_i les probits.

II _ Représentation graphique

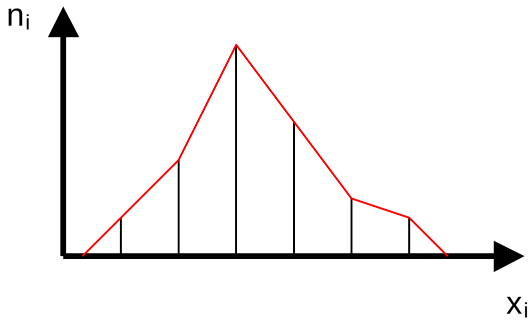
Les représentation se font dans un plan muni d'un repère Cartésien.

1. Diagramme en bâtons

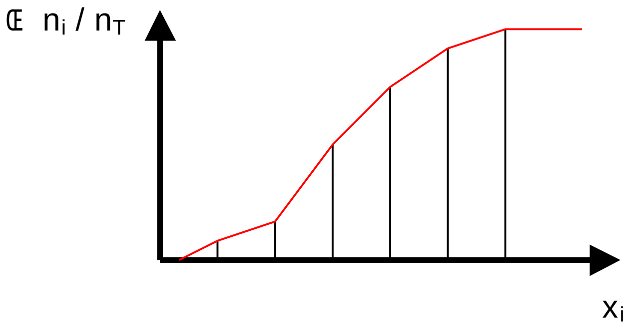
Soit (x_i ; n_i), une série statistique discrète.

Définition : Un diagramme en bâtons est constitué de l'ensemble des segments de longueurs proportionnelles à n_i et portés par les verticales issues de x_i.

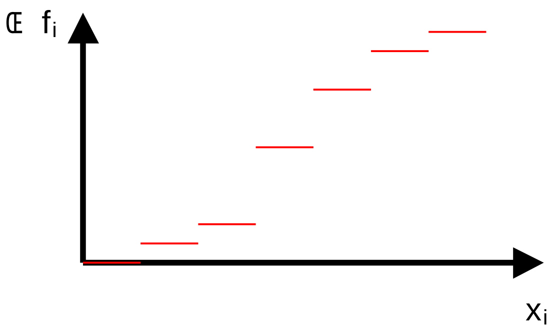
La ligne polygonale qui joint les sommets d'un diagramme en bâtons, est dite le polygone des fréquences.



Remarque : Le diagramme obtenu en substituant la fréquence f_i par la fréquence cumulée en x est dit le diagramme en bâtons des fréquences cumulées.



Définition : La fonction définie par : F(x) = Œ fi (x g þ) est dite la fonction de répartition de la série discrète.



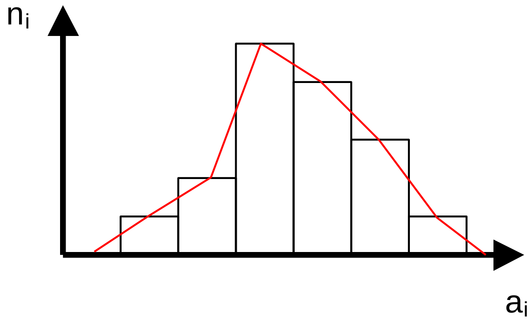
Remarque : par convention, $F(x) = 0$ sur $x < x_{min}$ et $F(x) = 1$ sur $x > x_{max}$. Il s'agit d'une fonction en escalier dont les paliers sont les fréquences cumulées.

2. Histogramme

Soit ( ] a_i ; a_{i+1}] ; n_i), une série statistique groupée

Définition : Un histogramme est constitué de l'ensemble des rectangles R_i de base [(a_i ; 0) ; (a_{i+1})] et d'aires proportionnelles à n_i.

Le polygone des fréquences d'un série subdivisée en classes de même amplitude est le même polygone joignant les milieux des faces supérieures des rectangles.



Remarque : En général on ajoute deux classes fictives d'amplitude élémentaire de part et d'autres de a_{min} et a_{max} et on prolonge le polygone des fréquences jusqu'au centre de ces deux classes.

On peut vérifier que la hauteur h_i de Ri est donnée par : $h_i = n_i$ / (Ç a / A) où A est l'amplitude élémentaire.

Définition : La fonction définie par g :] ai ; ai+1] à] 0 ; 1]
Et g(x) = Œ $f_j + (x - a_i) / (a_{i+1} - a_i) f_i$ est dite la fonction de répartition de la série groupée.

La courbe polygonale correspondant à la fonction g est dite le polynôme des fréquences cumulées. La courbe obtenue par lissage à partir de ce polygone est appelée la courbe de répartition.

Droite d'Henri : La droite d'Henri est donnée par : y = (x – m) / â . C'est une droite si les statistiques sont réparties selon une norme de probabilité (loi de Gauss).

III Paramètres de position

On distingue pour chaque définition deux cas :

à *cas 1* : (x_i ; n_i) est une série discrète

à *cas 2* : ( ]a_i ; a_{i+1}] ; ni) est une série groupée

A. Le mode

C'est la valeur qui est la plus présente (qui la plus grand effectif).

e Dans le *cas 1*, la valeur x_i d'effectif maximum est dite le mode.

- e Dans le *cas 2*, une classe est dite modale lorsque la hauteur du rectangle la représentant est maximale.

Remarque : le mode peut ne pas être unique.

B. La médiane

C'est la valeur centrale qui partage la population en deux.

- e Dans le *cas 1* :
  - _ si n_T est pair : M = ($x_{n/2}$ + $x_{n/2+1}$)/2
  - _ si n_T est impair : M = $x_{(n+1)/2}$
- e Dans le *cas 2*, on détermine le numéro des valeurs médianes comme précédemment et on en déduit dans quelle classe ils sont (c'est le groupe médian). Par interpolation linéaire, on en déduit la médiane.

C. La moyenne

Il s'agit de la valeur moyenne.

- e Dans le *cas 1* : x_{moy} = Œ $x_i f_i$
- e Dans le *cas 2* : x_{moy} = Œ (a_i + a_{i+1})/2 * f_i

IV _ Paramètres de dispersion

A. Les quantiles

Définition : le premier quartile est la valeur qui correspond à un quart de l'effectif cumulé croissant. Et ainsi de suit pour les déciles, centiles, milliles …

B. L'écart-type

Définition :

- e Dans le *cas 1*, l'écart-type est défini par : â = (Œ f_i(x_i – x_{moy})^2)^{½}
- e Dans le *cas 2*, l'écart-type s'obtient grâce à la formule de Keonig avec le centre des classes :

â ^{2} = Œ ($f_i x_i^2$) - x_{moy}^2

V _ Paramètres de concentration

A. Indice de Gini

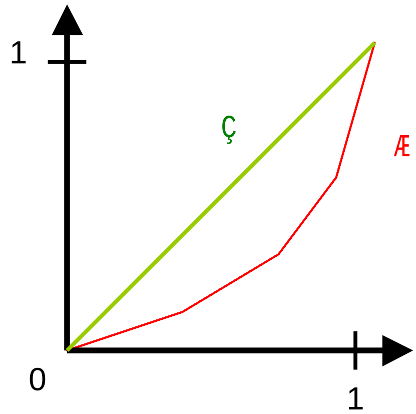
- e Dans le *cas 1* : x_i > 0, soit les points M_i(F_i ; [Œ $f_i x_i$] / [Œ $_{totale}$ $f_i x_i$]) où F_i est la fréquence cumulée et f_i la fréquence.

à On note Æ la ligne polygonale qui va de 0 en passant par tous les m_i, dans un repère orthonormé de 1 sur 1.

Définition : Æ est dite la courbe de Gini de la série statistique, et Ç la droite allant de (0;0) à (1;1).

Remarque : Si les écarts restent faibles, Æ reste proche de Ç

Si les écarts sont élevés et une partie de la distribution est concentrée, la courbe reste proche de l'axe horizontal.



<u>Définition</u> : l'indice de Gini est le <u>double de l'aire comprise entre Ç et Æ</u>.

Dans le *cas 1* : On a i_G = 2 * [aire du triangle – somme des trapèzes]

i_G = 2 * [½ - Œ ($F_{i+1} - F_i$) / 2($y_i + y_{i+1}$)]

$$i_G = 1 - \text{Œ}\ (F_{i+1} - F_i) / (y_i + y_{i+1})$$

B. Médiale

e Dans le *cas 1* : $x_i > 0$ et on pose $\text{î}_i = f_i x_i / x_{moy}$

<u>Définition</u> : La médiale de (x_i ; n_i) est la médiane de la série (xi ; î i).
<u>Remarques</u> : La médiale est une approximation de l'indice de Gini qui correspond à l'approximation de Æ sous forme d'arc de cercle.

e Dans le *cas 2*, on utilisera le centre des classes.