Notions fondamentales en statistique

Frédéric Bertrand et Myriam Maumy

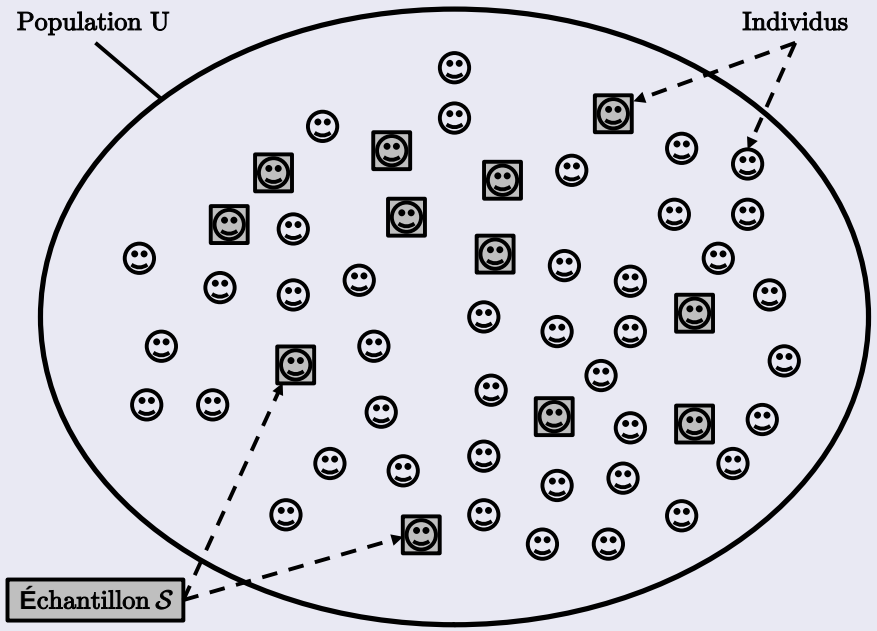
IRMA, UMR 7501, Université de Strasbourg

08 septembre 2011

1ère partie

La statistique

Les deux branches de la statistique



- **Statistique descriptive :** déterminer les caractéristiques d'une population.
- **Statistique inférentielle :** extrapoler les résultats numériques obtenus sur un échantillon à la population.

Objectif de la statistique descriptive

L'objectif de la statistique descriptive est de présenter et de décrire, c'est-à-dire de résumer numériquement et/ou de représenter graphiquement, les données disponibles quand elles sont nombreuses ou les données provenant d'un recensement.

Que trouvons-nous dans la statistique descriptive ?

- Le concept de population,
- le concept de résumés numériques, avec les trois sortes de caractéristiques : position, dispersion et forme.
- le concept de représentations graphiques, comme par exemple la boîte à moustaches ou l'histogramme.

Définition

L'ensemble sur lequel porte l'activité statistique s'appelle *la population*. Elle est généralement notée Ω. Ses éléments sont les *individus*.

Remarque

Ces individus peuvent être de natures très diverses : ensemble de personnes, mois d'une année, pièces produites par une usine, résultats d'expériences répétées un certain nombre de fois...

Définition

Les caractéristiques étudiées sur les individus d'une population sont appelées les *caractères*. Un *caractère* est donc une application χ d'un ensemble fini Ω (la population) dans un ensemble C (l'*ensemble des valeurs du caractère*), qui associe à chaque individu ω de Ω la valeur $\chi(\omega)$ que prend ce caractère sur l'individu ω.

Définition

La suite des valeurs $\chi(\omega)$ prises par χ s'appelle les *données brutes*. C'est une suite finie $(X_1, X_2, \ldots, X_N)$ de l'ensemble C.

Nous considérons plusieurs types de caractères :

1. les caractères qualitatifs
2. les caractères quantitatifs : leur détermination produit un nombre ou une suite de nombres. Nous distinguons
   1. les caractères simples : leur mesure sur un individu produit un seul nombre. L'ensemble de leurs valeurs est donc $\mathbb{R}$ ou une partie de $\mathbb{R}$.
   2. les caractères multiples : leur mesure sur un individu produit une suite finie de nombres. L'ensemble de leurs valeurs est donc $\mathbb{R}^n$ ou une partie de $\mathbb{R}^n$.

caractères qualitatifs

profession, adresse, situation de famille, sexe ...

caractères quantitatifs simples

taille, poids, salaire, température...

caractères quantitatifs multiples

relevé de notes d'un(e) étudiant(e), fiche de salaire,...

Remarque

Les caractères qualitatifs peuvent toujours être transformés en caractères quantitatifs par codage. C'est ce qui se fait le plus généralement. Mais un tel codage est purement conventionnel et n'a pas vraiment un sens quantitatif. Par exemple, on ne pourra pas calculer le sexe moyen.

Si X est un caractère quantitatif simple l'ensemble $X(\Omega) = \{X_1, X_2, \ldots, X_N\}$ des valeurs atteintes par le caractère (ou données brutes) est un ensemble fini $\{x_1, \ldots, x_n\}$. Nous supposerons que ces valeurs sont ordonnées :

$$x_1 < x_2 < \ldots < x_n.$$

Le fait que telle valeur soit relative à tel individu est un renseignement qui n'intéresse pas le statisticien. Seul l'ensemble des valeurs atteintes et le nombre de fois que chacune d'elle est atteinte est utile.

Définition

Nous appelons

- **effectif de la valeur x_i :** le nombre n_i de fois que la valeur x_i est prise, c'est-à-dire le cardinal de l'ensemble $X^{-1}(x_i)$;
- **effectif cumulé en x_i :** la somme $\sum_{j=1}^{i} n_j$;
- **fréquence de la valeur x_i :** le rapport $f_i = \frac{n_i}{N}$ de l'effectif de x_i à l'effectif total N de la population, c'est-à-dire le cardinal de Ω ou encore la somme des n_i ;
- **fréquence cumulée en x_i :** la somme $\sum_{j=1}^{i} f_j$.

Définition

Ces distributions statistiques sont qualifiées de *discrètes*.

Remarque

Lorsque le nombre des valeurs atteintes est important, nous préférons regrouper les valeurs en classes pour rendre la statistique plus lisible.
Nous partageons alors l'ensemble C des valeurs du caractère en classes $]a_i, a_{i+1}]$ avec $a_i < a_{i+1}$.
Nous parlons alors de statistique *groupée* ou *continue*.

Définition

Nous appelons

- **effectif de** $]a_i, a_{i+1}]$: le nombre n_i de valeurs prises dans $]a_i, a_{i+1}]$, c'est-à-dire $X^{-1}(]a_i, a_{i+1}])$;
- **effectif cumulé en** a_i : le nombre de valeurs prises dans $]-\infty, a_i]$;
- **fréquence de** $]a_i, a_{i+1}]$: le rapport $f_i = \dfrac{n_i}{N}$;
- **fréquence cumulée en** a_i : la somme $\displaystyle\sum_{j=1}^{i} f_j$.

Définition

La famille $(x_i, n_i)_{i=1,\dots,n}$ ou $(x_i, f_i)_{i=1,\dots,n}$ est encore appelée distribution statistique discrète.

Définition

De même, la famille $(]a_i, a_{i+1}], n_i)_{i=1,\dots,n}$ ou $(]a_i, a_{i+1}], f_i)_{i=1,\dots,n}$ est encore appelée distribution statistique groupée ou continue.

Définition

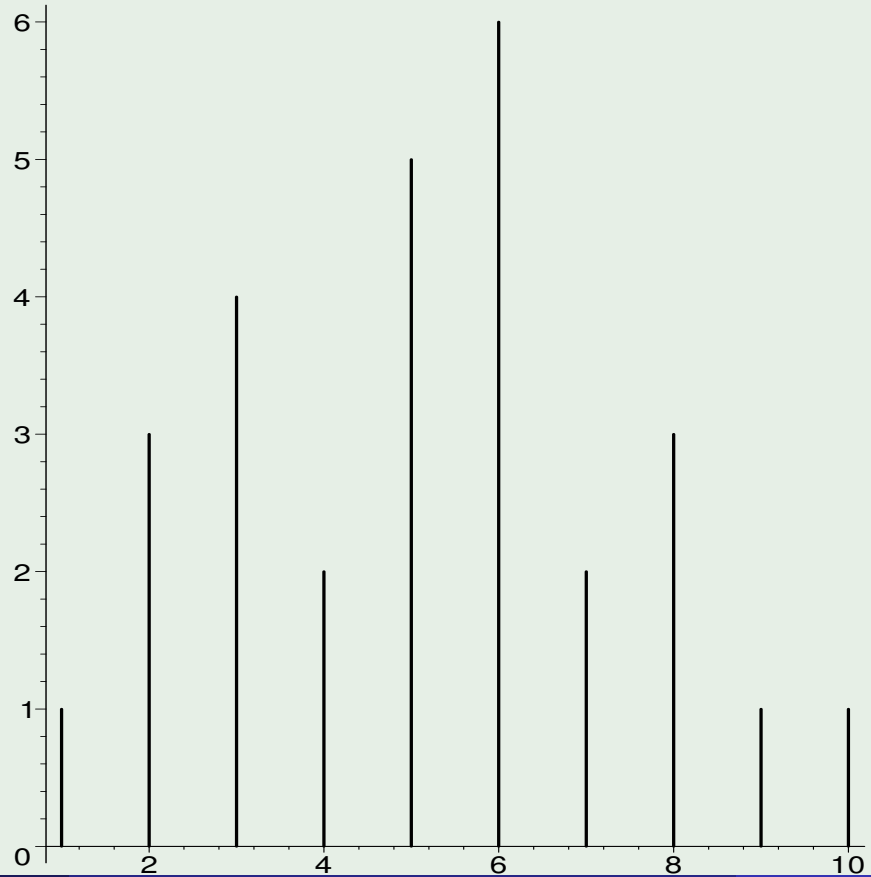
Le diagramme en bâtons d'une distribution statistique discrète est constitué d'une suite de segments verticaux d'abscisses x_i dont la longueur est proportionnelle à l'effectif ou la fréquence de x_i.

Exemple

La distribution suivante

$$(1,1),(2,3),(3,4),(4,2),(5,5),(6,6),(7,2),(8,3),(9,1),(10,1)$$

est représentée par le diagramme en bâtons de la figure 1



Définition

Le polygone des fréquences (resp. des effectifs) est obtenu à partir du diagramme en bâtons des fréquences (resp. des effectifs) en joignant par un segment les sommets des bâtons.

Remarque

Le graphique de la figure suivante superpose le polygone des effectifs et le diagramme en bâtons des effectifs de l'exemple précédent.

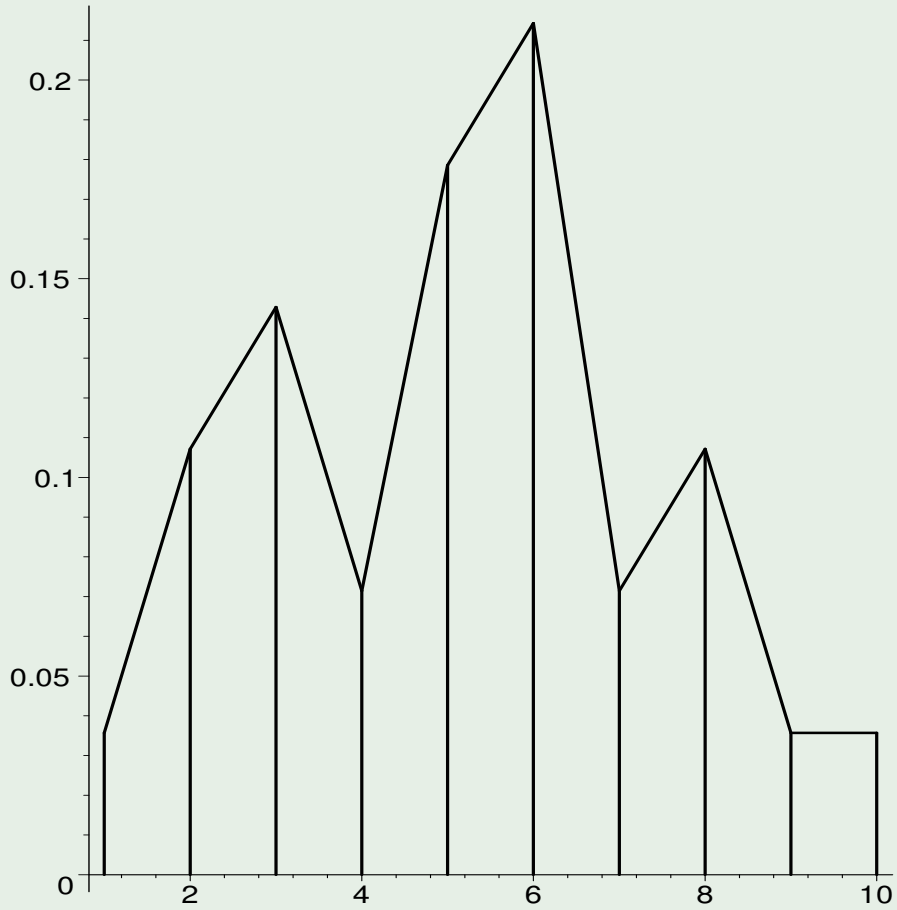


Figure: Diagramme en bâtons et polygone des effectifs

Définition

En remplaçant les fréquences (resp. les effectifs) par les fréquences cumulées (resp. les effectifs cumulés) on obtient le diagramme en bâtons et le polygone des fréquences cumulées (resp. des effectifs cumulés).

La figure 3 donne le diagramme en bâtons et le polygone des effectifs cumulés de l'exemple précédent.

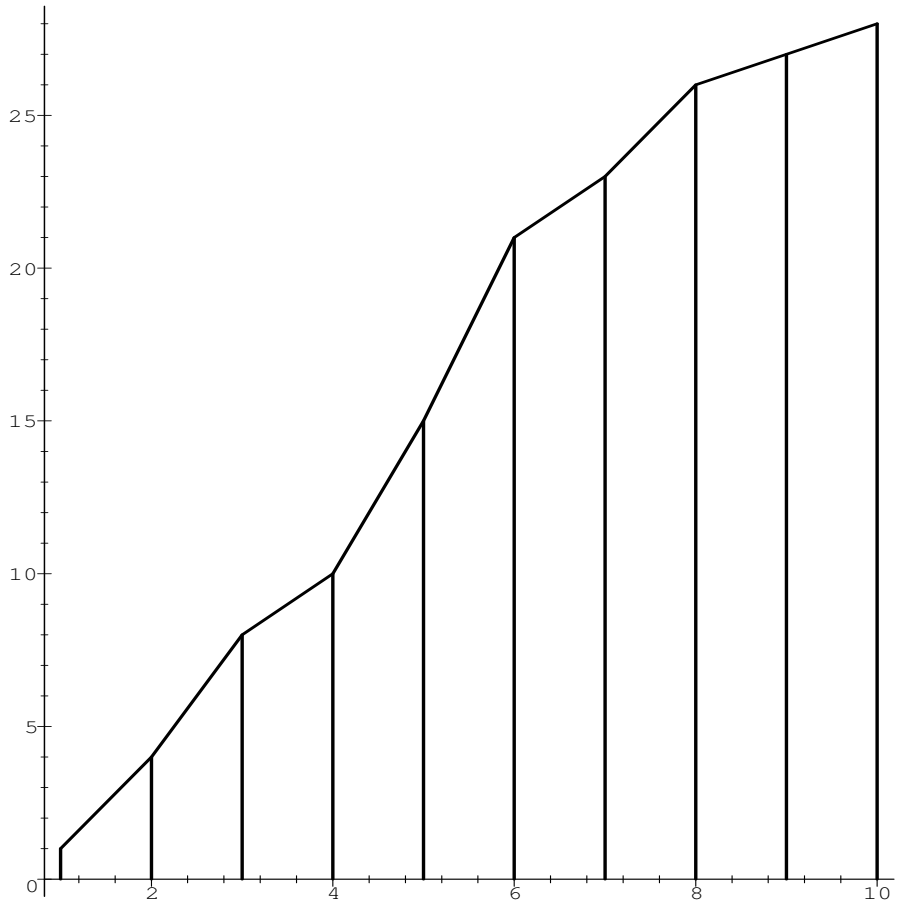


Figure: Diagramme en bâtons et polygone des effectifs cumulés

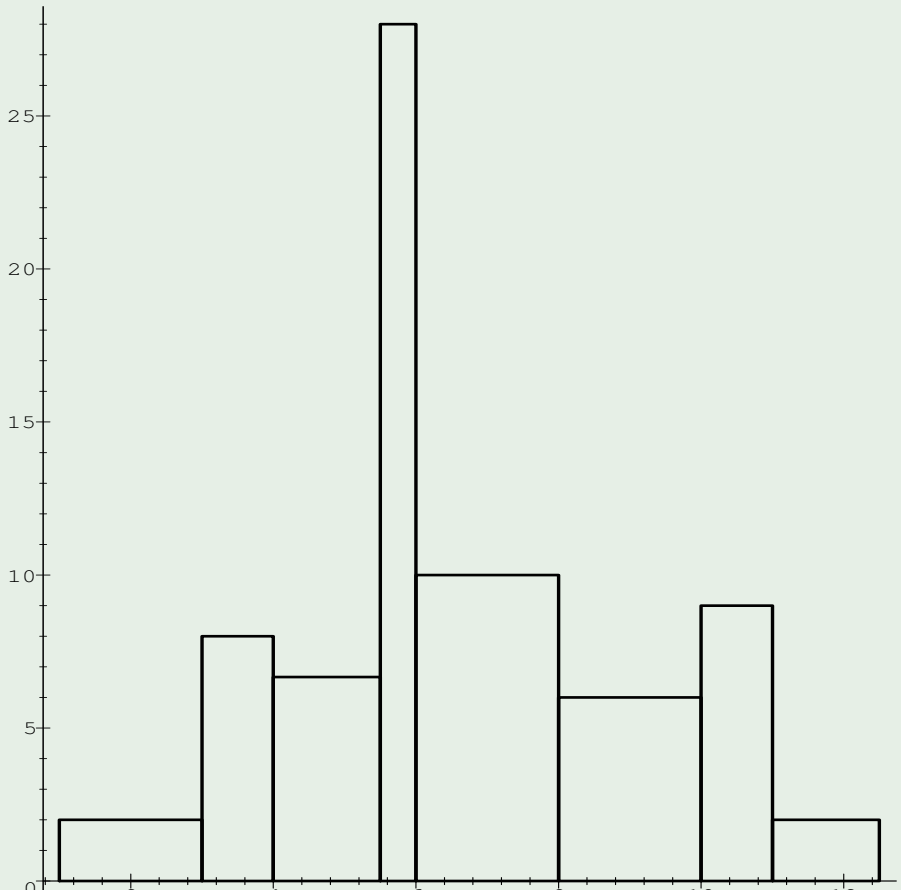
Définition

Nous appelons histogramme la représentation graphique d'une variable continue. Dans le cas où les amplitudes des classes sont égales, cet histogramme est constitué d'un ensemble de rectangles dont la largeur est égale à a, l'amplitude de la classe, et la hauteur égale à $K \times n_j$ où n_j est l'effectif de la classe et K est un coefficient arbitraire (choix d'une échelle), de sorte que l'aire totale sous l'histogramme est égale à $K \times N \times a$ où N est l'effectif total. Dans le cas de classes d'amplitudes $k_j \times a$ inégales, multiples entiers de l'une d'entre elles a, on convient, pour conserver le résultat précédent, de prendre pour hauteur du rectangle de la classe numéro j le quotient $\dfrac{K \times n_j}{k_j}$.

Exemple

En figure 4 nous donnons l'histogramme de la distribution suivante

$$(]1,3],4),\quad (]3,4],8),\quad (]4,5.5],10),\quad (]5.5,6],14),$$
$$(]6,8],20),\quad (]8,10],12),\quad (]10,11],9),\quad (]11,12.5],3).$$



Suite de l'Exemple

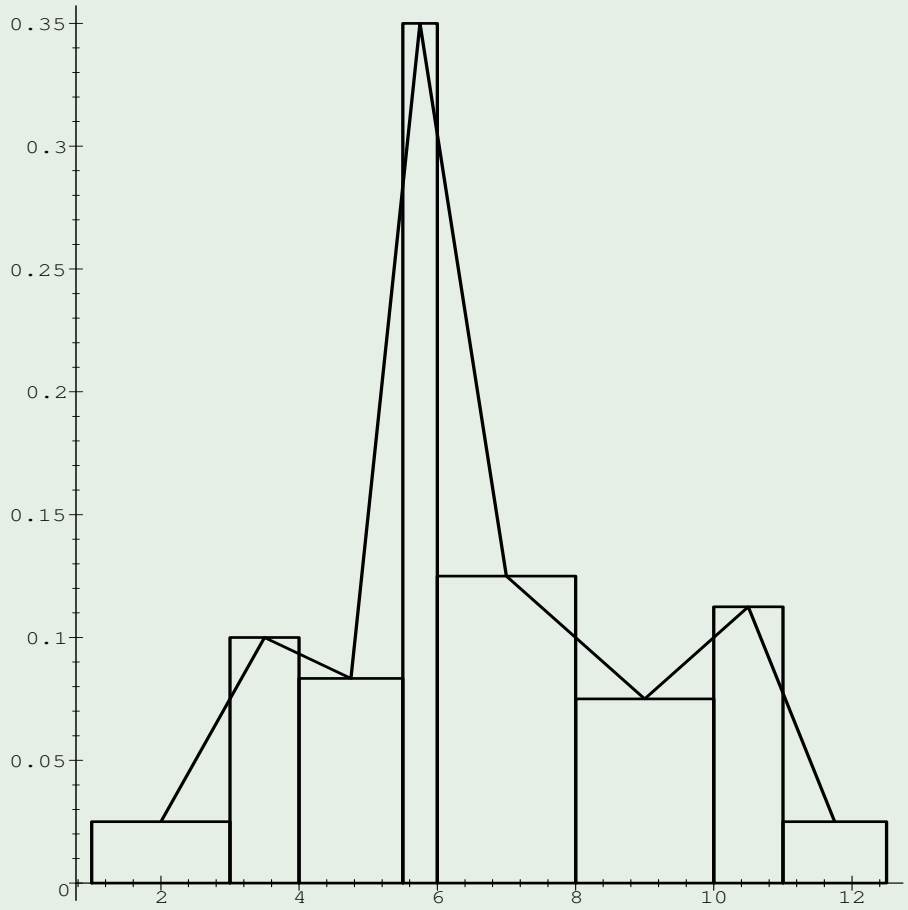


Figure: Histogramme et polygone des fréquences

Définition

Le polygone des effectifs ou des fréquences d'une distribution est obtenu en joignant dans l'histogramme de cette distribution les milieux des côtés horizontaux supérieurs.

Retour à l'Exemple.

La figure 5 superpose l'histogramme des fréquences de l'exemple précédent et son polygone des fréquences.

Définition

Le polygone des fréquences cumulées d'une distribution statistique groupée est la représentation graphique de la fonction définie par

$$f(x) = \sum_{j=1}^{i-1} f_j + \frac{x - a_i}{a_{i+1} - a_i} f_i$$

sur l'intervalle $]a_i; a_{i+1}]$.

Remarque

En particulier, remarquons que nous avons

$$f(a_i) = \sum_{j=1}^{i-1} f_j$$

et

$$f(a_{i+1}) = \sum_{j=1}^{i} f_j.$$

Retour à l'Exemple

Pour l'exemple précédent, nous obtenons le graphique de la figure 6.

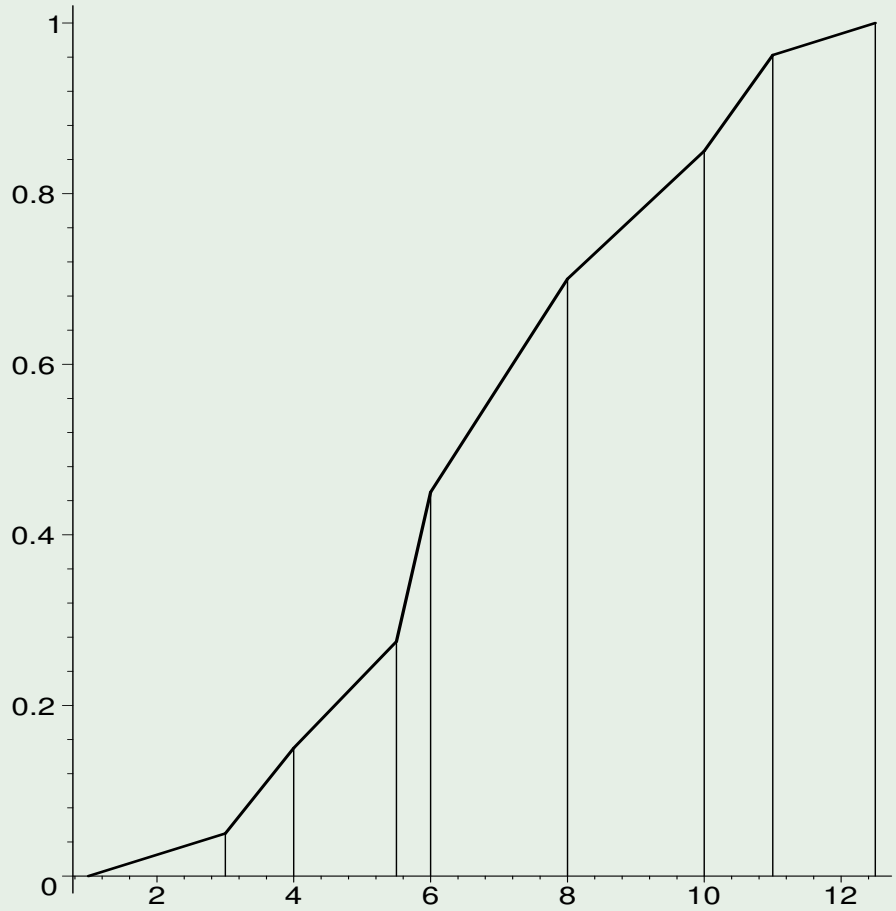


Figure: Polygone des fréquences cumulées d'une statistique groupée

Définition

Le **mode** est l'une des valeurs $x_1, x_2, \ldots, x_p$ dont la fréquence f_i est maximale.

Définition

La **classe modale** est une classe de densité, c'est-à-dire de rapport fréquence/longueur, maximale.

Définition

La distribution est unimodale si elle a un seul mode, si elle en a plusieurs elle est plurimodale (bimodale, trimodale, ...).

Remarque

Nous déterminons aisément les modes à partir des représentations graphiques.

Définition

Soit m et d les parties entière et décimale de $(N+1)/2$. La **médiane**, notée $\mathrm{Q}_2(x)$, est définie par

$$\mathrm{Q}_2(x) = x_{(m)} + d(x_{(m+1)} - x_{(m)})$$

où $x_{(m)}$ signifie la m-ième valeur lorsque la série des valeurs est classée par ordre croissant.
$x_{(m)}$ est aussi appelée la m-ième **statistique d'ordre**.

Définition

Pour tout nombre $\alpha \in]0;1[$, soit m et d les parties entière et décimale de $\alpha(N+1)$. Le **quantile d'ordre** α, noté $\mathrm{Q}_\alpha(x)$, est défini par :

$$\mathrm{Q}_\alpha(x) = x_{(m)} + d(x_{(m+1)} - x_{(m)}).$$

Définition

La moyenne d'une distribution statistique discrète $(x_i; f_i)_{i=1,\dots,p}$ est le nombre réel μ défini par

$$\mu = \sum_{i=1}^{p} x_i f_i = \frac{1}{N} \sum_{i=1}^{p} x_i n_i.$$

où N est l'effectif total de la population.

Remarque

Nous pouvons aussi la calculer directement à partir des données brutes par

$$\mu = \frac{1}{N} \sum_{j=1}^{N} X_j$$

c'est-à-dire en calculant le rapport entre la somme de toutes les valeurs relevée (avec répétitions éventuelles) et l'effectif total de la population.

Définition

Pour une statistique groupée $(]a_i; a_{i+1}], f_i)_{i=1,\dots,p}$ la moyenne se calcule par

$$\mu = \sum_{i=1}^{p} \frac{a_i + a_{i+1}}{2} f_i.$$

Cela revient à faire une hypothèse d'homogénéité en considérant les valeurs équidistribuées à l'intérieur d'une classe ou, au contraire, à supposer que toute la fréquence est concentrée au centre de la classe (ce qui revient au même : on remplace la distribution à l'intérieur de la classe par son barycentre).

Remarque

La moyenne de $X - a$ est $\mu - a$ et la moyenne de λX est $\lambda\mu$.

Remarque

Il existe d'autres moyennes :

- la moyenne géométrique,
- la moyenne harmonique,
- la moyenne arithmético-géométrique,
- ...

Il est à noter qu'il est intéressant de comparer les deux principaux paramètres de position que sont la médiane et la moyenne arithmétique. Les deux possèdent des avantages et des inconvénients.

1. Pour la médiane, nous avons
   - Avantage :
     - Peu sensible aux valeurs extrêmes (paramètre robuste).
   - Inconvénients :
     - Délicate à calculer (Rappelez-vous les différentes définitions que l'on peut rencontrer).
     - Ne se prête pas aux calculs algébriques.
2. Pour la moyenne arithmétique, nous avons
   - Avantages :
     - Facile à calculer.
     - Se prête bien aux calculs algébriques.
     - Répond au principe des moindres carrés.
   - Inconvénients :
     - Fortement influencée par les valeurs extrêmes.
     - Mauvais indicateur pour une distribution polymodale ou fortement asymétrique.

Quelques caractéristiques de dispersion

- La **variance**, notée $\sigma^2(x)$, est le nombre réel positif défini par

$$\sigma^2(x) = \sum_{i=1}^{p} (x_i - \mu(x))^2 f_i.$$

- L'**écart-type**, noté $\sigma(x)$, est la racine carrée de la variance. Il s'exprime dans la même unité que la moyenne.
- La **médiane des écarts absolus à la médiane**, notée $\mathrm{MAD}(x)$, d'une série statistique est le nombre réel défini par

$$\mathrm{MAD}(x) = \mathrm{Q}_2\left((|x_i - \mathrm{Q}_2(x)|)_{1 \leqslant i \leqslant n}\right).$$

- L'**intervalle inter-quartile**, noté $IIQ(x)$, est la différence entre le troisième quartile et le premier quartile.

Caractéristiques de forme

- Le **moment centré d'ordre** r est égal à ${}_{\mu}m_r(x) = \sum_{i=1}^{p}(x_i - \mu(x))^r f_i$.
- Le **coefficient d'asymétrie (skewness) de Fisher** est la quantité $\gamma_1(x)$ définie par

$$\gamma_1(x) = \frac{{}_{\mu}m_3(x)}{\sigma^3(x)} = \frac{{}_{\mu}m_3(x)}{({}_{\mu}m_2(x))^{3/2}}.$$

- Le **coefficient d'asymétrie de Pearson** est la quantité $\beta_1(x)$ définie par

$$\beta_1(x) = \frac{({}_{\mu}m_3(x))^2}{(\sigma^2(x))^3} = \frac{({}_{\mu}m_3(x))^2}{({}_{\mu}m_2(x))^3} = \gamma_1^2(x).$$

Caractéristiques de forme

- Le **coefficient d'aplatissement (kurtosis) de Fisher** est la quantité $\gamma_2(x)$ définie par

$$\gamma_2(x) = \frac{{}_\mu m_4(x)}{({}_\mu m_2(x))^2} - 3.$$

- Le **coefficient d'aplatissement de Pearson** est la quantité $\beta_2(x)$ définie par

$$\beta_2(x) = \frac{{}_\mu m_4(x)}{({}_\mu m_2(x))^2} = \frac{{}_\mu m_4(x)}{\sigma^4(x)}.$$