Notions fondamentales en statistique

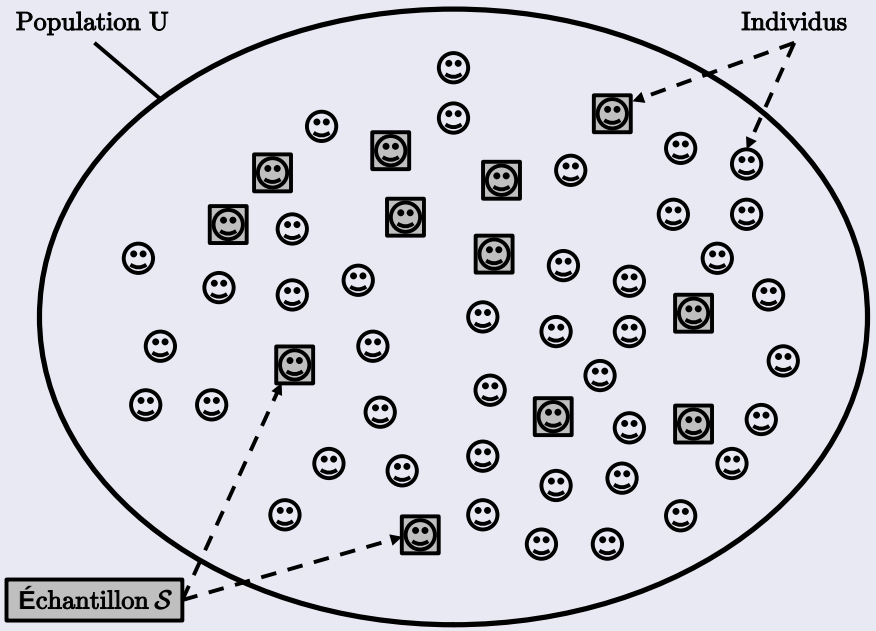
Frédéric Bertrand et Myriam Maumy-Bertrand

IRMA, UMR 7501, Université de Strasbourg

10 septembre 2013

Sommaire

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** :
  1. position,
  2. dispersion,
  3. forme.
- le concept de **représentations graphiques**, comme par exemple la boîte à moustaches ou l'histogramme.

Sommaire

Définition

L'ensemble sur lequel porte l'activité statistique s'appelle **la population**. Elle est généralement notée Ω, pour rappeler la notation des probabilités ou U, U comme Univers, notation souvent utilisée dans la théorie des sondages.
Le cardinal d'une population est en général noté N.

Définition

Les éléments de la population sont appelés les **individus** ou les **unités statistiques**.

Remarque

Les individus d'une population peuvent être de natures très diverses.

Exemples

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 de valeurs $(X_1, X_2, \ldots, X_N)$ de l'ensemble C.

Sommaire

Il existe deux types de caractères :

1. les **caractères qualitatifs** : les modalités ne sont pas mesurables et peuvent être décrites par un mot, un groupe de mots ou une phrase. Ils peuvent être de nature ordinale ou nominale,
2. les **caractères quantitatifs** : leur détermination produit un nombre ou une suite de nombres. Nous distinguons :
   - Les caractères discrets. Ils ne peuvent prendre que certaines valeurs particulières.
   - Les caractères continus. Ils peuvent prendre des valeurs réelles quelconques. Si nous prenons deux valeurs quelconques du caractère aussi rapprochées soient-elles, il existe toujours une infinité de valeurs comprises entre elles.

   Puis nous distinguons également :
   - 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 $\mathbb{R}$.
   - 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 nominaux

Profession, adresse, situation de famille, genre ...

Caractères qualitatifs ordinaux

Départements, ...

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, nous ne pourrons pas calculer le genre moyen.

Remarque

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 sont utiles.

Sommaire

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 $\sum_{j=1}^{i} f_j$.

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 $(x_i; n_i)_{i=1,\ldots,n}$ ou $(x_i; f_i)_{i=1,\ldots,n}$ est encore appelée **distribution statistique discrète**.

Définition

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

Sommaire

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
$(1,1),(2,3),(3,4),(4,2),(5,5),(6,6),(7,2),(8,3),(9,1),(10,1)$ est représentée par le diagramme en bâtons de la figure ci-dessous.

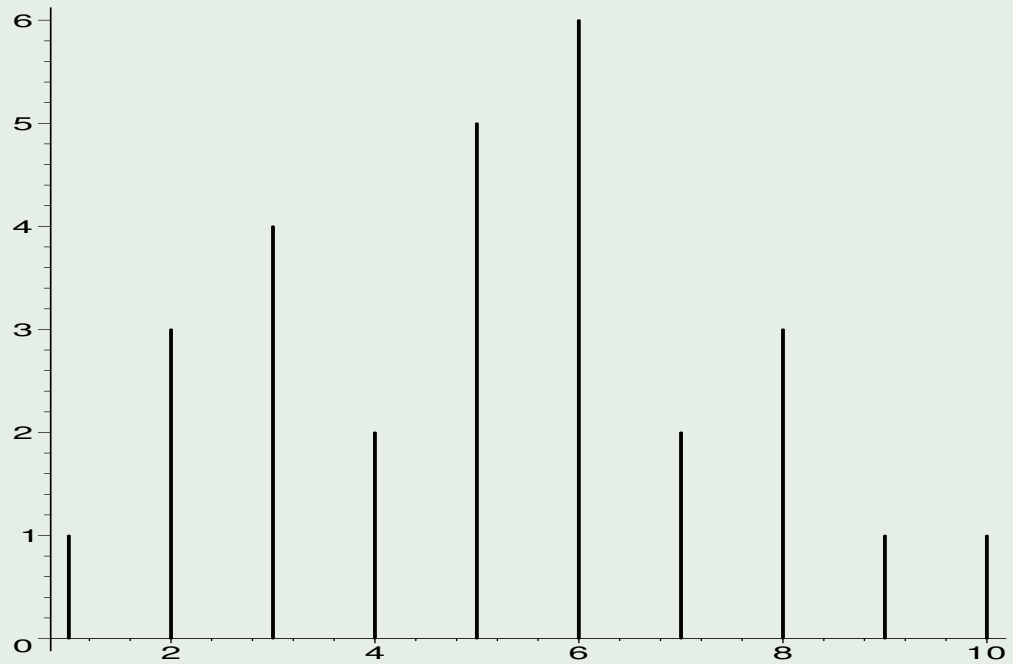


Figure : Diagramme en bâtons

Définition

Le **polygone des effectifs** (respectivement des fréquences) est obtenu à partir du diagramme en bâtons des effectifs (respectivement des fréquences) 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.

Suite de l'exemple

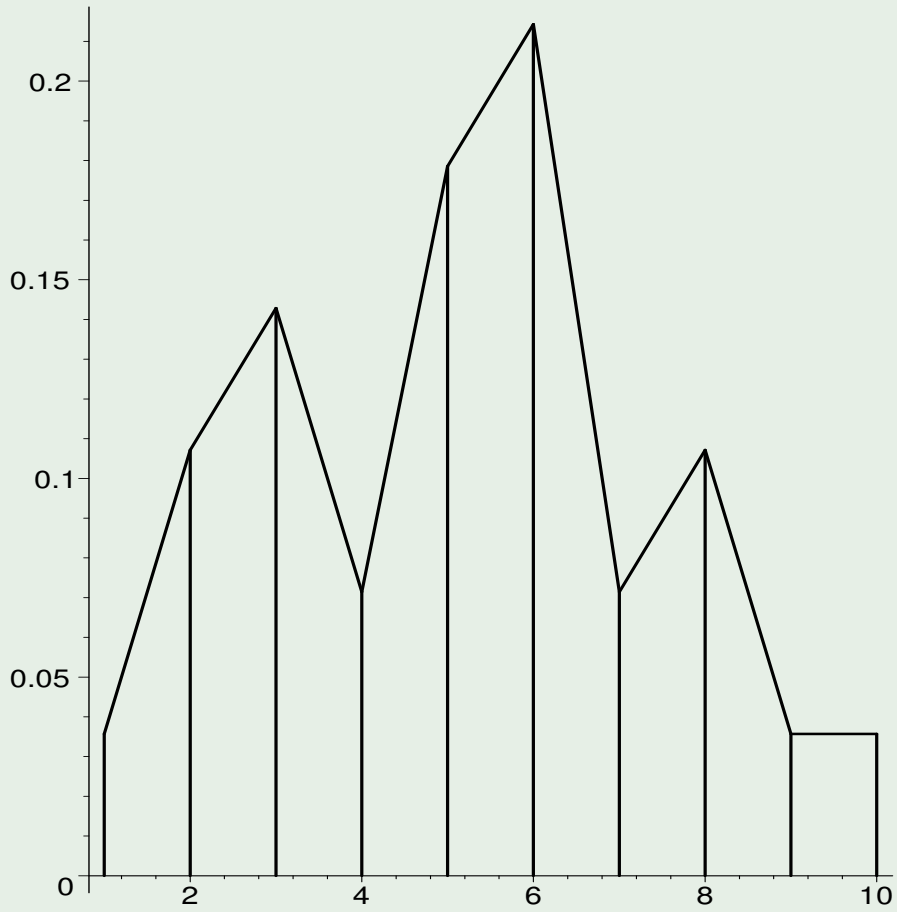


Figure : Diagramme en bâtons et polygone des fréquences

Définition

En remplaçant les effectifs (respectivement les fréquences) par les effectifs cumulés (respectivement les fréquences cumuléss) nous obtenons le diagramme en bâtons et le polygone des effectifs cumulés (respectivement des fréquences cumulées).

Suite de l'exemple

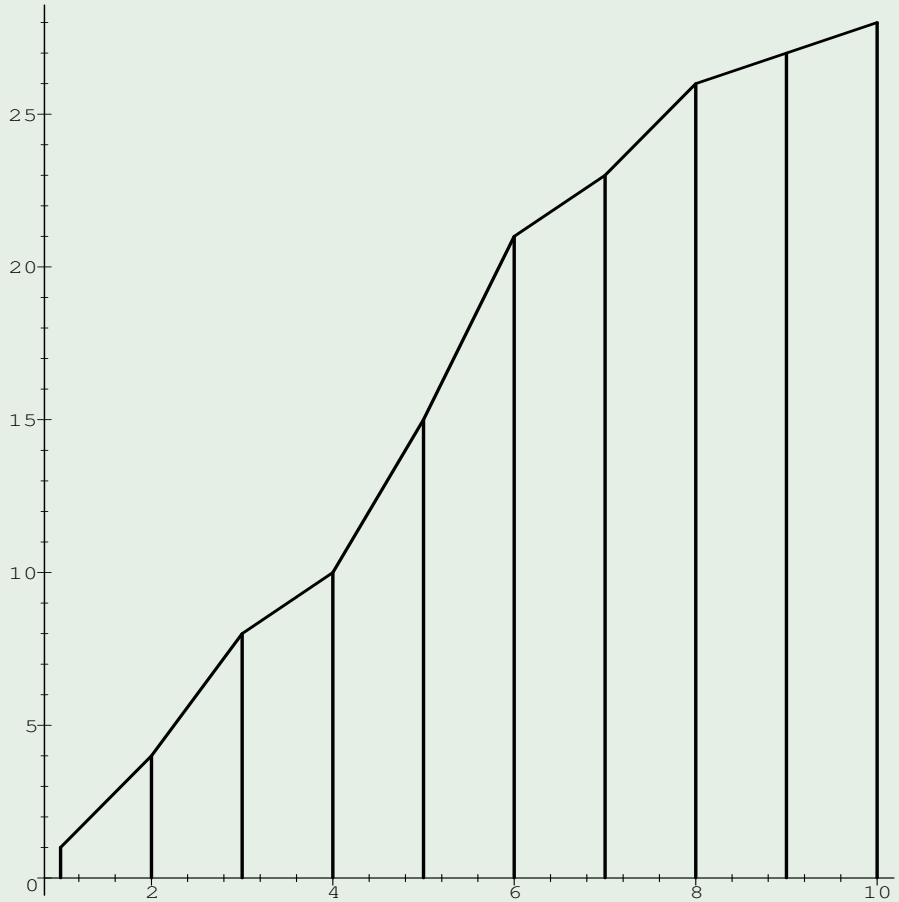


Figure : Diagramme en bâtons et polygone des effectifs cumulés

Définition

Nous appelons **histogramme** la représentation graphique d'un caractère quantitatif continu.

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 où les **classes sont d'amplitudes $k_j \times a$ inégales**, multiples entiers de l'une d'entre elles a, nous convenons, 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

Nous donnons l'histogramme de la distribution suivante :
$(]1;3],4),(]3;4],8),(]4;5,5],10),(]5,5;6],14),(]6;8],20),(]8;10],12),(]10;11],9),(]11;12,5],3)$.

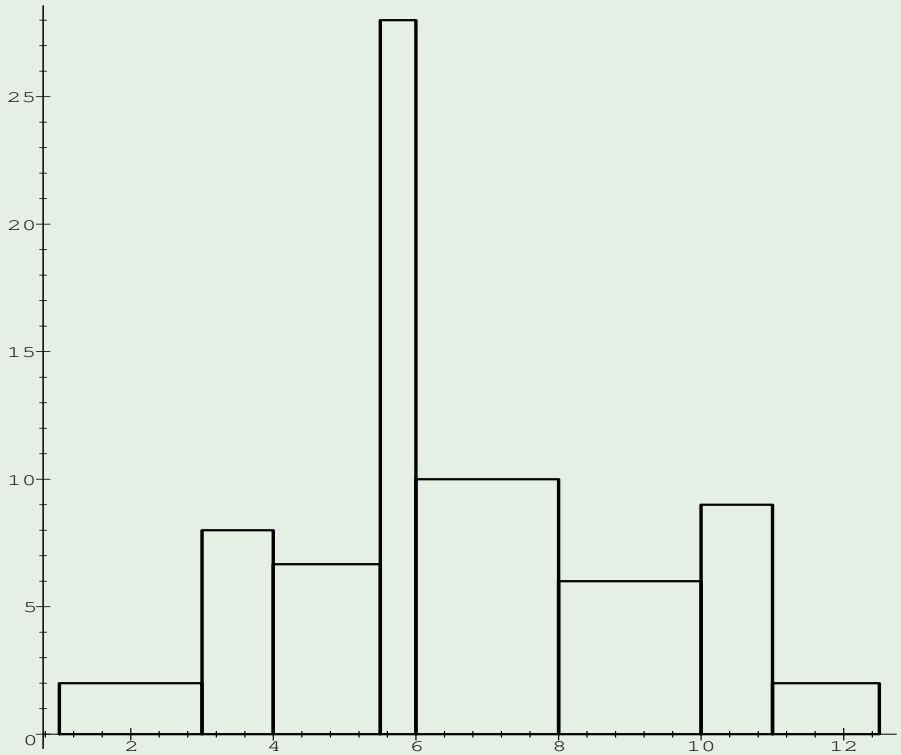


Figure : Histogramme

Définition

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

Suite de l'Exemple

La figure suivante superpose l'histogramme des fréquences de l'exemple précédent et le polygone des fréquences associé.

Suite de l'Exemple

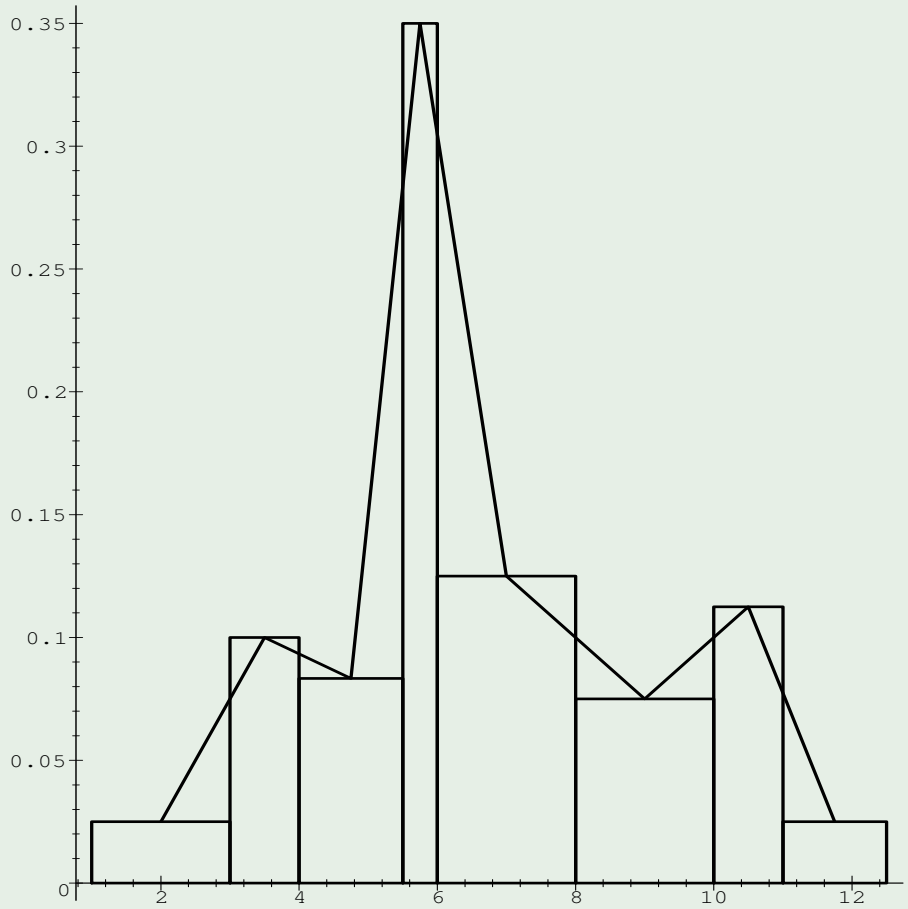


Figure : Histogramme et polygone des fréquences

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

Suite de l'Exemple

Pour l'exemple précédent, nous obtenons la figure ci-dessous.

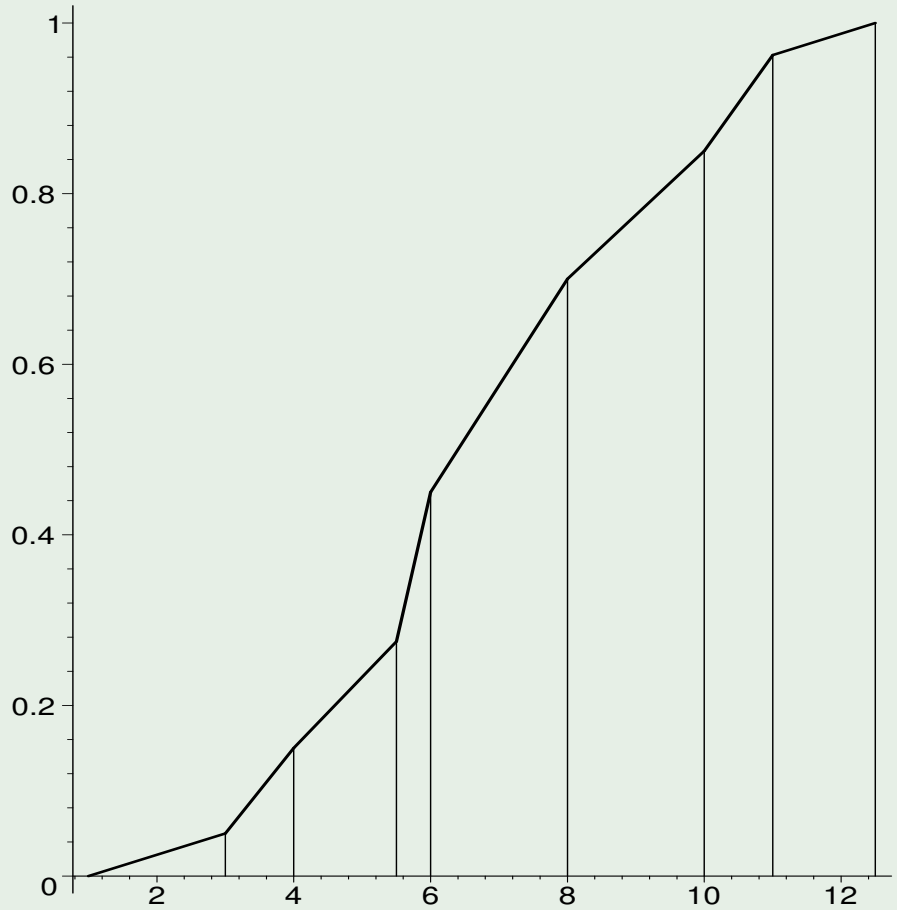


Figure : Polygone des fréquences cumulées d'une distribution statistique groupée

Sommaire

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

$$\mu(x) = \sum_{i=1}^{p} x_i f_i = \frac{1}{N} \sum_{i=1}^{p} x_i n_i,$$

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 distribution statistique groupée $(]a_i; a_{i+1}], f_i)_{i=1,\ldots,p}$ la moyenne arithmétique se calcule par :

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

Remarque

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 : nous remplaçons la distribution à l'intérieur de la classe par son isobarycentre).

Remarque

Il existe d'autres moyennes :

- la moyenne arithmétique pondérée,
- la moyenne géométrique,
- la moyenne quadratique,
- la moyenne harmonique,
- la moyenne arithmético-géométrique,
- ...

Remarque

Comparons les deux principales caractéristiques de position : la médiane et la moyenne arithmétique.

1. Pour la médiane,
   - Avantage :
     - Peu sensible aux valeurs extrêmes (caractéristique robuste).
   - Inconvénients :
     - Délicate à calculer (différentes définitions).
     - Ne se prête pas aux calculs algébriques.
2. Pour la moyenne arithmétique,
   - 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.

Exemple

Tableau - Nombre d'heures travaillées par semaine des personnes ayant un emploi à plein temps (2006).

Pays	Durée (heures)
Allemagne	41,7
Autriche	44,1
Belgique	41,0
Chypre	41,8
Danemark	40,5
Espagne	42,2
Estonie	41,5
Finlande	40,5

Exemple (suite)

Pays	Durée (heures)
France	$\mathbf{41,0}$
Grèce	$44,1$
Hongrie	$41,0$
Irlande	$40,7$
Italie	$41,3$
Lettonie	$43,0$
Lituanie	$39,8$
Luxembourg	$40,9$
Malte	$41,2$

Exemple (suite)

Pays	Durée (heures)
Pays-Bas	40,8
Pologne	42,9
Portugal	41,6
République Tchèque	42,7
Royaume-Uni	43,1
Slovaquie	41,6
Slovénie	42,5
Suède	41,1

Source : Eurostat.

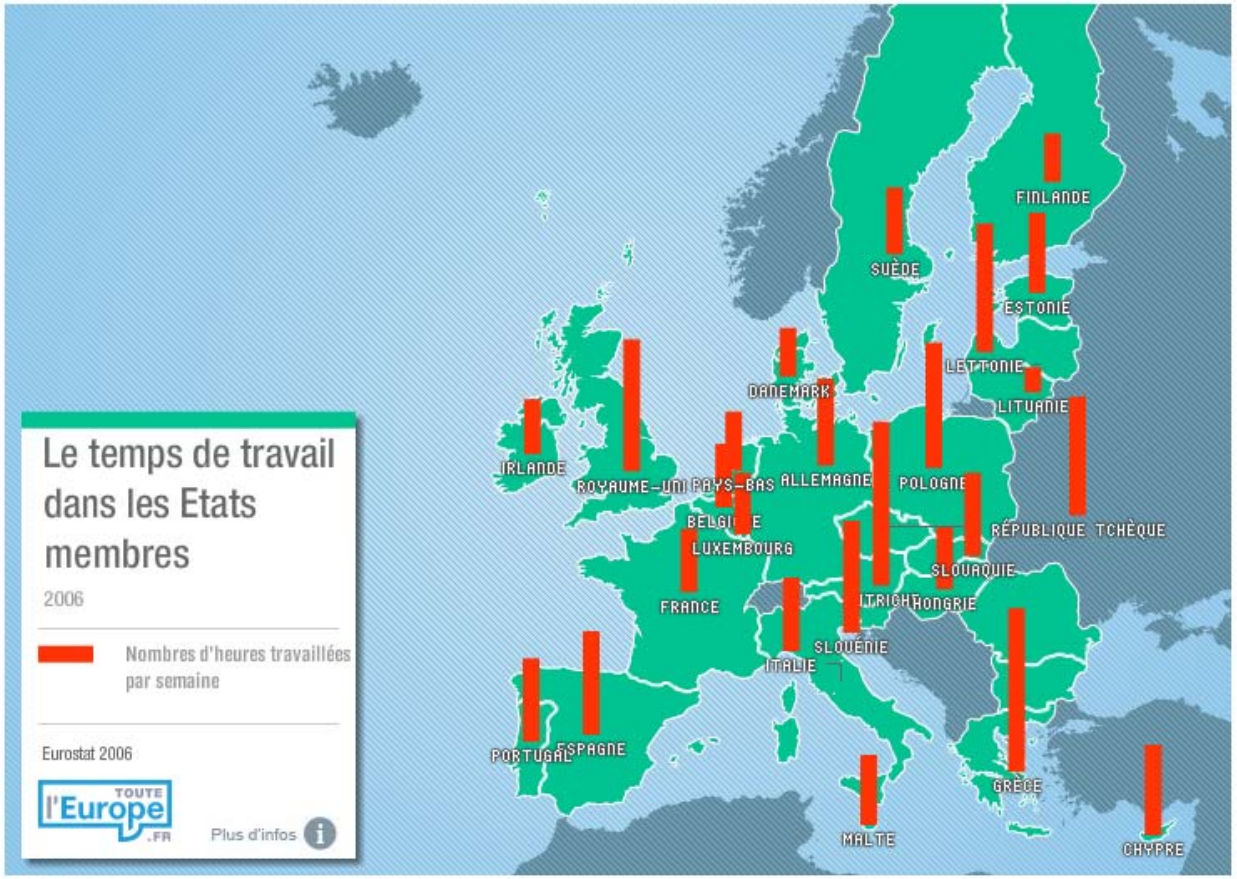


Figure : Carte des durées de travail

Exercice : c'est à vous

Calculer sur cet exemple

- le mode,
- la médiane,
- le premier quartile,
- le troisième quartile et
- la moyenne arithmétique.

Solutions

Statistiques descriptives :

N	Mode	Médiane	Q1	Q3	Moyenne
25	41,000	41,500	40,950	42,600	41,704

Sommaire

Définition

L'**étendue** est la différence entre le maximum et le minimum.

Définition

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

Définition

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

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

Définition

L'**écart-type**, noté $\sigma(x)$, est la racine carrée de la variance. Il s'exprime dans la même unité que la moyenne.

Définition

La **médiane des écarts absolus à la médiane**, notée $\mathrm{MAD}(x)$, d'une série statistique est le nombre réel défini par :

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

Exercice : c'est à vous

Calculer sur l'exemple d'Eurostat

- l'étendue,
- l'intervalle inter-quartile,
- l'étendue inter-quartile,
- la variance et
- l'écart-type.

Solutions

```
Statistiques descriptives :

Etendue                   IIQ         EIQ      Variance
  4,300      [40,950;42,600]       1,650         1,240

EcTyp
1,113
```

Sommaire

Définition

Le **moment centré d'ordre** r est égal à :

$$_{\mu}m_r(x) = \sum_{i=1}^{p}(x_i - \mu(x))^r f_i.$$

Définition

Le **coefficient d'asymétrie (skewness) de Fisher** est la quantité $\gamma_1(x)$ définie par :

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

Définition

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

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

Définition

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

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

Définition

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

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

Solutions

```
Statistiques descriptives :
Asymétrie de Fisher     Aplatissement de Fisher
                0,71                       -0,06
```

Sommaire

Objectif

La **boîte de distribution**, « box-plot » en anglais, ou encore « boîte à moustaches », « boîte de dispersion », « diagramme de Tukey » en français, fournit en un seul coup d'oeil les informations sur la tendance centrale, la dispersion, l'asymétrie et l'importance des valeurs extrêmes de la série de données que nous avons à explorer.

Elle est aussi particulièrement intéressante pour la comparaison de distributions sur plusieurs de ces critères.

Construction

Dans une **boîte de distribution** :

- la boîte représente l'intervalle interquartile ;
- à l'intérieur, la médiane sépare la boîte en deux parties ;
- les lignes qui partent du bord de la boîte s'étendent jusqu'aux valeurs les plus extrêmes qui ne sont pas considérées comme trop éloignées du reste de l'échantillon.

La plupart des logiciels de statistique note « valeur éloignée » les points situés à plus de 1,5 fois l'étendue interquartile par rapport aux bords de la boîte, et « valeur extrême », les points situés à plus de 3 fois l'étendue interquartile.

Construction (suite)

Anisi, la taille de la boîte représente l'étendue interquartile, la position de la médiane est un bon indicateur de la symétrie de la distribution, la taille des lignes de part et d'autre de la boîte traduit la dispersion, et les valeurs éloignées ou extrêmes sont immédiatement repérées.

Construction – détails

Nous représentons une **boîte de distribution** de la façon suivante :

1. Nous traçons un rectangle de largeur fixée à priori et de longueur $EIQ = Q_{0,75} - Q_{0,25}$.
2. Ensuite nous y situons la médiane par un segment positionné à la valeur $Q_{0,5}$, par rapport à $Q_{0,75}$ et $Q_{0,25}$. Nous avons alors la boîte.
3. Nous calculons $(Q_{0,75} + 1,5 \times EIQ)$ et nous cherchons la dernière observation x_h en deçà de la limite $(Q_{0,75} + 1,5 \times EIQ)$, soit

$$x_h = \max \{x_i / x_i \leqslant Q_{0,75} + 1,5 \times EIQ\}.$$

Construction – détails

4. Nous calculons $(Q_{0,25} - 1,5 \times EIQ)$ et nous cherchons la première observation x_b au delà de la limite $(Q_{0,25} - 1,5 \times EIQ)$, soit

$$x_b = \min \{x_i / x_i \geqslant Q_{0,25} - 1,5 \times EIQ\}.$$

5. Nous traçons deux lignes allant des milieux des largeurs du rectangle aux valeurs x_b et x_h.
6. Les observations qui ne sont pas comprises entre x_b et x_h sont représentés par des symboles spécifiques, généralement des étoiles.

Exemple

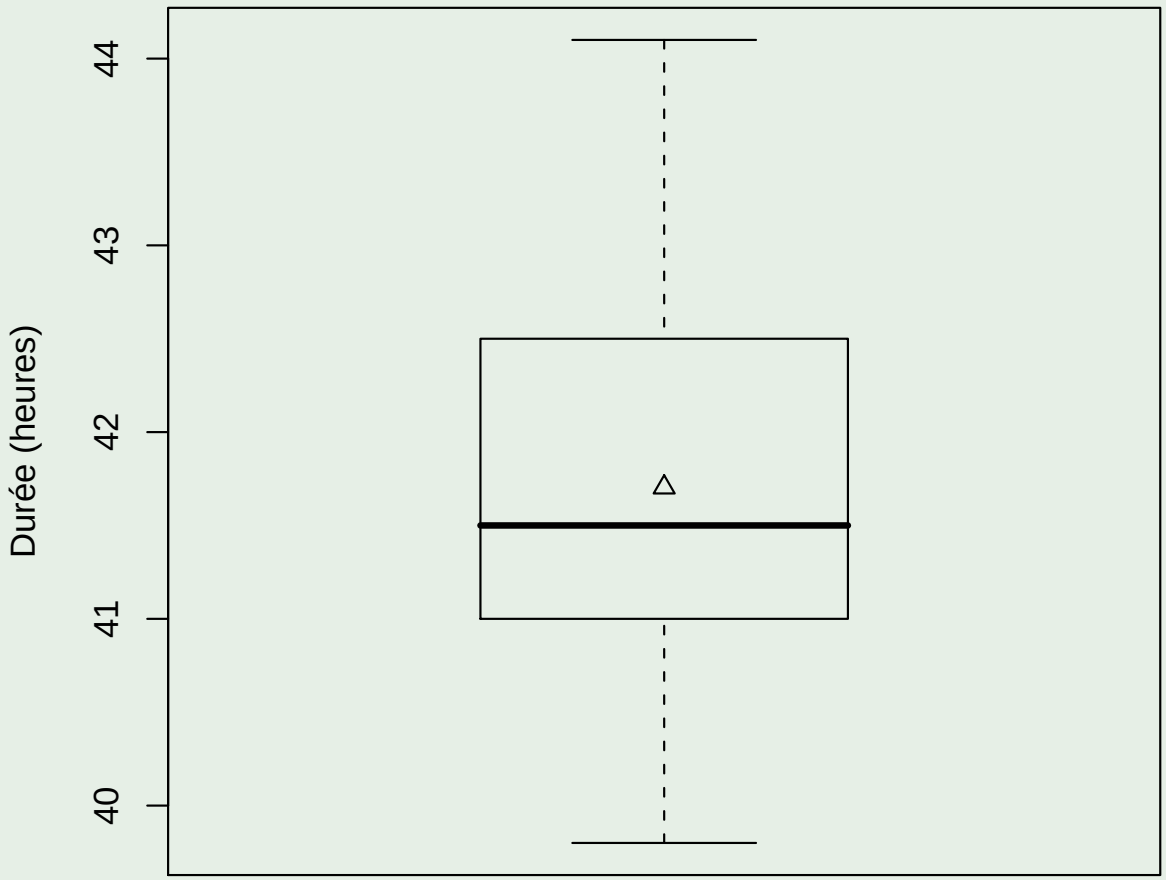


Figure : Boîte de distribution

Interprétation d'une boîte de distribution

Ce type de diagramme permet de comparer facilement plusieurs distributions en terme de médiane, quartiles et valeurs éloignées ou extrêmes.

Interprétation d'une boîte de distribution (suite)

Une **boîte de distribution** rend compte de la tendance centrale, de la dispersion, des valeurs éloignées ou extrêmes et de la forme de la distribution, même si d'autre modes de représentations peuvent apporter un complément d'information sur la forme.

Interprétation d'une boîte de distribution (suite)

Auparavant, nous avons mentionné l'importance du triplet (N, μ, σ). La **boîte de distribution** est un complément qui se révèle intéressant puisqu'elle permet de détecter l'asymétrie, les valeurs extrêmes, et de repérer la médiane et l'intervalle interquartile qui contient la moitié des observations.

Dans le cas d'une asymétrie, l'écart-type qui mesure la dispersion symétriquement par rapport à la moyenne n'est pas la mesure de dispersion la mieux adaptée, et peut-être complété par l'étendue interquartile. D'autre part, si la **boîte de distribution** indique des valeurs éloignées ou extrêmes, nous savons que la moyenne et l'écart-type sont particulièrement influencés par ces valeurs.