Statistique Univariée

Statistique descriptive

P. DOSSANTOS-UZARRALDE

Sommaire.

Sommaire.

Terminologie

- **Population (population statistique)** Ensemble (au sens mathématique du terme) concerné par une étude statistique (i.e. champ de l'étude). Ensemble des objets ou individus statistiques étudiés.
- **Individu (unité statistique)** Tout élément de la population.
- **Echantillon** Sous ensemble de la population sur lequel sont effectivement réalisées les observations.
- **Enquête (statistique)** Opération consistant à observer (ou mesurer, ou questionner) l'ensemble des individus d'un échantillon.
- **Recensement** Enquête dans laquelle l'échantillon observé est la population tout entière (enquête exhaustive).
- **Sondage** Enquête dans laquelle l'échantillon observé est un sous-ensemble strict de la population (enquête non exhaustive).
- **Variable (statistique)** : définie sur la population et observée sur l'échantillon. Mathématiquement, application définie sur l'échantillon. Variable à valeurs dans $\mathcal{R}$ (ou une partie de $\mathcal{R}$, ou un ensemble de parties de $\mathcal{R}$) = **quantitative** (âge, salaire, taille. . .) ; sinon **qualitative** (sexe, catégorie socioprofessionnelle. . .).
- **Données (statistiques)** : ensemble des individus observés (échantillon), des variables considérées, et des observations de ces variables sur ces individus. Présentées sous forme de tableaux (individus en lignes et variables en colonnes) et stockées dans un fichier informatique. Un tableau comportant que des nombres (valeurs des variables quantitatives ou codes associés aux variables qualitatives) correspond à la notion mathématique de matrice.

Variables qualitatives (1)

- Variable représentée par des qualités (caractéristiques).
- Modalités : les valeurs possibles de cette variable.
- Ensemble des modalités : $E = \{e_1, \ldots, e_n\}$.

On dit que X est une variable qualitative si les conditions suivantes sont remplies :

- Ses valeurs ne sont pas numériques
- Ses valeurs expriment des états uniques
- Ses valeurs peuvent être classifiées
- Ses valeurs ne sont pas exploitables mathématiquement
- Ses valeurs sont dénombrables

Variables qualitatives (2)

Exemple

On demande aux élèves d'une classe composée de dix élèves quelle est leur chaîne de télévision préférée. L'ensemble des modalités est E = TF1, F2, F3, C+, M6.
Les données brutes : suite du type : TF1, M6, M6, TF1, ..., C+, C+.

Définitions

Variables nominales et ordinales

• Variables qualitatives nominales :
Une variable X est dite qualitative nominale si ses valeurs ne sont pas hiérarchisables. Les variables correspondent à des noms. Il n'y a aucun ordre précis (sexe, langues parlées, ...)

• Variables qualitatives ordinales :
Une variable X est dite qualitative ordinale si ses valeurs sont hiérarchisables. Les variables contiennent un ordre (degré de satisfaction, notation : A, A+, ...).

Fréquences absolues et relatives

• Fréquence absolue (ou effectif) de la modalité e_j = le nombre total n_j d'individus de l'échantillon pour lesquels la variable a pris la modalité e_j :

$$n_j = \sum_{i=1}^{n} \mathbf{1}_{e_j}(x_i).$$

Variables qualitatives (3)

• Fréquence relative de la modalité e_j = le pourcentage n_j/n d'individus de l'échantillon pour lesquels la variable a pris la modalité e_j.

Exemple

Chaîne TV	TF1	F2	F3	C+	M6
fréquences absolues	4	3	0	1	2
fréquences relatives	0,4	0,3	0	0,1	0,2

Exemple : élections européennes de 2009

Les individus sont les n = 42 millions d'électeurs et la variable est la personne ou la liste pour laquelle l'individu a voté. La suite des 42 millions de votes ne présente aucun intérêt. Le résultat est exprimé directement sous forme du tableau des fréquences relatives.

Listes	NPA	LO	FrGauche	PS	EurEco	EcoInd	Modem	DivD	UMP	Libertas	FN	Aut
Voix (%)	4.9	1.2	6.0	16.5	16.3	3.6	8.4	1.8	27.8	4.6	6.3	2.

Représentations graphiques

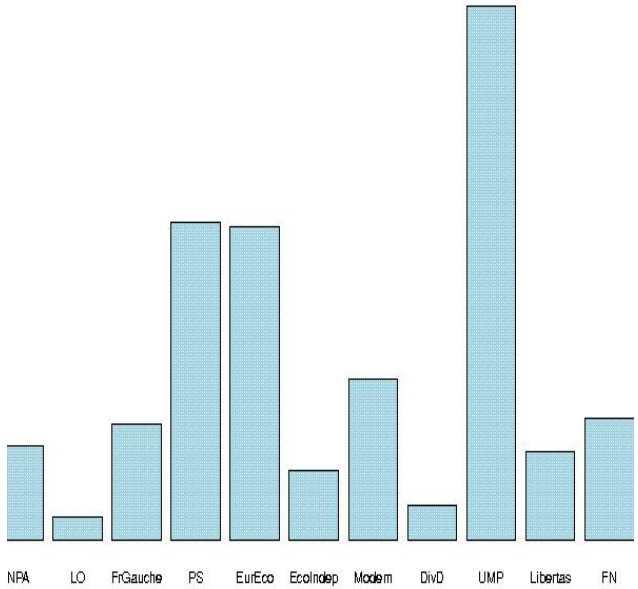
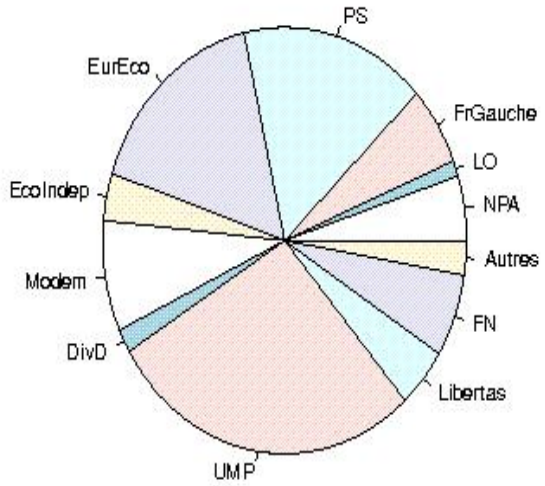
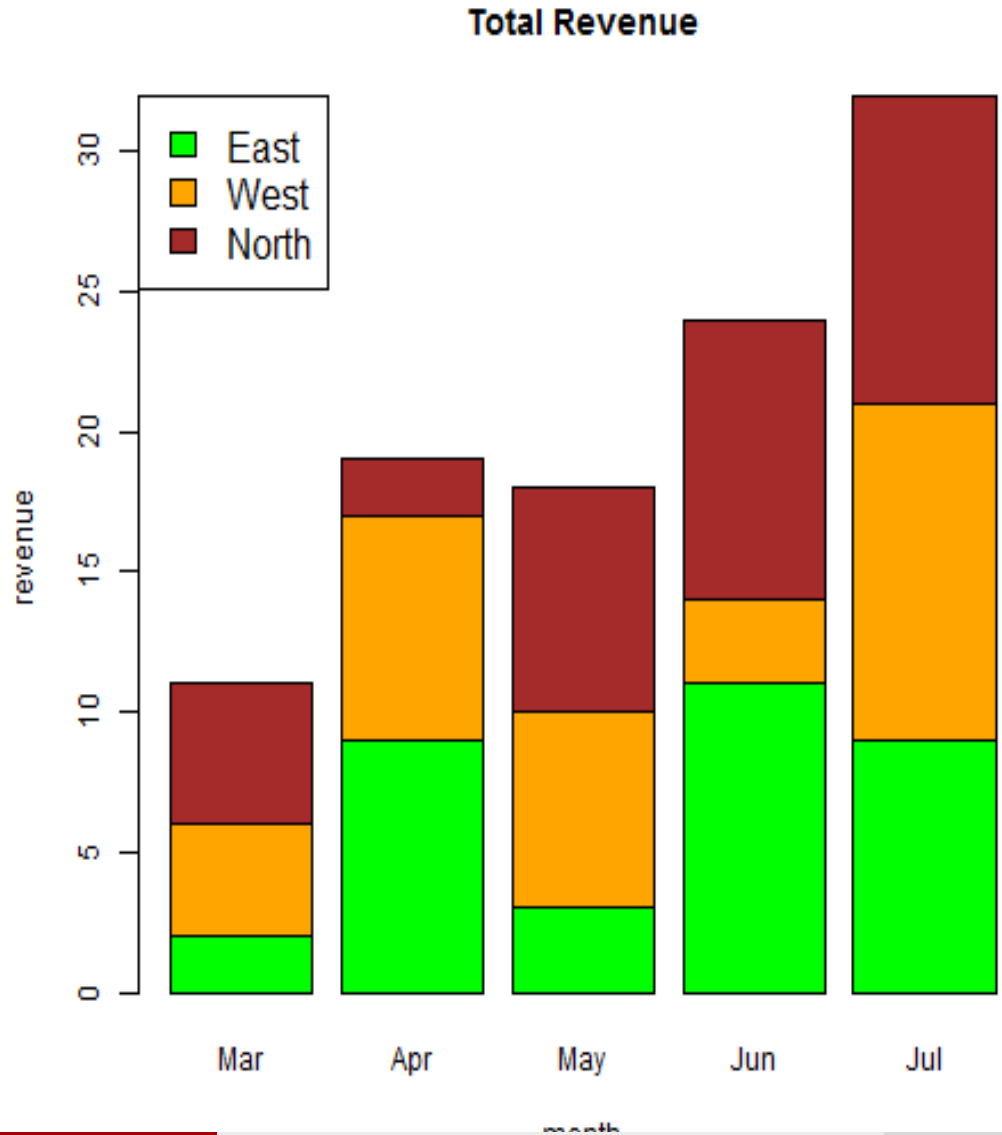
Diagrammes en colonnes (ou en bâtons)	Diagrammes en secteurs (ou camemberts)
	
Rectangle vertical Hauteur proportionnelle $\propto$ fréquence relative (modalité)	Secteur de disque Aire (ou l'angle au centre) $\propto$ fréquence relative (modalité)

Diagramme en colonne : exemple



Variables quantitatives

Définitions

1. **Variable quantitative :** une variable X est dite quantitative si les conditions suivantes sont remplies :
   - ses valeurs sont numériques
   - ses valeurs sont exploitables mathématiquement

2. **Variable discrète :** variable à valeurs dans un ensemble fini ou dénombrable. L'ensemble des valeurs prises par cette variable dans un échantillon de taille n est forcément fini.

3. **Variable continue :** elle prend ses valeurs dans un ensemble infini ou non dénombrable. Les observations qui lui sont associées ne sont pas des valeurs précises mais des intervalles réels.
   Dans ce cas, le sous-ensemble de $\mathbb{R}$ des valeurs possibles de la variable étudiée a été divisé en r intervalles contigus appelés classes.

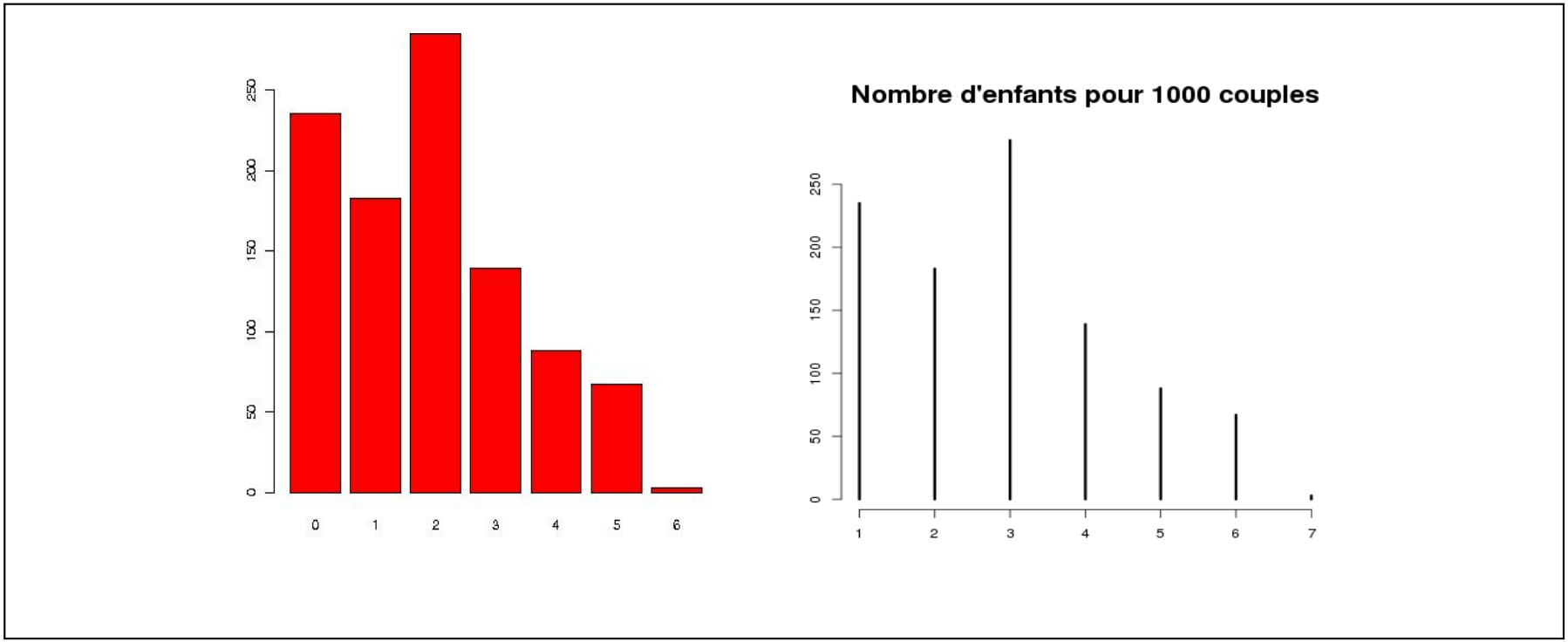
Variables quantitatives discrètes

Exemple une enquête en vue de la réduction du montant des allocations familiales a été réalisée auprès 1000 couples en leur demandant leur nombre d'enfants :

Nombre d'enfants	0	1	2	3	4	5	6	> 6
fréquence absolue	235	183	285	139	88	67	3	0
fréquence relative	23.5%	18.3%	28.5%	13.9%	8.8%	6.7%	0.3%	0

Représentations graphiques

Diagramme en barres (en bâtons)



Fonction cumulative

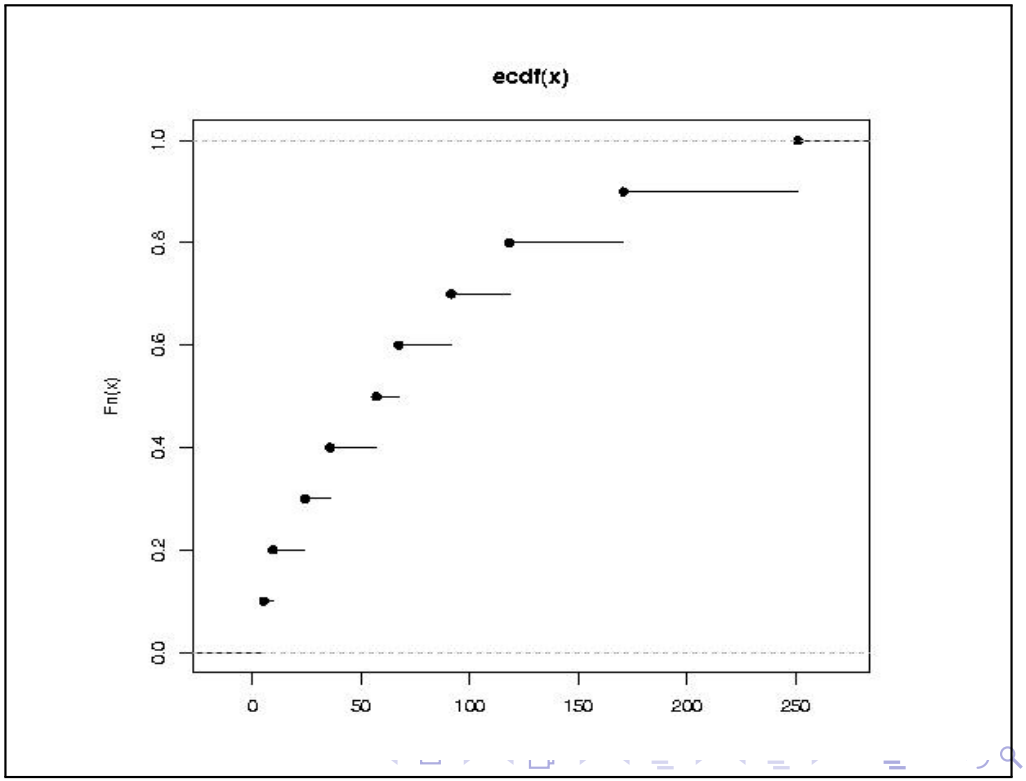
Exemple on fait fonctionner en parallèle et indépendamment les unes des autres $n = 10$ ampoules identiques, dans les mêmes conditions expérimentales, et on relève leurs durées de vie. Admettons que l'on obtienne les durées de vie (h) suivantes :

Durée de vie	91.6	35.7	251.3	24.3	5.4	67.3	170.9	9.5	118.4	57.1

L' échantillon ordonné devient :

Durée de vie	5.4	9.5	24.3	35.7	57.1	67.3	91.6	118.4	170.9	251.3

$x_1 = 91.6$ = durée de vie de la première ampoule.
$x_1^* = min(x_1, ..., x_n) = 5.4$ = plus petite des durées de vie des 10 ampoules.



Variables quantitatives continues

Une variable quantitative continue est à valeurs réelles. Elle prend un trop grand nombre de valeurs pour qu'on puisse toutes les recenser.

1. Découpage en classes :
   Soit a_{min} la plus petite valeur prise par la variable et a_{max} la plus grande valeur ; on se donne une série d'intervalles appelés classes de la forme $]a,\ b]$ couvrant l'ensemble des valeurs de la variable :

$$]a_0, a_1] \cup]a_1, a_2] \cup \ldots \cup]a_{k-1}, a_k] \subset [amin; amax]$$

   Exemple : On a demandé aux 10 élèves d'une classe la durée (en minutes) du trajet domicile-lycée.
   Données individuelles : 6 ; 6 ; 7 ; 10 ; 12 ; 13 ; 20 ; 23 ; 30 ; 36 $\Rightarrow a_{\text{min}} = 6\,;\, a_{\text{max}} = 36.$
   On se donne le découpage en classes]5,15],]15,30] et]30,40] On appelle amplitude de la classe $]a, b]$ la valeur de la différence $b - a$ (amplitude de]5,15] = 15 - 5 = 10).
   Remarque : lors de ce découpage, les classes peuvent être de même amplitude ou d'amplitudes différentes (trois classes de même amplitude :]5,20] ;]20,35] ;]35,50])
   Le principe de base est que les observations sont réparties uniformément au sein de chaque classe.
2. Pour chaque classe $]a,\ b]$ on compte le nombre d'individus pour lesquels la variable prend une valeur strictement supérieure à a et inférieure ou égale à b : On appelle n_i l'effectif de la i-ème classe

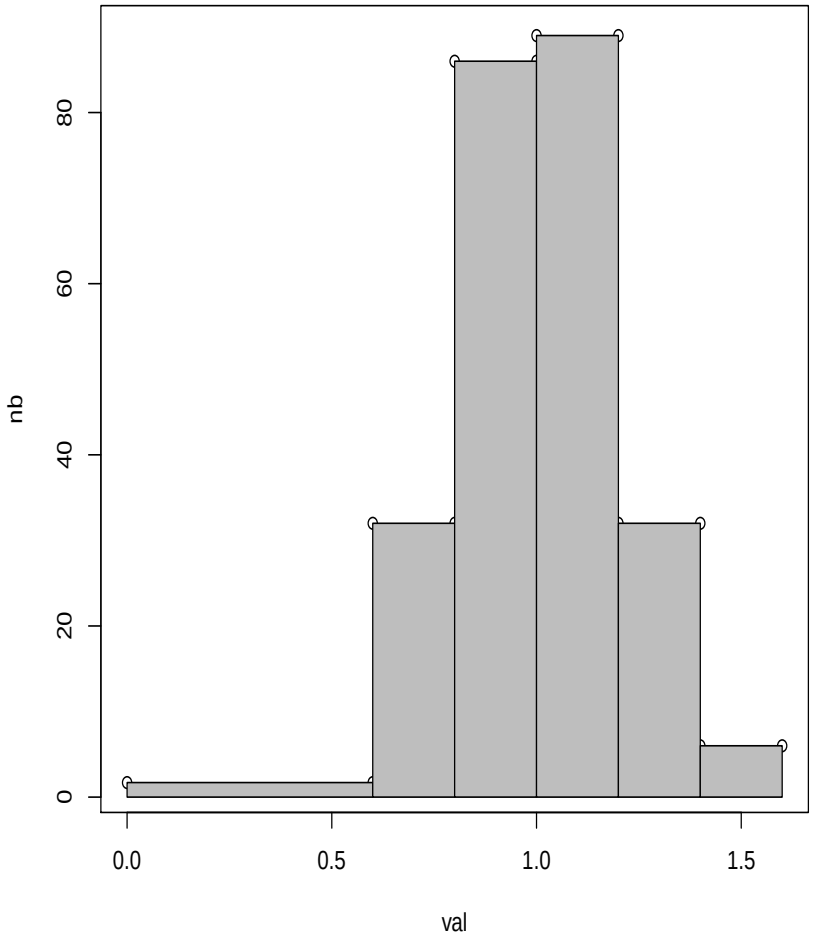
Variables quantitatives continues : Histogrammes

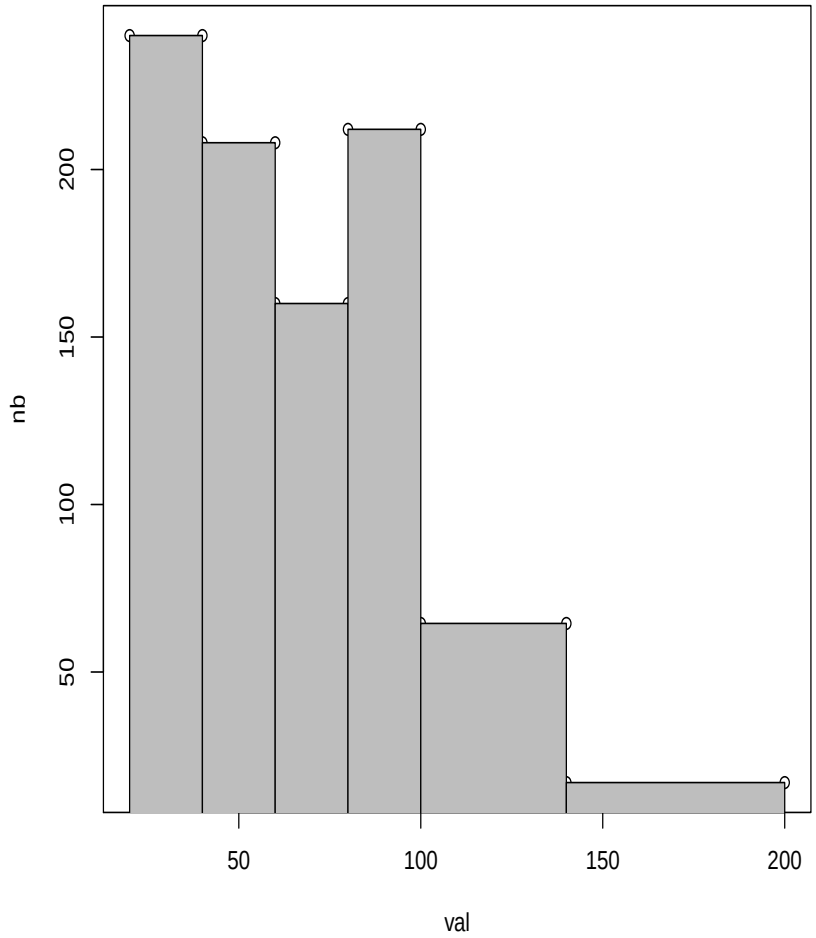
- Après un découpage en classes des observations d'une variable continue, ce graphique sert à représenter les distributions des fréquences
- La surface de chaque barre est proportionnelle à la fréquence de la classe.
- Il est constitué d'un ensemble de rectangles adjacents, dont chacune des bases coïncide avec un intervalle de classe
- Chacune des surfaces mesure la fréquence de la classe correspondante
- Pour des classes d'amplitude égale, la hauteur de chaque barre est proportionnelle à la fréquence.

Remarques :

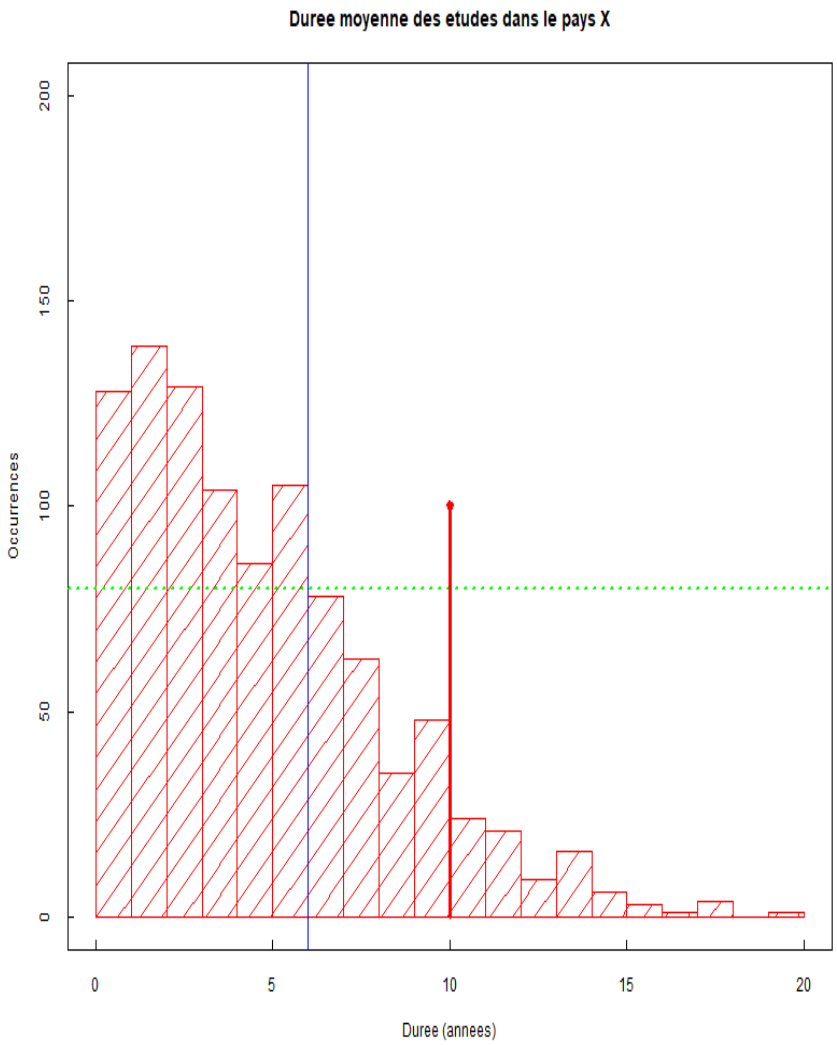
1. Une certaine analogie à la courbe de densité d'une variable aléatoire,
2. Une approximation assez pauvre d'une fonction densité.

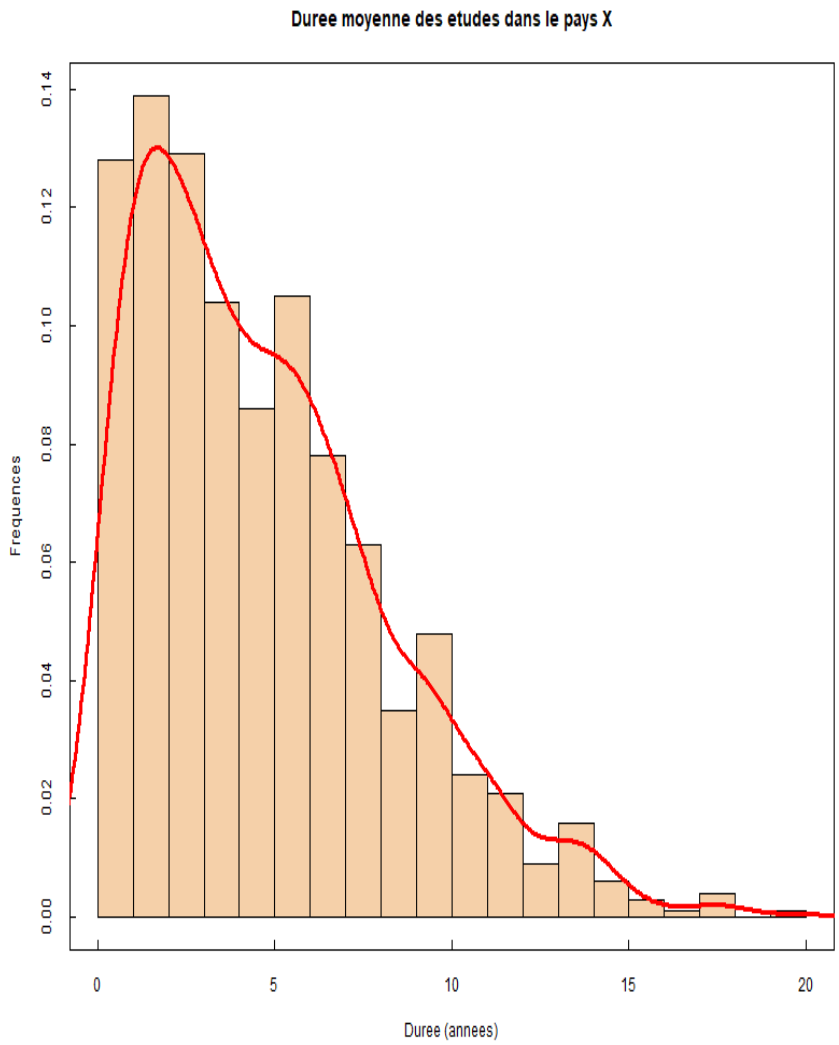
Histogrammes : exemples





Histogrammes : exemples





L'estimation proposée par l'histogramme dépend à la fois des bornes inférieure et supérieure a_0 et a_k, du nombre et de la largeur des classes.
Plusieurs histogrammes peuvent être dessinés à partir des mêmes données et avoir des allures assez différentes, pouvant donner lieu à des interprétations trompeuses. En pratique :

- Calcul du nombre de classes pour un échantillon de taille n.

  La règle de Sturges :

$$\textbf{nb. classes} = 1 + \log_2 n = 1 + \frac{\ln n}{\ln 2}$$

  La règle de Yule :

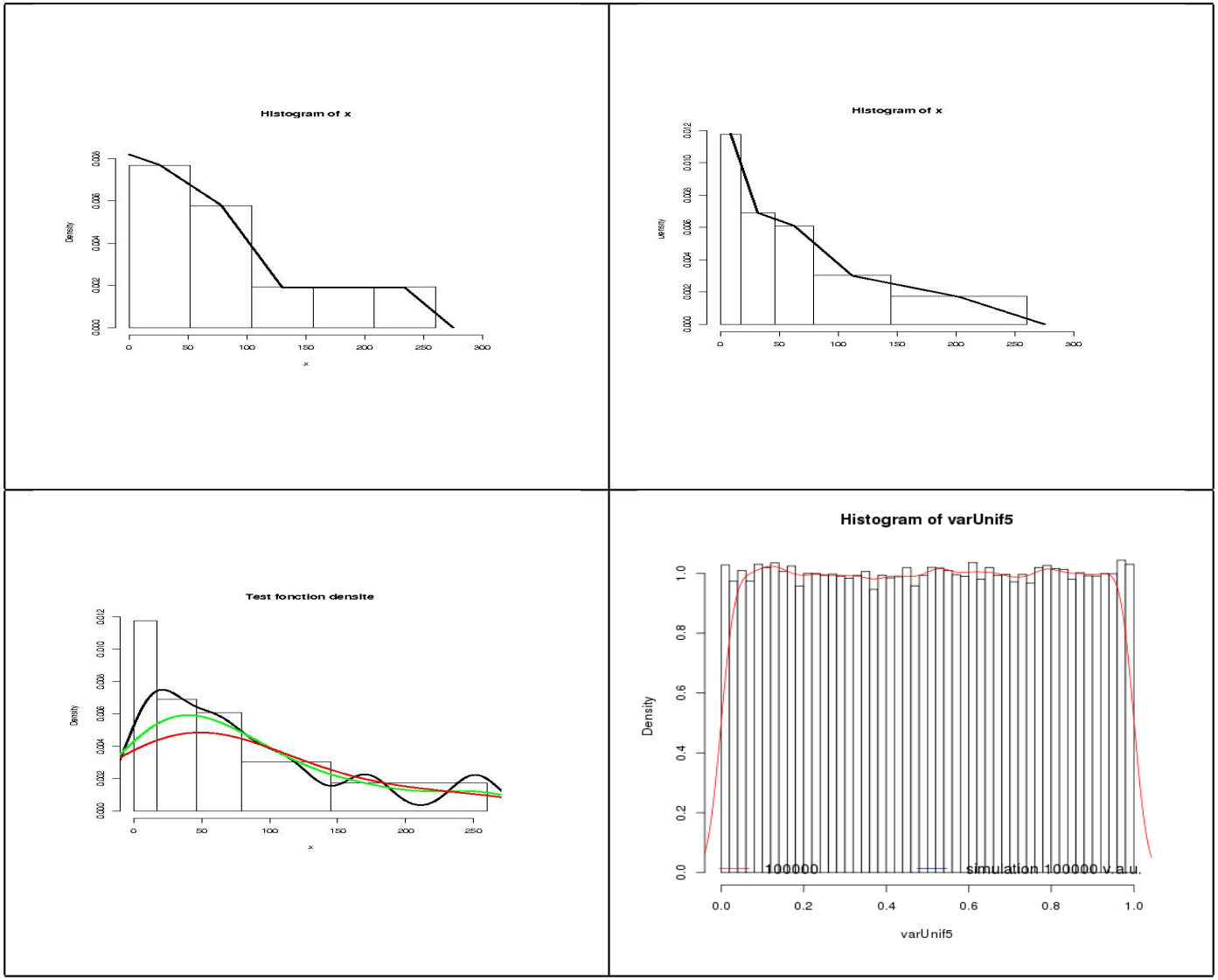
$$\textbf{nb. classes} = 2.5 \sqrt[4]{n}.$$

- Choix des bornes a_0 et a_k fait de façon à respecter une certaine homogénéité des largeurs de classes. Un choix fréquent est :

$$a_0 = x_1^* - 0.025(x_n^* - x_1^*) \; et \; a_k = x_n^* + 0.025(x_n^* - x_1^*).$$

Le choix le plus fréquent est celui de l'histogramme à pas fixe (classes de même largeur) $h = (a_k - a_0)/k$. La hauteur d'un rectangle est proportionnelle à l'effectif de sa classe.
Le polygone des fréquences cumulées est une estimation de la fonction de répartition des observations. La fonction de répartition empirique en est une autre, de meilleure qualité.

Histogrammes - Exemples

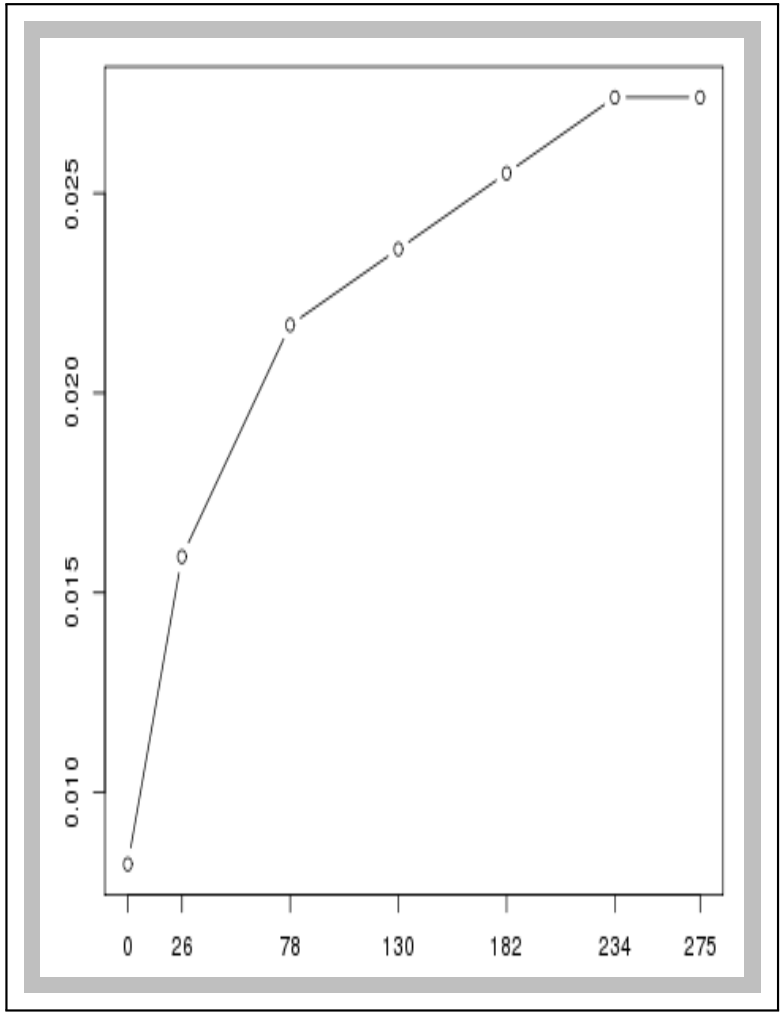


La fonction cumulative

La fonction cumulative (ou fonction de répartition) notée par F donne, pour tout nombre réel t, le pourcentage, noté par $F(t)$, des individus de la population pour lesquels on a observé une valeur de la variable X plus petite ou égale à t.
Propriétés importantes de la fonction cumulative F :

- Elle est croissante, c.-à-d. que pour tous nombres réels t_1 et t_2 , vérifiant $t_1 \leq t_2$, on a $F(t_1) \leq F(t_2)$.
- Elle est nulle pour tout nombre réel t inférieur à x_0 , où x_0 désigne la borne de gauche de la première classe c.-à-d. $[x_0, x_1]$.
- Elle est égale à 1 pour tout nombre réel t supérieur à x_n , où x_n désigne la borne de droite de la dernière classe c.-à-d. $]x_{n-1}, x_n]$.

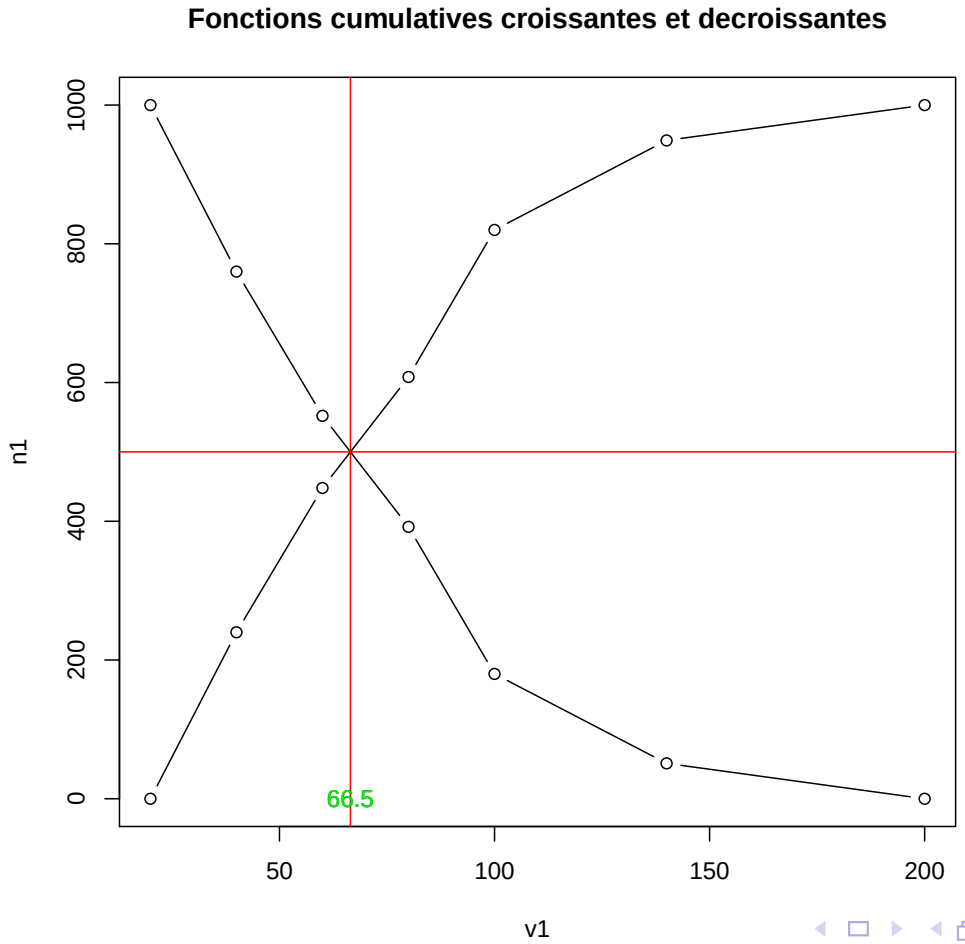
Si X est une variable continue, F n'est connue qu'aux valeurs de X égales aux extrémités des classes c.-à-d. pour $t = x_0, t = x_1, \ldots, t = x_k$. On considére que F est linéaire entre ces valeurs, parce qu'on suppose que les classes forment des entités homogènes.



Exemple

On considère la série statistique suivante :

x_i	$[20,\ 40[$	$[40,\ 60[$	$[60,\ 80[$	$[80,\ 100[$	$[100,\ 140[$	$[140,\ 200[$
n_i	240	208	160	212	129	51



Courbe de concentration

La courbe de concentration de Lorenz est un moyen de représenter la fonction de répartition d'une variable X. Elle est notamment utilisée en économie pour mesurer les inégalités de possession de richesse (on supposera donc que X représente un certain bien possédé par les individus de la population).

Soit x une valeur prise par X :

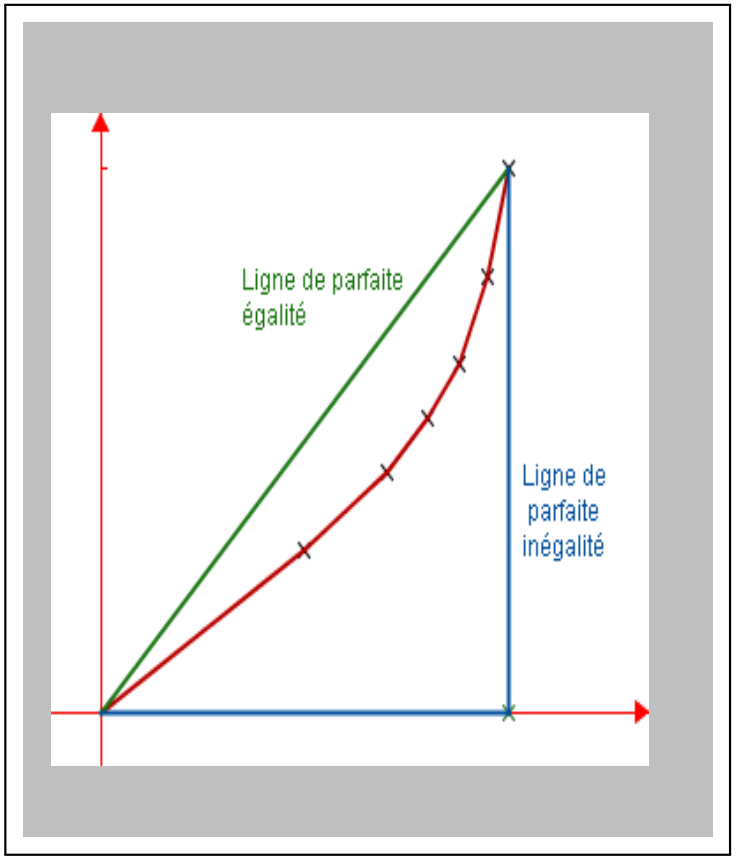
- On note $F(x)$ la proportion de la population pour laquelle $X < x$ (F = la fonction de répartition de X).
- On note $FQ(x)$ la proportion du bien possédé par ces individus par rapport au bien total. Alors la courbe de Lorenz est la courbe joignant tous les points $(F(x), FQ(x))$.

Exemple :
Dans le cas de l'analyse des revenus des ménages, soit le pourcentage ou le nombre x des ménages les moins riches qui détient telle part en valeur ou en pourcentage y du revenu de l'ensemble des ménages, la part des ménages, classée par ordre de revenu individuel croissant, est figurée en abscisse, et la part du revenu en ordonnée.

Courbe de concentration

La courbe de Lorenz joint toujours le point (0,0) au point (1,1).
Elle est située sous le segment joignant ces deux points.

Dans une société, on dira que la distribution des revenus est parfaitement égalitaire si tous les ménages reçoivent le même revenu. Alors la part x des ménages les moins riches reçoit une part $y = x$ du revenu global. Une répartition égalitaire est donc représentée par la première bissectrice (équation $y = x$). Cette droite est appelée **la ligne d'égalité parfaite**.
A l'inverse, on parlera de distribution parfaitement inégalitaire si dans la société considérée, un ménage accapare le revenu total (global). Dans ce cas, la fonction associée prend la valeur $y = 0$ pour tout $x < 100\%$, et $y = 100\%$ quand $x = 100\%$. La courbe de Lorenz correspondant à cette situation est appelée **la ligne de parfaite inégalité**. Plus la courbe "colle" à la ligne de parfaite égalité, plus la société est égalitaire.
Le coefficient de Gini permet de quantifier cela.



COURS 2

Sommaire.

Indicateurs statistiques

Les représentations graphiques présentées dans la section précédente ne permettent qu'une analyse visuelle de la répartition des données. L'objectif est de caractériser la distribution de la série à l'aide de grandeurs résumant de façon suffisamment complète l'ensemble de ses valeurs. Ces indicateurs faciliteront la comparaison d'échantillons.

Généralement trois types d'indicateurs :

- indicateur de localisation (position, tendance centrale),
- indicateur de dispersion,
- indicateur de forme.

Indicateur statistique : le mode

Variable quantitative discrète (non classée)
Le mode correspond à la valeur de la variable pour laquelle l'effectif (ou la fréquence) est le plus grand

Variable quantitative continue
La classe modale est la classe dont la fréquence par unité d'amplitude est la plus élevée ; cette classe correspond donc au rectangle le plus haut de l'histogramme des fréquences.
Signalons au passage que certaines variables peuvent avoir plusieurs classes modales.

Exemple

Soit $S = \{2,\ 3,\ 0,\ 2,\ 1,\ 3,\ -1,\ 3\}$. Le mode $M_o = 3$.

Exemple

On considère la série statistique suivante :

x_i	$[20,\ 40[$	$[40,\ 60[$	$[60,\ 80[$	$[80,\ 100[$	$[100,\ 140[$	$[140,\ 200[$
n_i	**240**	208	160	212	129	51

Indicateurs statistiques : illustration

Pour des variables quantitatives

Exemple 4

- Population : 25 poudres de lait
- Variable **MAT - MST**

Teneur en protéine / Matière sèche

Poudre	MAT-MST
17	82,79
22	82,96
14	83,17
21	83,92
11	84,57
20	84,65
25	85,02
19	85,14
13	85,34
12	85,62
16	85,68
24	85,7
23	85,77
15	86,73
9	87,4
8	87,97
10	88,24
1	88,44
7	89,06
6	89,63
3	89,88
2	90,17
4	91,64
5	92,21
18	97,06

Moyenne arithmétique

Le but est de donner un ordre de grandeur général des observations, un nombre unique qui résume au mieux les données.

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

La moyenne est le centre de gravité des données affectées où chaque individu a le même poids.
Elle peut être considérée comme une valeur centrale, même si elle n'est pas égale à une des modalités.

Exemple des ampoules : La durée de vie moyenne d'une ampoule est de $\overline{X} = 83.15$.

- la moyenne est très sensible aux valeurs extrêmes dites valeurs aberrantes. Si une des observations est extrêmement grande, elle va tirer la moyenne vers le haut.

Moyenne arithmétique pondérée

On pondère chacune des valeurs distinctes de X par la fréquence correspondante

$$\overline{X}_n = \frac{1}{N} \sum_{i=1}^{n} n_i \, x_i = \sum_{i=1}^{n} f_i \, x_i$$

- Y est une variable quantitative continue, dont l'intervalle de variation a été divisé en k classes jointives $[y_0, y_1],]y_1, y_2], \ldots,]y_{k-1}, y_k]$;
- Y est une variable discrète classée dont les classes sont $[y_0, y_1],]y_1, y_2], \ldots,]y_{k-1}, y_k]$. Alors, $\overline{Y}$ la moyenne arithmétique de Y , est définie comme la moyenne arithmétique des centres des classes de Y pondérées par les fréquences correspondantes ; plus précisément :

$$\overline{Y} = \frac{1}{N} \sum_{i=1}^{n} n_i \left[\frac{y_{i-1} + y_i}{2}\right] = \sum_{i=1}^{n} f_i \left[\frac{y_{i-1} + y_i}{2}\right]$$

où, pour tout i, f_i et n_i désignent respectivement la fréquence et l'effectif de la i-ème classe, $N = \sum_{i=1}^{n} n_i$ étant l'effectif total.

Moyenne quadratique, harmonique et géométrique

Moyenne quadratique

Elle est notée par m_2 :

$$m_2 = \sqrt{\sum_{i=1}^{n} f_i \, x_i^2}$$

Moyenne harmonique

Elle est notée par m_{-1} :

$$m_{-1} = \frac{f_i}{\sum_{i=1}^{n} \frac{f_i}{x_i}}$$

Utilisée chaque fois qu'il est possible d'attribuer un sens réel aux inverses des données (taux d'équipement, pouvoir d'achat, calcul d'indice, . . .).

Moyenne géométrique

Si les observations $x_1, \ldots, x_N$ sont toutes des nombres réels positifs, la moyenne géométrique notée par m_g est définie par :

$$m_g = \sqrt{\prod_{i=1}^{n} x_i^{f_i}}$$

Pour les statisticiens, elle est moins sensible que la moyenne arithmétique aux valeurs les plus élevées d'une série de données. Elle donne, par conséquent, une autre et meilleure estimation de la tendance centrale des données dans le cas d'une distribution à longue traîne à l'extrémité supérieure de la courbe.

Indicateur statistique : la médiane

La médiane d'une variable quantitative, notée Me, $\tilde{x}_n$ ou $\tilde{x}_{1/2}$, est un réel qui partage la population étudiée en deux parties de même effectif. La moitié des observations sont inférieures à $\tilde{x}_n$ et l'autre moitié lui sont supérieures. Il y a donc une chance sur deux pour qu'une observation soit inférieure à la médiane, et évidemment une chance sur deux pour qu'une observation soit supérieure à la médiane.

Quand la distribution est symétrique, moyenne et médiane empiriques sont proches (pour une variable aléatoire de loi symétrique, l'espérance et la médiane théoriques sont égales).
On constate donc que la moyenne et la médiane empiriques sont deux résumés de l'échantillon dont la connaissance simultanée peut être riche d'enseignements.

Indicateur statistique : la médiane

Variable quantitative discrète

- Si n est impair, la médiane empirique est la valeur située au centre de l'échantillon ordonné : $\tilde{x}_n = \tilde{q}_{n,1/2} = x_{\frac{(n+1)}{2}}$.
- Si n est pair, n'importe quel nombre compris entre $x_{\frac{n}{2}}$ et $x_{\frac{n}{2}+1}$ vérifie la définition de la médiane. Par convention, on prend en général le milieu de cet intervalle :
  $\tilde{x}_n = \tilde{q}_{n,1/2} = \left(x_{(\frac{n}{2})} + x_{(\frac{n}{2})+1}\right)/2.$
  L'expression de la médiane montre bien que c'est un indicateur qui n'est pas sensible aux valeurs aberrantes.

Indicateur statistique : la médiane

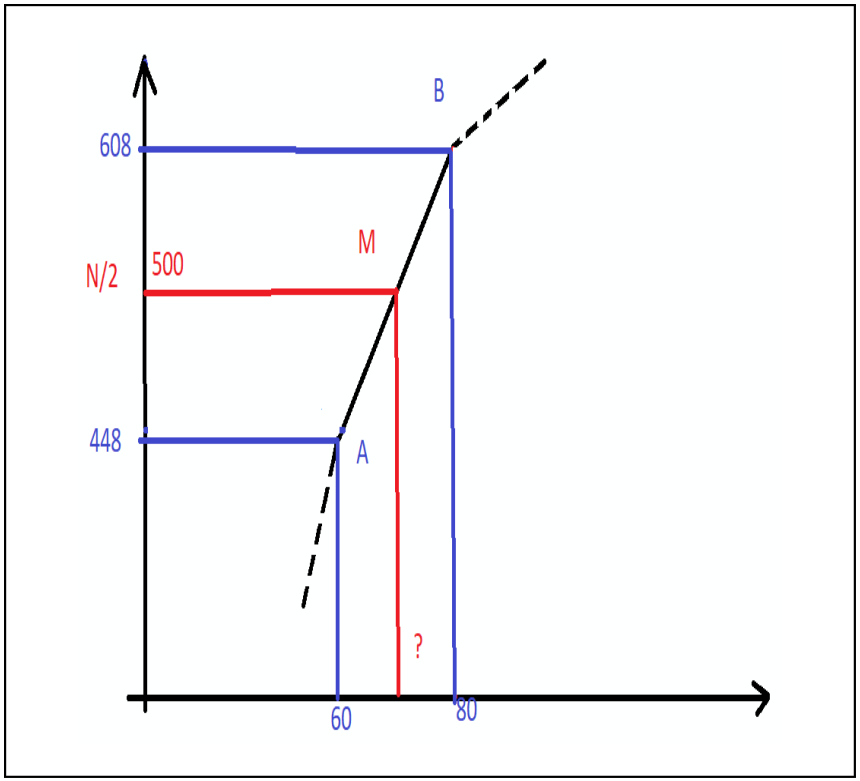
Variable quantitative continue

x_i	$[20,\ 40[$	$[40,\ 60[$	$[60,\ 80[$	$[80,\ 100[$	$[100,\ 140[$	$[140,\ 200[$
x_i^c	30	50	70	90	120	170
n_i	240	208	160	212	129	51
$\sum_i n_i$	240	448	608	820	949	1000

$N/2 = 500, \Rightarrow\ 500 \in [60,\ 80[$.
On cherche l'équation de la droite $y = ax + b$ qui passe par $A(60, 448)$ et $B(80,\ 608)$.

448	=	60 a + b
608	=	80 a + b
160	=	20 a

$a = 8;\ \Rightarrow\ b = -32$ donc $y = 8x - 32$.
Pour $y = N/2$, on a $500 = 8\,x_{Q2} - 32$
$x_{Q2} = 66.5$



Les quantiles

Les quantiles très utilisés pour décrire des phénomènes concernant les extrémités des échantillons :

- En finance, la value at risk (VaR) est la plus utilisée des mesures de risque de marché. Elle représente la perte potentielle maximale d'un investisseur sur la valeur d'un portefeuille d'actifs, compte-tenu d'un horizon de détention et d'un niveau de confiance donnés. Par exemple, quand on dit qu'un portefeuille a une VaR de -3 Meuros à 95% pour un horizon mensuel, cela signifie que l'on estime que ce portefeuille a 95% de chances de ne pas se déprécier de plus de 3 Meuros en un mois. La VaR est donc ici le quantile d'ordre 5% de la distribution des rendements de ce portefeuille en un mois.

- Dans l'industrie pétrolière, les réserves sont classées en 3 catégories P10, P50 et P90, selon la probabilité qu'elles ont de pouvoir être exploitées dans le futur. Cela correspond aux quantiles d'ordre 10%, 50% et 90% de la loi du débit de pétrole du puits.

Les quantiles

Le quartile trois valeurs du caractère qui partage la série statistique en quatre groupes de même effectif :

- le 1^{er} quartile est la valeur du caractère à partir de laquelle la fréquence cumulée atteint ou dépasse 0.25 : $\tilde{x}_{n,1/4} = \tilde{q}_{n,1/4}$.
- le $2^{ème}$ quartile est confondu avec la médiane $\tilde{x}_{n,1/2} = \tilde{q}_{n,1/2}$.
- le $3^{ème}$ quartile est la valeur du caractère à partir de laquelle la fréquence cumulée atteint ou dépasse 0.75 : $\tilde{x}_{n,3/4} = \tilde{q}_{n,3/4}$.

Le **décile** (déciles) :

- le 1^{er} décile est la valeur du caractère à partir de laquelle la fréquence cumulée atteint ou dépasse 0.1.
- le $2^{ème}$ décile est la valeur du caractère à partir de laquelle la fréquence cumulée atteint ou dépasse 0.2. etc...
- le $9^{ème}$ décile est la valeur du caractère à partir de laquelle la fréquence cumulée atteint ou dépasse 0.9.

Le **centile** (99 centiles) :

- le 1^{er} centile est la valeur du caractère à partir de laquelle la fréquence cumulée atteint ou dépasse 0.01. etc...
- le $99^{ème}$ centile est la valeur du caractère à partir de laquelle la fréquence cumulée atteint ou dépasse 0.99.

Le **quantile d'ordre p** (p un réel de [0,1[), est la valeur du caractère à partir de laquelle

Dispersion

Les indicateurs de dispersion :

- expriment les caractéristiques d'un échantillon,
- complétent les indicateurs de localisation
- mesurent la variabilité des données.

Exemple Températures mensuelles moyennes, en degrés Celsius, à New-York et à San Francisco, calculées sur une période de 30 ans.

	J	F	M	A	M	J	J	A	S	O	N	D
New-York	0	1	5	12	17	22	25	24	20	14	8	2
San Francisco	9	11	12	13	14	16	17	17	18	16	13	9

La température annuelle moyenne est de 12.5 degrés à New-York et de 13.7 à San Francisco. En se basant uniquement sur ces moyennes, on pourrait croire que les climats de ces deux villes sont similaires. Or il est clair que la différence de température entre l'hiver et l'été est beaucoup plus forte à New-York qu'à San Francisco. Pour le déceler, il suffit de calculer un indicateur qui exprime la variabilité des observations.
$e = \frac{1}{n} \sum_{i=1}^{n} d(x_i, c)$, e exprime bien la variabilité de l'échantillon autour de c. On pourra donc construire des indicateurs de dispersion à partir de e en considérant différentes distances.

Indicateur de dispersion : l'étendue

Une série statistique de n valeurs $(x_1, x_2, \ldots, x_i, \ldots, x_n)$ L'étendue est la différence entre la plus grande valeur et la plus petite valeur prise par la variable :

L'étendue = amplitude = $e_n = x_n - x_1$.

- Le plus simple et le plus intuitif
- Moins riche que la variance empirique
- Très sensible aux valeurs aberrantes
- Employé couramment en contrôle de qualité, notamment pour détecter ces valeurs aberrantes.

Indicateur de dispersion : La variance - l'écart-type

$$Var(X) = \frac{1}{N}\sum_{i=1}^{n} n_i\,(x_i - \bar{x}_n)^2 = \sum_{i=1}^{n} f_i\,(x_i - \bar{x}_n)^2 = \frac{1}{N}\sum_{i=1}^{n} n_i\,x_i^2 - \bar{x}_n^2\,.$$

- $Var(X)$ est appelée **variance** de la variable et mesure l'écart quadratique moyen de la variable à sa moyenne.

- $s = \sqrt{Var(X)}$ est **L'écart-type** de la variable. Il s'exprime dans la même unité que les données, ce qui rend son interprétation plus facile que celle de la variance.
- Dans le cas d'une série regroupée par classe, on utilise le centre des classes.
- Une variance est toujours positive !
- Ne pas oublier que la variance est la moyenne des écarts au carré. C'est d'ailleurs à partir de ce fait que l'on interprète la variance comme un indicateur de dispersion

Variance, écart-type empiriques

Si on choisit la distance euclidienne ($c = \bar{x}_n$), l'indicateur de dispersion correspondant est

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}_n^2.$$

Il est appelé **variance empirique** de l'échantillon, et mesure l'écart quadratique moyen de l'échantillon à sa moyenne. $s_n = \sqrt{s_n^2}$ est **L'écart-type empirique** de l'échantillon.

Exemple

L'écart-type des températures annuelles est de 8.8 degrés à New-York et de 3 degrés à San Francisco, ce qui exprime bien la différence de variabilité des températures entre les deux villes.

Le coefficient de variation

Mesure de la dispersion relative

“ Ecart-type normalisé / standardisé”
“ Ecart-type en pourcentage de la moyenne”

$$C_V = \frac{s_x}{\overline{X}} 100$$

Quel intérêt ?

- Comparer des dispersions entre elles
- Le CV permet de comparer la dispersion de variables ayant :
  - des unités de mesure différentes
  - des moyennes différentes
- Il faut qu'il soit le plus faible possible ($< 15\%$ en pratique).

L'écart absolu moyen

- L'écart absolu moyen à la moyenne de la variable quantitative X est la moyenne arithmétique des valeurs absolues des écarts à la moyenne arithmétique :

$$e_{\overline{x}} = \frac{1}{n} \sum_{i=1}^{n} |x_i - \overline{X}|$$

$$e_{\overline{x}} = \frac{1}{N} \sum_{i=1}^{n} n_i \, |x_i - \overline{X}| = \sum_{i=1}^{n} f_i \, |x_i - \overline{X}|$$

- L'écart absolu moyen à la médiane de la variable quantitative X est la moyenne arithmétique des valeurs absolues des écarts à la médiane M_e .

$$e_{M_e} = \frac{1}{n} \sum_{i=1}^{n} |x_i - M_e|$$

$$e_{M_e} = \frac{1}{N} \sum_{i=1}^{n} n_i \, |x_i - M_e| = \sum_{i=1}^{n} f_i \, |x_i - M_e|$$

L'écart interquartile

L'écart interquartile ou étendue interquartile (EIQ) est une mesure de dispersion qui s'obtient en faisant la différence entre le troisième et le premier quartile :

$$EI = \tilde{q}_{n,3/4} - \tilde{q}_{n,1/4}.$$

L'EI est un estimateur statistique robuste insensible aux valeurs aberrantes. $[\tilde{q}_{n,1/4}, \tilde{q}_{n,3/4}]$ est un intervalle qui contient la moitié la plus centrale des observations.
On définit de la même manière des distances inter-déciles, inter-centiles,...

L'écart interquartile sert à apprécier la dispersion de X, de façon absolue, ou bien par comparaison avec une autre variable quantitative, à condition que cette dernière soit exprimée dans la même unité que X. En effet, les valeurs Q_1 et Q_3 délimitent une plage au sein de laquelle 50% des valeurs de X sont concentrées. Plus EIQ est grand, plus X est dispersée.

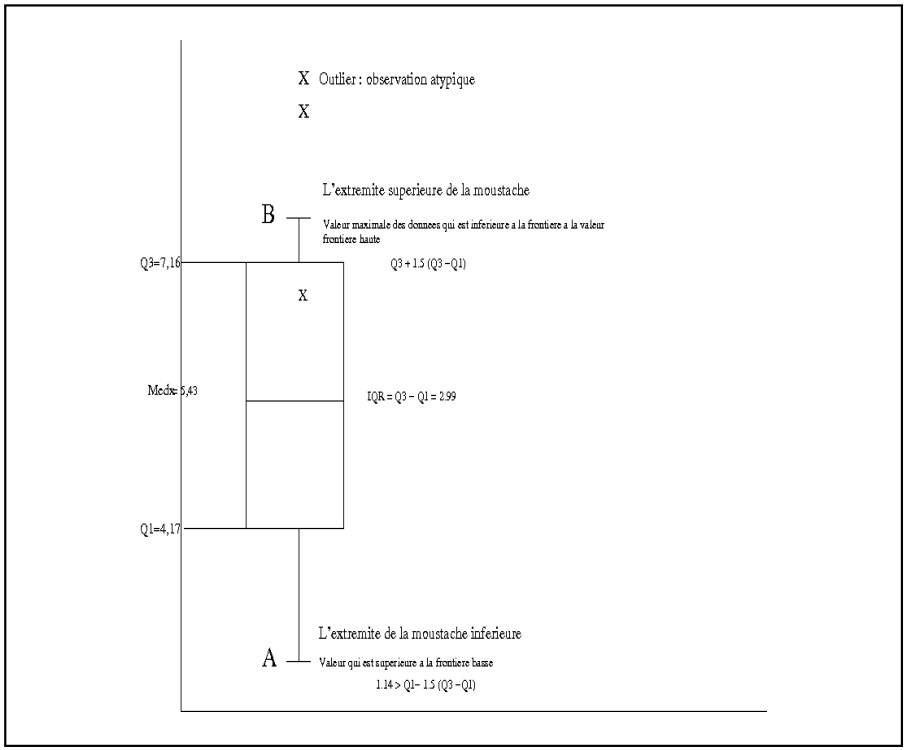
Exemple

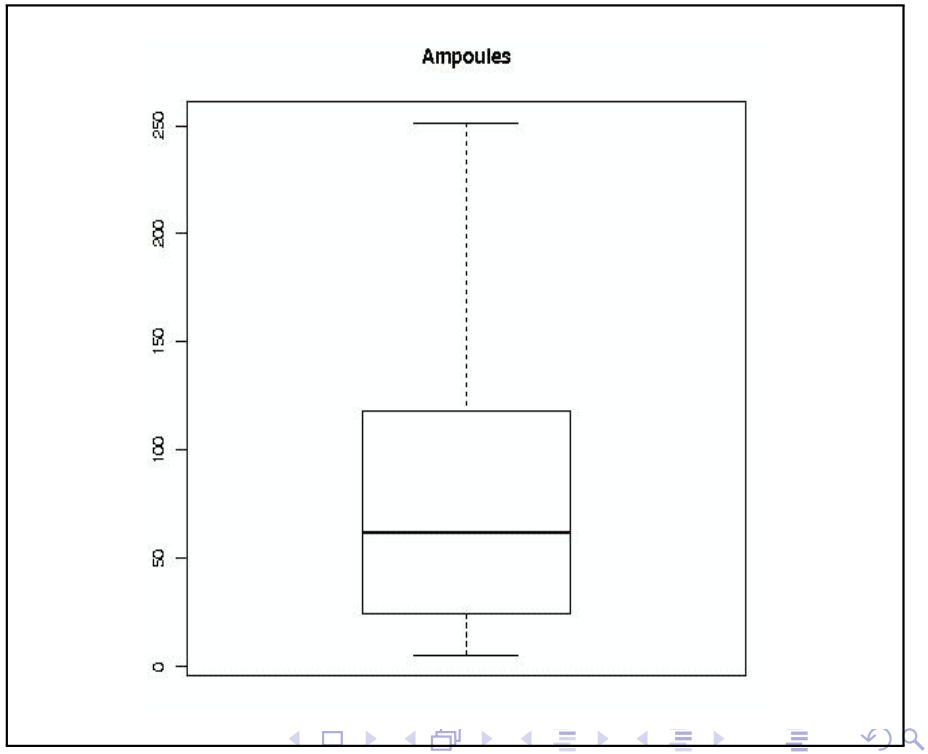
Dans l'exemple des ampoules, distance inter-quartiles : 94.1 h.

Boxplot ou boîte à moustaches

met en évidence un ou plusieurs paramètres de tendance centrale et un paramètre de dispersion :

- La valeur du premier quartile (Q1) : trait inférieur de la boîte.
- La valeur du troisième (Q3) : trait supérieur de la boîte.
- La valeur de second quartile (Q2 la médiane) : trait horizontal au sein de la boîte.
- Les moustaches inférieure et supérieure : traits verticaux de chaque côté de la boîte.
- La moyenne peut être représentée (ex par un +).
- Les valeurs extrèmes (valeurs atypiques, outliers) : cercle ou *.





Conclusion

Ces quelques outils permettent déjà de se faire une première idée d'un jeu de données mais surtout, en préalable à toute analyse, ils permettent de s'assurer de la fiabilité des données, de repérer des valeurs extrêmes atypiques, éventuellement des erreurs de mesures ou de saisie, des incohérences de codage ou d'unité.
Les erreurs, lorsqu'elles sont décelées, conduisent naturellement et nécessairement à leur correction ou à l'élimination des données douteuses mais d'autres problèmes pouvant apparaître n'ont pas toujours de solutions évidentes.

- Faut-il supprimer les individus incriminés ou les variables ?
- Faut-il compléter, par une modélisation et prévision partielles, les valeurs manquantes ?
- Les solutions dépendent-elles du taux de valeurs manquantes, de leur répartition (sont-elles aléatoires) et du niveau de tolérance des méthodes qui vont être utilisées ?
- La présence de valeurs atypiques peut influencer sévèrement des estimations de méthodes peu robustes car basées sur le carré d'une distance. Ces valeurs sont-elles des erreurs ?
- Sinon faut-il les conserver en transformant les variables ou en adoptant des méthodes robustes basées sur des écarts absolus ?

Remarque :
Même sans hypothèse explicite de normalité des distributions, il est préférable d'avoir à faire à des distributions relativement symétriques. Une transformation des variables par une fonction monotone (log, puissance) est hautement recommandée afin d'améliorer la symétrie de leur distribution ou encore pour linéariser (nuage de points) la nature d'une liaison.