Chapitre 3: Réduction des données

Arnout Van Messem

Bachelier en Sciences informatiques

LIÈGE université

Introduction

Soit $S = \{x_1, x_2, \ldots, x_n\}$ la série brute,

ou $S = \{(x_j, n_j) : j = 1, \ldots, J\}$ la série recensée

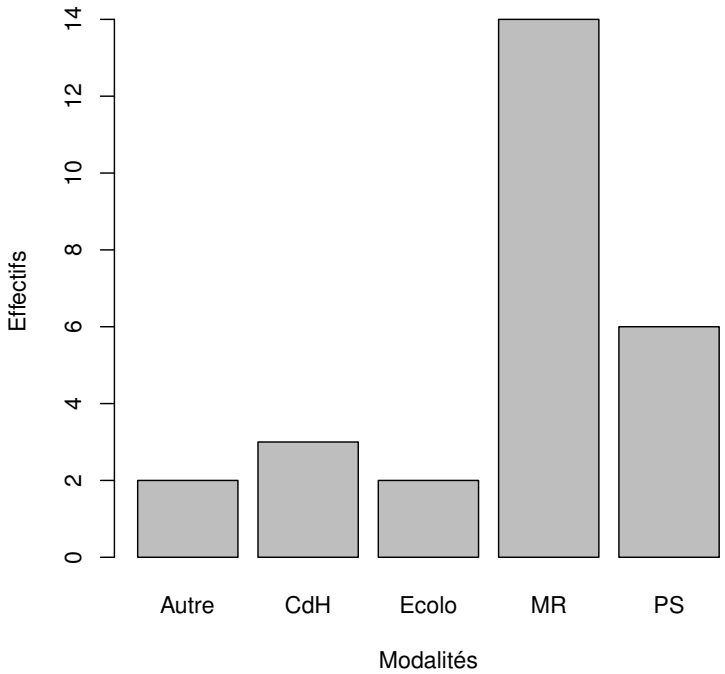
ou $S = \{(c_j, n_j) : j = 1, \ldots, J\}$ la série groupée

On désire résumer l'information disponible par des indicateurs précisant où se trouve le centre de la série, comment les observations se dispersent autour de ce centre,...

LIÈGE université

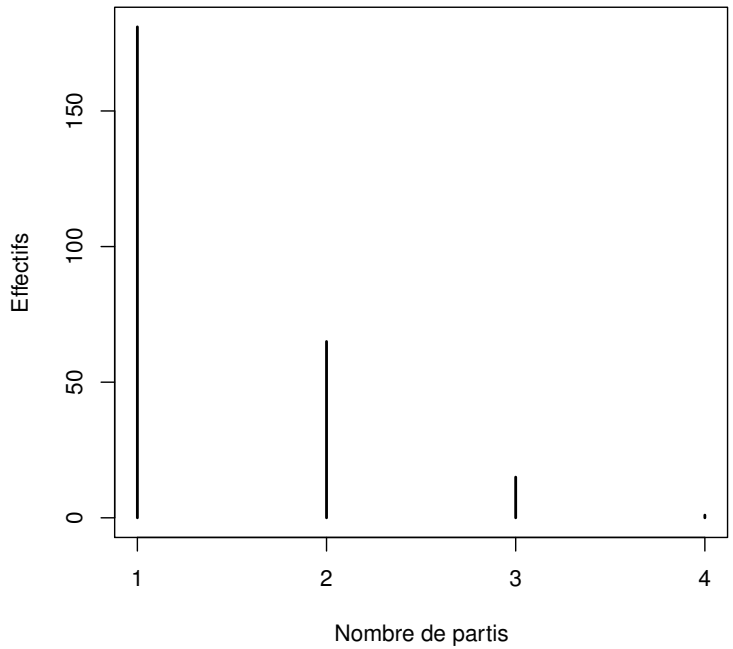
Variable qualitative `PartiBourgmestre` (Brabant W.)

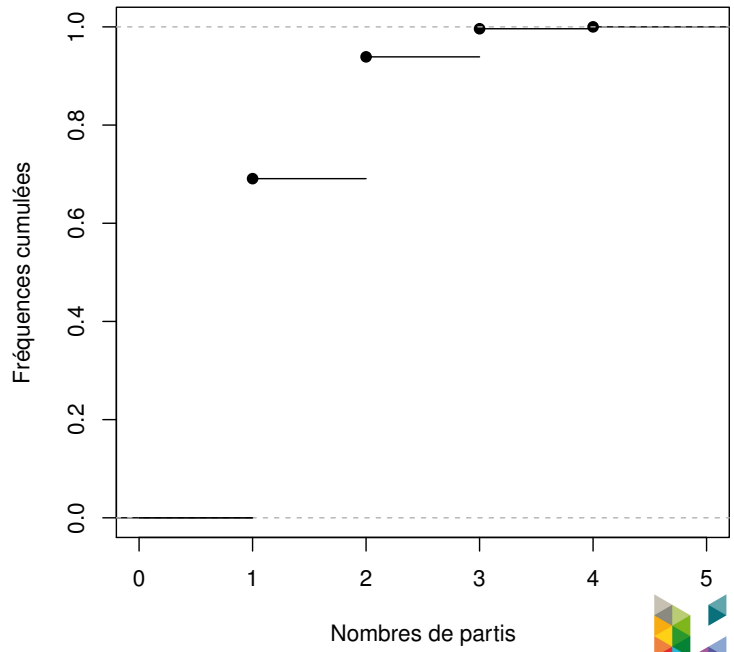
Partis	Effectifs	Fréquences
PS	6	0.222
CdH	3	0.111
Ecolo	2	0.074
MR	14	0.519
Autre	2	0.074
Total	27	1



Variable quantitative discrète `NombrePartis`

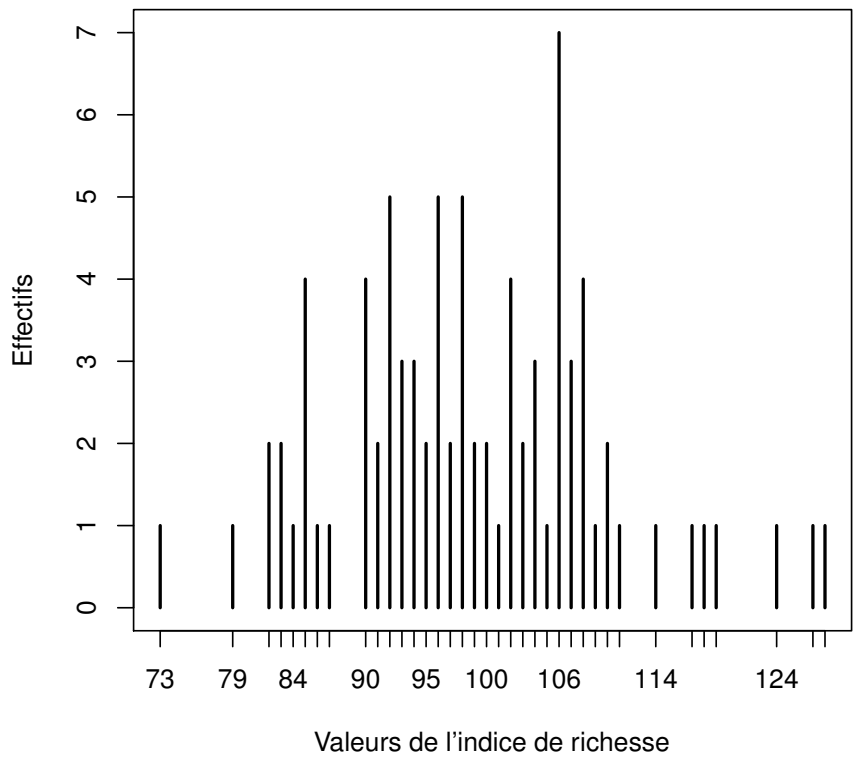
Valeurs	Effectifs	Fréquences	Effectifs cumulés	Fréquences cumulées
1	181	0.691	181	0.691
2	65	0.248	246	0.939
3	15	0.057	261	0.996
4	1	0.004	262	1.000
Total	262	1	x	x





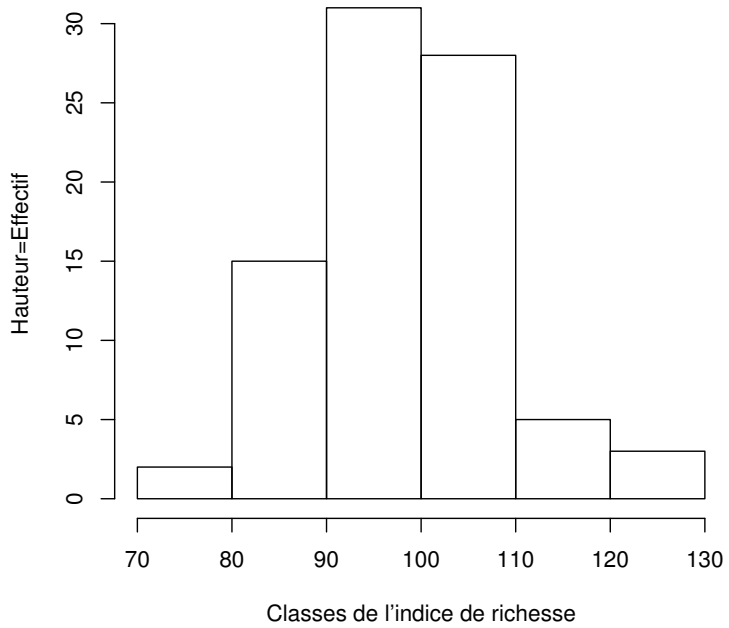
Variable quantitative continue `IndiceRichesse` (Liège)

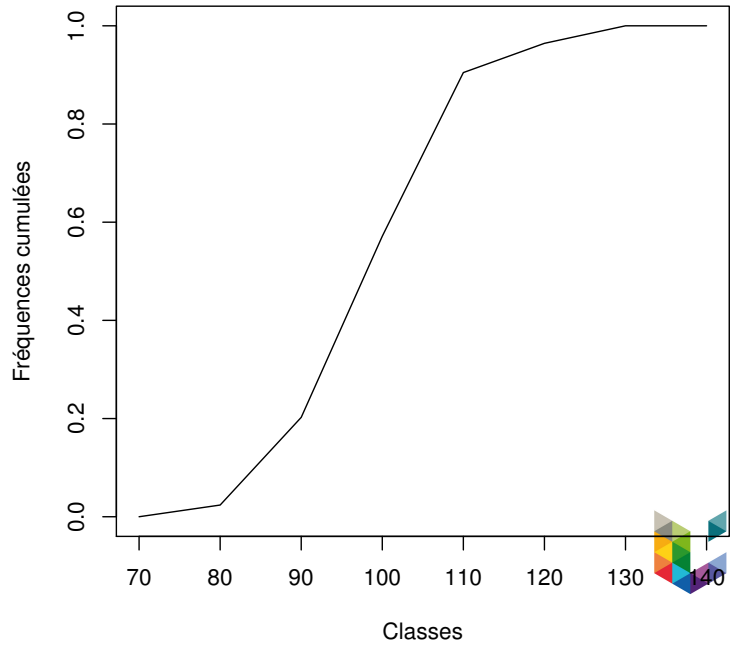
73	79	82	82	83	83	84	85	85	85	85	86	87	90
90	90	90	91	91	92	92	92	92	92	93	93	93	94
94	94	95	95	96	96	96	96	96	97	97	98	98	98
98	98	99	99	100	100	101	102	102	102	102	103	103	104
104	104	105	106	106	106	106	106	106	106	107	107	107	108
108	108	108	109	110	110	111	114	117	118	119	124	127	128



IndiceRichesse (Liège): classes d'amplitude 10

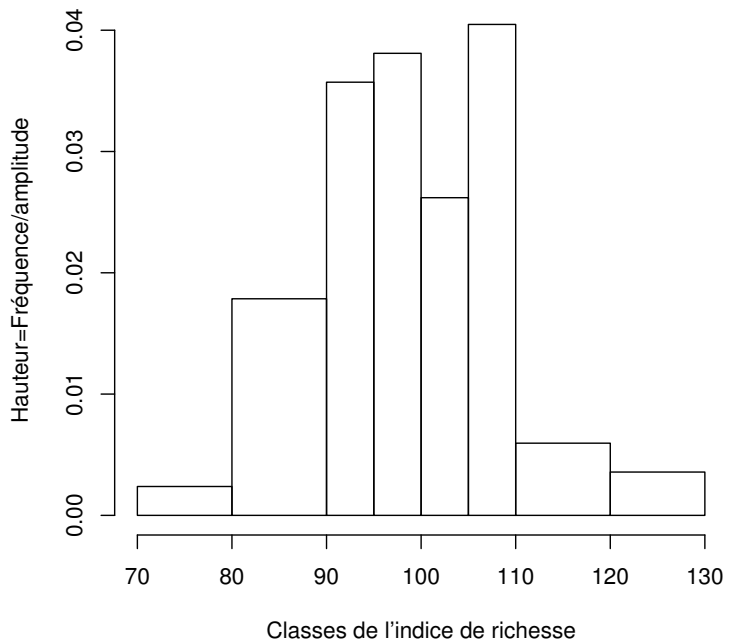
Classes	Effectifs	Fréquences	Eff. cum.	Fréq. cum.
$[70, 80]$	2	0.02	2	0.02
$]80, 90]$	15	0.18	17	0.20
$]90, 100]$	31	0.37	48	0.57
$]100, 110]$	28	0.33	76	0.90
$]110, 120]$	5	0.06	81	0.96
$]120, 130]$	3	0.04	84	1.00
Total	84	1	x	x

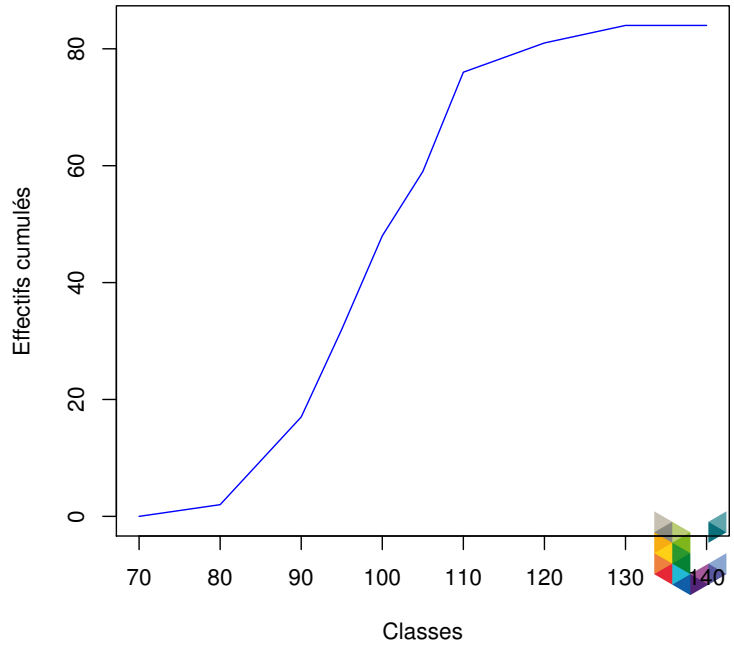




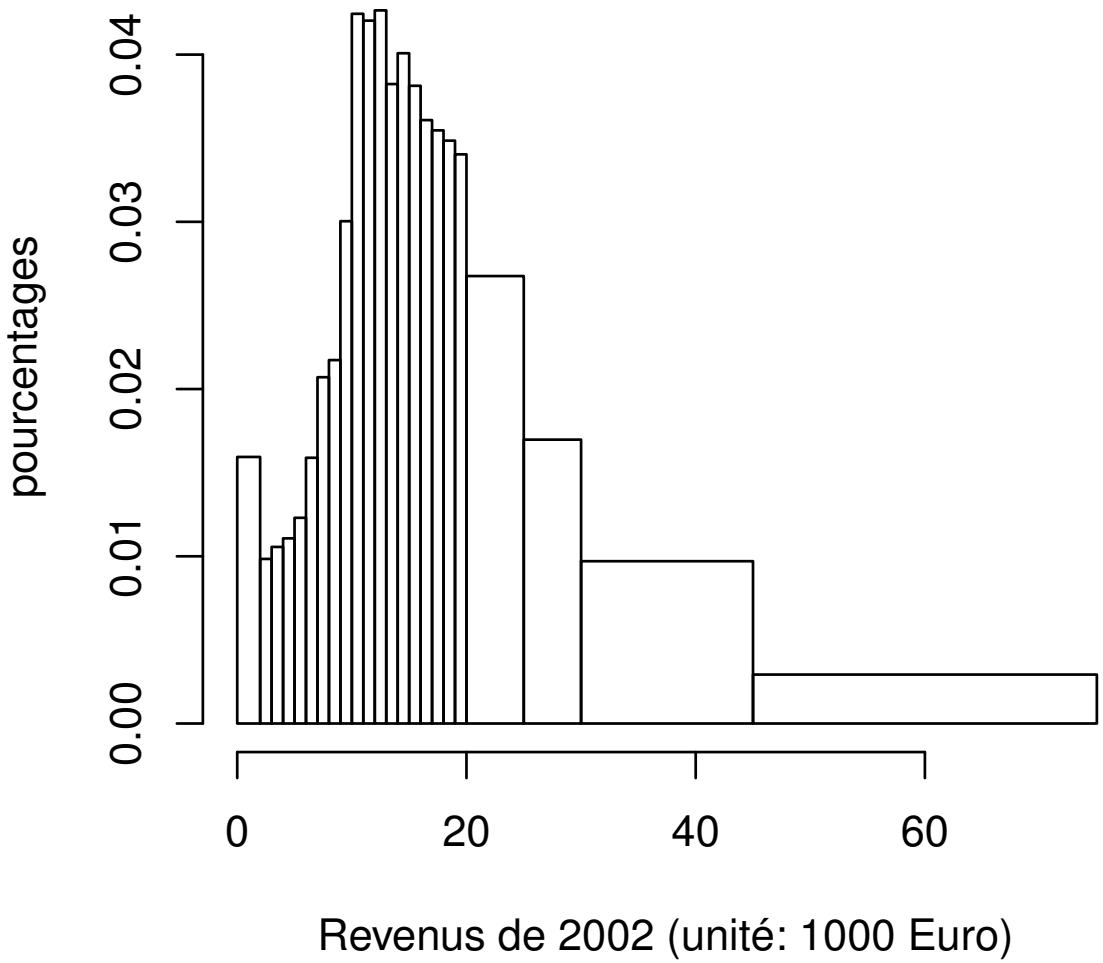
IndiceRichesse (Liège): classes d'amplitude 5 ou 10

Classes	Effectifs	Fréquences	Eff. cum.	Fréq. cum.
[70; 80]	2	0.02	2	0.02
]80; 90]	15	0.18	17	0.20
]90; 95]	15	0.18	32	0.38
]95; 100]	16	0.19	48	0.57
]100; 105]	11	0.13	59	0.70
]105; 110]	17	0.20	76	0.90
]110; 120]	5	0.06	81	0.96
]120; 130]	3	0.04	84	1
Total	84	1	x	x





Distributions des revenus des contribuables belges en 2002



Paramètres de position

1. La moyenne
2. La médiane et les quantiles
3. le mode

La moyenne arithmétique

$$\bar{x} = \frac{\text{somme des observations}}{n}.$$

Remarques:

1. La série doit être numérique.
2. L'ordre des observations n'a pas d'importance.
3. La moyenne arithmétique est rarement égale à une valeur observée et peut même ne pas être une valeur *observable* de la variable.

Exemples:

- Soit un ensemble de 5 tailles: 172cm, 176cm, 179cm, 186cm et 188cm. La moyenne vaut 180.2cm.
- Le nombre moyen de partis dans les majotirés communales est 1.4.

LIÈGE université

Calcul de la moyenne

A partir de la série brute $S = \{x_1, \ldots, x_n\}$

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}.$$

A partir de la série recensée:

$$\bar{x} = \frac{\sum_{j=1}^{J} n_j x_j}{n},$$

ou

$$\bar{x} = \sum_{j=1}^{J} \frac{n_j}{n} x_j = \sum_{j=1}^{J} f_j x_j.$$

Exemple: moyenne du nombre de partis

Nombres x_j	Effectifs n_j	Fréquences	Effectifs cumulés	Fréquences cumulées	$n_j \times x_j$
1	181	0.691	181	0.691	181
2	65	0.248	246	0.939	130
3	15	0.057	261	0.996	45
4	1	0.004	262	1.000	4
Total	262	1	x	x	360

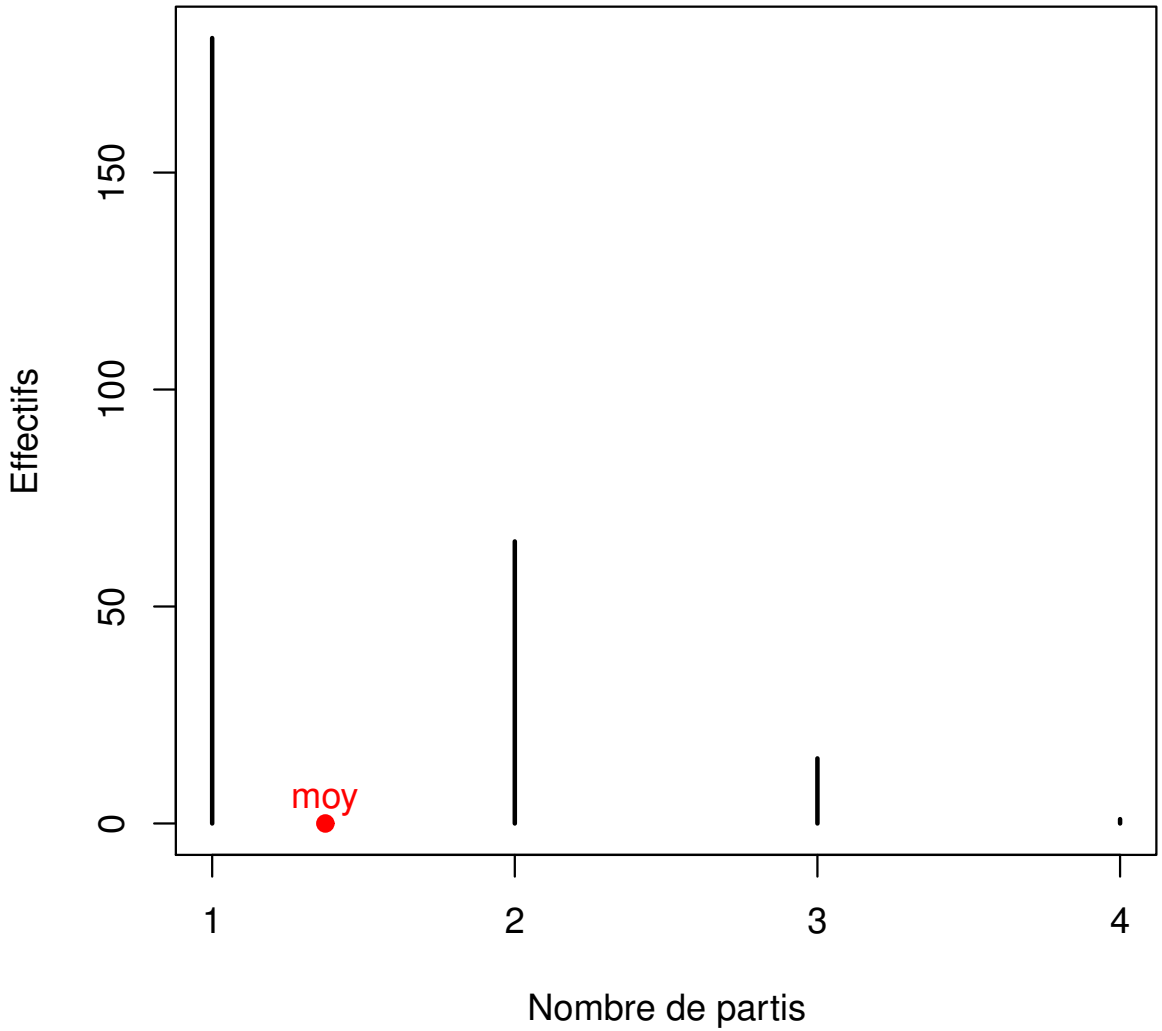
$$\bar{x} = \frac{\sum_{j=1}^{4} n_j x_j}{n} = \frac{360}{262} = 1.37.$$

ou

Nombres x_j	Effectifs	Fréquences f_j	Effectifs cumulés	Fréquences cumulées	$f_j \times x_j$
1	181	0.691	181	0.691	0.691
2	65	0.248	246	0.939	0.496
3	15	0.057	261	0.996	0.171
4	1	0.004	262	1.000	0.004
Total	262	1	x	x	1.36

$$\bar{x} = \sum_{j=1}^{5} f_j x_j = 1.36.$$

Illustration du “centre” de la série des nombres de partis



Calcul de la moyenne (suite)

A partir de la série groupée:

Calcul exact **si** les moyennes des observations de chaque classe, $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_J$, sont connues où

$$\bar{x}_j = \frac{\text{somme des observations de } \mathcal{C}_j}{n_j}$$

$\Leftrightarrow$ somme des observations de $\mathcal{C}_j = n_j \ \bar{x}_j$.

On a donc

$$\bar{x} = \frac{n_1 \ \bar{x}_1 + n_2 \ \bar{x}_2 + \ldots + n_J \ \bar{x}_J}{n} = \frac{\sum_{j=1}^{J} n_j \bar{x}_j}{n}.$$

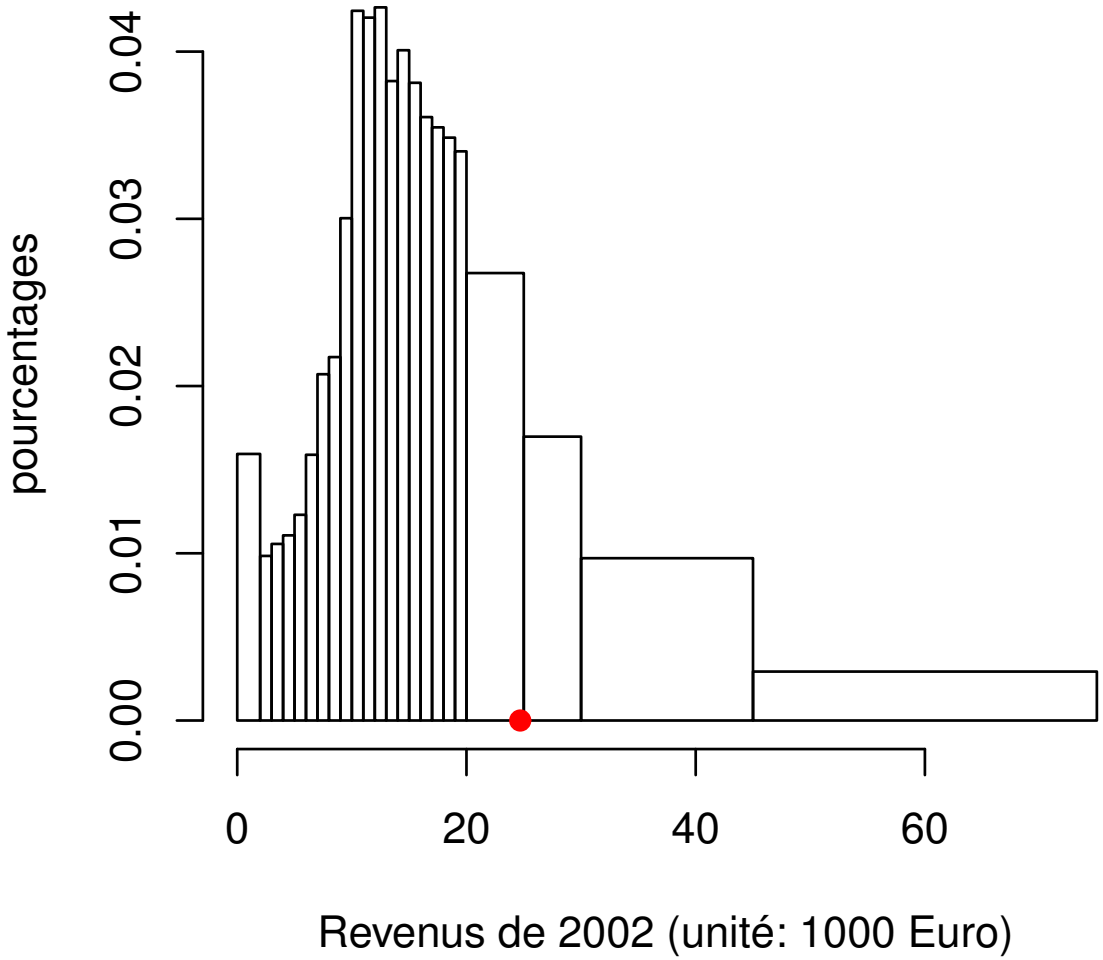
Moyenne des revenus en Belgique en 2002

Classes		Nombres de déclarations	%	Montant des revenus	%
1	$[0, 2[$	152 812	3.11	143 477 244	0.11
2	$[2, 3[$	47 179	0.96	117 727 370	0.10
3	$[3, 4[$	50 414	1.03	177 581 318	0.15
4	$[4, 5[$	52 955	1.08	238 124 842	0.20
5	$[5, 6[$	58 689	1.20	323 655 391	0.27
6	$[6, 7[$	75 877	1.55	496 101 333	0.41
7	$[7, 8[$	99 202	2.02	745 566 907	0.62
8	$[8, 9[$	103 927	2.12	884 436 706	0.73
9	$[9, 10[$	143 691	2.93	1 373 325 220	1.13
10	$[10, 11[$	203 232	4.14	2 145 593 991	1.77
11	$[11, 12[$	201 342	4.10	2 313 775 629	1.91
12	$[12, 13[$	203 870	4.16	2 546 638 256	2.10
13	$[13, 14[$	182 769	3.73	2 468 146 816	2.04
14	$[14, 15[$	191 786	3.91	2 781 779 980	2.30
15	$[15, 16[$	182 325	3.72	2 824 336 576	2.33
16	$[16, 17[$	172 746	3.52	2 849 732 579	2.35
17	$[17, 18[$	169 881	3.46	2 972 416 927	2.45
18	$[18, 19[$	166 977	3.40	3 088 535 702	2.55
19	$[19, 20[$	162 825	3.32	3 174 523 620	2.62
20	$[20, 25[$	639 957	13.05	14 279 109 275	11.79
21	$[25, 30[$	405 970	8.28	11 103 299 769	9.16
22	$[30, 45[$	697 043	14.20	25 435 294 910	20.98
23	$[45, 75[$	420 780	8.56	23 449 489 989	19.38
24	$[75, +\infty[$	119 374	2.42	15 195 811 336	12.54
Total		4 905 623	100	121 128 481 686	100

D'où, $\bar{x} = 24691.76$.

LIÈGE université

Illustration du “centre” de la série des revenus



Approximation de la moyenne en cas de données groupées

Si les moyennes des classes ne sont pas connues, on ne peut obtenir qu'une *approximation* de la moyenne.

Celle-ci s'obtient en supposant que $\bar{x}_j = c_j$.

Dans ce cas,

$$\bar{x} = \frac{\sum_{j=1}^{J} n_j c_j}{n}.$$

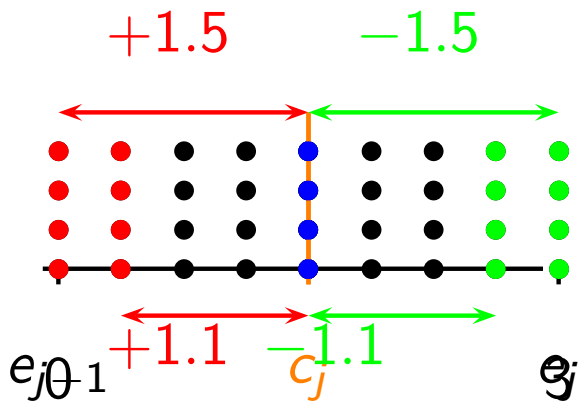
Plus exactement, $\bar{x} \approx \frac{\sum_{j=1}^{J} n_j c_j}{n}$.

Justification de l'approximation $\bar{x}_j = c_j$

$$\bar{x}_j \approx c_j \Leftrightarrow \frac{\Sigma \text{ des obs. de } \mathcal{C}_j}{n_j} \approx c_j \Leftrightarrow \Sigma \text{ des obs. de } \mathcal{C}_j \approx n_j \; c_j$$

Sous l'hypothèse de répartition uniforme, l'égalité est exacte:
Σ des obs. de $\mathcal{C}_j = n_j \; c_j$ car des écarts positifs/négatifs par rapport au centre sont compensés par des écarts égaux de signe contraire.

En effet, soient les 36 observations suivantes dans $\mathcal{C}_j$:



Dans la pratique

L'estimation de la moyenne à l'aide des centres des classes est assez bonne car

- Les sur- et sous-estimations ont tendance à se compenser au sein d'une classe;
- L'erreur positive ou négative obtenue dans une classe peut ensuite être compensée par une erreur de signe contraire faite dans une autre classe.

Bien entendu, l'estimation dépend de la répartition choisie et si la borne inférieure de la première classe et/ou la borne supérieure de la dernière classe ne sont pas connues ou représentatives de la situation, l'estimation peut devenir problématique.

LIÈGE université

Indice de richesse moyen à partir de la série groupée

Classes	Centres c_j	Effectifs n_j	Fréquences	Eff. cum.	Fréq. cum.	$n_j\ c_j$
$[70, 80]$	75	2	0.02	2	0.02	150
$]80, 90]$	85	15	0.18	17	0.20	1275
$]90, 100]$	95	31	0.37	48	0.57	2945
$]100, 110]$	105	28	0.33	76	0.90	2940
$]110, 120]$	115	5	0.06	81	0.96	575
$]120, 130]$	125	3	0.04	84	1	375
Total		84	1	x	x	8260

D'où, $\bar{x} = \frac{8260}{84} = 98.33$

Alors que la moyenne exacte est égale à 98.90.

Et, à partir de la 2ème répartition en classes (voir notes), on obtient 98.50.

Illustration du centre estimé de la série des indices de richesse

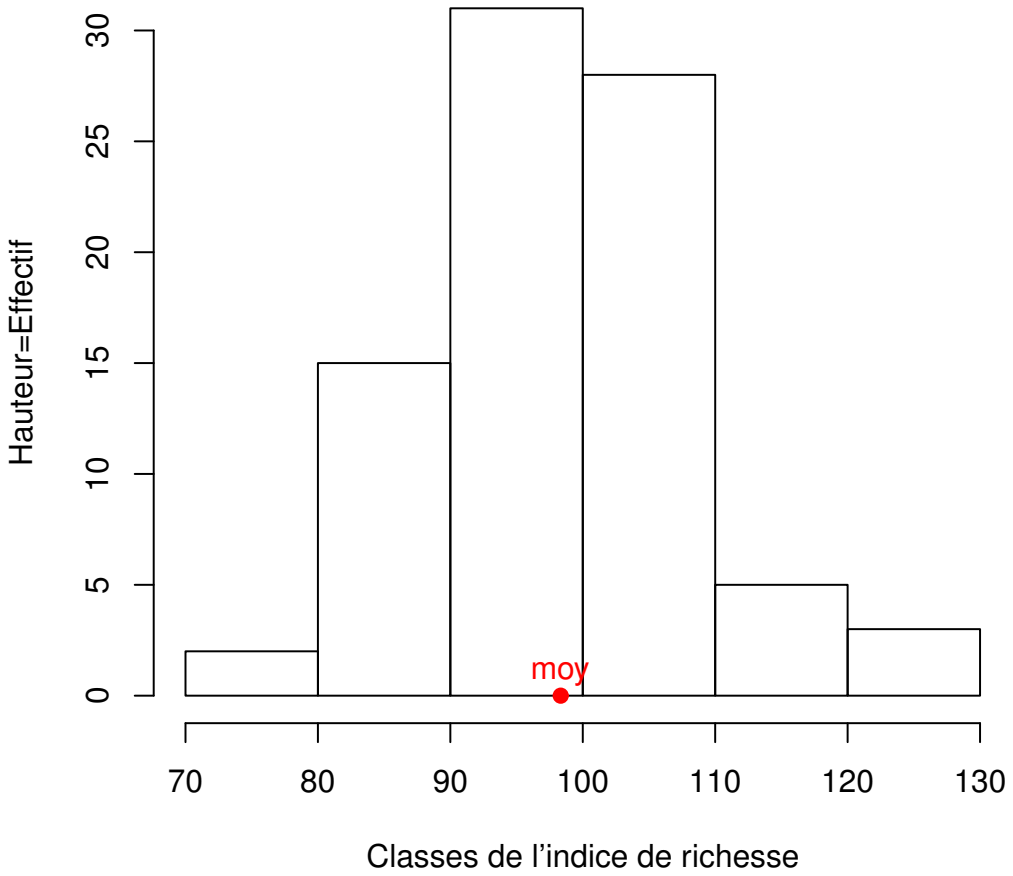
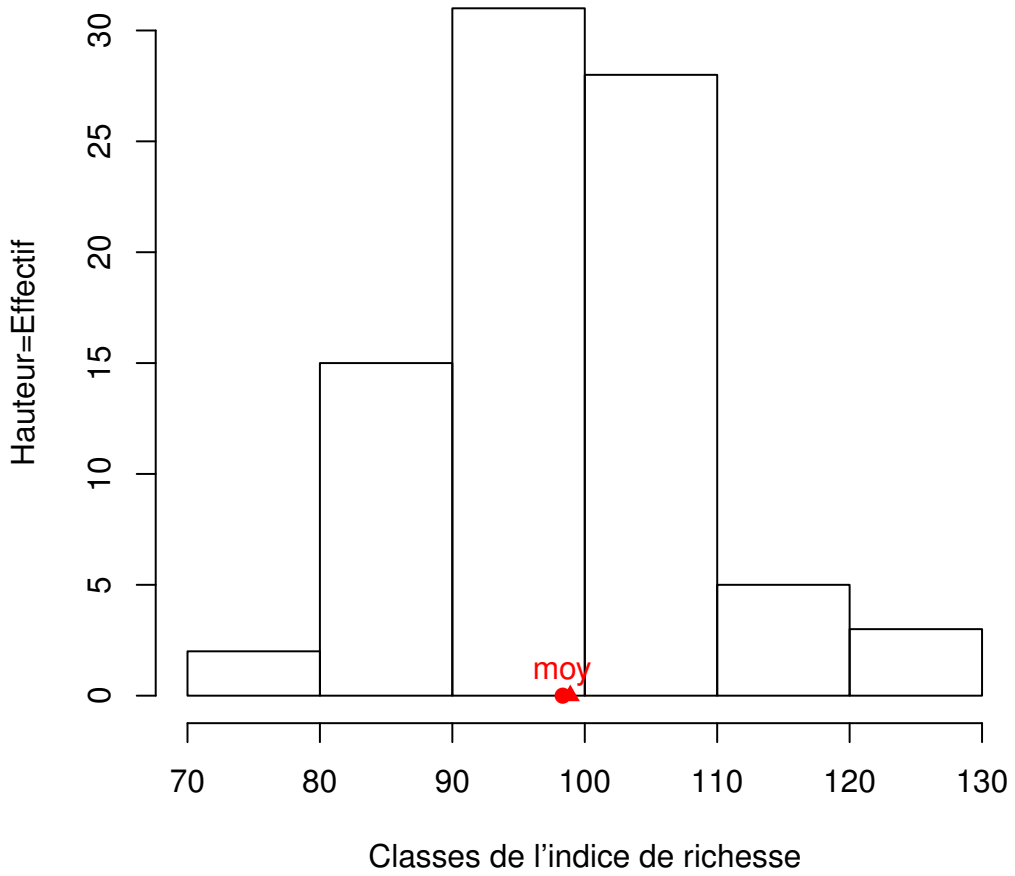


Illustration du centre exact et du centre approximatif de la série des indices de richesse



Propriétés mathématiques de la moyenne arithmétique

1) Effet sur la moyenne d'un changement d'échelle et/ou d'origine effectué sur les observations $x_1, \ldots, x_n$

$$x_i \xrightarrow[\text{d'échelle: } \times a]{\text{changement}} ax_i \xrightarrow[+b]{\text{translation}} x'_i = ax_i + b$$

pour obtenir la nouvelle série $S' = \{x'_1, \ldots, x'_n\}$, où a et b sont des constantes réelles,

Exemples de transformations $ax_i + b$

- Si x_i est la taille de l'individu i en centimètre, alors la valeur x'_i obtenue par la transformation

$$x_i \xrightarrow{\times \frac{1}{100}} x'_i = \frac{xi}{100}$$

est la taille de l'individu en mètre.

- Il existe plusieurs échelles de température:

	Degré Fahrenheit		Degré Celsius
Congélation de l'eau	32 °F	$\longleftrightarrow$	0 °C
Ebulition de l'eau	212 °F	$\longleftrightarrow$	100 °C

Si x_i désigne la température en °C pour le jour i, alors la valeur x'_i obtenue par la transformation

$$x_i \xrightarrow{\times \frac{9}{5}} \frac{9}{5}x_i \xrightarrow{+32} x'_i = \frac{9}{5}x_i + 32$$

est la température du même jour en degrés Fahrenheit.

Propriétés mathématiques de la moyenne arithmétique

Proposition

Si un changement d'échelle et/ou d'origine est effectué sur $x_1, \ldots, x_n$

$$x_i \xrightarrow[\text{d'échelle: } \times a]{\text{changement}} ax_i \xrightarrow[+b]{\text{translation}} x_i' = ax_i + b$$

pour obtenir la nouvelle série $S' = \{x_1', \ldots, x_n'\}$*, où* a *et* b *sont des constantes réelles, alors la moyenne arithmétique* $\bar{x}'$ *de la série* S' *est donnée par*

$$\bar{x}' = a\bar{x} + b,$$

où $\bar{x}$ *est la moyenne arithmétique de* $S = \{x_1, \ldots, x_n\}$.

1) En d'autres termes, la moyenne subit la même transformation:

$$\bar{x} \xrightarrow[\text{d'échelle: } \times a]{\text{changement}} a\bar{x} \xrightarrow[+b]{\text{translation}} \bar{x}' = a\bar{x} + b$$

Propriétés mathématiques de la moyenne arithmétique

1) $\bar{x}' = a\bar{x} + b$ si $x'_i = ax_i + b$
2) La série des valeurs centrées est $S_c = \{x_1 - \bar{x}, \ldots, x_n - \bar{x}\}$

Proposition

La série des valeurs centrées a une moyenne nulle.

Démonstration

La série centrée correspond à la transformation particulière

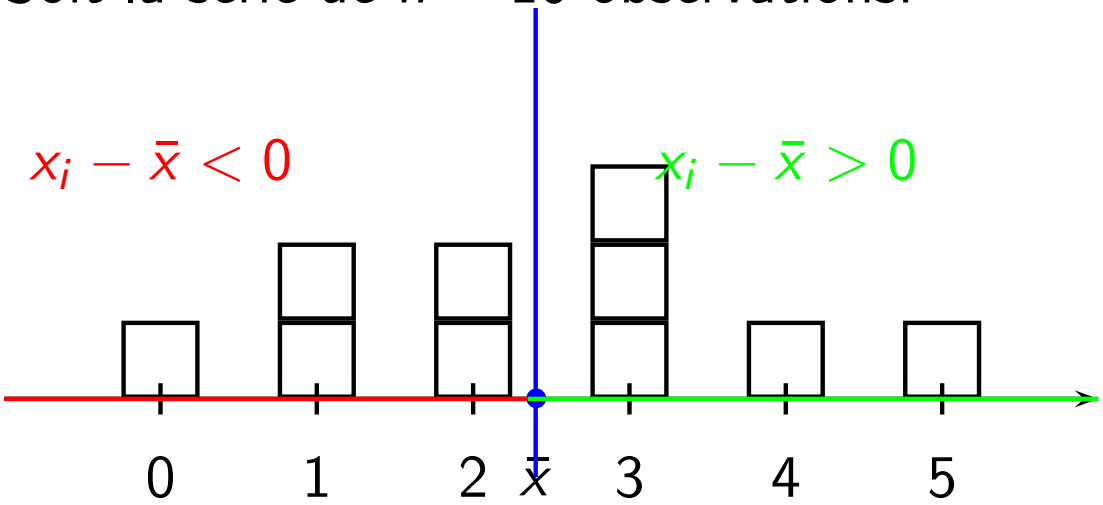
$$x_i \xrightarrow{-\bar{x}} x'_i = x_i - \bar{x}$$

c'est-à-dire $a = 1$ et $b = -\bar{x}$. Par la propriété 1, $\bar{x}_c = \bar{x} - \bar{x} = 0$

Interprétation de la moyenne nulle de la série centré

$$\bar{x}_c = 0 \Leftrightarrow \sum_{i=1}^{n}(x_i - \bar{x}) = 0 \Leftrightarrow \sum_{x_i<\bar{x}}(x_i - \bar{x}) + \sum_{x_i>\bar{x}}(x_i - \bar{x}) = 0$$

Soit la série de $n = 10$ observations:

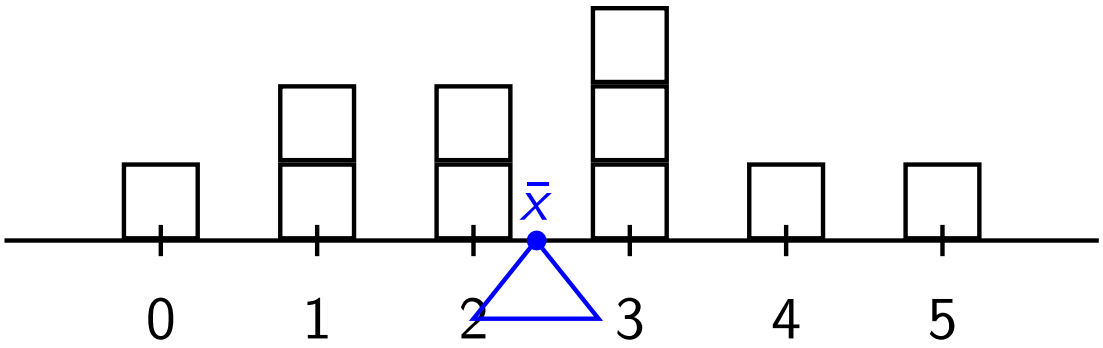


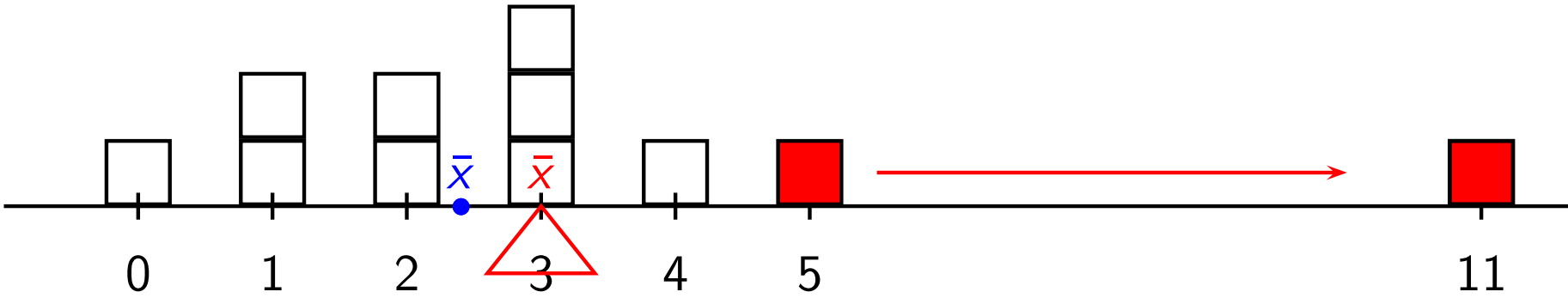
$$\sum_{x_i<\bar{x}}(x_i - \bar{x}) + \sum_{x_i>\bar{x}}(x_i - \bar{x}) = 0$$

ou, en mots, la somme des valeurs centrées positives est compensée par la somme des valeurs centrées négatives.

Propriétés mathématiques de la moyenne arithmétique

3) La moyenne arithmétique est le *centre de gravité de la série*.
Soit la série de $n = 10$ observations:





Cette propriété explique la sensibilité de la moyenne à la présence d'observations atypiques.

Propriétés mathématiques de la moyenne arithmétique

4) Propriété d'optimalité de la moyenne:

Proposition

La somme des carrés des écarts des éléments d'une série par rapport à la moyenne arithmétique de la série est inférieure ou égale à la somme des carrés des écarts par rapport à toute autre valeur $a \in \boldsymbol{R}$, c'est-à-dire

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 \leq \sum_{i=1}^{n}(x_i - a)^2 \ \forall \ a \in \boldsymbol{R}.$$

Démonstration (Outils: recherche du minimum d'une fonction sur $\boldsymbol{R}$)

La moyenne est donc le paramètre de tendance centrale optimal selon le critère des moindres carrés.

Propriétés mathématiques de la moyenne arithmétique (fin)

5) Moyenne d'une population P répartie en k sous-populations $P_1, \ldots, P_k$ d'effectifs $n_1, \ldots, n_k$ (avec $n_1 + \ldots + n_k = n$)

Proposition

Soient $\bar{x}_i$, $1 \leq i \leq k$, les moyenne des sous-populations et $p_i = n_i/n$, $1 \leq i \leq k$, les proportions d'individus dans chaque sous-population, la moyenne globale de P, notée $\bar{x}$, est égale à

$$\bar{x} = \frac{\sum_{i=1}^{k} n_i \bar{x}_i}{\sum_{i=1}^{k} n_i} = \sum_{i=1}^{k} p_i \bar{x}_i$$

Démonstration

Outils: exploitation de la définition de la moyenne (cfr moyenne d'une série groupée en classes)

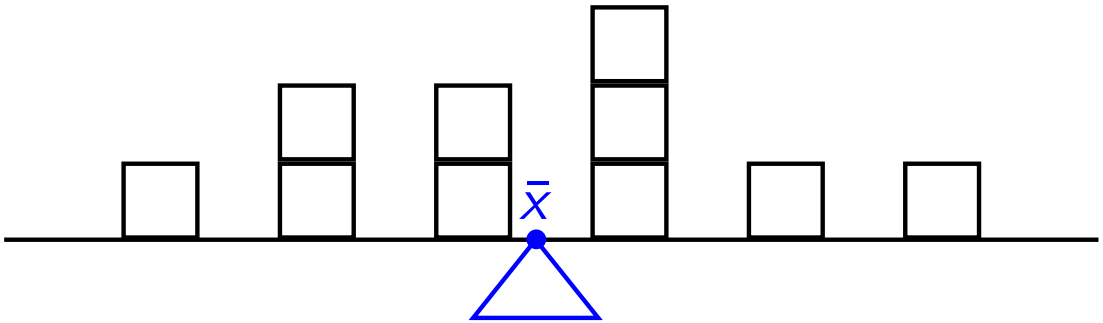
somme des observations de $P_j = n_j \ \bar{x}_j$.

On a donc $\bar{x} = \frac{\sum_{j=1}^{k} n_j \bar{x}_j}{\sum_{j=1}^{k} n_j}$.

Généralisation de la moyenne

Par définition:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{\sum_{i=1}^{n} 1 \; x_i}{\sum_{i=1}^{n} 1}$$



Dans certaines applications, il est utile de pouvoir attribuer des poids différents aux observations.

La moyenne arithmétique pondérée

A chaque observation x_i, on attribue un poids w_i, positif ou nul.

La moyenne arithmétique pondérée est alors définie par

$$\bar{x}_w = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}.$$

Application : Calcul de taux de change moyen

Déplacement 1: q_1 £ à t_1 €/£
Déplacement 2: q_2 £ à t_2 €/£
⋮
Déplacement n: q_n £ à t_n €/£

Globalement, une quantité $Q = \sum_{i=1}^{n} q_i$ £ a été achetée pour un coût total de $C = \sum_{i=1}^{n} q_i \ t_i$ en €.

Le taux de change moyen est le taux $\bar{t}$ vérifiant

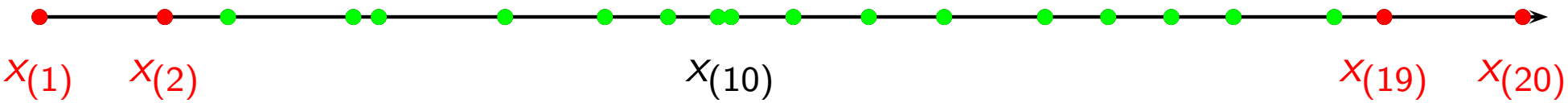
$$Q\bar{t} = C \Leftrightarrow \bar{t} = \frac{C}{Q} = \frac{\sum_{i=1}^{n} q_i \ t_i}{\sum_{i=1}^{n} q_i}$$

Cas particulier: la *moyenne tronquée au seuil* α

Soit $0 \leq \alpha < \frac{1}{2}$.

La *moyenne tronquée au seuil* α, $\bar{x}_\alpha$, consiste à attribuer un poids nul aux $\alpha \times n$ plus petites observations ainsi qu'aux $\alpha \times n$ plus grandes observations et un poids égal à 1 aux autres observations.

Soient les $n = 20$ observations suivantes

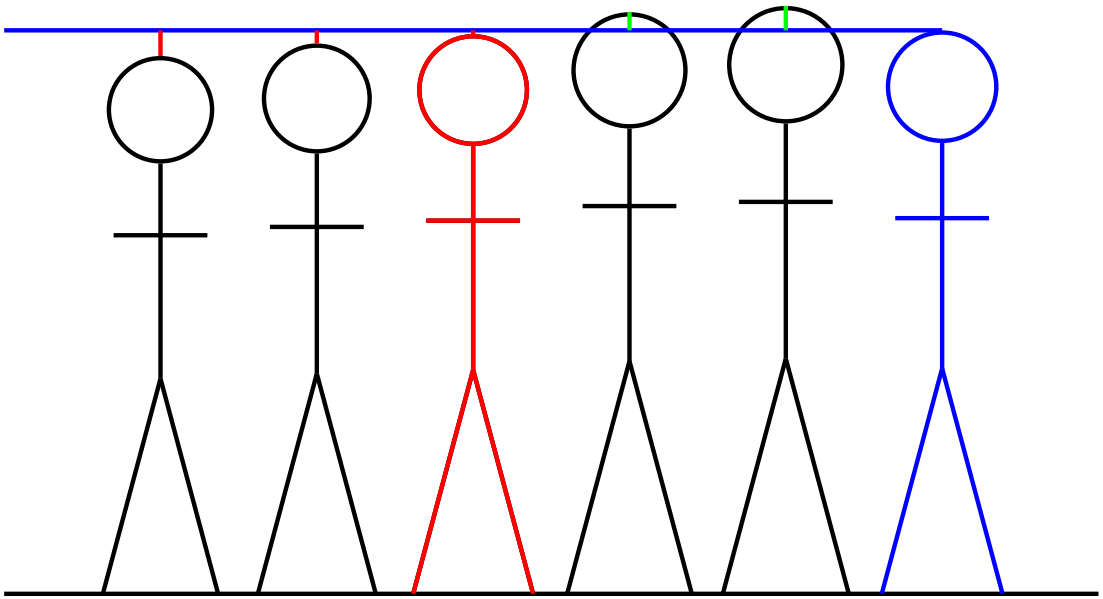


Prenons $\alpha = 0.1$. Dans ce cas, $\alpha n = 0.120 = 2$.

$$\bar{x}_\alpha = \frac{0 \times x_{(1)} + 0 \times x_{(2)} + \sum_{i=3}^{18} 1 \times x_{(i)} + 0 \times x_{(19)} + 0 \times x_{(20)}}{16}.$$

$$= \frac{\sum_{i=[\alpha n]+1}^{n-[\alpha n]} x_{(i)}}{n - 2[\alpha n]} \text{ où } [z] \text{ est le plus grand entier} \leq z.$$

Moyenne $\longleftrightarrow$ médiane



$\tilde{x}$ = valeur située "au milieu" de la série ordonnée de telle sorte qu'approximativement 50 % des observations aient une valeur inférieure et 50 % une valeur supérieure.

Dès qu'une notion d'ordre est disponible, une médiane peut être calculée.

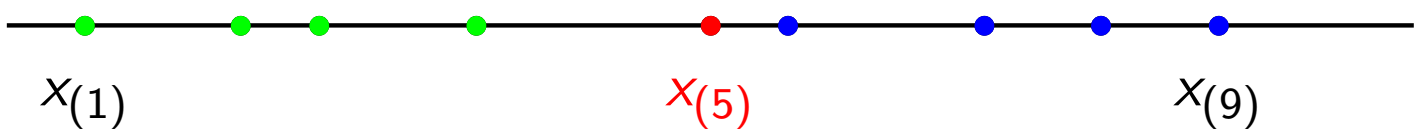
La définition précise de la médiane dépend de la forme dans laquelle la série se présente.

Médiane d'une série brute

Supposons $x_{(1)} < x_{(2)} < \ldots < x_{(n)}$

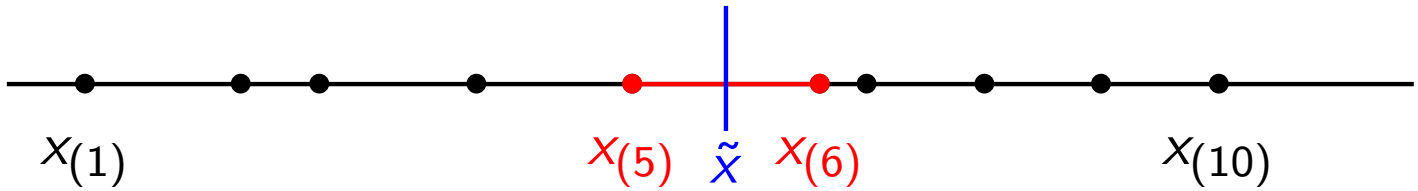
La position du milieu va dépendre de la parité de n

- n impair: $n = 2k + 1$



$\Rightarrow \tilde{x} = x_{(k+1)}$

- n pair: $n = 2k$



$\Rightarrow$ Intervalle médian $=]x_{(k)}, x_{(k+1)}[$

Convention: $\tilde{x} = \frac{x_{(k)} + x_{(k+1)}}{2}$.

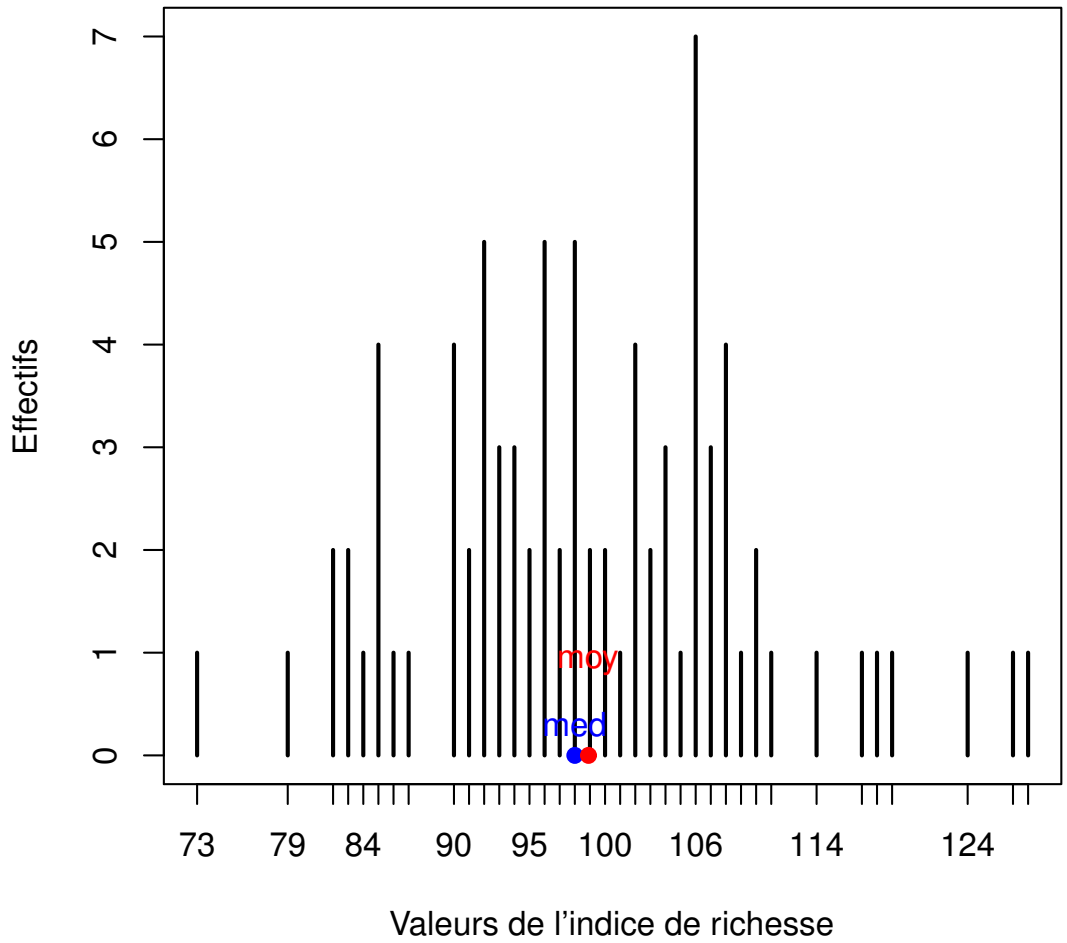
Indice de richesse médian en province de Liège

Même si les données ne sont pas strictement ordonnées, la définition de la médiane reste d'application.

73	79	82	82	83	83	84	85	85	85	85	86	87	90
90	90	90	91	91	92	92	92	92	92	93	93	93	94
94	94	95	95	96	96	96	96	96	97	97	98	98	98
98	98	99	99	100	100	101	102	102	102	102	103	103	104
104	104	105	106	106	106	106	106	106	106	107	107	107	108
108	108	108	109	110	110	111	114	117	118	119	124	127	128

$$\tilde{x} = \frac{x_{(42)} + x_{(43)}}{2} = 98$$

Visualiation du centre



Médiane d'une série recensée

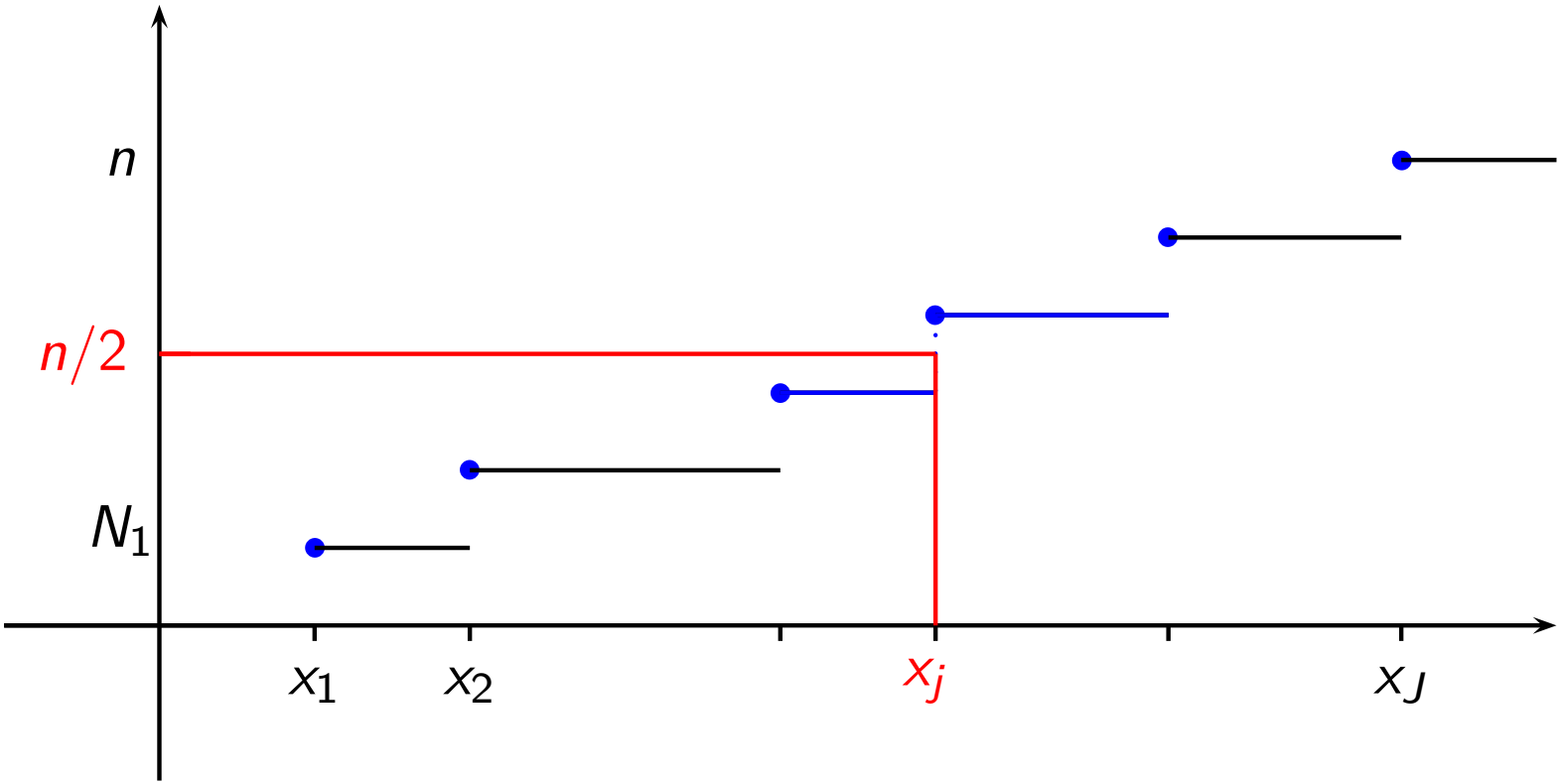
Par définition, la médiane devrait se positionner là où l'effectif cumulé passe par $n/2$.

Valeurs	Effectifs	Effectifs cumulés	
x_1	n_1	N_1	$< n/2$?
x_2	n_2	N_2	$< n/2$?
$\vdots$			
x_{j-1}	n_{j-1}	N_{j-1}	$< n/2$
x_j	n_j	N_j	$> n/2 = n/2$
x_{j+1}	n_{j+1}	N_{j+1}	
$\dots$			
x_J	n_J	N_J	
	n		

Deux cas possibles:

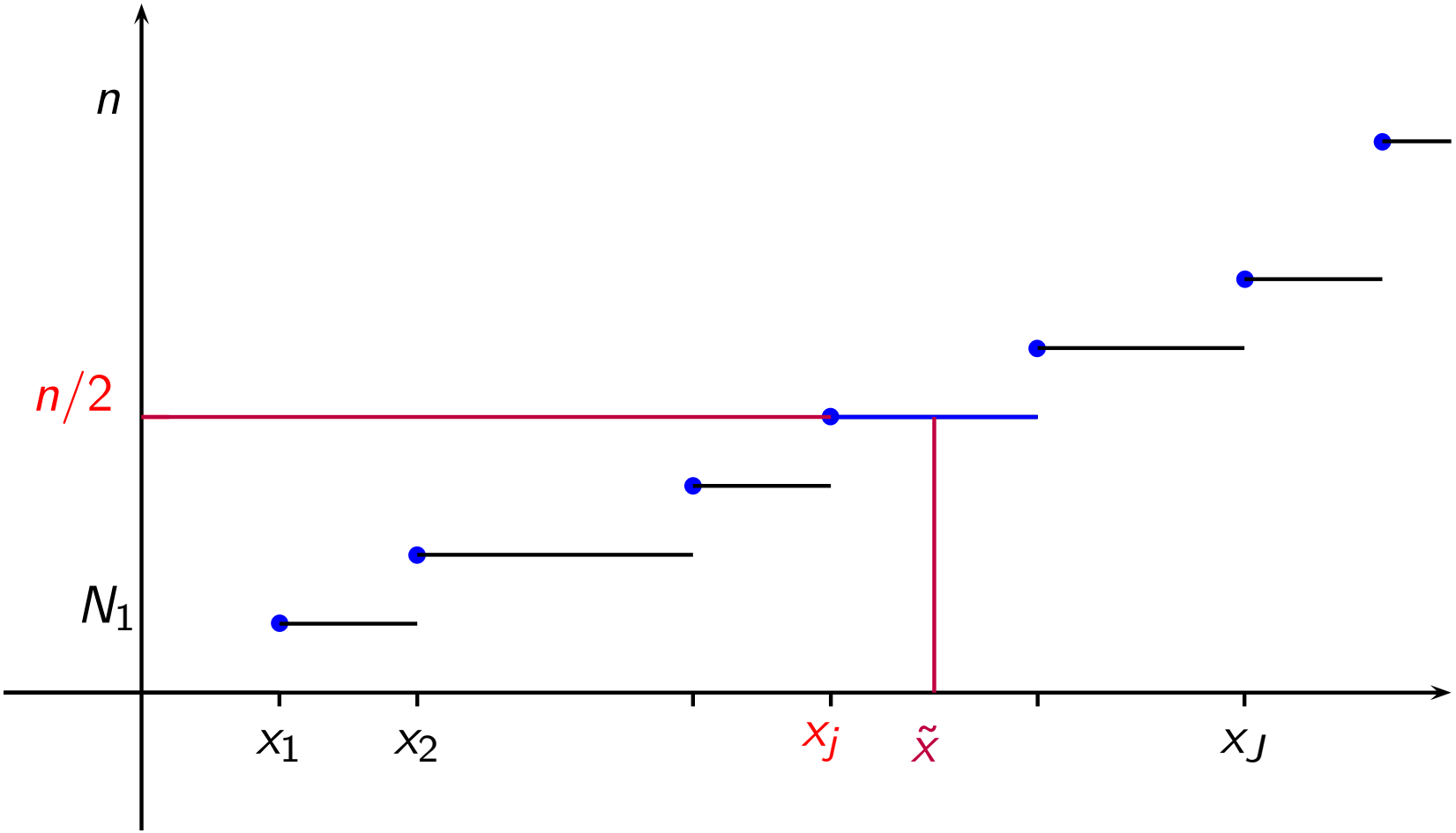
- Il existe j tel quel $N_{j-1} < \frac{n}{2} < N_j \Rightarrow \tilde{x} = x_j$
- Il existe j tel quel $N_j = \frac{n}{2} \Rightarrow \tilde{x} = \frac{x_j + x_{j+1}}{2}$

A partir de la courbe cumulative des effectifs cumulés



- Soit il existe j tel quel $N_{j-1} < \frac{n}{2} < N_j \Rightarrow \tilde{x} = x_j$

A partir de la courbe cumulative des effectifs cumulés

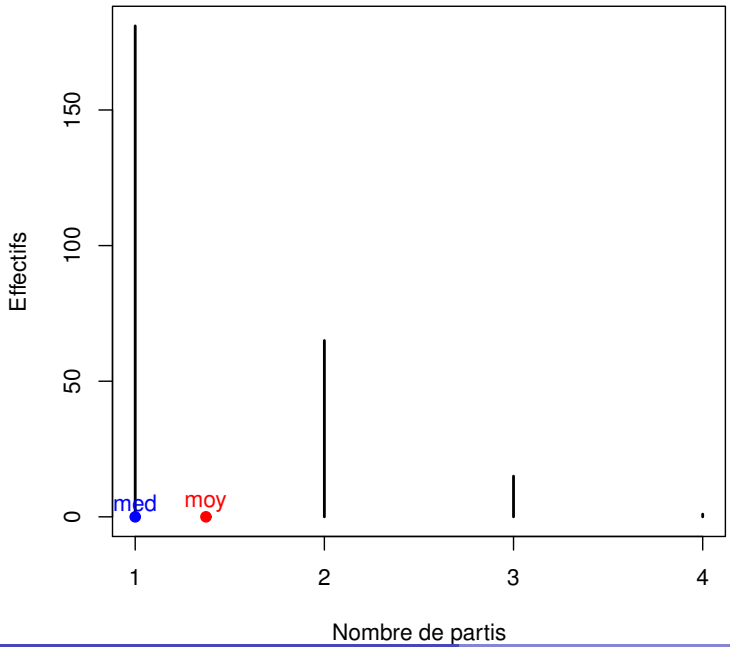


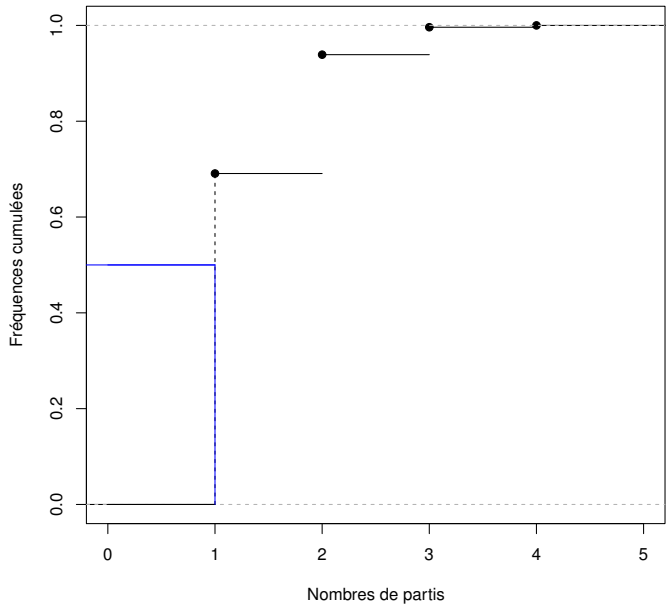
- soit il existe j tel quel $N_j = \frac{n}{2} \Rightarrow \tilde{x} = \frac{x_j + x_{j+1}}{2}$

La même démarche peut être suivie avec la distribution des fréquences cumulées!

Nombre médian de partis

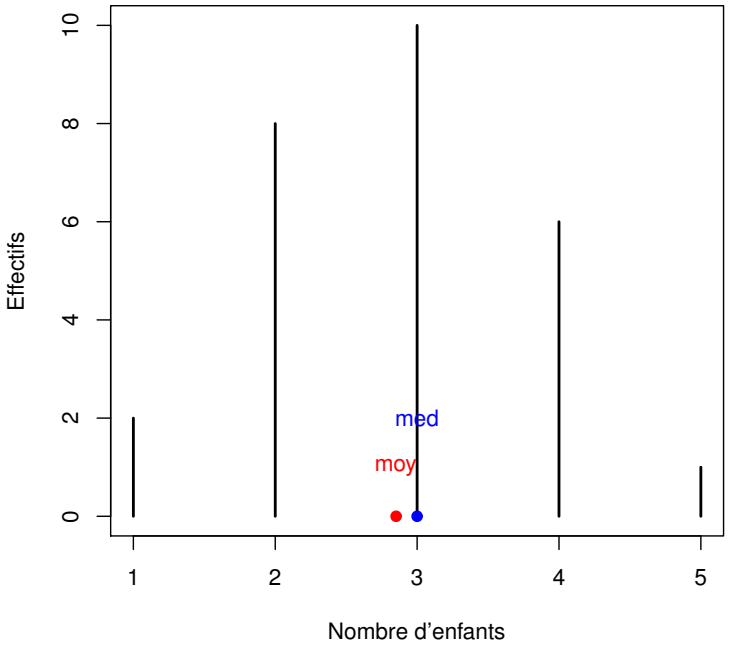
Valeurs	Effectifs	Fréquences	Effectifs cumulés	Fréquences cumulées
1	181	0.691	181	0.691
2	65	0.248	246	0.939
3	15	0.057	261	0.996
4	1	0.004	262	1.000
Total	262	1	x	x

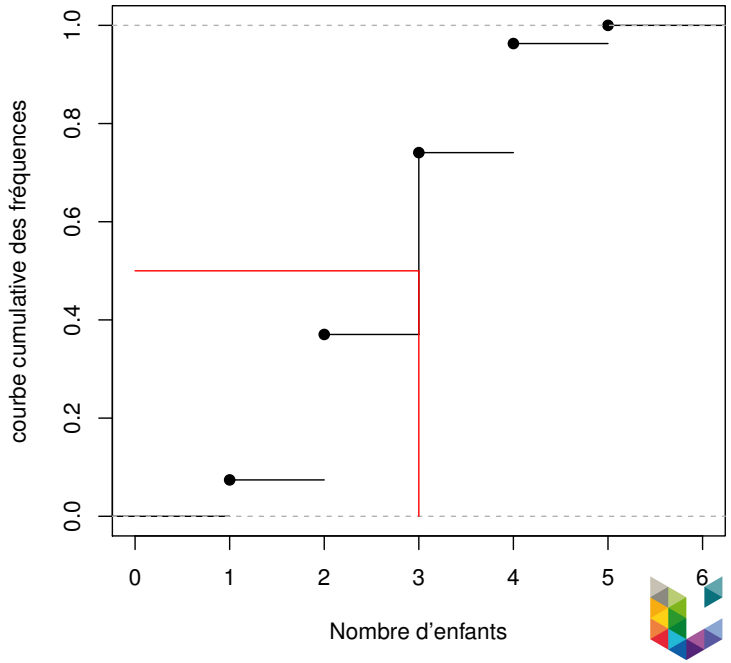




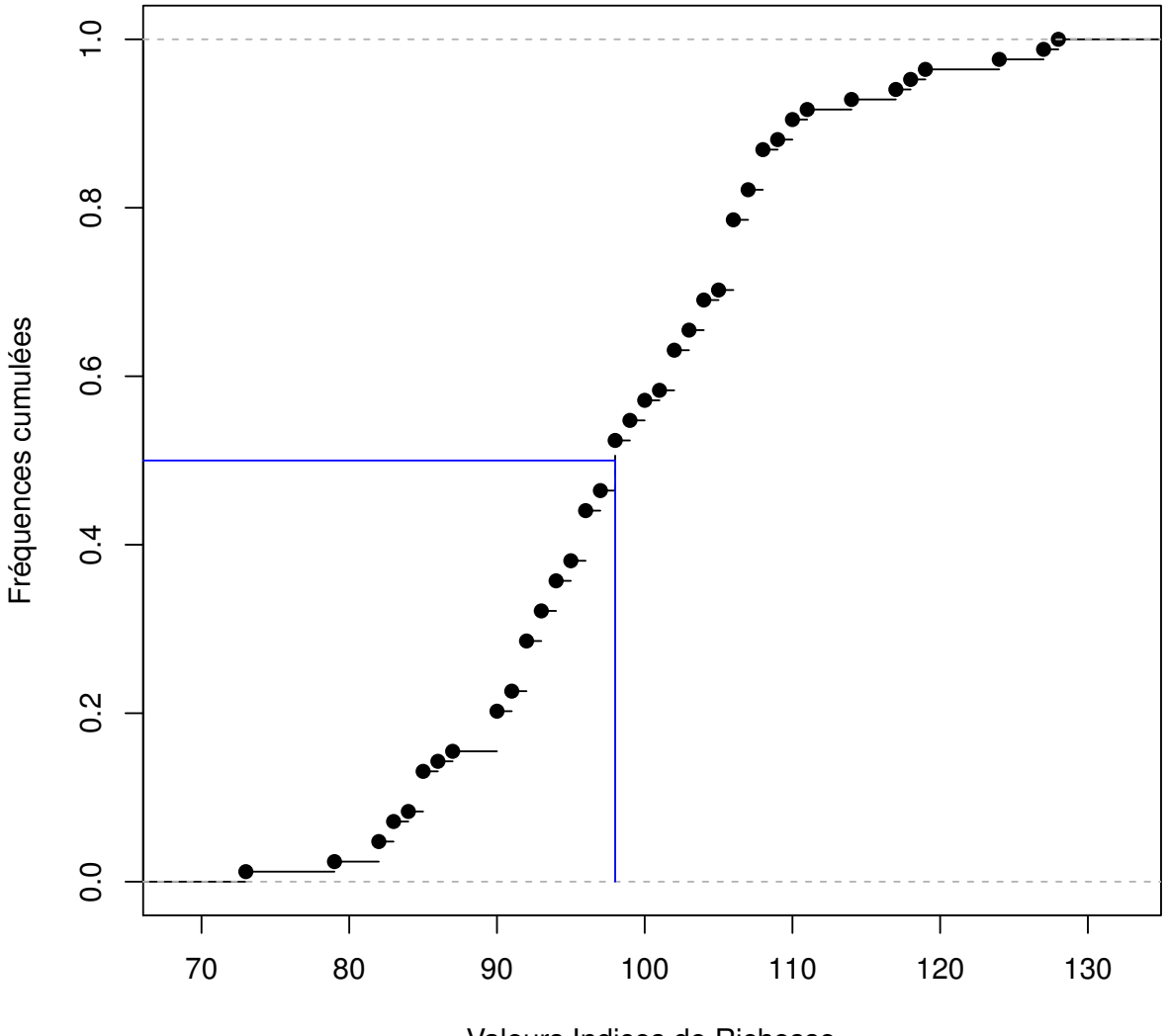
Autre exemple: nombre d'enfants dans 27 familles

Nombres	Effectifs	Fréquences	Effectifs cumulés	Fréquences cumulées
1	2	0.07	2	0.07
2	8	0.30	10	0.37
3	10	0.37	20	0.74
4	6	0.22	26	0.96
5	1	0.04	27	1.00
Total	27	1	x	x





A partir de la fonction de répartition pour les indices de richesse



Définition de la médiane par le chat de Geluck

Médiane d'une série groupée

La fonction $y = N(x)$ donne le nombre d'observations de la série dont la valeur est inférieure ou égale à x.

D'où, $\tilde{x}$ vérifie

$$N(\tilde{x}) = \frac{n}{2}.$$

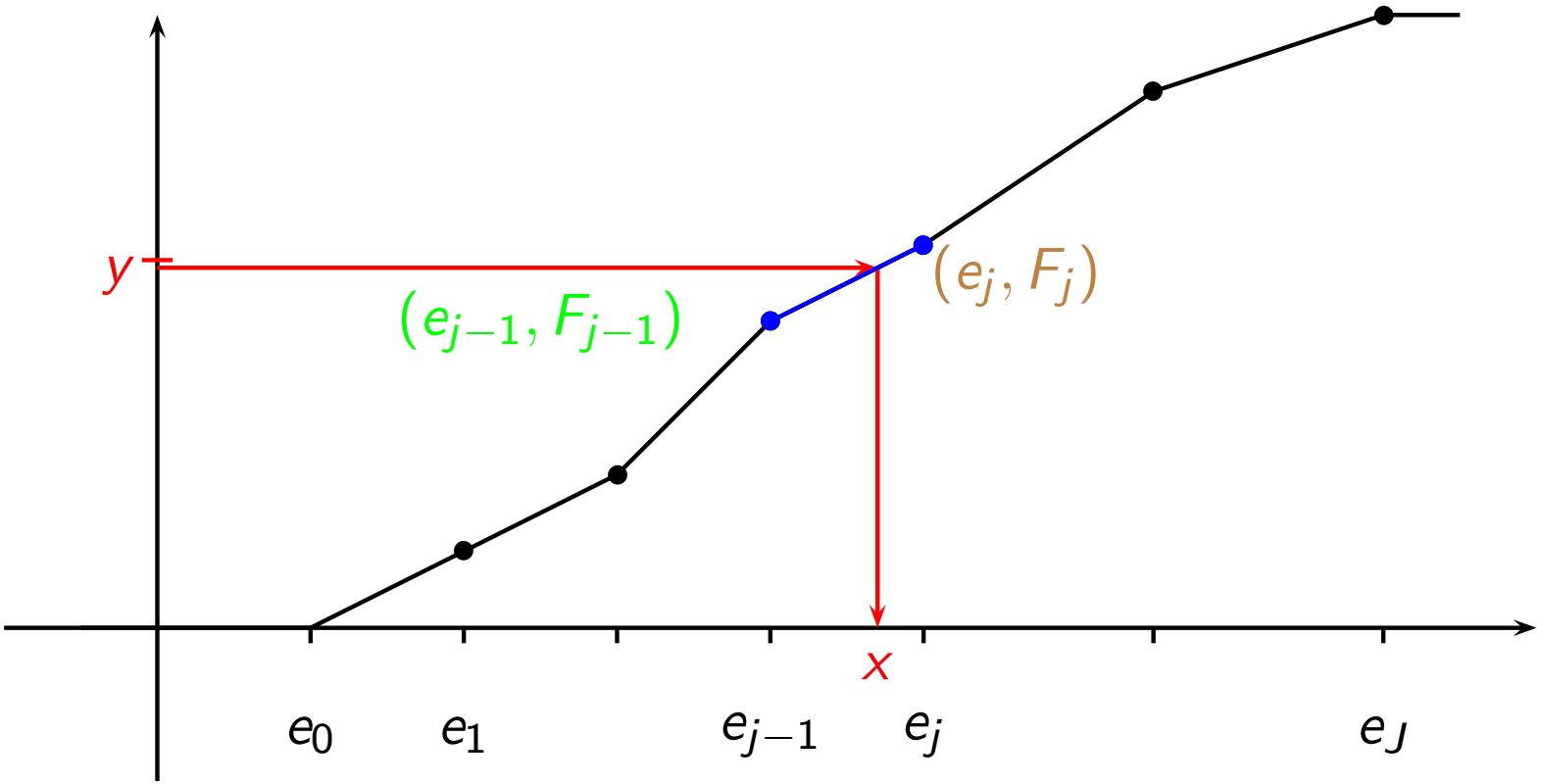
De la même manière, $\tilde{x}$ vérifie

$$F(\tilde{x}) = 0.5$$

où F est l'ogive des fréquences cumulées.

Rappel: exploitation de l'ogive $y = F(x)$

Soit $y \in \ ]0, 1[$, il y a une valeur $x \in \boldsymbol{R}$ telle que la proportion d'observations inférieures ou égales à x est égale à y.



Pour calculer x: Equation de la droite:

$$y - F_{j-1} = \frac{F_j - F_{j-1}}{e_j - e_{j-1}}(x - e_{j-1}) \Leftrightarrow x = e_{j-1} + \frac{e_j - e_{j-1}}{F_j - F_{j-1}}(y - F_{j-1})$$

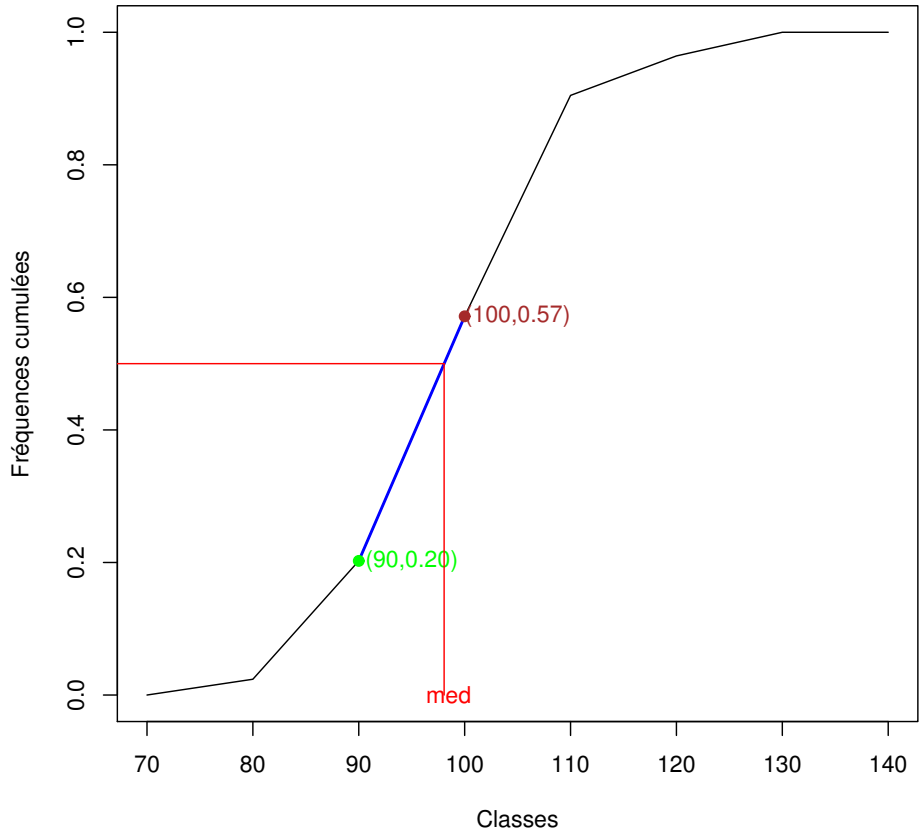
Calcul de la médiane

La classe médiane, $\mathcal{C}_m$, est la première classe dont la fréquence cumulée est supérieure ou égale à $1/2$.

A partir de l'ordonnée $1/2$ sur l'ogive des fréquences cumulées, on obtient la médiane par interpolation linéaire.

Attention: il s'agit d'une approximation basée sur l'hypothèse de répartition uniforme et qui varie en fonction de la répartition en classes effectuée!

Médiane des indices de richesse groupés (classes de 10)



$$\tilde{x} = 90 + 10 \times \frac{0.5 - 0.20}{0.369} = 98.06.$$

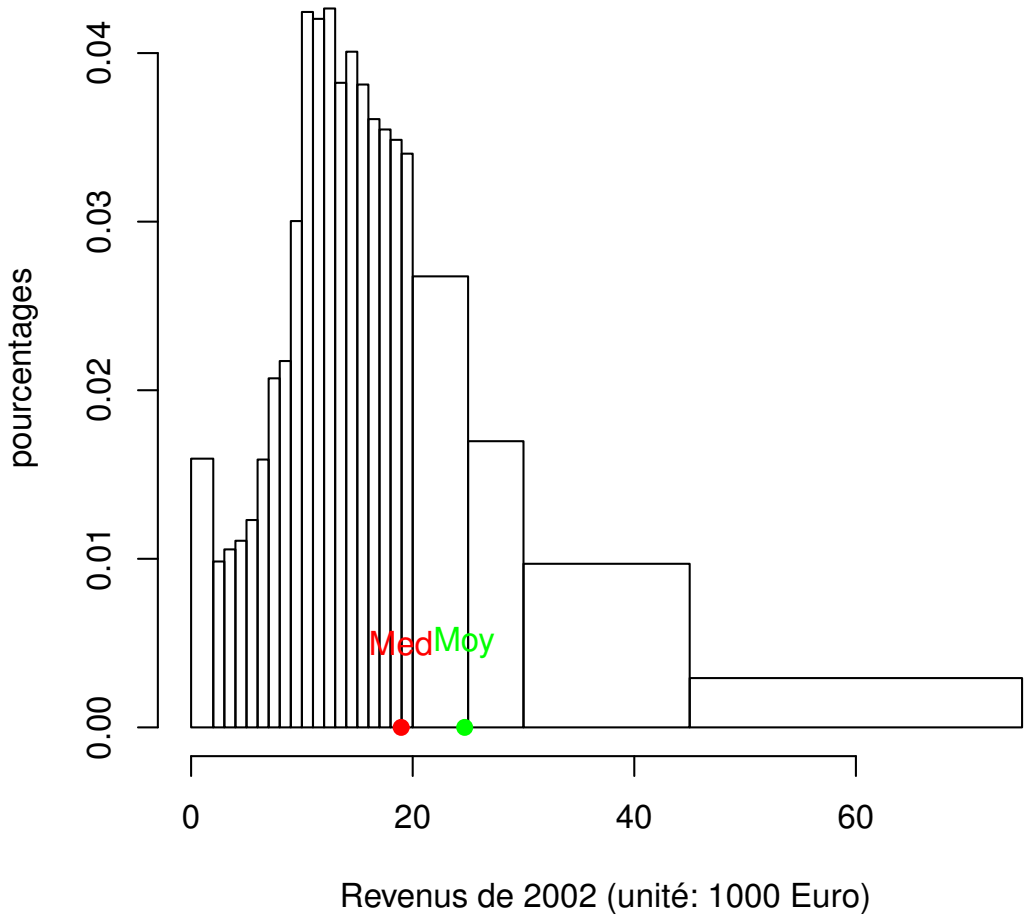
Revenu médian

Classes	f_j en %	F_j en %
[0, 2[	3.11	3.11
[2, 3[	0.96	4.07
[3, 4[	1.03	5.10
[4, 5[	1.08	6.18
[5, 6[	1.20	7.38
[6, 7[	1.55	8.93
[7, 8[	2.02	10.95
[8, 9[	2.12	13.07
[9, 10[	2.93	16.00
[10, 11[	4.14	20.14
[11, 12[	4.10	24.24
[12, 13[	4.16	28.40
[13, 14[	3.73	32.13
[14, 15[	3.91	36.04
[15, 15[	3.72	39.76
[16, 17[	3.52	43.28
[17, 18[	3.46	46.74
[18, 19[	3.40	50.14
[19, 20[	3.32	53.46
[20, 25[	13.05	66.51
[25, 30[	8.28	74.79
[30, 45[	14.20	88.99
[45, 75[	8.56	97.55
[75, $+\infty$[	2.42	100.00

$$\tilde{x} = 18 + 1 \times \frac{0.5 - 0.4674}{0.034} = 18.958 \times 1000€.$$

contre un salaire moyen de 24 691,76 €.

Moyenne et médiane des revenus de 2002



Propriétés de la médiane

1) La médiane peut se calculer sur une variable qualitative ordinale.

2) Effet d'un changement d'échelle et d'origine sur les observations

Proposition

Si une transformation affine est appliquée aux observations:

$$x_1, \ldots, x_n \quad \longrightarrow \quad x'_1, \ldots, x'_n$$

avec $x'_i = ax_i + b$ *où* a *et* b *sont des constantes réelles, alors la médiane subit la même transformation:*

$$\tilde{x} \quad \longrightarrow \quad \tilde{x}' = a\tilde{x} + b$$

Démonstration; Outils: propriétés des inégalités

A partir de la série $\{x_1, \ldots, x_n\}$, on a

$$x_{(1)} \leq \ldots \leq x_{(j)} \leq x_{(j+1)} \leq \ldots \leq x_{(n)}$$

D'où, pour tout $a > 0$ et tout b

$$x_{(1)} \leq \ldots \leq x_{(j)} \leq x_{(j+1)} \leq \ldots \leq x_{(n)}$$

$$ax_{(1)} \leq \ldots \leq ax_{(j)} \leq ax_{(j+1)} \leq \ldots \leq ax_{(n)}$$

$$ax_{(1)}+b \leq \ldots \leq ax_{(j)}+b \leq ax_{(j+1)}+b \leq \ldots \leq ax_{(n)}+b$$

On a donc $x'_{(j)} = ax_{(j)} + b$ pour tout j.

En particulier:

Si $n = 2k + 1$, l'observation centrale, $x'_{(k+1)}$, vérifie

$$x'_{(k+1)} = ax_{(k+1)} + b \Rightarrow \tilde{x}' = a\tilde{x} + b$$

Et si $n = 2k$, les deux observations du milieu, $x'_{(k)}$ et $x'_{(k+1)}$, sont telles que

$$\tilde{x}' = \frac{x'_{(k)} + x'_{(k+1)}}{2} = \frac{ax_{(k)} + b + ax_{(k+1)} + b}{2} = a\frac{x_{(k)} + x_{(k+1)}}{2} + b$$

Démonstration

Et si $a < 0$?

A partir de

$$x_{(1)} \leq \ldots \leq x_{(j)} \leq x_{(j+1)} \leq \ldots \leq x_{(n)}$$

Il vient

$$ax_{(1)} \geq \ldots \geq ax_{(j)} \geq ax_{(j+1)} \geq \ldots \geq ax_{(n)}$$

Ou encore

$$ax_{(1)}+b \geq \ldots \geq ax_{(j)}+b \geq ax_{(j+1)}+b \geq \ldots \geq ax_{(n)}+b$$

L'ordre est inversé: $x'_{(j)} = ax_{(n-j+1)} + b$ mais le milieu reste au milieu:
Si $n = 2k + 1$

$$ax_{(1)}+b \geq \ldots \geq \underbrace{ax_{(k+1)}+b}_{=x'(k+1)} \geq \ldots \geq ax_{(n)}+b$$

Si $n = 2k$

$$ax_{(1)}+b \geq \ldots \geq \underbrace{ax_{(k)}+b}_{} \geq \underbrace{ax_{(k+1)}+b}_{} \geq \ldots \geq ax_{(n)}+b$$

Démonstration du chat de Geluck

Propriétés de la médiane (suite)

3) Interprétation:

- Dans le cas discret: ☹

Valeurs	Effectifs	Fréquences	Effectifs cumulés	Fréquences cumulées
1	181	0.691	181	0.691
2	65	0.248	246	0.939
3	15	0.057	261	0.996
4	1	0.004	262	1.000
Total	262	1	x	x

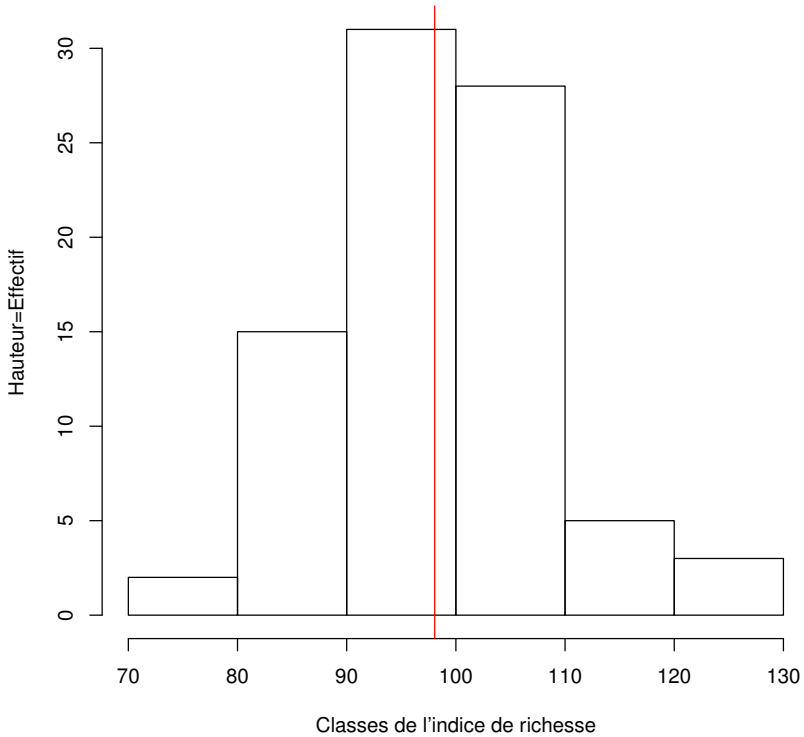
Nombres	Effectifs	Fréquences	Effectifs cumulés	Fréquences cumulées
1	2	0.07	2	0.07
2	8	0.30	10	0.37
3	10	0.37	20	0.74
4	6	0.22	26	0.96
5	1	0.04	27	1.00
Total	27	1	x	x

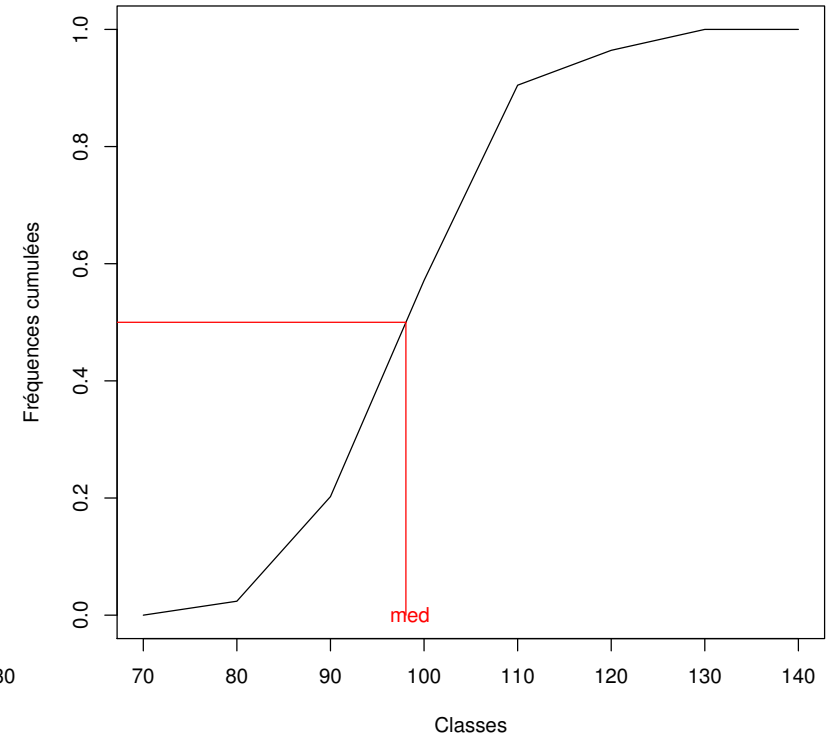
LIÈGE université

Propriétés de la médiane (suite)

3) Interprétation:

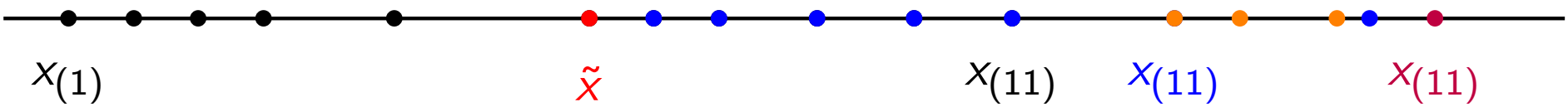
- Dans le cas continu: ☺





Propriétés de la médiane (suite)

4) Résistance aux observations atypiques



La médiane peut résister jusqu'à 50% de contamination dans les données!

Propriétés de la médiane (fin)

5) Propriété d'optimalité:

Proposition

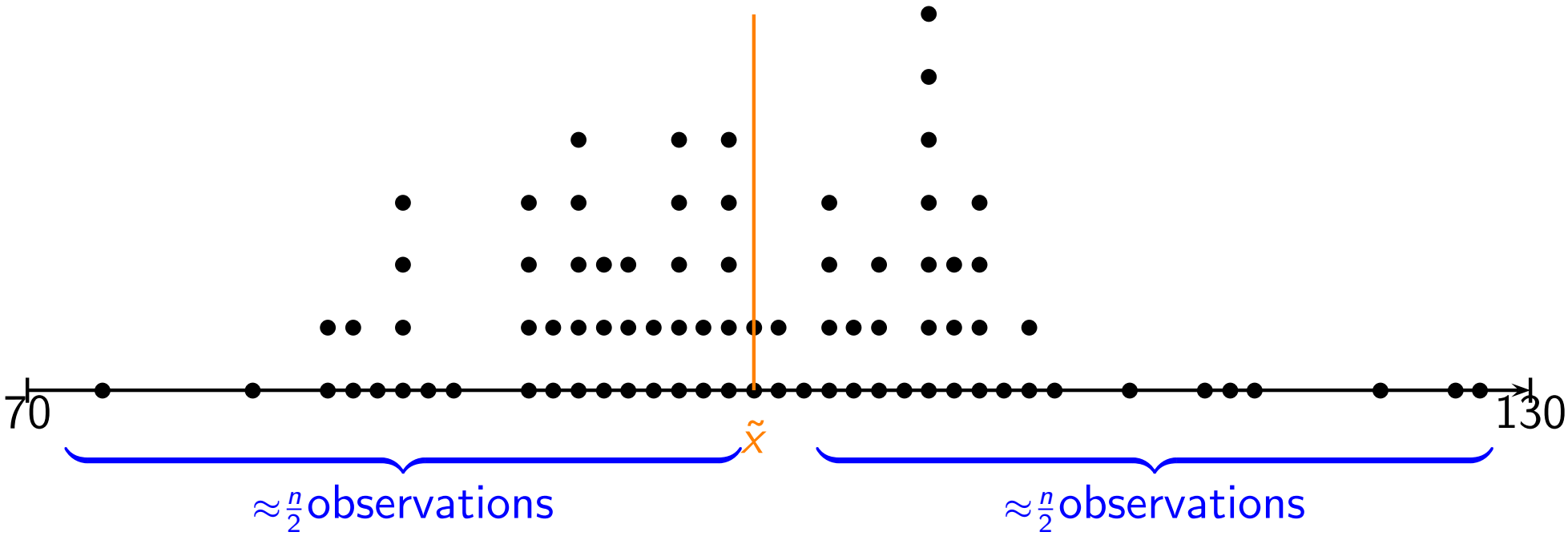
La somme des écarts absolus des éléments d'une série par rapport à la médiane de la série est inférieure ou égale à la somme des écarts absolus par rapport à toute autre valeur $a \in \boldsymbol{R}$.

$$\sum_{i=1}^{n} |x_i - \tilde{x}| \leq \sum_{i=1}^{n} |x_i - a| \ \forall \ a \in \boldsymbol{R}.$$

Résultat admis

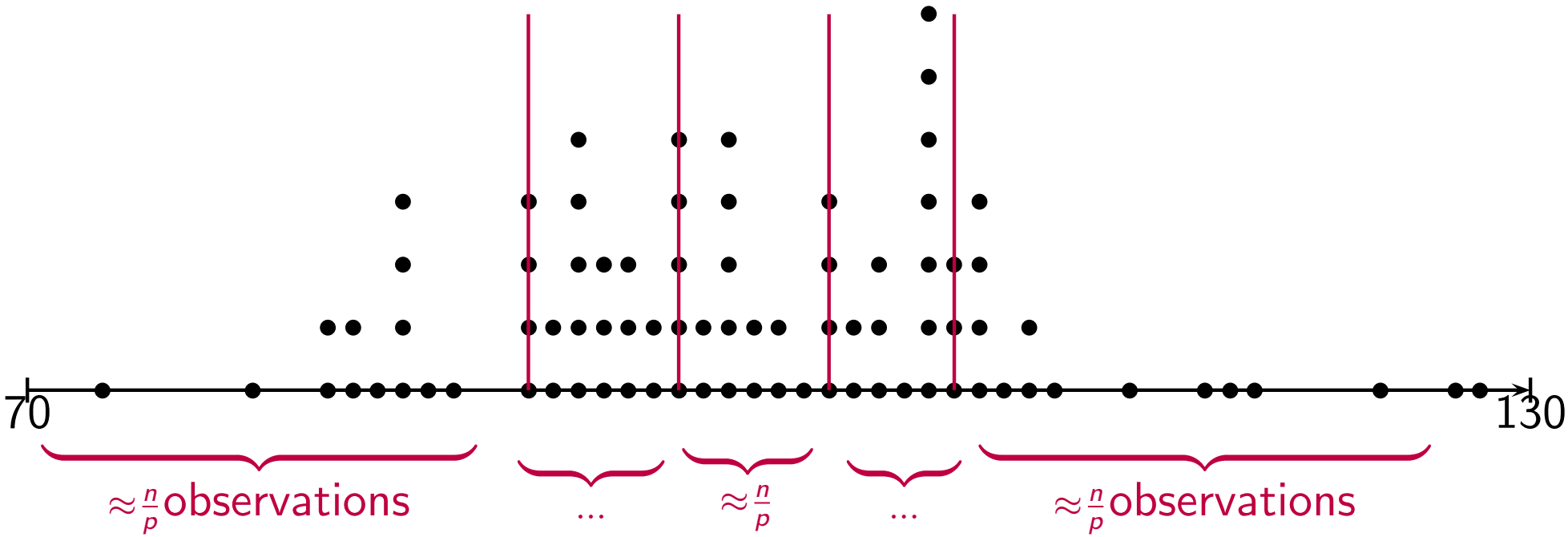
La médiane est donc le paramètre de tendance centrale optimal en ce qui concerne le critère des moindres valeurs absolues (critère L_1).

Généralisation: quantiles d'ordre p, $p \in \mathbb{N}$



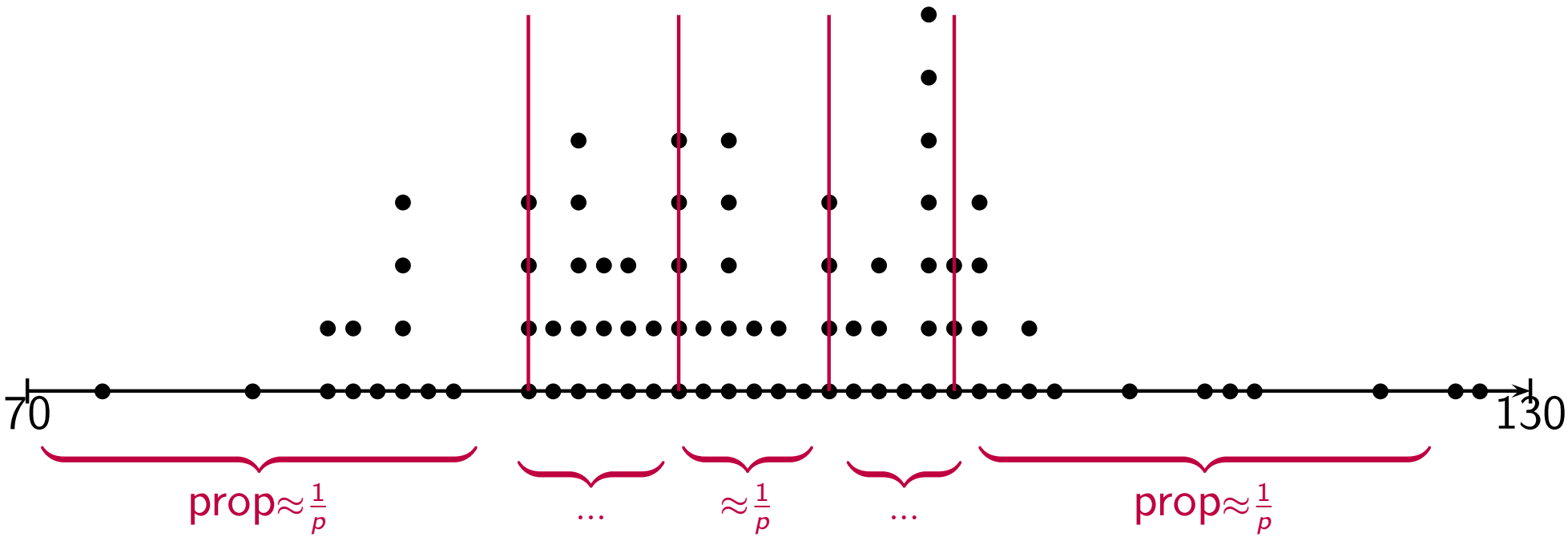
Généralisation: quantiles d'ordre p, $p \in \mathbb{N}$

Les $p-1$ quantiles d'ordre p, avec $p \in \mathbb{N}$, découpent la série en p parties d'approximativement n/p observations

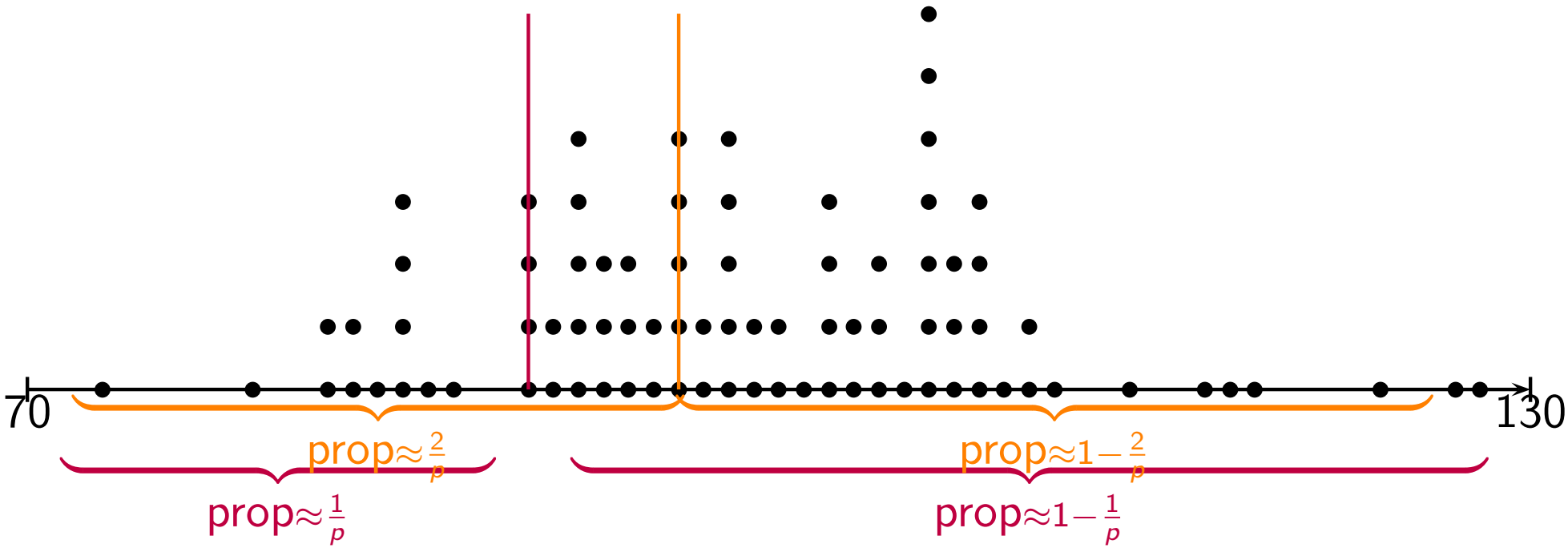


Généralisation: quantiles d'ordre p, $p \in \mathbb{N}$

Les $p-1$ quantiles d'ordre p, avec $p \in \mathbb{N}$, découpent la série en p parties de masse $1/p$



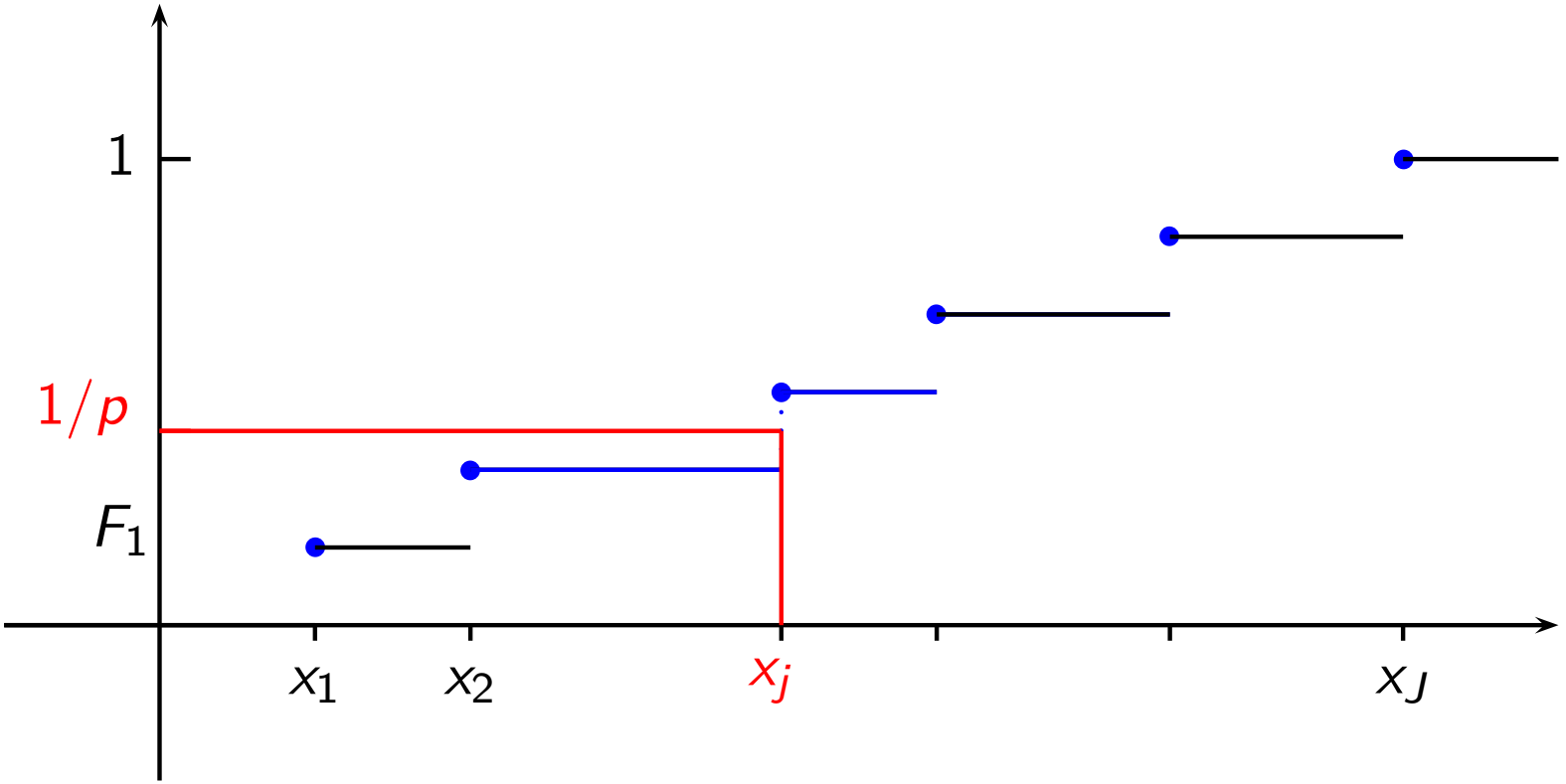
Calcul d'un quantile d'ordre p



$\Rightarrow$ Même démarche que pour le calcul de la médiane en remplaçant la masse de $1/2$ par $1/p$, $2/p$, ...

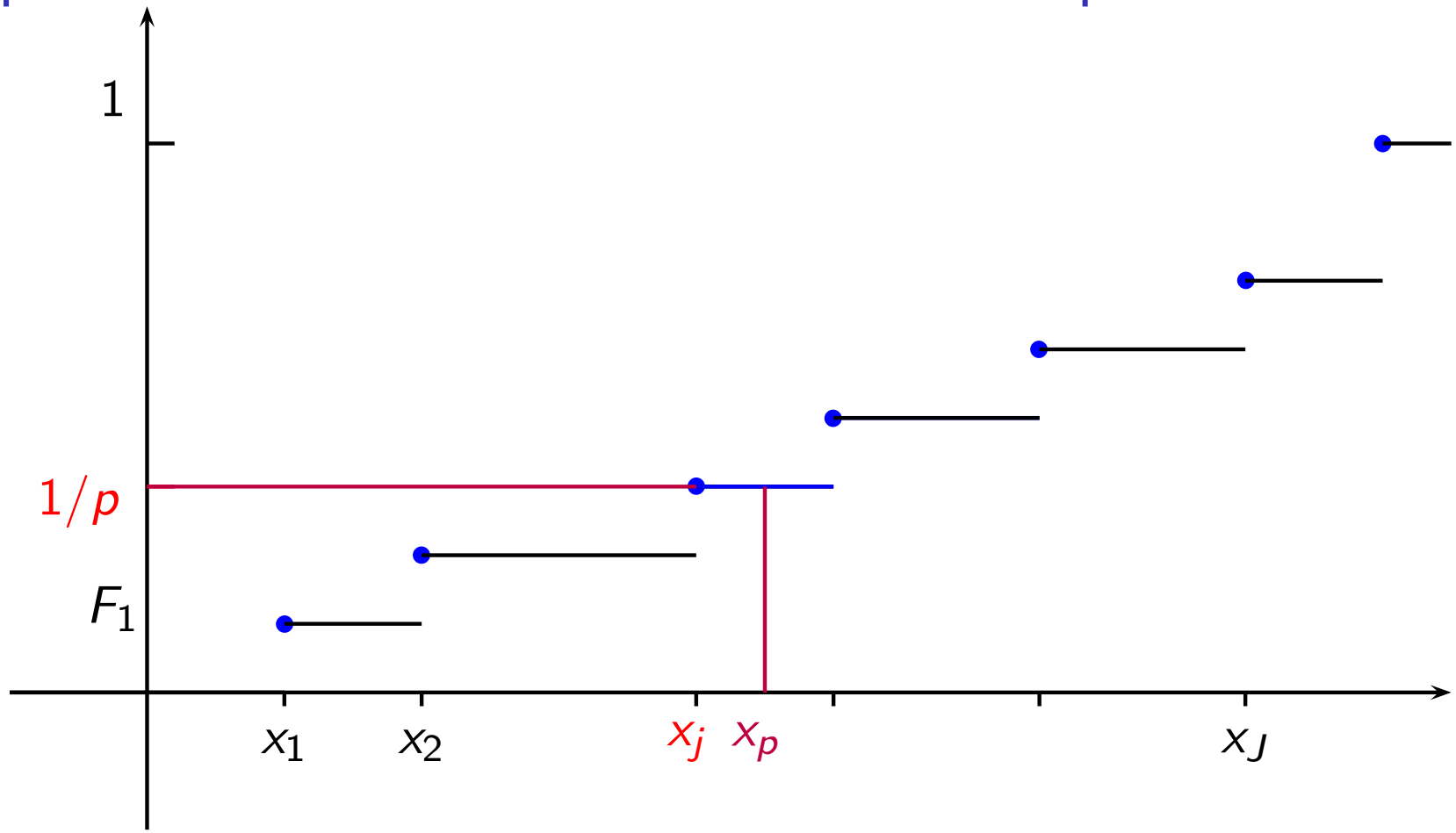
... tout en traitant la série brute de façon recensée (car la position exacte du quantile d'ordre p dépend du reste de la division de n par p!)

Série recensée: calcul du quantile d'ordre p à partir de la courbe cumulative des fréquences cumulées



- Soit il existe j tel quel $F_{j-1} < \frac{1}{p} < F_j \Rightarrow x_p = x_j$

Série recensée: calcul du "premier" quantile d'ordre p à partir de la courbe cumulative des fréquences cumulées

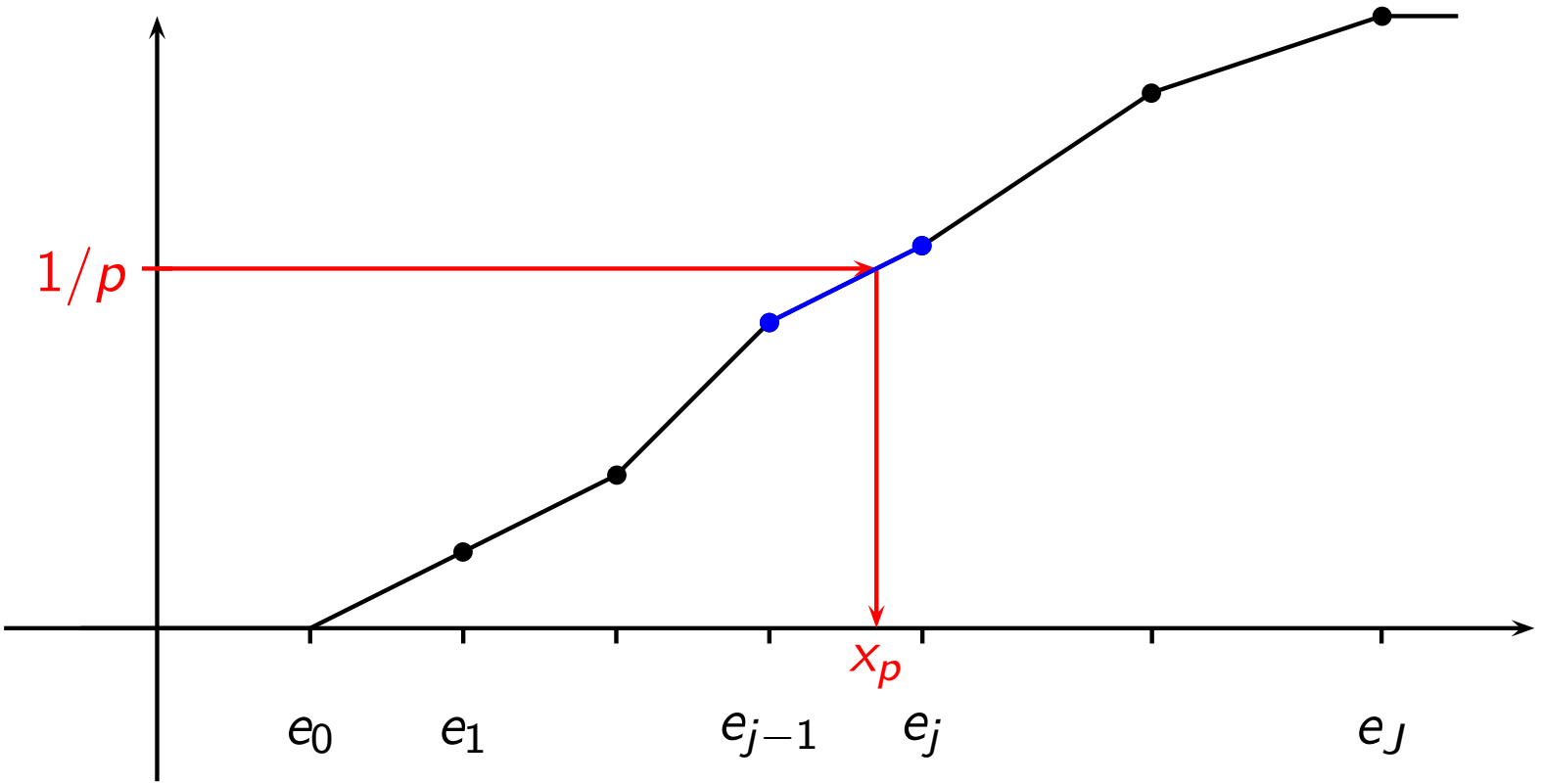


- soit il existe j tel quel $F_j = \frac{1}{p} \Rightarrow x_p = \frac{x_j + x_{j+1}}{2}$

La même démarche peut être suivie avec la distribution des effectifs cumulés!

Série groupée: calcul du "premier" quantile d'ordre p à partir de l'ogive des fréquences cumulées

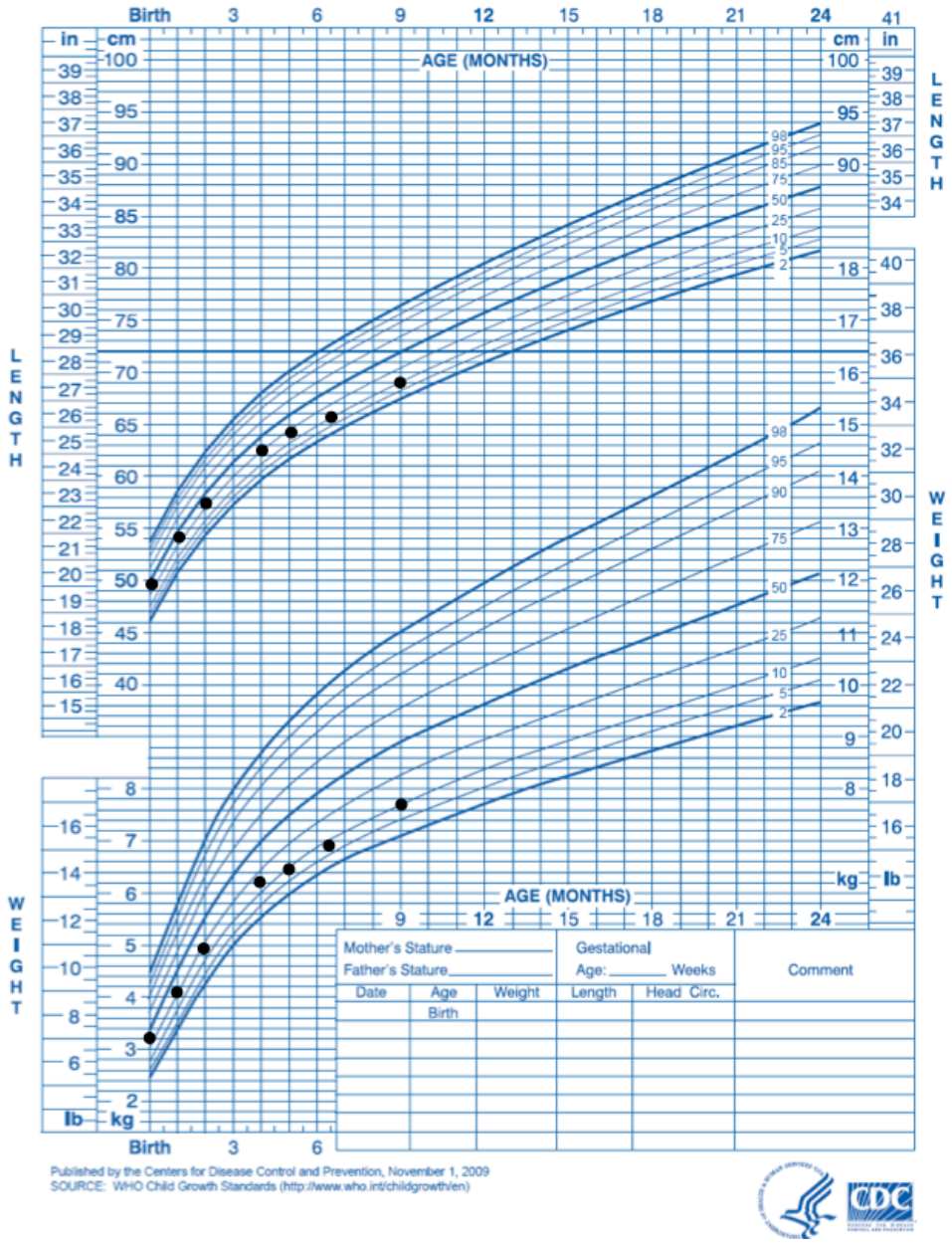
La classe du premier quantile d'ordre p est la première classe dont la fréquence cumulée est supérieure ou égale à $1/p$.



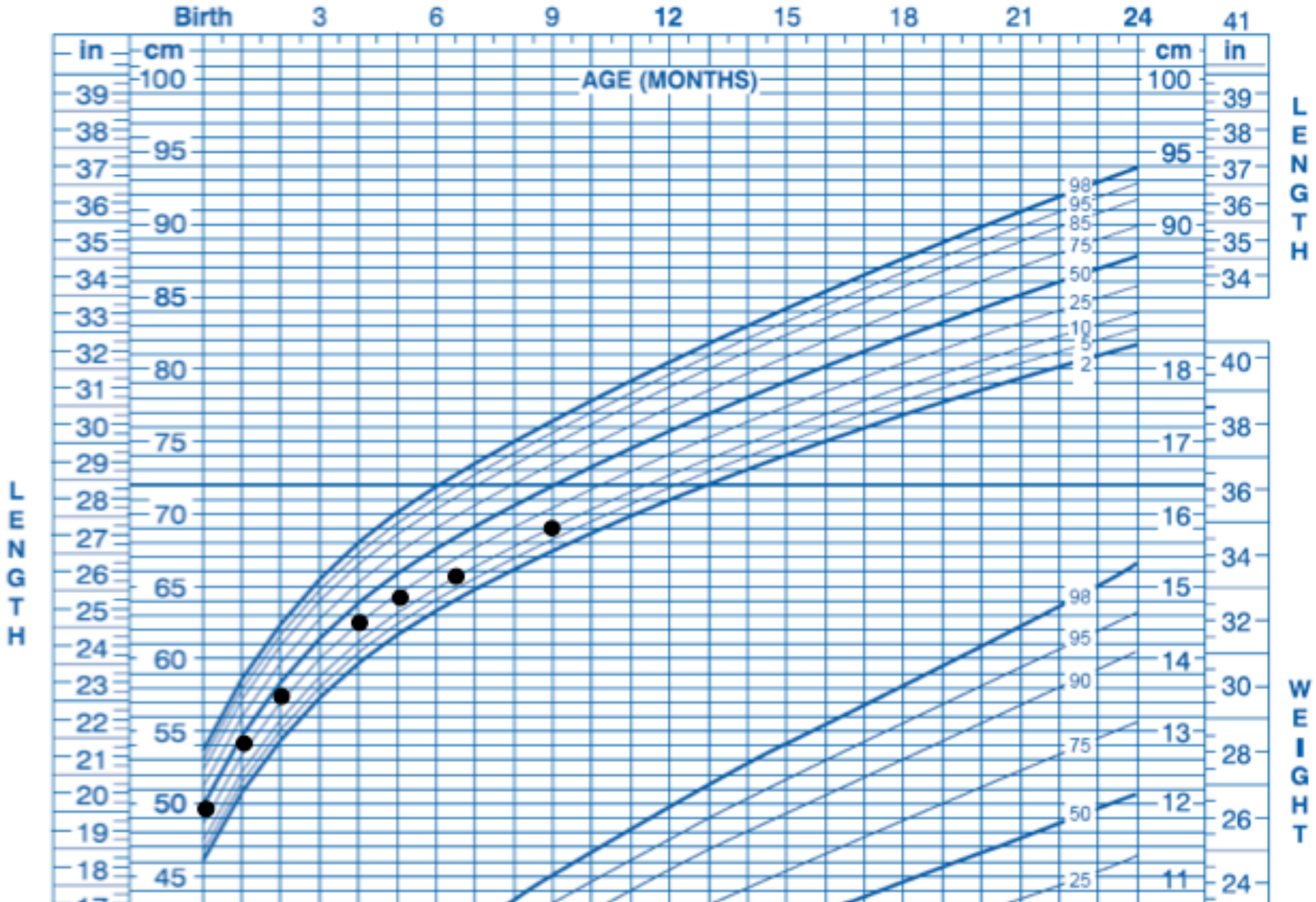
Quantiles classiques

- Les centiles ($p = 100$)
- Les déciles ($p = 10$)
- Les quartiles ($p = 4$)

Centiles des tailles et des poids des bébés (ONE,...)



Centiles des tailles et des poids des bébés (ONE,...); zoom



Les quantiles dans la presse (Le Soir, 8/2/17)

Le Soir Mercredi 8 février 2017

4 LABELGIQUE

Le niveau de vie baisse pour les plus défavorisés

BIEN-ÊTRE Niveau de vie, travail, santé : les plus pauvres sont toujours plus mal lotis

▶ Le rapport du Bureau fédéral du Plan a été présenté mardi.

▶ Il n'est pas dénué de bonnes nouvelles, sauf pour les plus pauvres.

Depuis 2014, le Bureau fédéral du Plan est chargé par le gouvernement de mesurer la qualité de la vie en Belgique. Pour ce faire, il lui a été demandé d'établir une série d'indicateurs complémentaires au PIB qui, s'il s'intéresse au niveau de production d'un pays, reste assez inutile quand il s'agit d'évaluer le bien-être de sa population.

Cette année, le Bureau du Plan a ventilé les résultats de treize de ces indicateurs selon le revenu. Et le tableau ainsi dressé est peu réjouissant.

Si l'écart de bonheur entre les 20 % de Belges les plus pauvres et les 20 % les plus riches (autrement dit entre les quintiles 1 et 5, ainsi qu'on le verra dans l'infographie ci-

disposer d'un poste de télévision (alors qu'elle le souhaite), ne pas avoir de téléphone (alors qu'elle le souhaite), ne pas avoir de machine à

plus roses. L'obésité, si elle concerne 13,7 % de la population, touche environ 17 % du premier quintile et un peu moins de 10 %

ciale augmente de 40 % tandis que le pourcentage de ménages devant reporter ou annuler des soins pour des raisons financières a quadruplé

Population obèse selon la catégorie de revenu

Pourcentage de la population de 18 ans et plus

Chaque **quintile** représente 1/5 de la population globale, ils sont divisés ici, sur base de leurs revenus. Ainsi, le **premier quintile** représente les 20 % de Belges au revenu le plus faible. Le **second quintile** représente la moyenne basse. Le **troisième** groupe dispose d'un revenu médian. Le **quatrième** représente les 20 % de population disposant d'un revenu plus élevé enfin, le **cinquième quintile** représente les 20 % de population les plus aisés

Espérance de vie à 50 ans selon le sexe et le niveau d'éducation

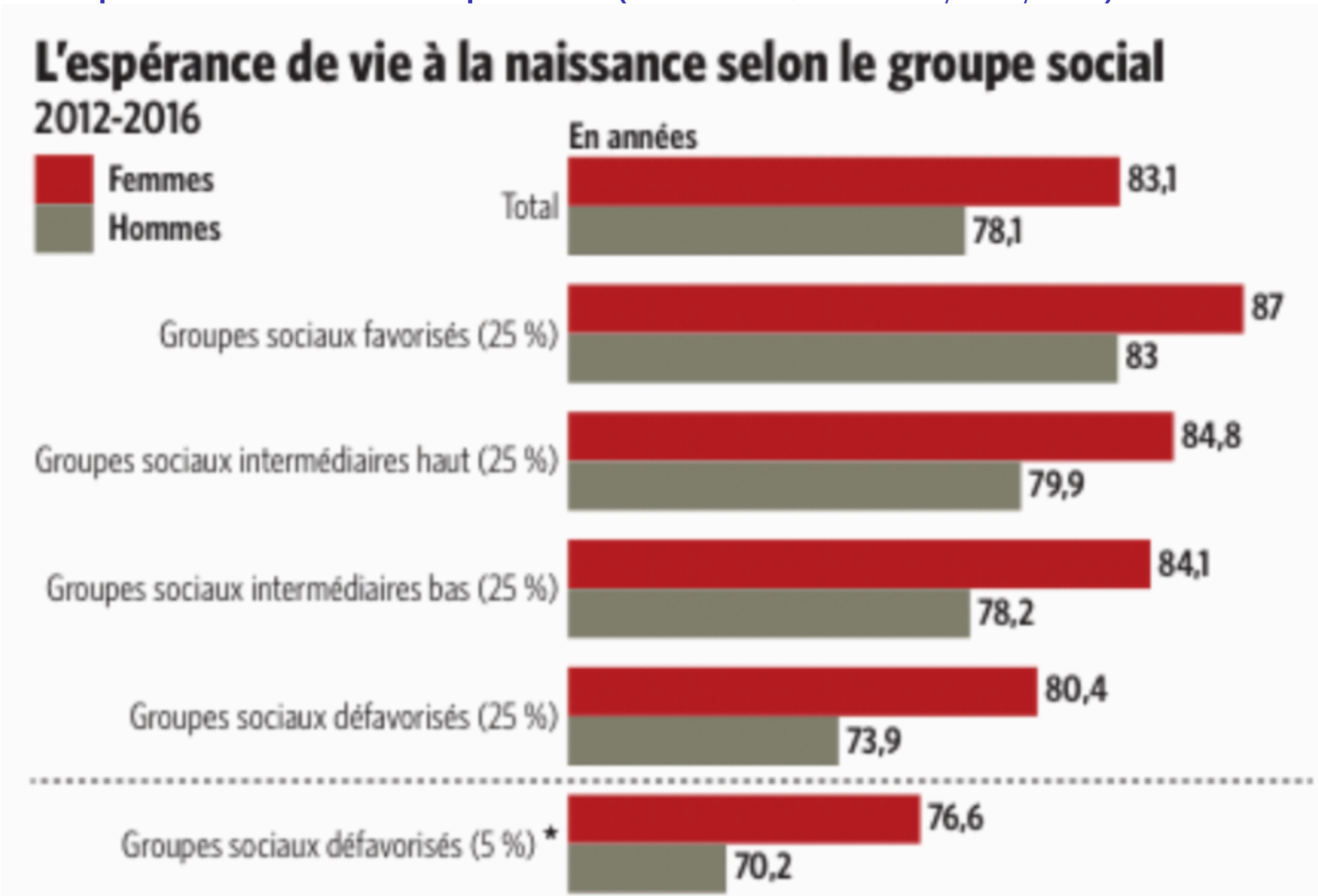
En années de vie après 50 ans — Primaire — Secondaire inférieur — Secondaire supérieur — Supérieur

Ecart salarial entre les hommes et les femmes

Pourcentage de différence des femmes par rapport aux hommes en salaire annuel moyen

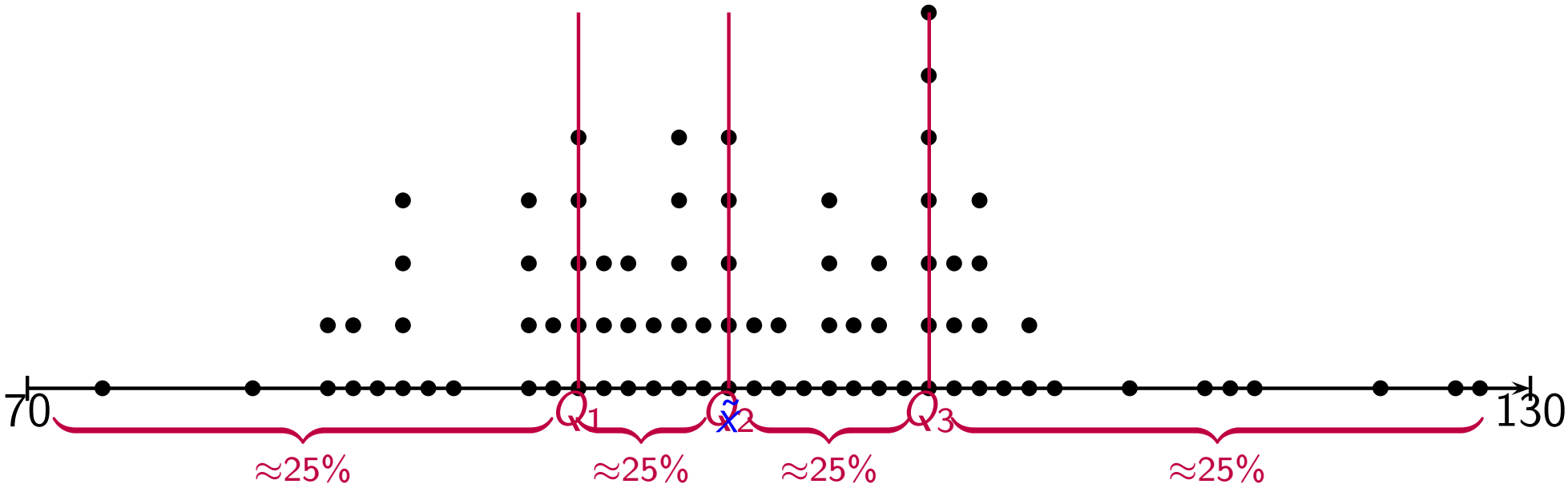
Les inégalités hommes-femmes se réduisent

Les quantiles dans la presse (Le Soir, 24-25/12/18)



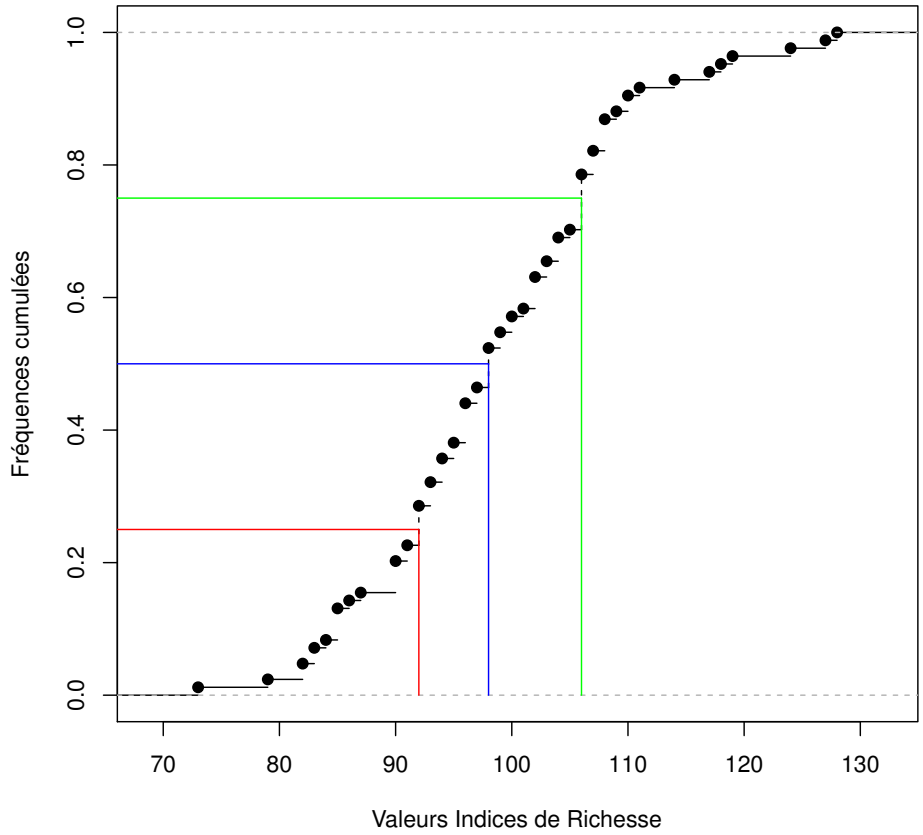
Les quartiles et le résumé à 5 valeurs de Tukey

Les 3 quartiles découpent la série en 4 parties contenant approximativement 25% des observations



Résumé à 5 valeurs: $x_{(1)} \leq Q_1 \leq \tilde{x} \leq Q_3 \leq x_{(n)}$.

Résumé à 5 valeurs des indices de richesse (Liège)



$\tilde{x} = 98$; $Q_1 = 92$; $Q_3 = 106$

Le résumé à cinq valeurs est donc:

$x_{(1)} = 73 \leq 92 \leq 98 \leq 106 \leq 128 = x_{(n)}$

Résumé à cinq valeurs des nombres de partis

Valeurs	Effectifs	Fréquences	Effectifs cumulés	Fréquences cumulées
1	181	0.691	181	0.691
2	65	0.248	246	0.939
3	15	0.057	261	0.996
4	1	0.004	262	1.000
Total	262	1	x	x

$Q_1 = \tilde{x} = 1$ et $Q_3 = 2$

Richesse limitée du résumé à cinq valeurs dans le cas discret...

Résumé à cinq valeurs des revenus de 2002

Classes	f_j en %	F_j en %
[0, 2[	3.11	3.11
[2, 3[	0.96	4.07
[3, 4[	1.03	5.10
[4, 5[	1.08	6.18
[5, 6[	1.20	7.38
[6, 7[	1.55	8.93
[7, 8[	2.02	10.95
[8, 9[	2.12	13.07
[9, 10[	2.93	16.00
[10, 11[	4.14	20.14
[11, 12[	4.10	24.24
[12, 13[	4.16	28.40
[13, 14[	3.73	32.13
[14, 15[	3.91	36.04
[15, 15[	3.72	39.76
[16, 17[	3.52	43.28
[17, 18[	3.46	46.74
[18, 19[	3.40	50.14
[19, 20[	3.32	53.46
[20, 25[	13.05	66.51
[25, 30[	8.28	74.79
[30, 45[	14.20	88.99
[45, 75[	8.56	97.55
[75, $+\infty$[	2.42	100.00

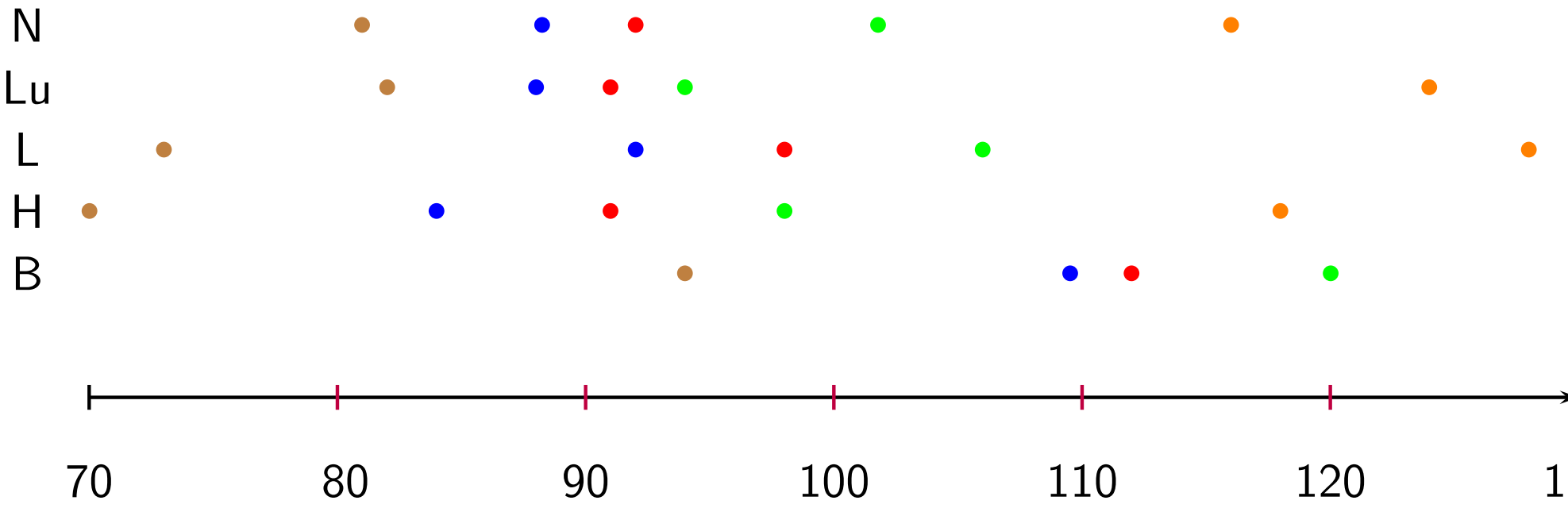
$\tilde{x} = 18.958 \times 1000€$.

$Q_1 = 12 + \frac{1}{0.0416}(0.25 - 0.2424) = 12.182 \times 1000€$

$Q_3 = 30 + \frac{15}{0.1420}(0.75 - 0.7479) = 30.222 \times 1000€$

Malheureusement, les revenus $x_{(1)}$ et $x_{(n)}$ ne sont pas connus...

Utilité: comparaison de distributions



Min: 94, 70, 73, 82, et 81
Q_1: 109.5, 84, 92, 88 et 88.5
$\tilde{x}$: 112, 91, 98, 91 et 92
Q_3: 120, 98, 106, 94 et 102
Max: 138, 118, 128, 124 et 116

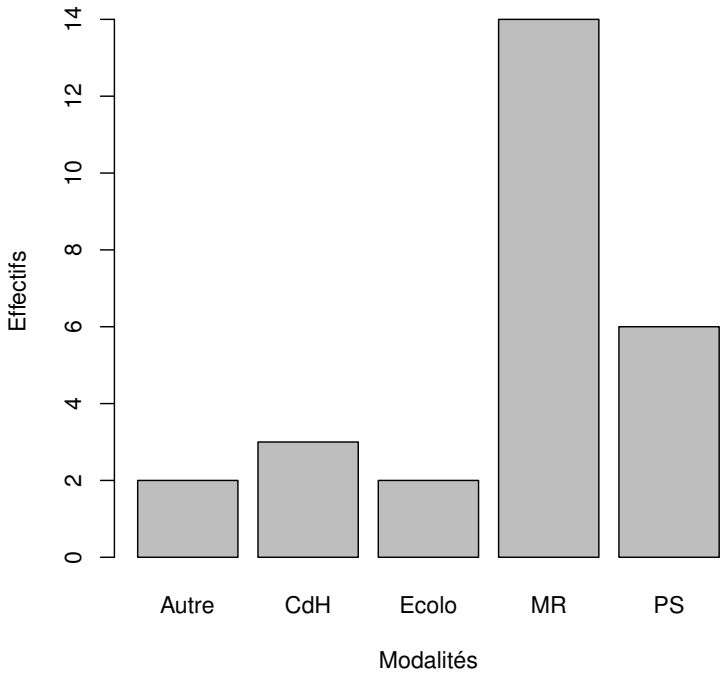
Le Mode

x_M = valeur ou modalité observée le plus souvent (c'est-à-dire dont l'effectif ou la fréquence est le ou la plus grand(e)).

- Paramètre simple, ne demandant pas de calcul,
- Il s'obtient pour tous les types de données (même les données qualitatives),
- Graphiquement, il correspond au plus grand bâton du diagramme en bâtons ou à la plus grande barre du diagramme en barres

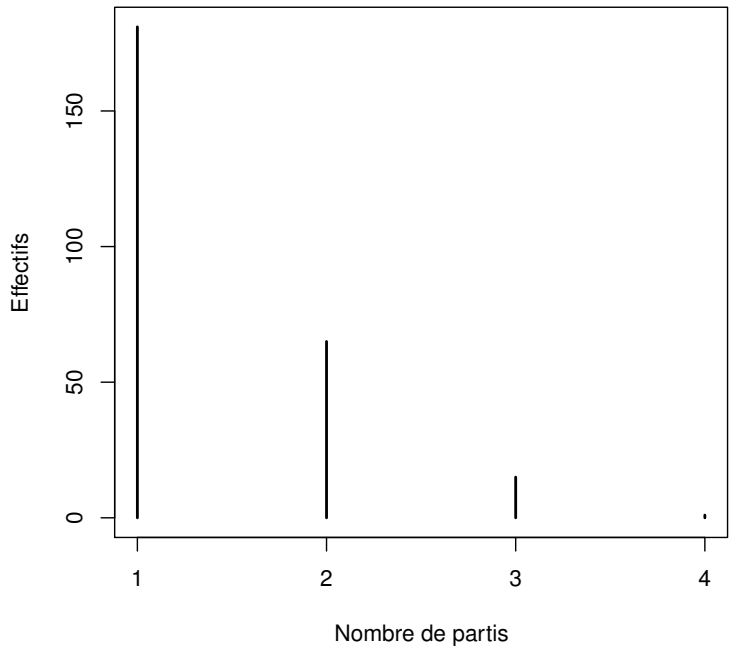
Variable qualitative `PartiBourgmestre` (Brabant W.)

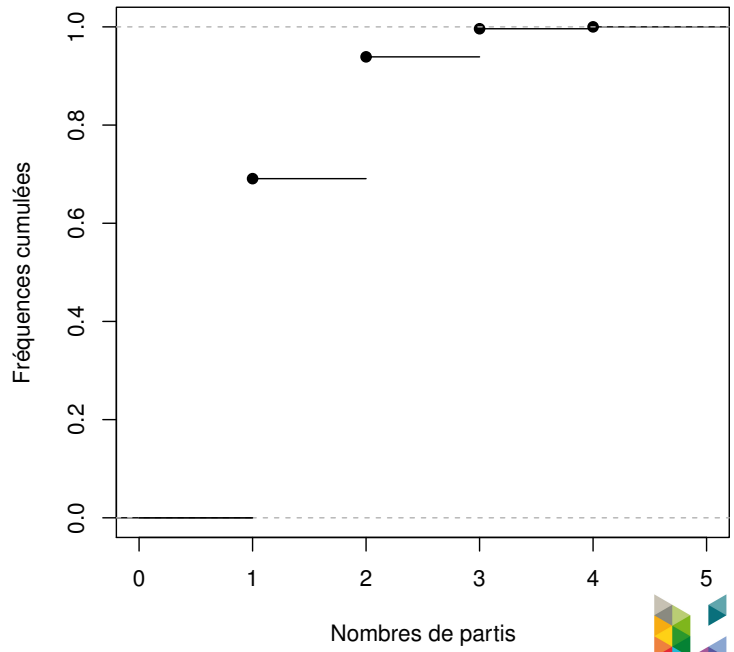
Partis	Effectifs	Fréquences
PS	6	0.222
CdH	3	0.111
Ecolo	2	0.074
MR	14	0.519
Autre	2	0.074
Total	27	1



Variable quantitative discrète `NombrePartis`

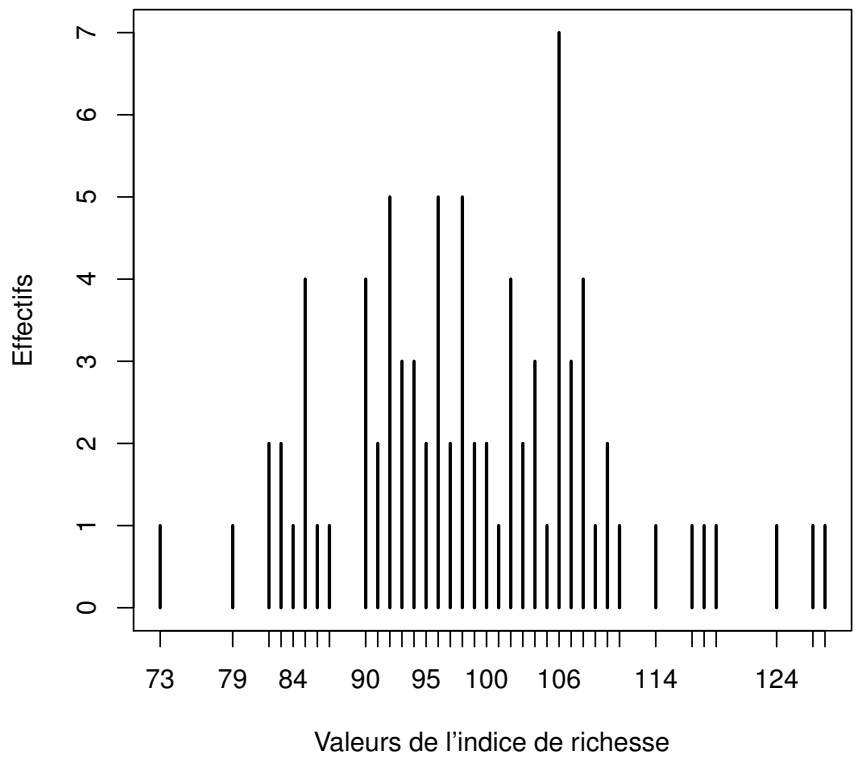
Valeurs	Effectifs	Fréquences	Effectifs cumulés	Fréquences cumulées
1	181	0.691	181	0.691
2	65	0.248	246	0.939
3	15	0.057	261	0.996
4	1	0.004	262	1.000
Total	262	1	x	x





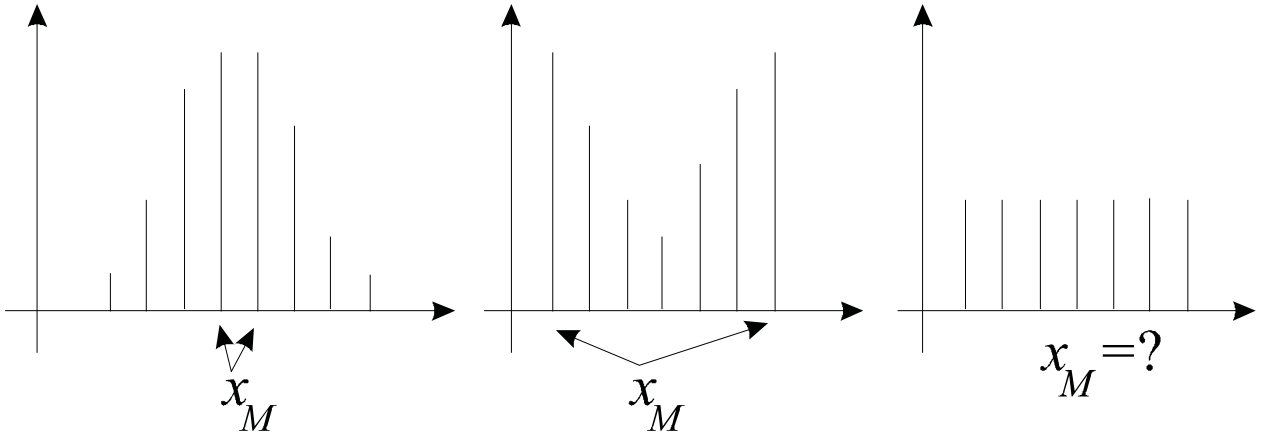
Variable quantitative continue `IndiceRichesse` (Liège)

73	79	82	82	83	83	84	85	85	85	85	86	87	90
90	90	90	91	91	92	92	92	92	92	93	93	93	94
94	94	95	95	96	96	96	96	96	97	97	98	98	98
98	98	99	99	100	100	101	102	102	102	102	103	103	104
104	104	105	106	106	106	106	106	106	106	107	107	107	108
108	108	108	109	110	110	111	114	117	118	119	124	127	128



Quelques difficultés tout de même

- Ce paramètre n'est pas nécessairement unique.



  Dans certains cas, on considère même qu'il n'est pas défini!

- Le mode n'est pas nécessairement au *centre* de la distribution.

Difficultés: suite

- Son interprétation peut être perturbée lorsque plusieurs valeurs ou modalités sont observées un grand nombre de fois.

  On introduit alors la notion de *modes relatifs:*

$$x_j \text{ est un mode relatif si } n_j > n_{j-1} \text{ et } n_j > n_{j+1}$$

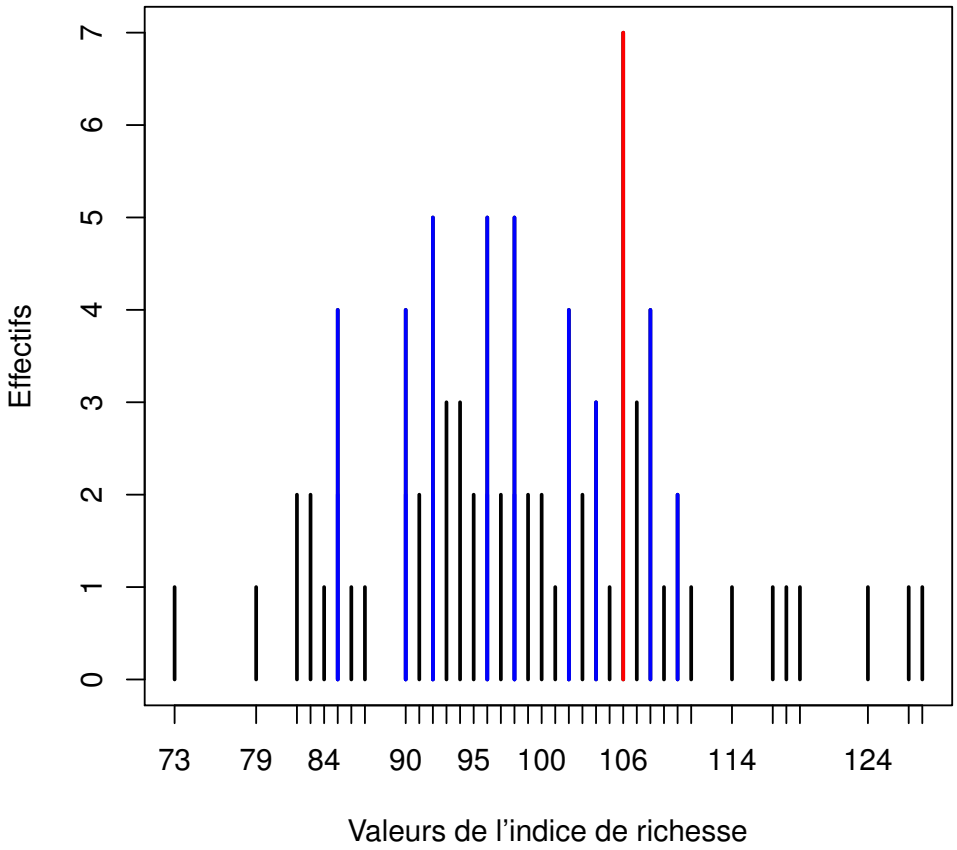
  et on distingue:

  - Les séries *unimodales* qui ne possèdent qu'un seul mode relatif.
  - Les séries *plurimodales* qui possèdent plusieurs modes relatifs.

  Le mode relatif dont l'effectif est le plus élevé est qualifié de mode absolu.

LIÈGE université

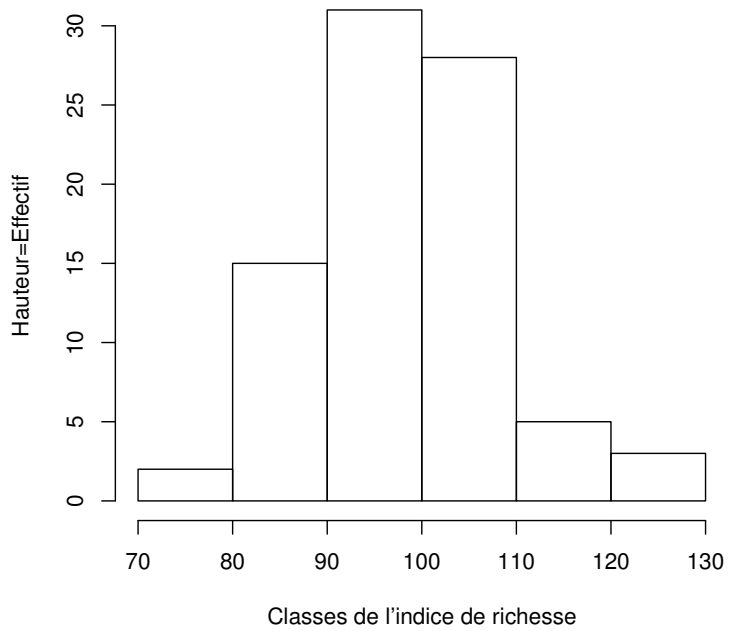
Modes relatifs ou absolus



- Lorsque la variable est quantitative continue, il y a habituellement beaucoup de modes relatifs.

Mode d'une série groupée

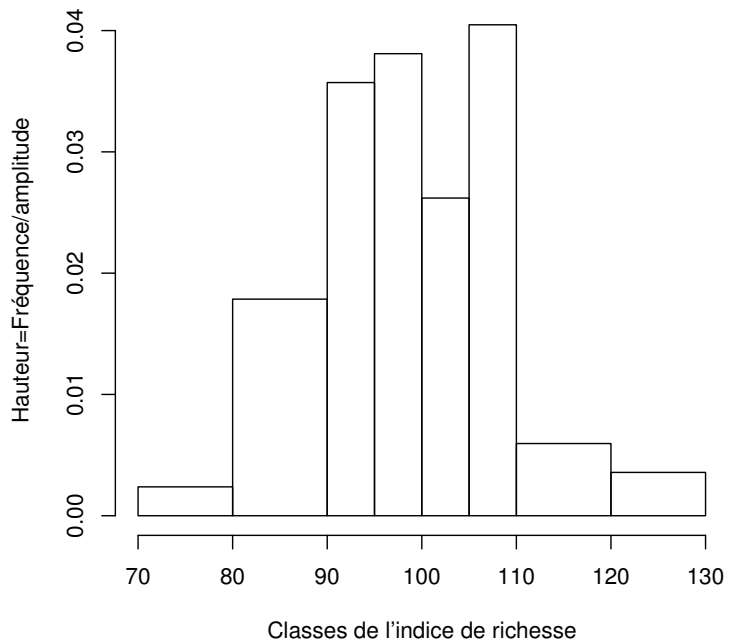
- Si toutes les classes ont la même amplitude, la classe modale est celle dont l'effectif associé est le plus élevé.



Classe modale: $]90, 100] \not\ni$ le vrai mode

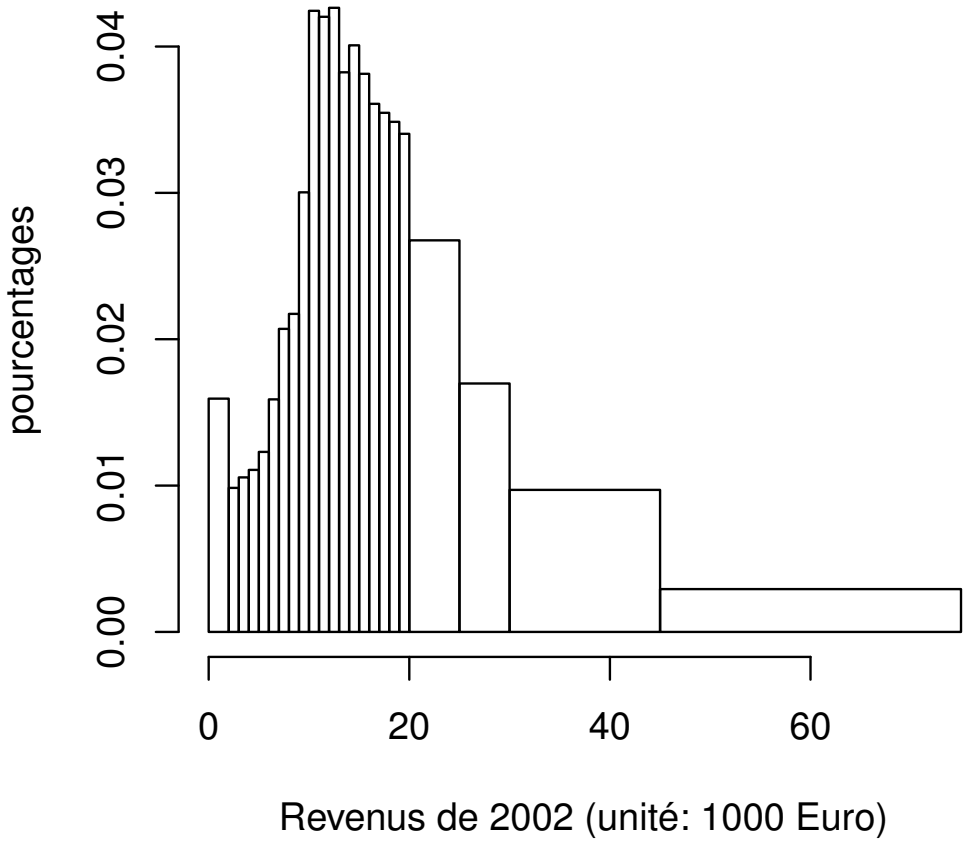
Mode d'une série groupée

- Si les classes ont des amplitudes différentes, on effectue la comparaison des effectifs ajustés par rapport aux amplitudes de classes.



Classe modale: $]105, 110] \ni$ le vrai mode

Distributions des revenus des contribuables belges en 2002

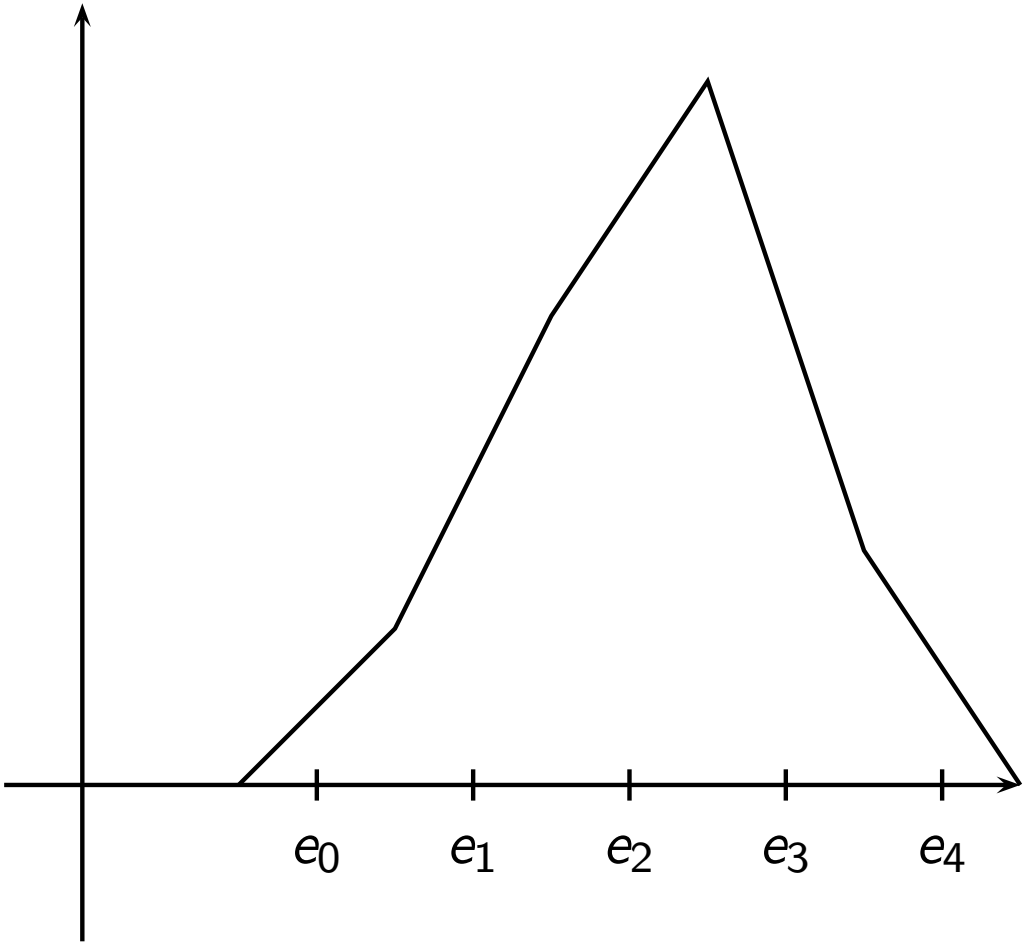


Classe modale: $[12, 13[$

Valeur approchée du mode

La classe modale dépend de la répartition en classe effectuée et, de plus, il s'agit d'un intervalle de valeurs.

$\Rightarrow$ Approximation de Yule et Kendall, valable pour des séries dont le polygone des effectifs est en forme de cloche (plus ou moins symétrique):



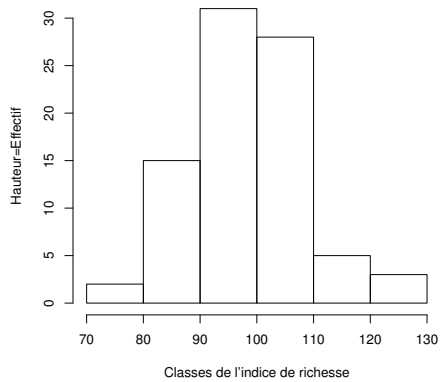
Valeur approchée du mode

Sous cette hypothèse, la relation suivante se vérifie empiriquement entre $\bar{x}, \tilde{x}$ et le mode:

$$\bar{x} - x_M \approx 3(\bar{x} - \tilde{x}).$$

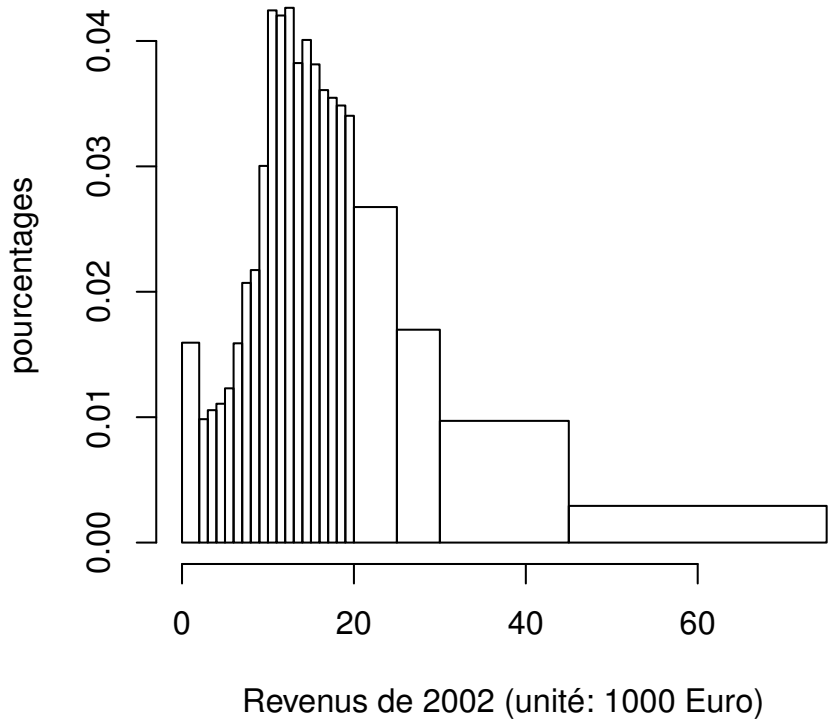
D'où $x_M \approx 3\tilde{x} - 2\bar{x}$.

Ex: Indice de richesse à Liège



Moyenne et médiane estimées: $\bar{x} = 98.33$, $\tilde{x} = 98.06$
D'où $x_M \approx 97.52 \in$ à la classe modale!

Distributions des revenus des contribuables belges en 2002



Moyenne et médiane estimées: $\bar{x} = 24.691,76$ Eur, $\tilde{x} = 18.958,82$ Eur

D'où $x_M \approx 7.492,94$ Eur

Dans ce cas-ci, le mode estimé ne se trouve pas dans la classe modale...

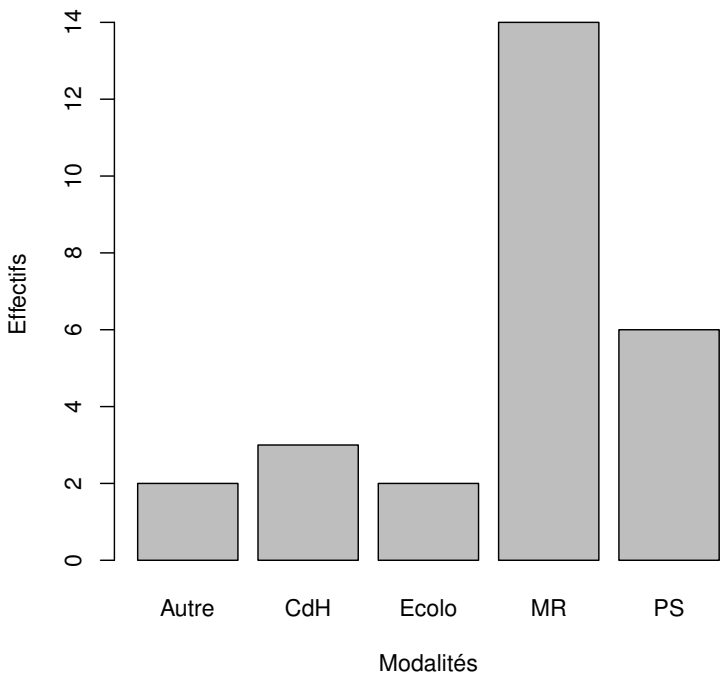
Quel paramètre de tendance centrale choisir ?

Dans certains cas, l'utilisation de l'un ou l'autre paramètre est suggérée par le contexte de l'étude statistique.

LIÈGE université

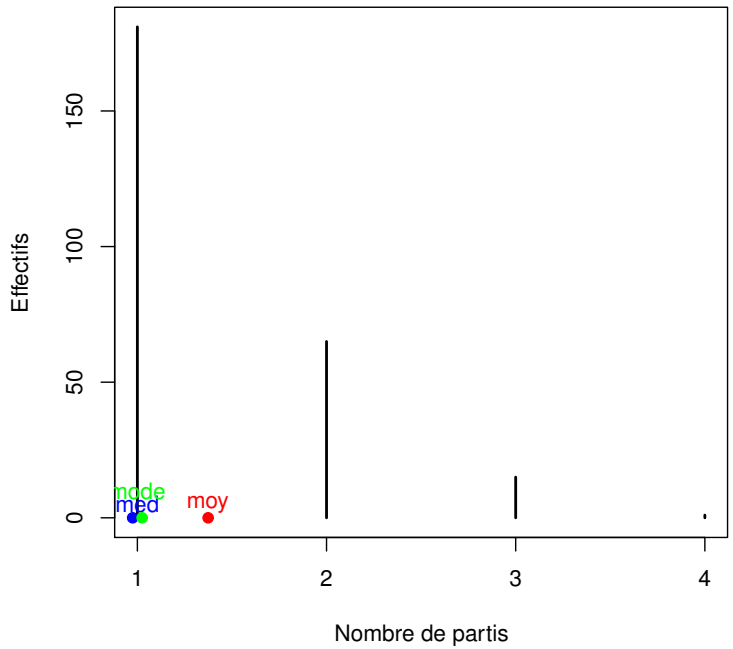
Variable qualitative `PartiBourgmestre` (Brabant W.)

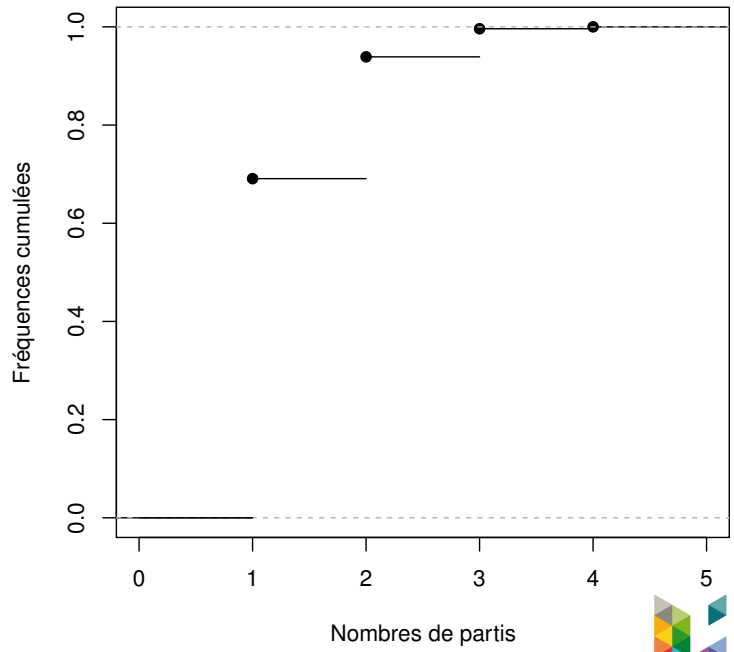
Partis	Effectifs	Fréquences
PS	6	0.222
CdH	3	0.111
Ecolo	2	0.074
MR	14	0.519
Autre	2	0.074
Total	27	1



Variable quantitative discrète `NombrePartis`

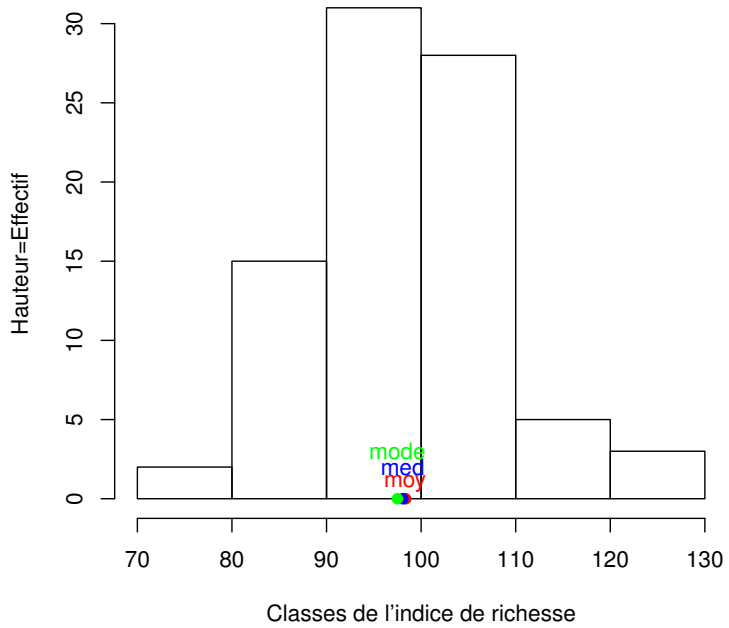
Valeurs	Effectifs	Fréquences	Effectifs cumulés	Fréquences cumulées
1	181	0.691	181	0.691
2	65	0.248	246	0.939
3	15	0.057	261	0.996
4	1	0.004	262	1.000
Total	262	1	x	x

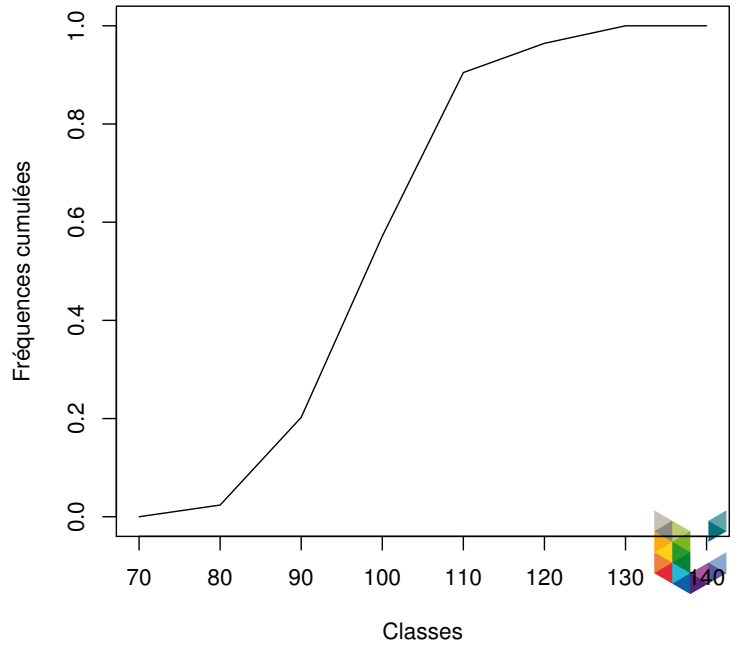




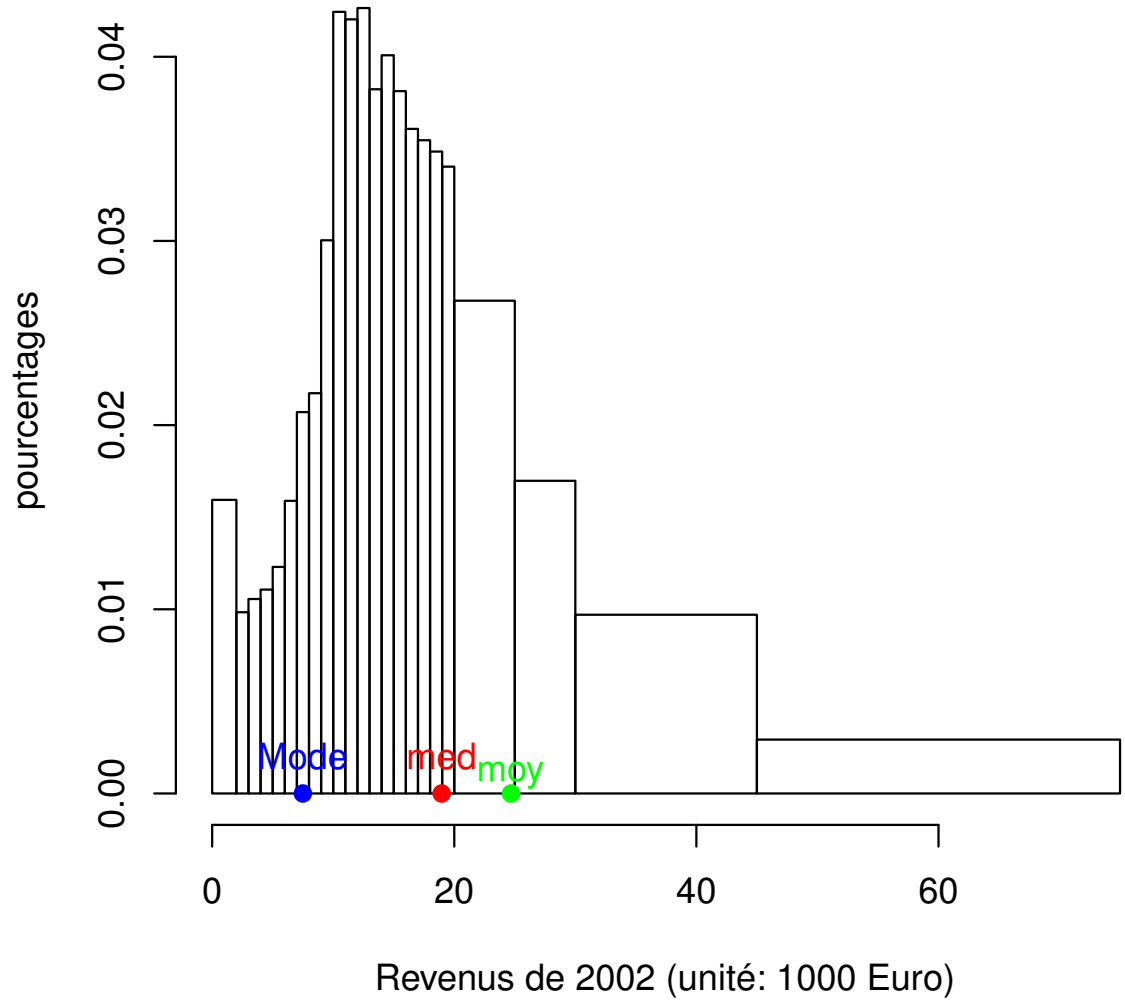
IndiceRichesse (Liège): classes d'amplitude 10

Classes	Effectifs	Fréquences	Eff. cum.	Fréq. cum.
$[70, 80]$	2	0.02	2	0.02
$]80, 90]$	15	0.18	17	0.20
$]90, 100]$	31	0.37	48	0.57
$]100, 110]$	28	0.33	76	0.90
$]110, 120]$	5	0.06	81	0.96
$]120, 130]$	3	0.04	84	1.00
Total	84	1	x	x



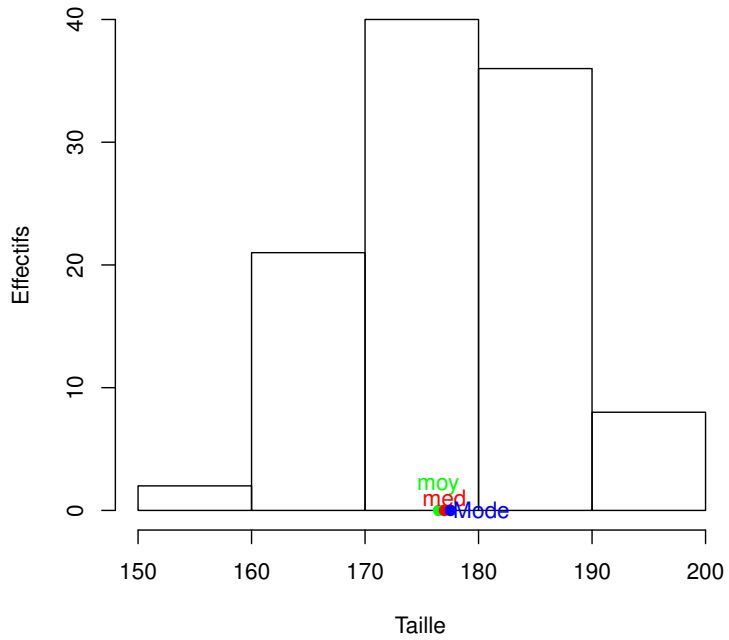


Distributions des revenus des contribuables belges en 2002



Quelques généralités

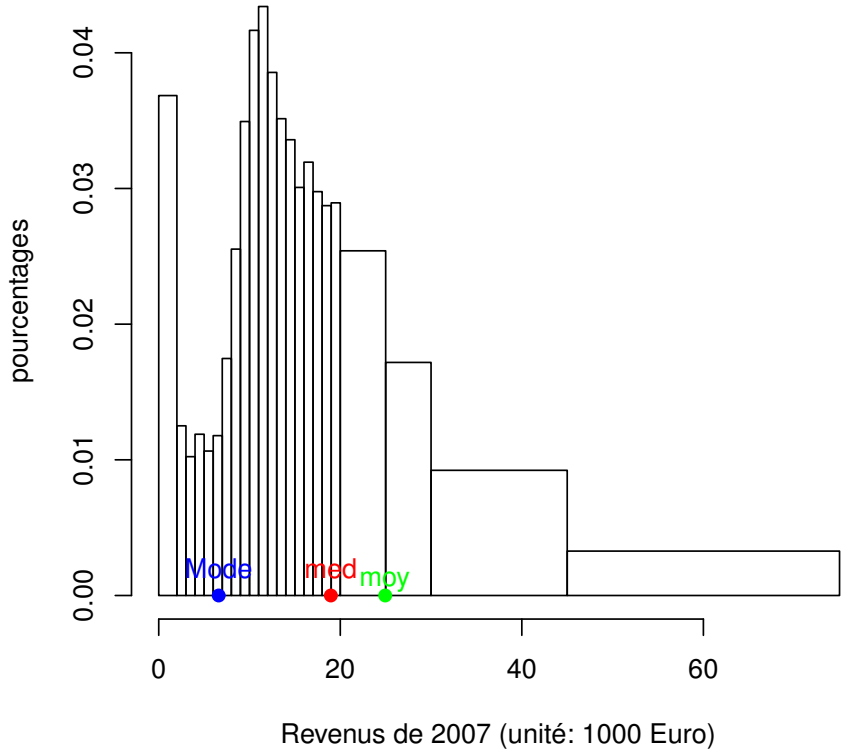
Sous l'hypothèse de symétrie,



$$\bar{x} \approx \tilde{x} \approx x_M$$

Quelques généralités

En cas d'étalement (à droite par exemple)



$$x_M < \tilde{x} < \bar{x}$$

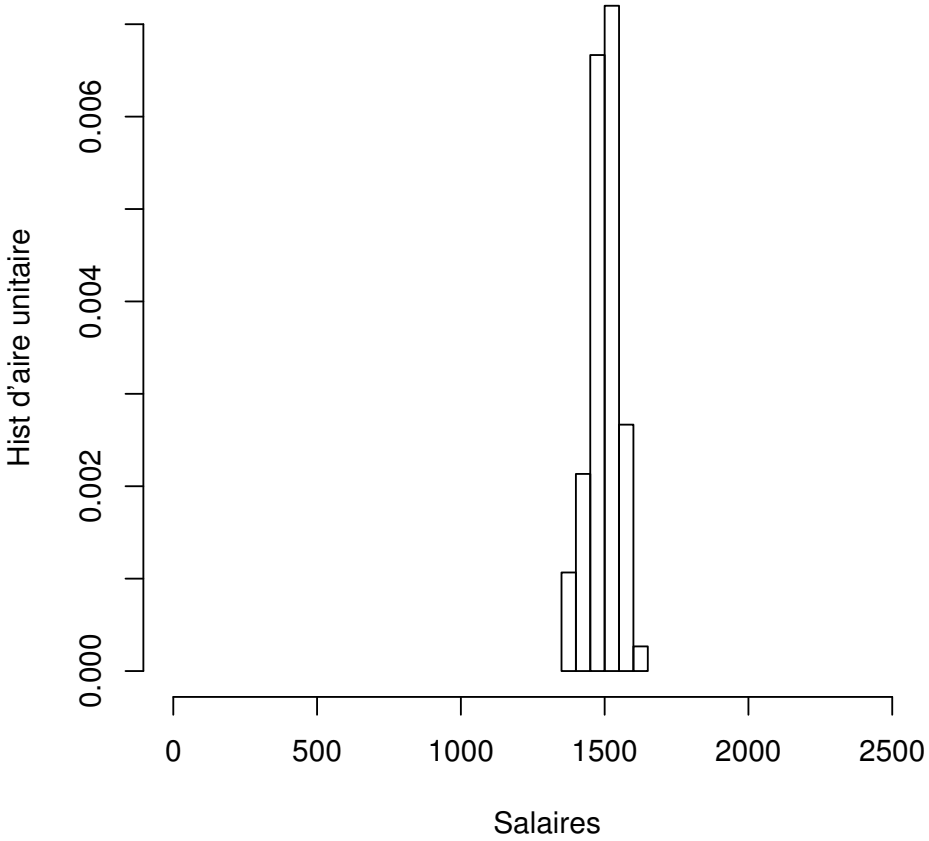
Paramètres de dispersion

Soit $S = \{x_1, x_2, \ldots, x_n\}$ la série brute,
ou $S = \{(x_j, n_j) : j = 1, \ldots, J\}$ la série recensée
ou $S = \{(c_j, n_j) : j = 1, \ldots, J\}$ la série groupée

Connaître le centre de la série n'est pas suffisant, il faut aussi caractériser la dispersion des observations (dispersion globale ou dispersion autour de ce centre).

LIÈGE université

Illustration à l'aide de deux histogrammes



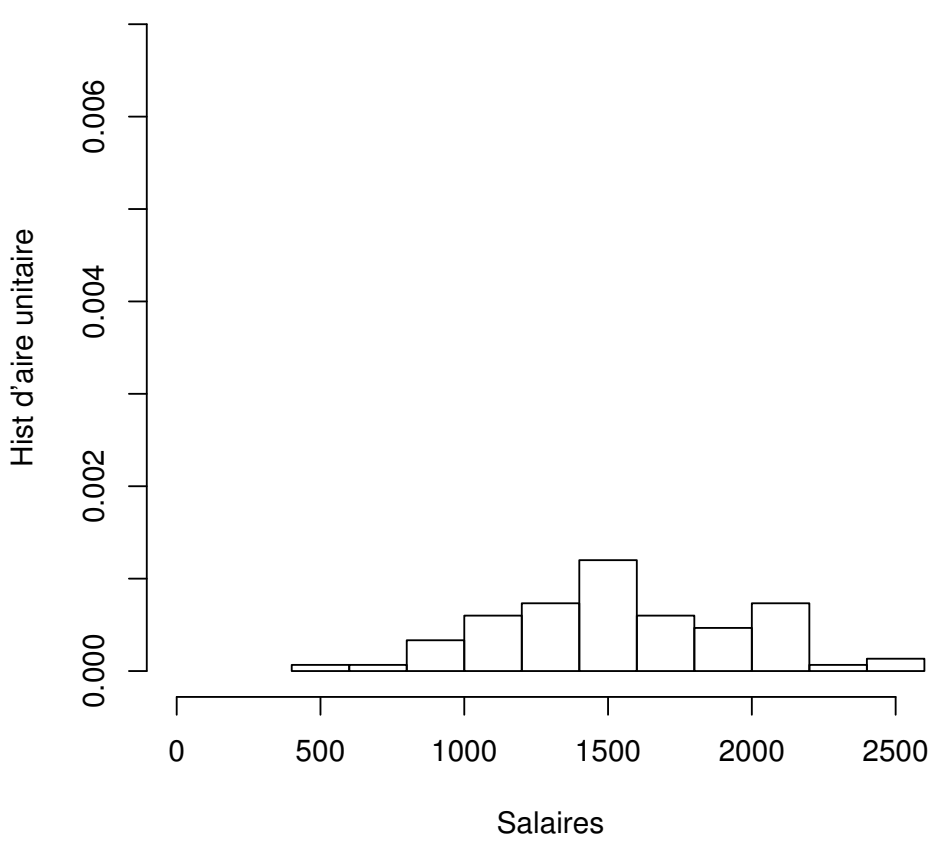
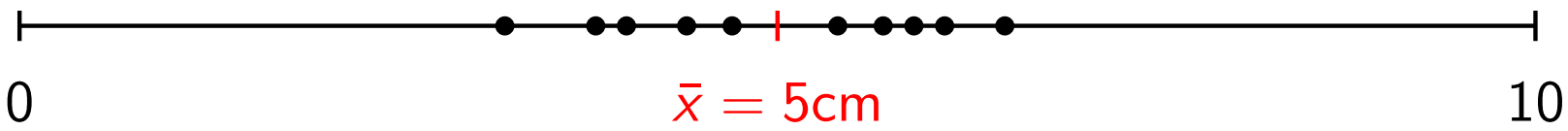
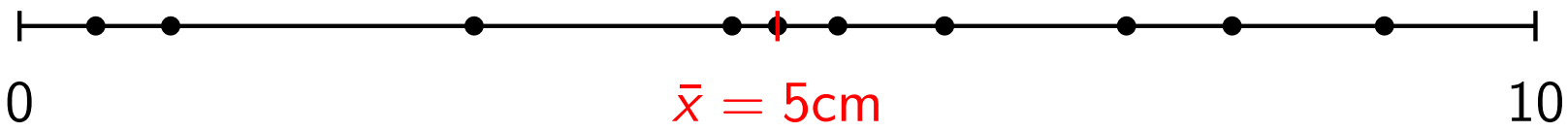


Illustration à l'aide de deux séries de tailles (mesurées en cm)





Paramètres de dispersion

1. Etendue
2. Ecart interquartile
3. Variance et écart-type
4. Coefficient de variation

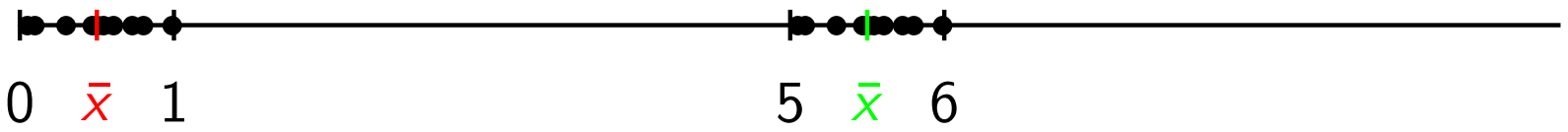
Comme exemple, les données fictives exprimées en cm et les séries d'indices de richesse dans les différentes provinces vont être considérées.

Mais avant de donner les définitions, interrogeons-nous sur le comportement adéquat d'un paramètre de dispersion lorsqu'une transformation affine est appliquée aux données.

LIÈGE université

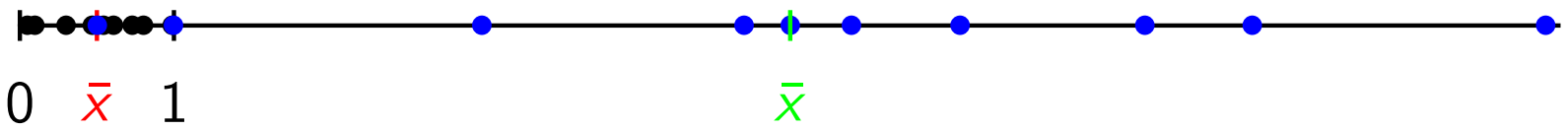
Quel est l'effet "naturel" d'un changement d'origine et/ou d'échelle sur un paramètre de dispersion?

Soient les 10 observations suivantes ; translatons-les de 5 unités



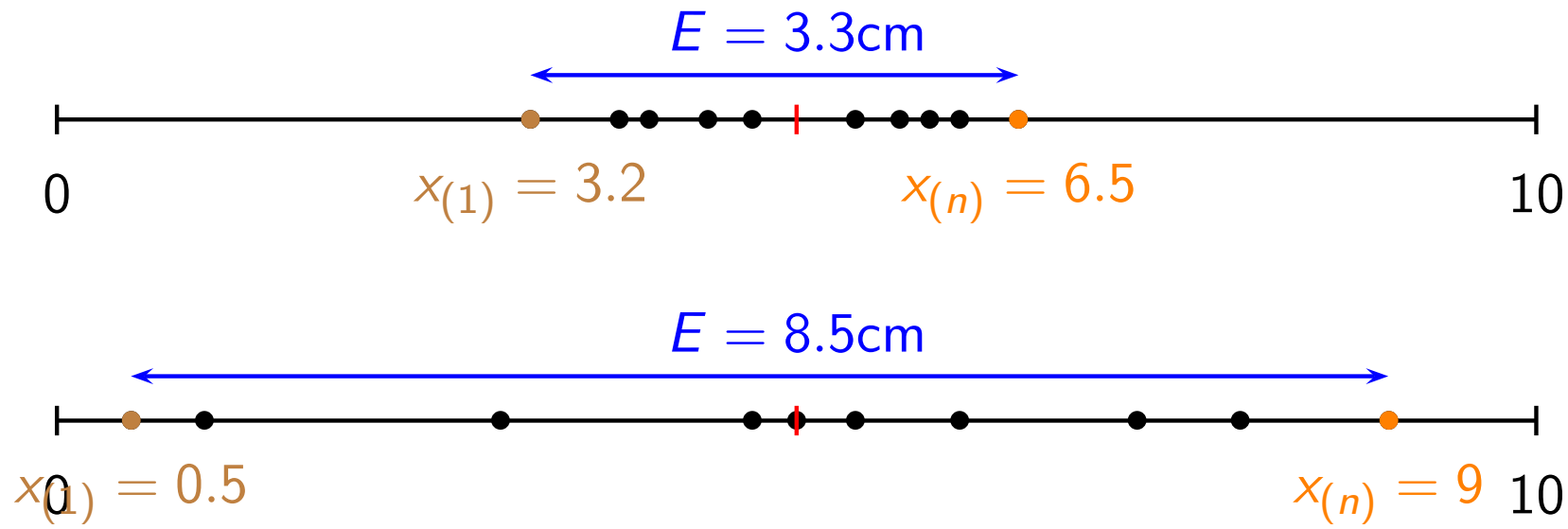
Dans ce cas, on sait que la moyenne est translatée de la même manière mais, visiblement, la dispersion reste inchangée!
Un paramètre de dispersion adéquat devrait donc être insensible à toute translation mais devrait s'adapter au changement d'unité!

Reprenons les 10 observations de départ, et multiplions-les par 10



La moyenne suit le mouvement (elle est mutlipliée par 10) et, visiblement, la dispersion a également été modifiée...

L'étendue $E = x_{(n)} - x_{(1)}$



Indices de richesse

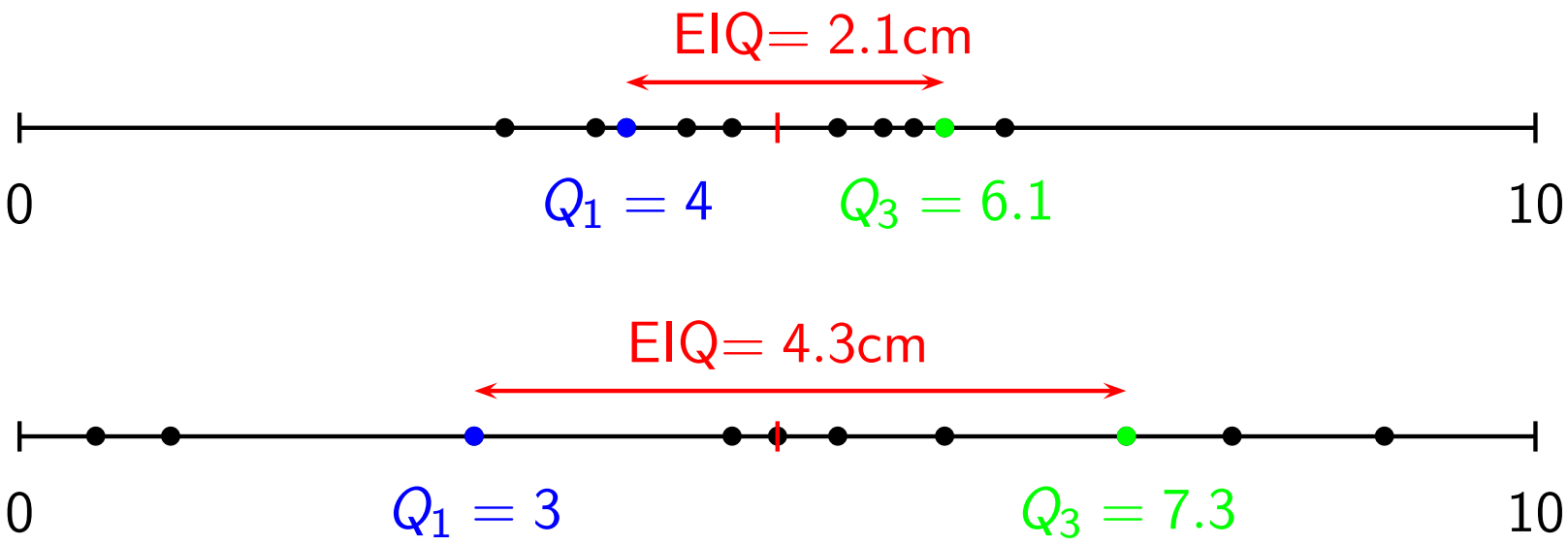
Provinces	Min	Max	Etendue
Brabant wallon	94	138	44
Hainaut	70	118	48
Liège	73	128	55
Luxembourg	82	124	42
Namur	81	116	35

Trois remarques:

- E se transforme naturellement si les données sont translatées/changées d'unité.
- E est très sensible à la présence de valeurs extrêmes.
- E pour des données groupées?

L'écart interquartile et les boîtes à moustaches

$$EIQ = Q_3 - Q_1.$$



Indices de richesse

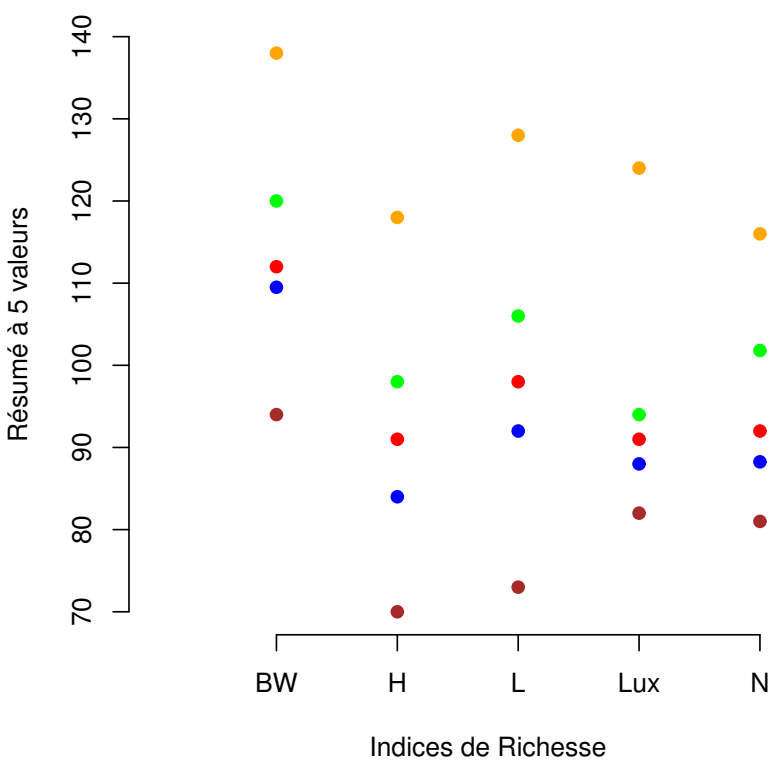
Provinces	Etendue	Q_1	Q_3	EIQ
Brabant wallon	44	109.5	120	10.5
Hainaut	48	84	98	14
Liège	55	92	106	14
Luxembourg	42	88	94	6
Namur	35	88.5	102	13.5

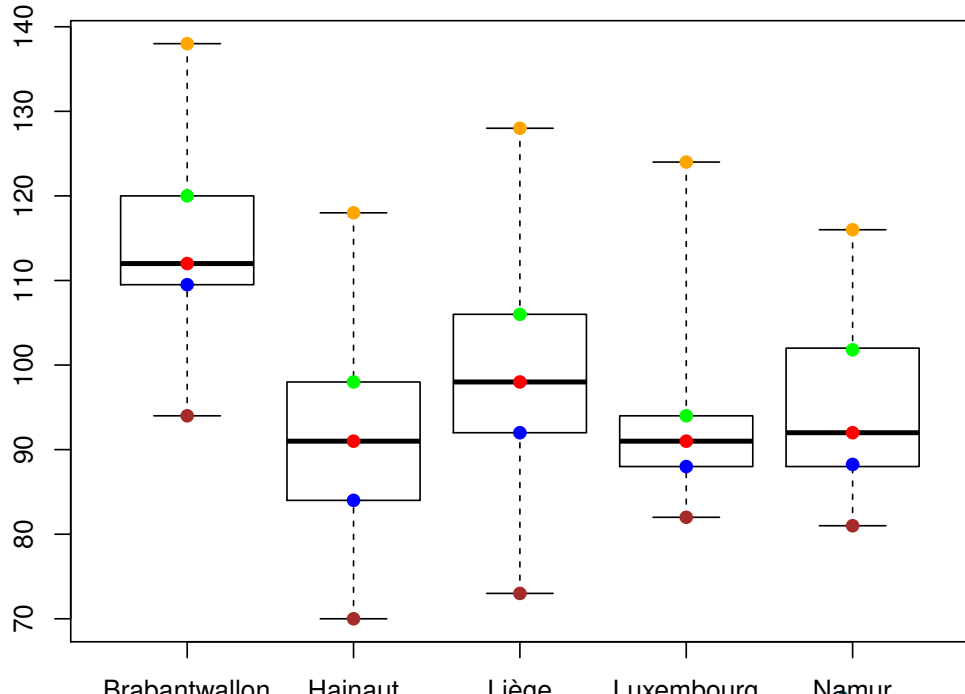
Trois remarques:

- Se transforme naturellement si les données sont translatées/changées d'unité.
- Insensible à la présence de valeurs extrêmes.
- Peut être approximé pour des données groupées.

Les boîtes à moustaches (ou diagramme en boîte, en anglais, *box-and-whisker plot*, Tukey, 1977)

Représentation graphique mettant en valeur le résumé à 5 valeurs:
$x_{(1)}, Q_1, \tilde{x}, Q_3, x_{(n)}$



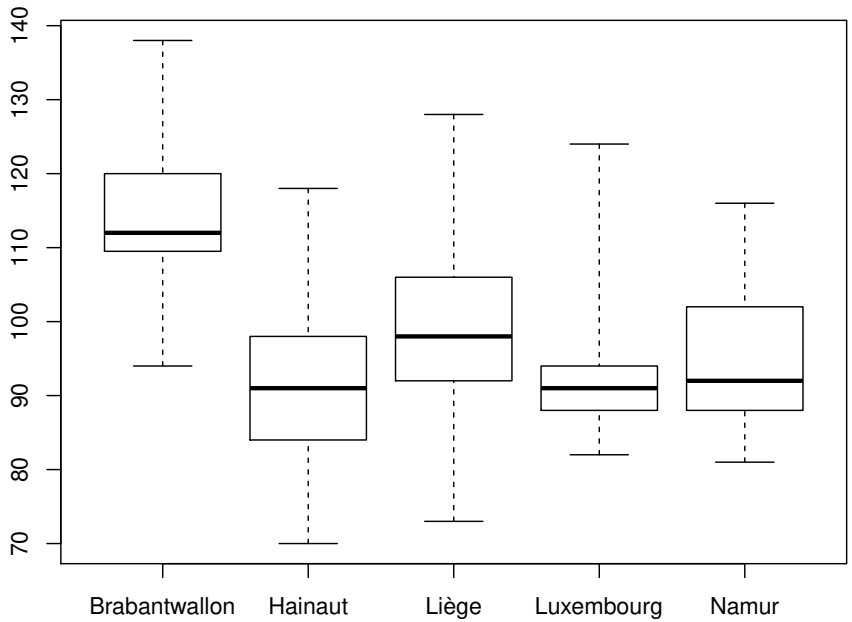


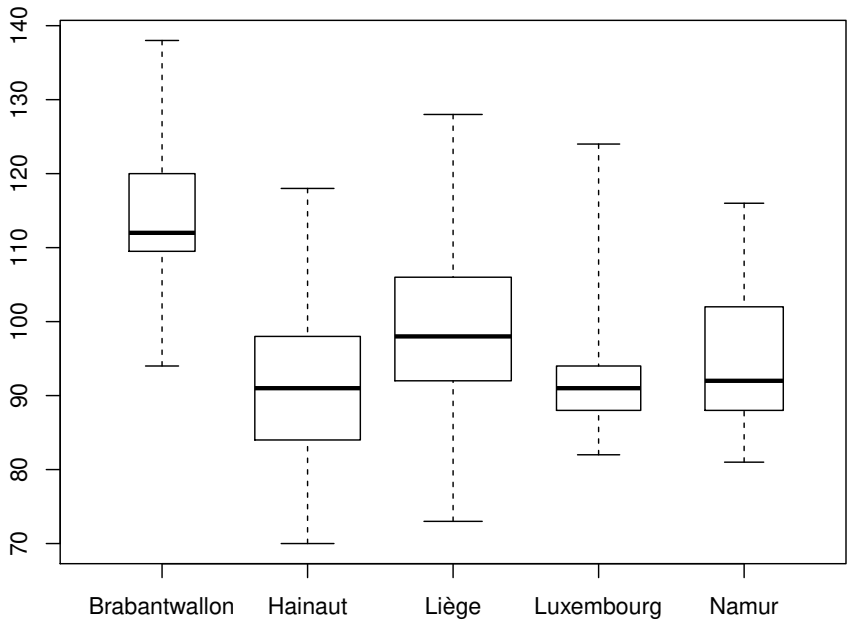
Utilité

- La médiane renseigne sur la valeur centrale de la série.
- Les longueurs des deux parties de la boîte rendent compte de la dispersion et de la symétrie des valeurs situées au centre de la série.
- Les longueurs des moustaches indiquent la dispersion et la symétrie présentes parmi les plus petites et les plus grandes observations (chaque moustache représente le comportement d'environ 25% des observations).

LIÈGE université

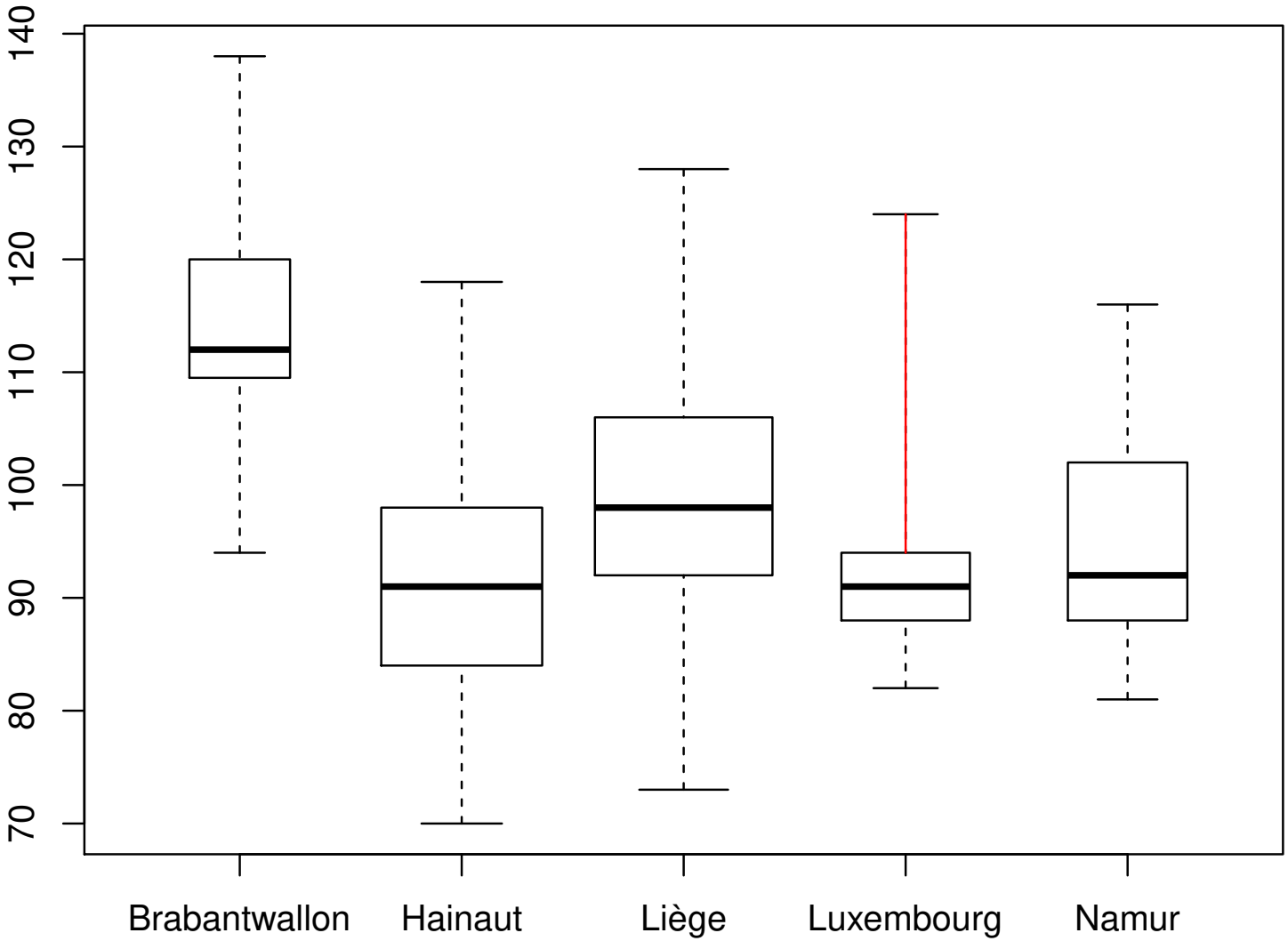
Variation de la taille pour tenir compte des effectifs différents





Effectifs des séries: 27, 69, 84, 44, 38

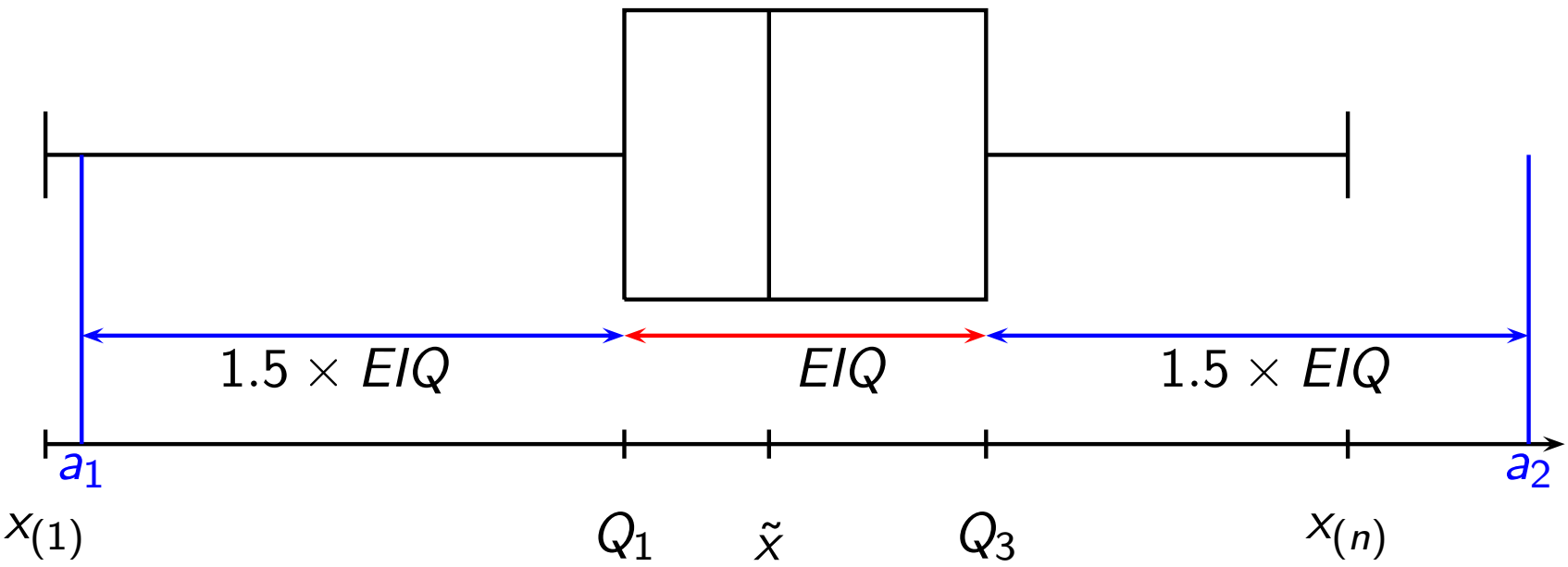
Moustaches trop longues?



Boîte à moustaches modifiée

1) Les valeurs pivots a_1 et a_2:

$$\begin{aligned} a_1 &= Q_1 - 1,5EIQ \\ a_2 &= Q_3 + 1,5EIQ \end{aligned}$$

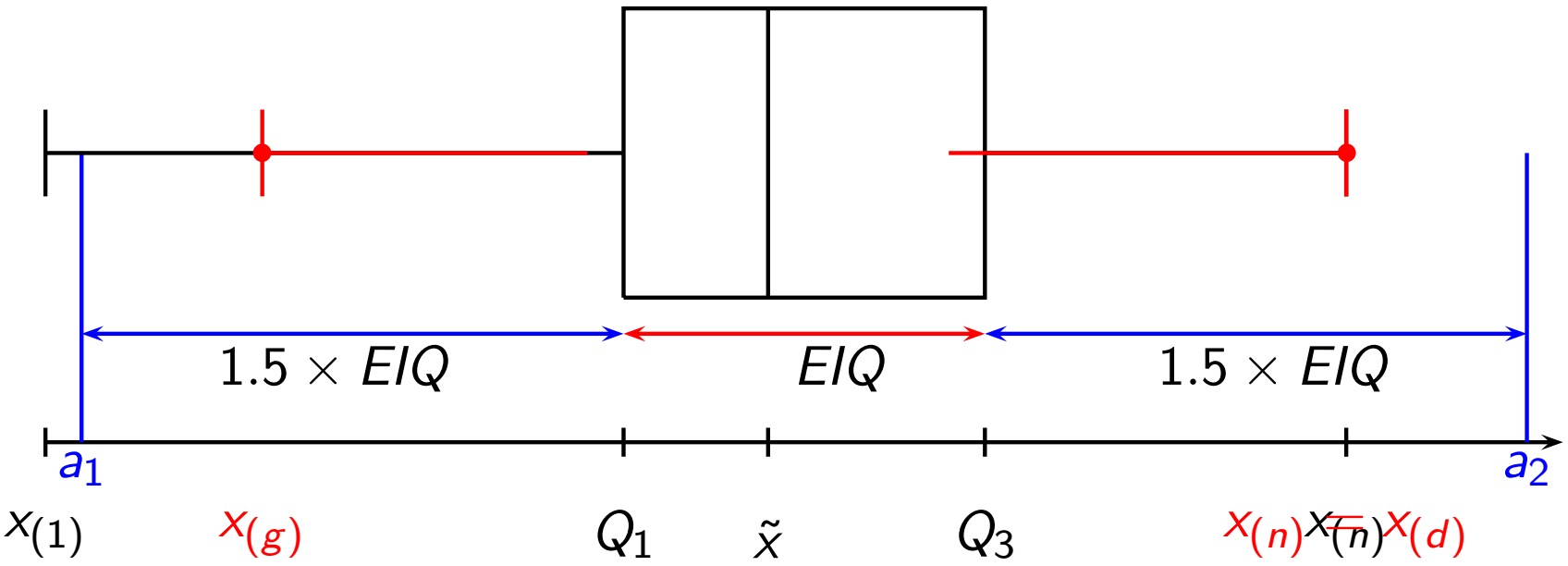


La plupart des séries ne contenant pas de valeurs aberrantes ont leurs observations comprises dans l'intervalle $[a_1, a_2]$.

Boîte à moustaches modifiée

2) Les valeurs adjacentes sont les observations de rang g et d telles que

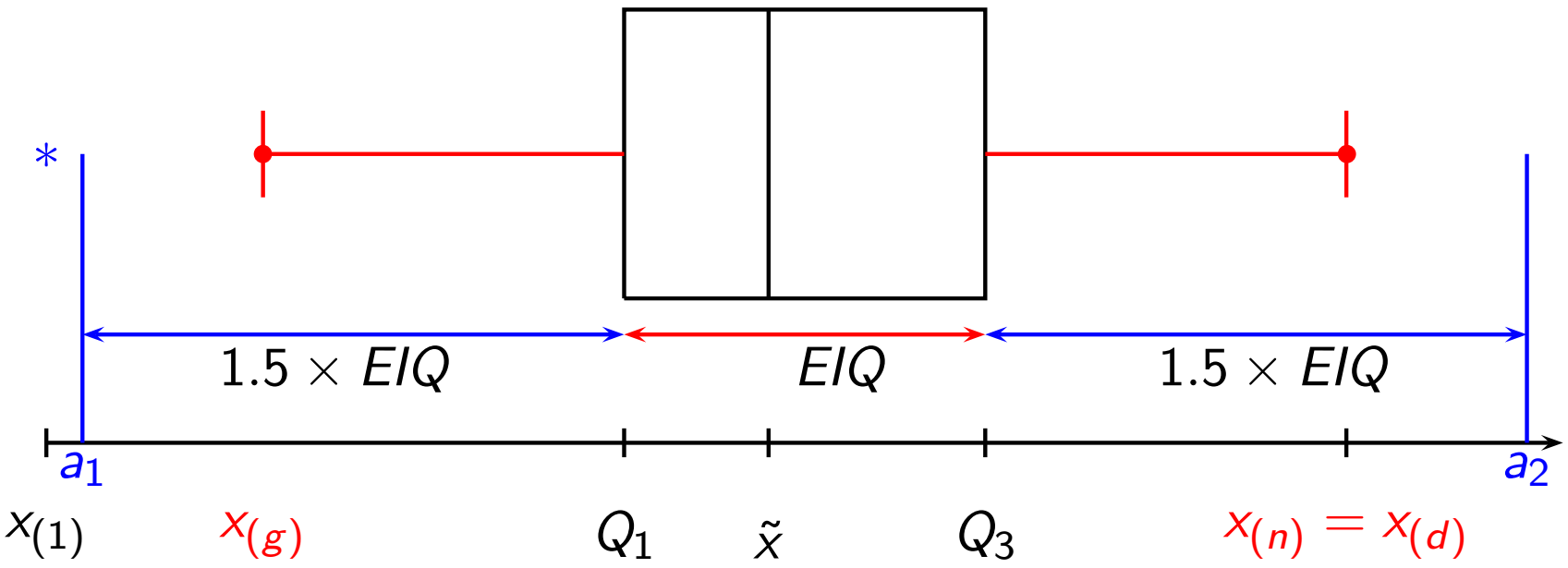
$$x_{(g)} = \text{la plus petite observation supérieure ou égale à } a_1$$
$$x_{(d)} = \text{la plus grande observation inférieure ou égale à } a_2$$



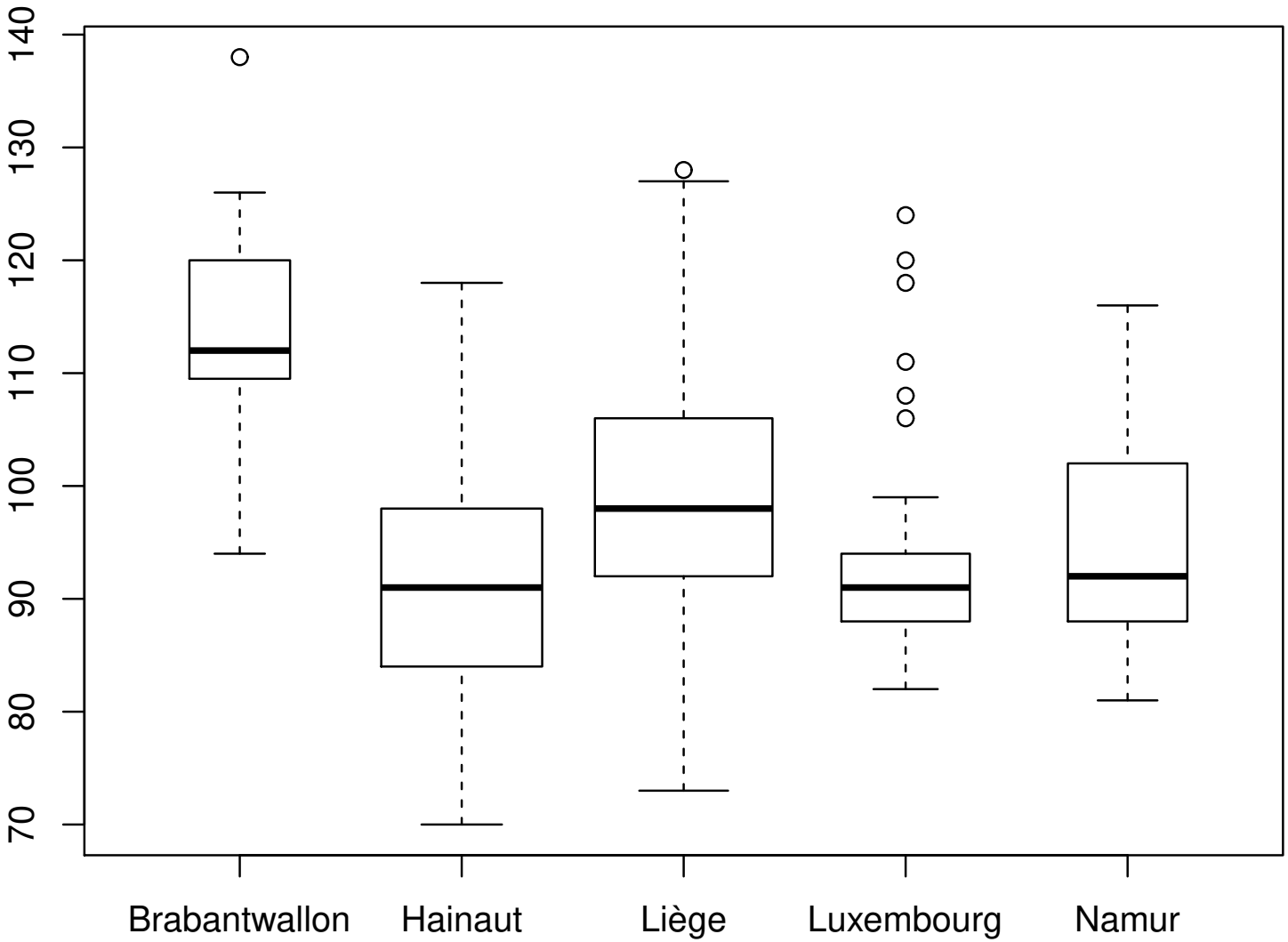
Les moustaches de la boîte à moustaches modifiée relient les quartiles à ces valeurs qui correspondent aux valeurs observées les plus proches des valeurs pivots tout en appartenant à l'intervalle $[a_1, a_2]$.

Boîte à moustaches modifiée

3) Les valeurs extérieures sont les observations situées en dehors de $[a_1, a_2]$.



Boîtes à moustaches modifiée des indices de richesse

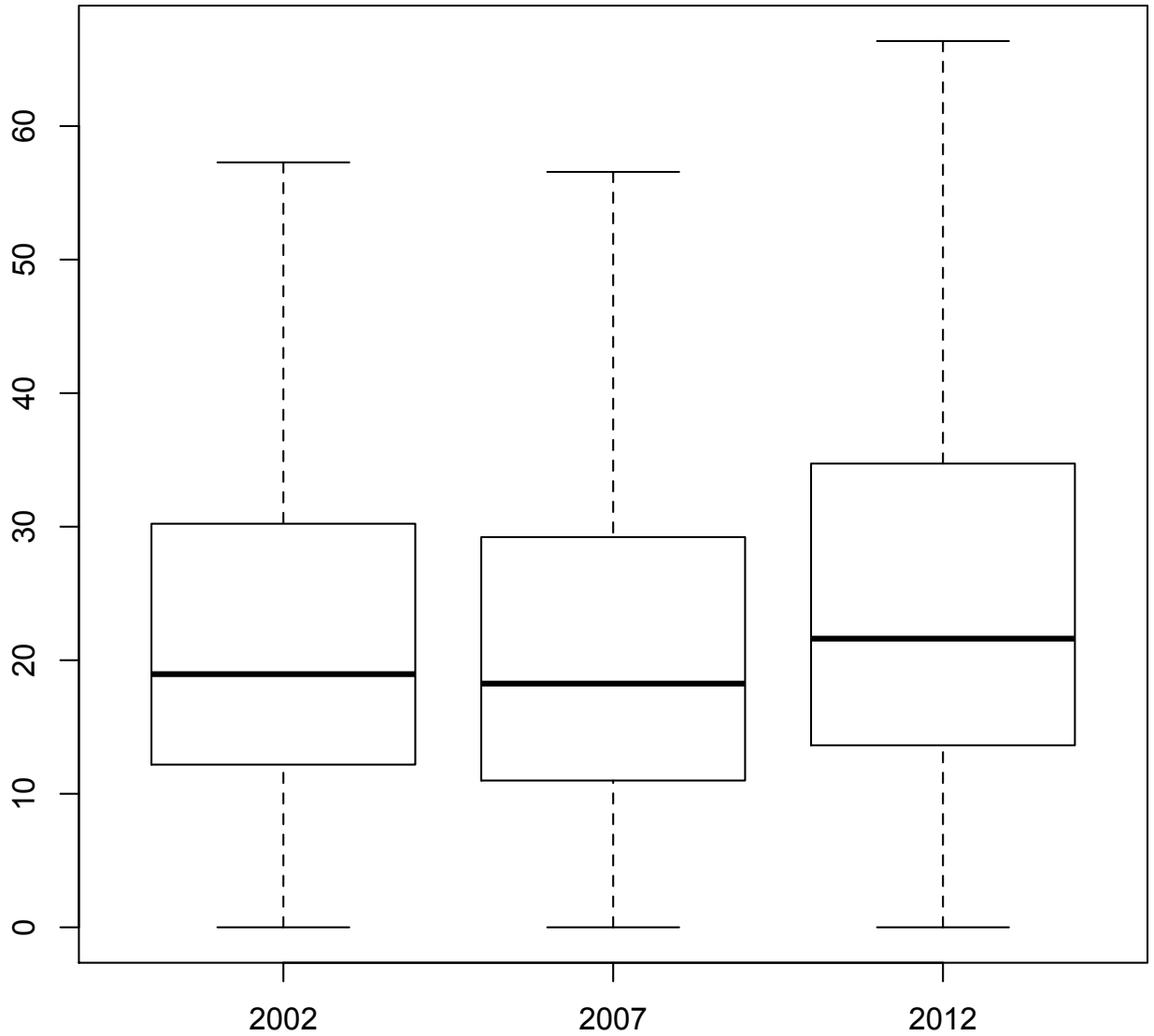


Boîte à moustaches de séries groupées

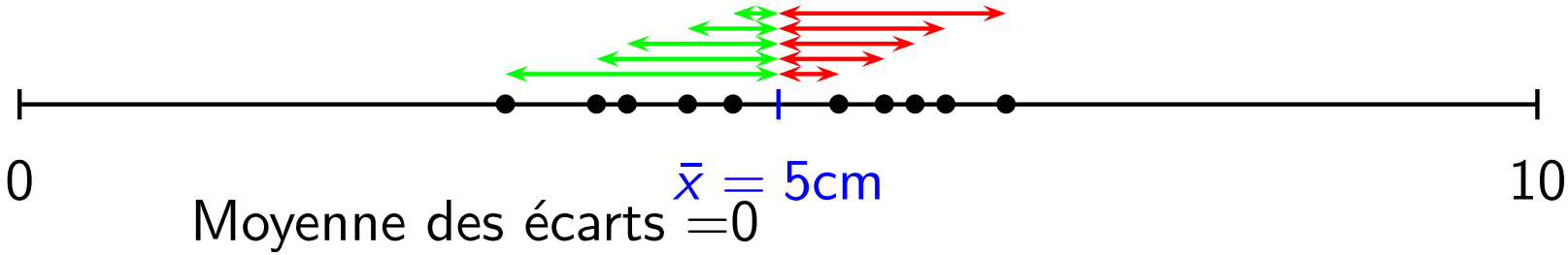
Les observations individuelles sont généralement manquantes, ce qui ne permet pas de représenter exactement les moustaches de la boîte.

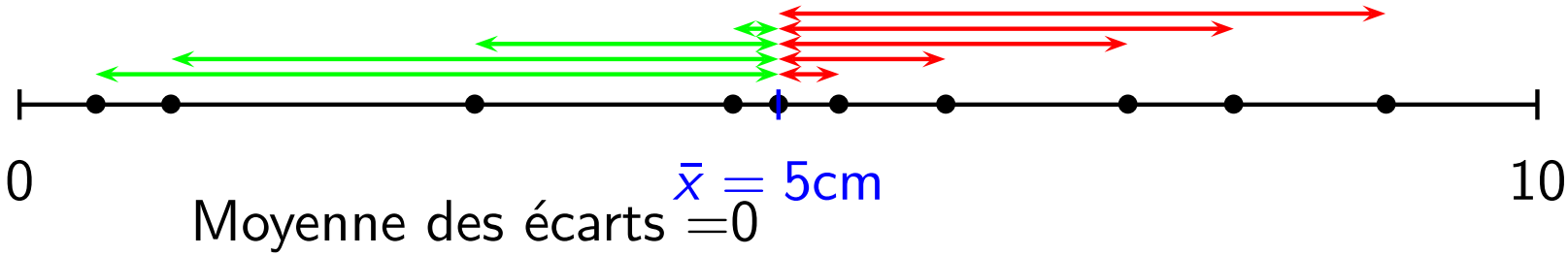
On se contente alors de dessiner des moustaches allant jusqu'aux valeurs pivots tandis que les autres paramètres $(Q_1, \tilde{x}, Q_3)$ sont estimés par interpolation linéaire.

Boîtes à moustaches de la série des revenus



Mesure de dispersion basée sur les écarts $x_i - \bar{x}$





Prendre en compte l'ensemble de tous les écart $x_i - \bar{x}$ devrait permettre de caractériser la dispersion des données!

Comment combiner ces écarts?

Il faut éliminer le signe des écarts avant de calculer une moyenne!

Ecarts *absolus* ou écarts *au carré?*

1. Première option: éliminer le signe en prenant la valeur absolue

$$\text{Ecart absolu moyen: } E_{\bar{x}} = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|$$

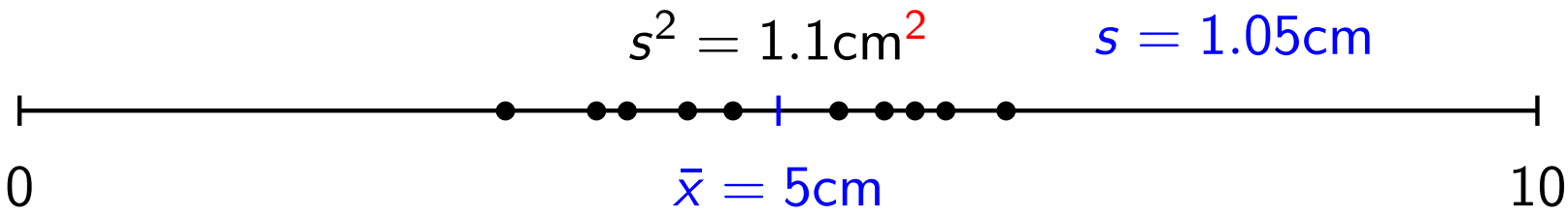
2. Deuxième option: éliminer le signe en prenant le carré

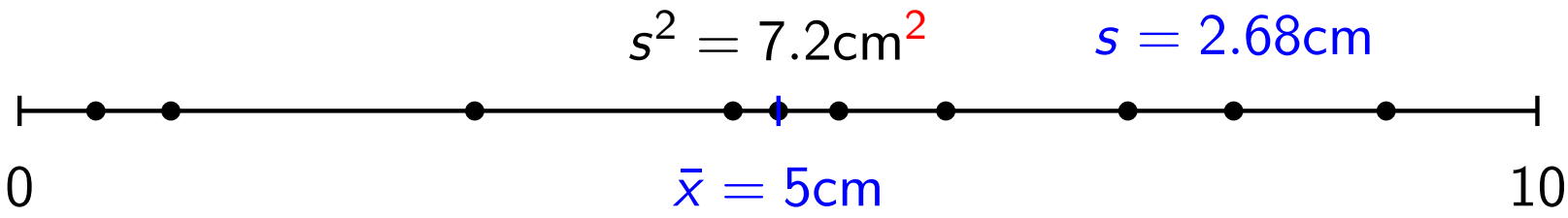
$$\text{Variance: } s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

LIÈGE université

La variance et l'écart-type

La variance est donc définie par la moyenne des carrés des écarts entre les observations et leur moyenne $\bar{x}$: $s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$.





L'écart-type est défini par la racine carrée de la variance

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

Variances et écarts-types des indices de richesse

Provinces	Moyennes	Ecarts-types
Brabant wallon	114.1	8.63
Hainaut	91.7	10.56
Liège	98.9	10.66
Luxembourg	93.4	9.55
Namur	94.2	8.87

Variance d'une série recensée ou groupée

- Si $S = \{(x_j, n_j), j = 1, \ldots, J\}$ avec $x_1 < x_2 < \ldots < x_J$

$$s^2 = \frac{1}{n} \sum_{j=1}^{J} n_j (x_j - \bar{x})^2,$$

avec $\bar{x} = \frac{1}{n} \sum_{j=1}^{J} n_j x_j$.

- Si la série est groupée en J classes de centres $c_1, \ldots, c_J$ et d'effectifs $n_1, \ldots, n_J$, une valeur approximative de s^2 est donnée par

$$s^2 \approx \frac{1}{n} \sum_{j=1}^{J} n_j (c_j - \bar{x})^2,$$

avec $\bar{x} = \frac{1}{n} \sum_{j=1}^{J} n_j c_j$.

Correction de Sheppard (pour des amplitudes constantes):

$$s^2 \approx \frac{1}{n} \sum_{j=1}^{J} n_j (c_j - \bar{x})^2 - \frac{a^2}{12}$$

Illustration: calcul de la variance des indices de richesse de Liège à partir de la série groupée

On sait que l'estimation de la moyenne est $\bar{x} = 98.33$.

Classes	Centres c_j	Effectifs n_j	Fréquences	Eff. cum.	Fréq. cum.	$n_j(c_j - \bar{x})^2$
$[70, 80]$	75	2	0.02	2	0.02	1088.6
$]80, 90]$	85	15	0.18	17	0.20	2665.3
$]90, 100]$	95	31	0.37	48	0.57	343.8
$]100, 110]$	105	28	0.33	76	0.90	1245.7
$]110, 120]$	115	5	0.06	81	0.96	1389.4
$]120, 130]$	125	3	0.04	84	1	2133.9
Total		84	1	x	x	8866.7

D'où, $s^2 = \frac{8866.7}{84} = 105.6$

Alors que la variance exacte est égale à 115.

Appliquer la correction, dans ce cas-ci, donnerait une sous-estimation de la variance.

Deuxième exemple: calcul de la variance de 107 tailles à partir de la série groupée

La moyenne estimée sur la série groupée ci-dessous vaut $\bar{x} = 177.52$.

Classes	Centres	Eff	Fréq	Eff. cum.	Fréq. cum.	$n_j\ (c_j - \bar{x})^2$
$[150; 160[$	155	2	0.02	2	0.02	1015
$[160; 170[$	165	21	0.20	23	0.21	3294
$[170; 180[$	175	40	0.37	63	0.59	255
$[180; 190[$	185	36	0.34	99	0.93	2012
$[190; 200]$	195	8	0.07	107	1	2443
Total	x	107	1	x	x	9019

D'où $s^2 = \frac{9019}{107} = 84.28$.
Or, la variance calculée sur la série brute vaut 76.32.
En appliquant la correction: $84.28 - \frac{10^2}{12} = 75.95$

Propriétés de la variance

1) Définition plus générale:

$$s^2(a) = \frac{1}{n}\sum_{i=1}^{n}(x_i - a)^2$$

Proposition

Théorème de König-Huygens: La moyenne des carrés des écarts entre les observations d'une série et un paramètre a se décompose de la façon suivante:

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

où s^2 est la variance de la série.

Outils pour la démonstration:

- $(a+b)^2 = a^2 + b^2 + 2ab$
- Distributivité et mise en évidence;
- Propriété de la série centrée.

Démonstration du TKH: $\frac{1}{n}\sum_{i=1}^{n}(x_i-a)^2 = s^2+(\bar{x}-a)^2$

Pour tout a, on a

$$\sum_{i=1}^{n}(x_i-a)^2 = \sum_{i=1}^{n}(x_i-\bar{x}+\bar{x}-a)^2 = \sum_{i=1}^{n}\left((x_i-\bar{x})+(\bar{x}-a)\right)^2$$

D'où

$$\sum_{i=1}^{n}(x_i-a)^2 = \sum_{i=1}^{n}\left((x_i-\bar{x})^2+(\bar{x}-a)^2+2(x_i-\bar{x})(\bar{x}-a)\right)$$

Il vient donc

$$\sum_{i=1}^{n}(x_i-a)^2 = \sum_{i=1}^{n}(x_i-\bar{x})^2+\sum_{i=1}^{n}(\bar{x}-a)^2+2\sum_{i=1}^{n}(x_i-\bar{x})(\bar{x}-a)$$

Et finalement

$$\sum_{i=1}^{n}(x_i-a)^2 = \sum_{i=1}^{n}(x_i-\bar{x})^2+\underbrace{\sum_{i=1}^{n}(\bar{x}-a)^2}_{n\ (\bar{x}-a)^2}+2(\bar{x}-a)\underbrace{\sum_{i=1}^{n}(x_i-\bar{x})}_{=0}$$

En divisant par n, on a $\frac{1}{n}\sum_{i=1}^{n}(x_i-a)^2 = \underbrace{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}_{s^2}+(\bar{x}-a)^2$

Propriétés de la variance

2) Formule équivalente:

Proposition

La moyenne des carrés des écarts entre les observations et leur moyenne peut se calculer comme suit:

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

Démonstration: Application directe de TKH avec $a = 0$.

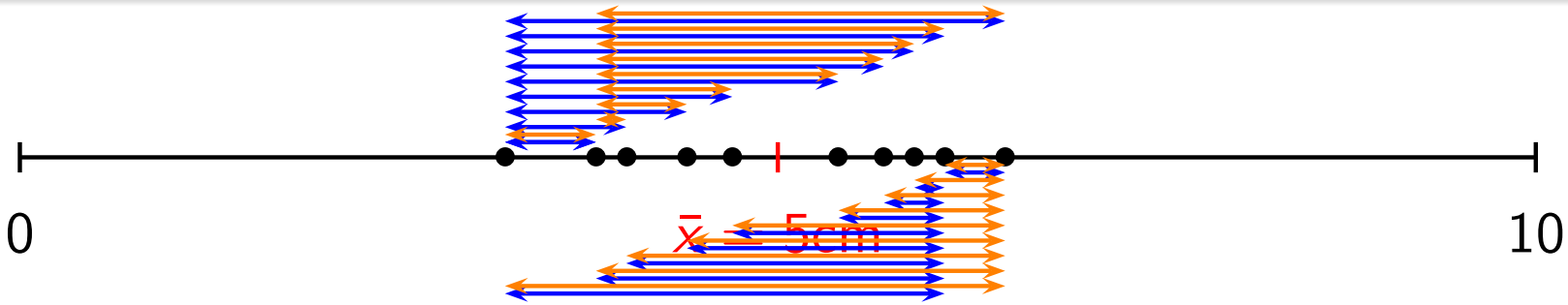
Propriétés de la variance (suite)

3) Formule équivalente:

Proposition

La moyenne des carrés des écarts entre les observations et leur moyenne peut se calculer comme suit:

$$s^2 = \frac{1}{2n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} (x_i - x_j)^2.$$



Démonstration:

Outils: $(a+b)^2$ et distributivité.

Développer le carré, puis distribuer les doubles signes sommatoires; simplifier.

Propriétés de la variance (suite)

4) Transformation affine:

Proposition

Si $S' = \{x'_1, \ldots, x'_n\}$, *avec* $x'_i = ax_i + b$, $a, b \in \mathbb{R}$, *alors la variance* s'^2 *de* S' *est donnée par*

$$s'^2 = a^2 s^2$$

où s^2 *est la variance de* S. *Pour les écarts-types, la relation devient* $s' = |a|s$.

Démonstration:

Outils: Définition de s^2, propriété de $\bar{x}$ et mise en évidence.

Par déf, $s'^2 = \frac{1}{n}\sum_{i=1}^{n}(x'_i - \bar{x}')^2 = \frac{1}{n}\sum_{i=1}^{n}(ax_i + b - (a\bar{x} + b))^2$

D'où, $s'^2 = \frac{1}{n}\sum_{i=1}^{n}(ax_i + b - a\bar{x} - b)^2$

En mettant a en évidence, on obtient

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

Propriétés de la variance (suite)

5) *Propriété de Tchebychev*:

Pafnouti Tchebychev
(1821–1894)
Mathématicien russe (probabilité – statistique – théorie des nombres)

Propriétés de la variance (suite)

5) *Propriété de Tchebychev*

Proposition

Soit $S = \{x_1, \ldots, x_n\}$ une série de moyenne $\bar{x}$ et d'écart-type s.
La proportion d'observations s'écartant d'au moins t écarts-types de la moyenne est inférieure ou égale à $\frac{1}{t^2}$.

Démonstration:
Outils:

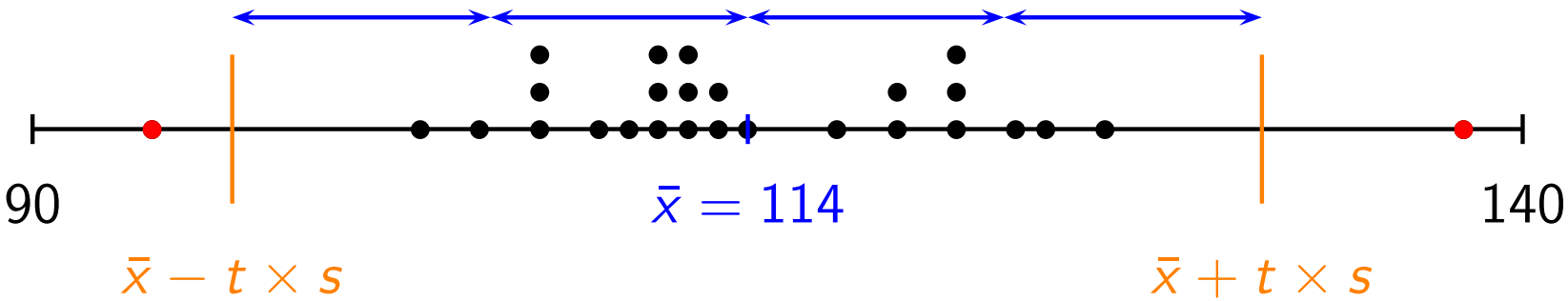
- Définition de s^2;
- Exploitation des valeurs absolues;
- Majoration de sommes.

Au tableau.

Propriétés de la variance (suite)

5) Interprétation de la propriété de Tchebychev: La proportion d'observations s'écartant d'au moins t écarts-types de la moyenne est inférieure ou égale à $\frac{1}{t^2}$.

Considérons l'indice de richesse dans le Brabant ($n = 27$)



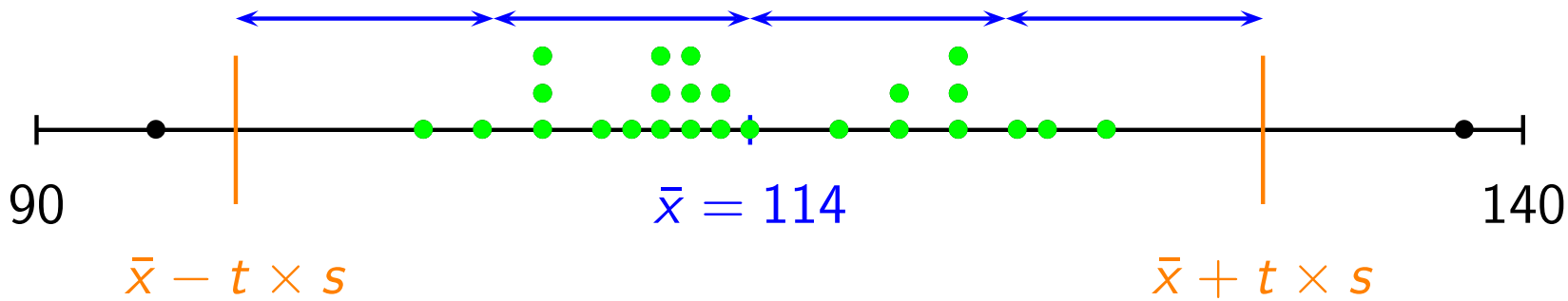
On sait que $s = 8.63$. Prenons $t = 2$

Il y a 2 observations s'écartant d'au moins 2 écarts-types de la moyenne;

La proportion observée est donc $\frac{2}{27} = 0.074 \leq 0.25 = \frac{1}{t^2}$

De façon équivalente

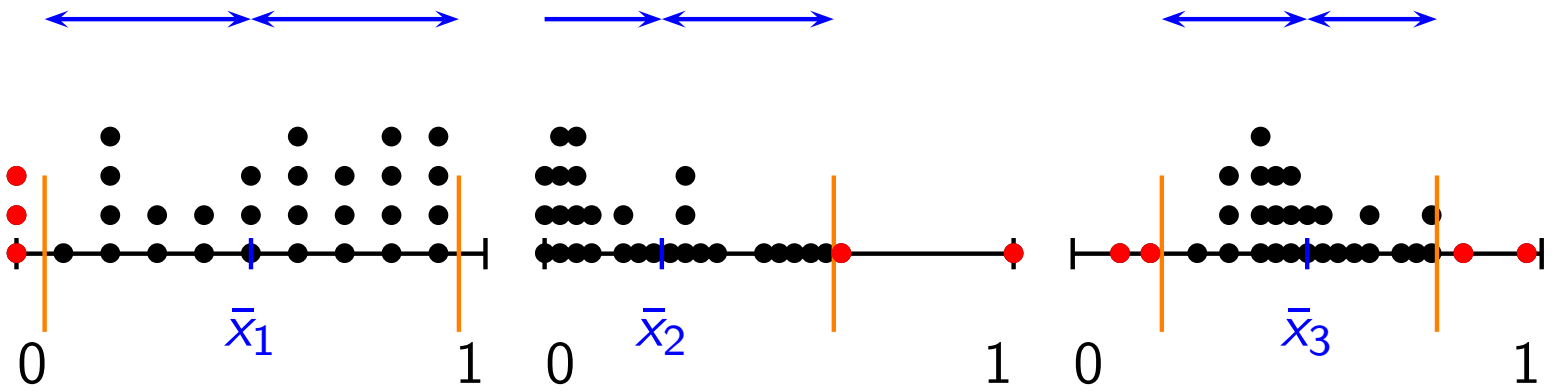
5) *Propriété de Tchebychev*: La proportion d'observations situées dans l'intervalle$]\bar{x} - t\ s, \bar{x} + t\ s[$ vaut au moins $1 - \frac{1}{t^2}$.



Il y a 25 observations dans l'intervalle;Ce qui donne une proportion observée égale à 0.925$>$ 0.75

Force et faiblesse de la propriété de Tchebychev

- Aucune hypothèse n'est imposée à la forme de la distribution;
- L'inégalité est aussi valide dans un espace probabilisé;
- La borne est peu contraignante...



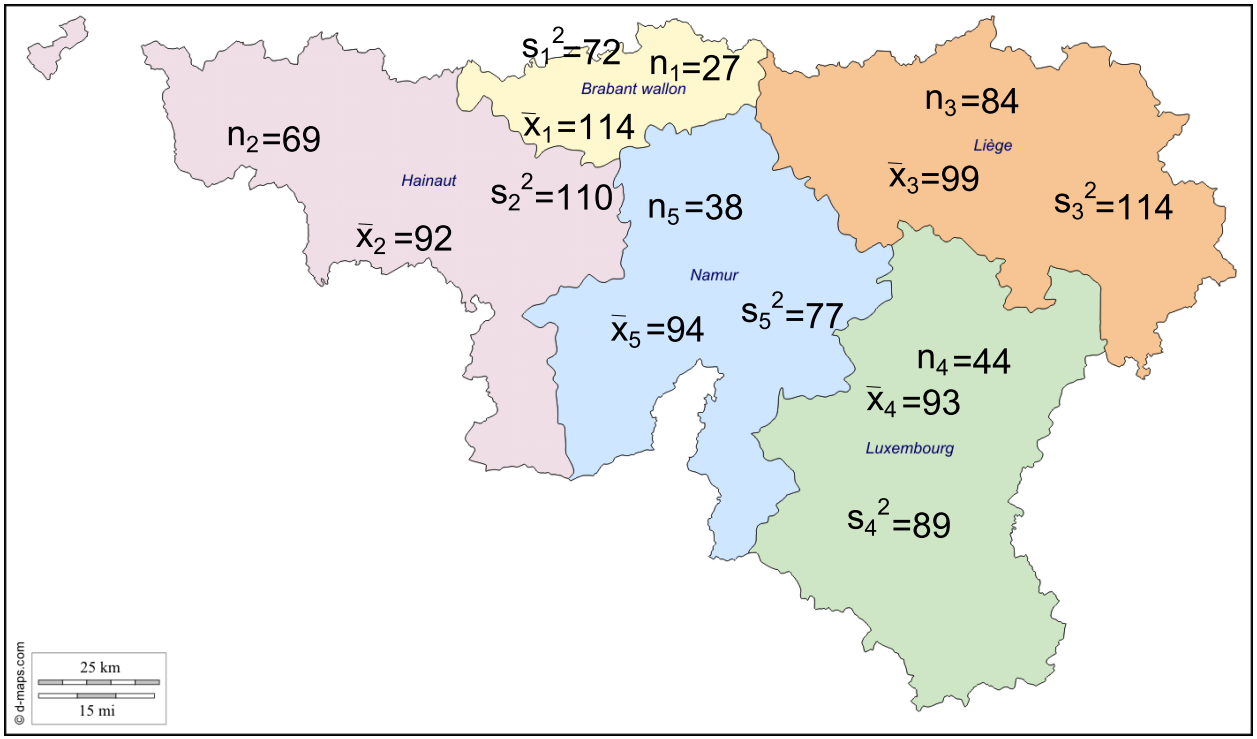
Prenons $t = 1.5$ (et notons que $s_1 = 0.29, s_2 = 0.25$ et $s_3 = 0.19$)

Il y a resp. 3, 2 et 4 observations sur 30 en dehors de l'intervalle, menant à des proportions égales à 0.1, 0.066 et 0.133, alors que la proportion maximale est de $1/1.5^2 = 0.44$!

Propriétés de la variance (suite)

6) Décomposition de la variance: Soit une population P de n individus partagée en k sous-populations $P_1, \ldots, P_k$ d'effectifs $n_1, \ldots, n_k$ (avec $\sum_{i=1}^{k} n_i = n$), de moyennes $\bar{x}_1, \ldots, \bar{x}_k$ et de variances $s_1^2, \ldots, s_k^2$.

Considérons les 262 communes de Wallonie



On sait que la moyenne de P, $\bar{x}$, vaut $\frac{\sum_{i=1}^{k} n_i \times \bar{x}_i}{n}$.

Qu'en est-il de la variance de P?

Propriétés de la variance (suite)

6) Décomposition de la variance:

Proposition

Soit une population P de n individus décomposée en k sous-populations $P_1, \ldots, P_k$ d'effectifs $n_1, \ldots, n_k$ (avec $\sum_{i=1}^{k} n_i = n$), de moyennes $\bar{x}_1, \ldots, \bar{x}_k$ et de variances $s_1^2, \ldots, s_k^2$,. La variance globale s^2 de la population est donnée par

$$s^2 = \frac{\sum_{i=1}^{k} n_i s_i^2}{n} + \frac{\sum_{i=1}^{k} n_i (\bar{x}_i - \bar{x})^2}{n},$$

où $\bar{x}$ est la moyenne globale de P.

Démonstration:

Outils:

- Définition de s^2;
- Application du théorème de König-Huygens;
- Manipulation des sommes.

Au tableau.

Interprétation

$$s^2 = \frac{\sum_{i=1}^{k} n_i s_i^2}{n} + \frac{\sum_{i=1}^{k} n_i (\bar{x}_i - \bar{x})^2}{n}$$

Moyenne pondérée des variances des sous-populations: "variance dans les groupes".

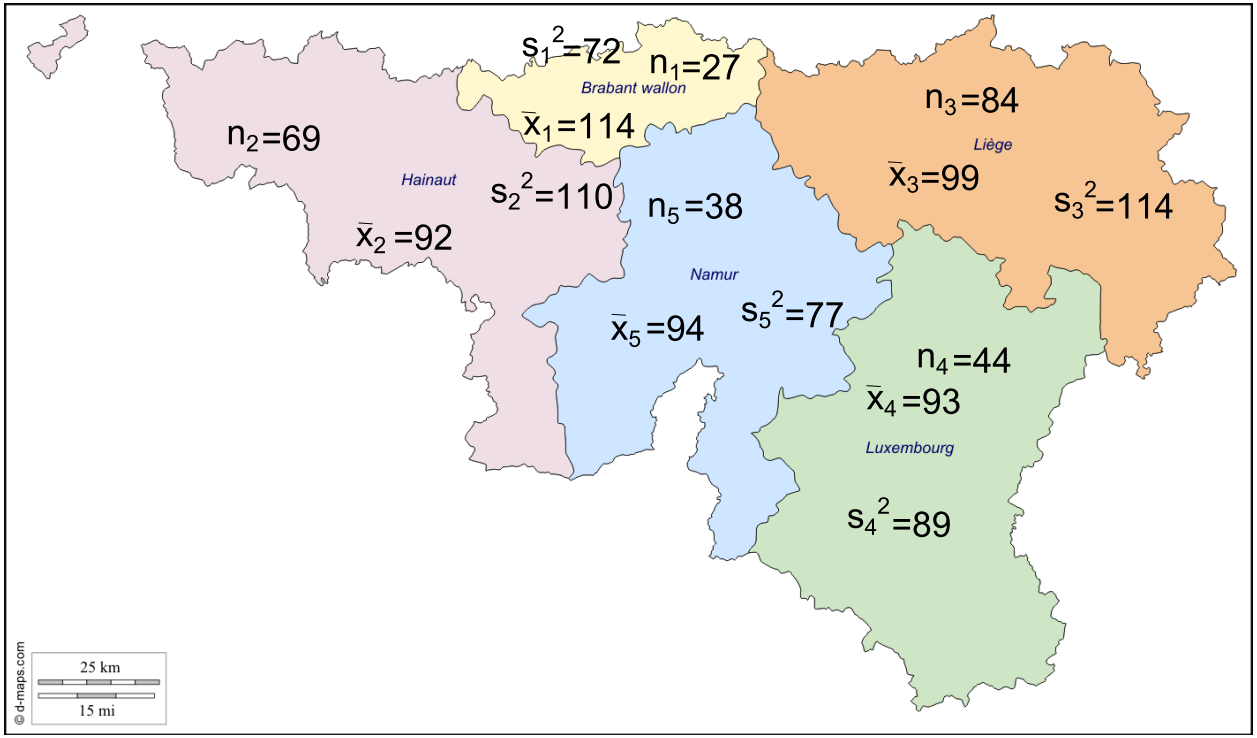
Variance des moyennes $\bar{x}_i$: "variance entre les groupes"

En résumé:

$$\begin{array}{ccccc} \text{Variance} & = & \text{Variance dans} & + & \text{Variance entre} \\ \text{globale} & & \text{les groupes} & & \text{les groupes} \end{array}$$

Si la source principale de variabilité est entre les groupes, cela signifie que les groupes sont constitués d'individus homogènes mais que les groupes sont différents. Si la variabilité est pluôt confinée dans les groupes, ceux-ci sont hétérogènes.

Illustration sur les indices de richesse



Sachant que $\bar{x} = 97$, on obtient la décomposition suivante:

Variance globale	=	Variance *dans* les groupes	+	Variance *entre* les groupes
141	=	99	+	42

La variance dans les groupes représente 70% de la variabilité totale.

Le coefficient de variation

Le coefficient de variation constitue une mesure relative de dispersion puisqu'il est défini par

$$CV = \frac{s}{\bar{x}}.$$

Il s'agit d'un nombre "pur" (c'est-à-dire sans unité) que l'on exprime généralement en %. Il permet de comparer plusieurs séries dépendant d'unités différentes.

Indice de richesse à Liège: $\bar{x} = 98.9$ et $s = 10.66$

Chômage: $\bar{x} = 12.08\%$ et $s = 4.85$ %

D'òu:
$CV = 0.11$ pour l'indice de richesse et $CV = 0.40$ pour le chômage.

Choix d'un paramètre de dispersion

La mesure de la dispersion dans un ensemble de données est primordiale puisqu'elle est liée au caractère d'homogénéité ou d'hétérogénéité d'un groupe.

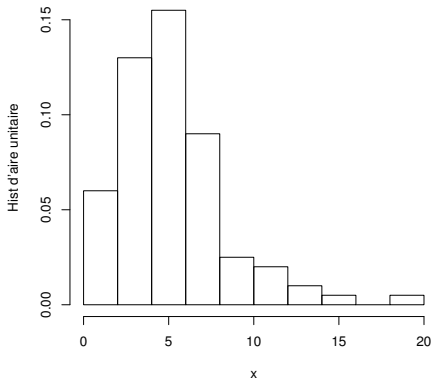
La variance, l'écart-type et le coefficient de variation sont les paramètres les plus courants.

La boîte à moustaches est habituellement très utile et facile à construire. Elle offre des informations complémentaires par rapport aux autres paramètres.

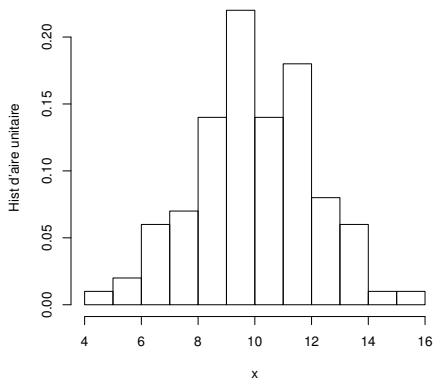
LIÈGE université

Les paramètres de forme

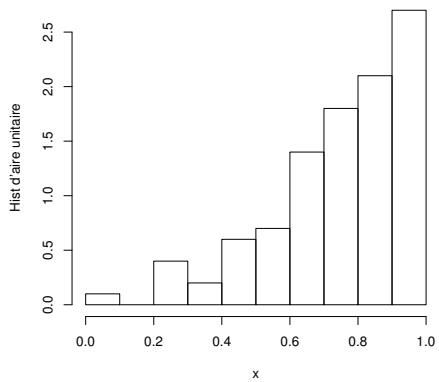
- les paramètres de dissymétrie



Dissymétrie à gauche



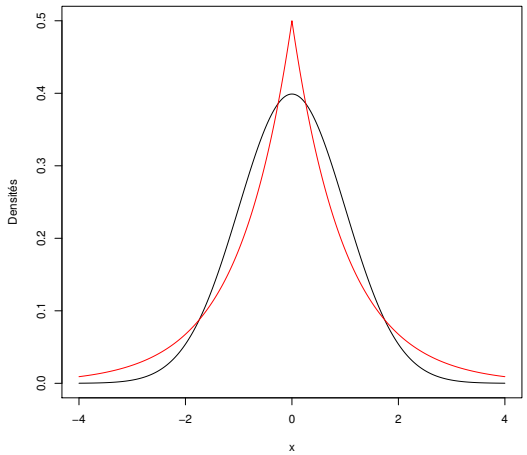
Symétrie



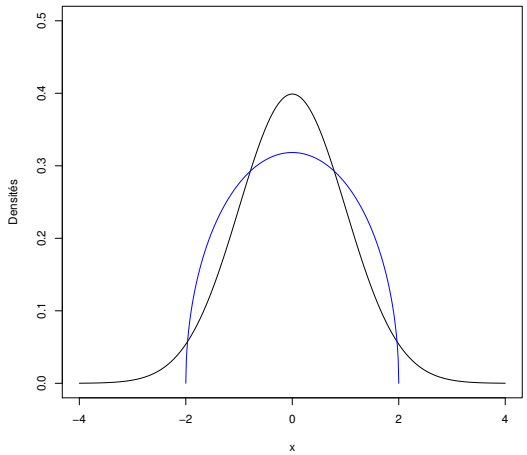
Dissymétrie à droite

Les paramètres de forme

- les paramètres d'aplatissement: ils caractérisent l'aplatissement (et aussi la grosseur des queues) d'une série.



Distribution leptokurtique



Distribution platikurtique

Cette mesure est réalisée en comparant le polygone de la distribution (courbe lisse "à la limite") par rapport à la courbe de Gauss qui, pour une moyenne nulle et une variance égale à 1, est donnée par

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$$

Les moments d'une série statistique

La plupart des paramètres de dissymétrie ou d'aplatissement sont définis à partir des moments centrés de la série.

Le *moment centré d'ordre k* d'une série $S = \{x_1, \ldots, x_n\}$ est donné par

$$m_k = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^k.$$

1. Les moments d'ordre 1 et 2 sont $m_1 = 0$ et $m_2 = s^2$.
2. Si $x'_i = ax_i + b, i = 1, \ldots, n$, alors $m'_k = a^k m_k$.
3. Les moments non centrés sont définis par $\mu_k = \frac{1}{n} \sum_{i=1}^{n} x_i^k$.

Utilisation des moments et autres paramètres pour caractériser la dissymétrie

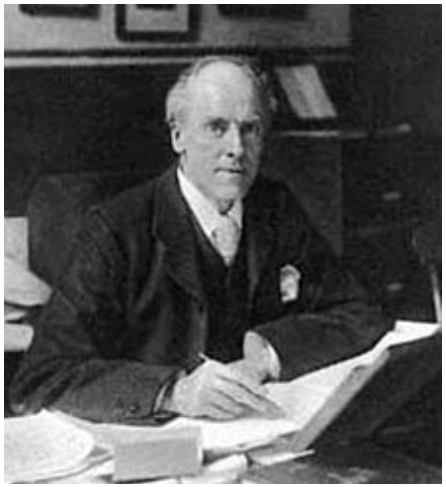
Le signe de m_3 ainsi que le signe de l'écart entre $\bar{x}$ et x_M renseignent sur le type de dissymétrie présente dans les données:

- Dissymétrie à gauche (ou étalement à droite): $m_3 > 0, \bar{x} - x_M > 0$
- Symétrie: $m_3 \approx 0, \bar{x} - x_M \approx 0$
- Dissymétrie à droite (ou étalement à gauche): $m_3 < 0, \bar{x} - x_M < 0$

LIÈGE université

Principaux acteurs

Ronald Aylmer Fisher
(1890–1962)
Biologiste et statisticien anglais



Karl Pearson
(1857–1936)
Mathématicien anglais

Les paramètres de dissymétrie

(1) Le coefficient de dissymétrie de Fisher:

$$\gamma_1 = \frac{m_3}{s^3}.$$

- Insensible "en valeur absolue" aux changements d'origine et d'échelle (mais $a < 0$ change le signe de γ_1)
- Sensible aux observations extrêmes

Coefficients de dissymétrie empiriques

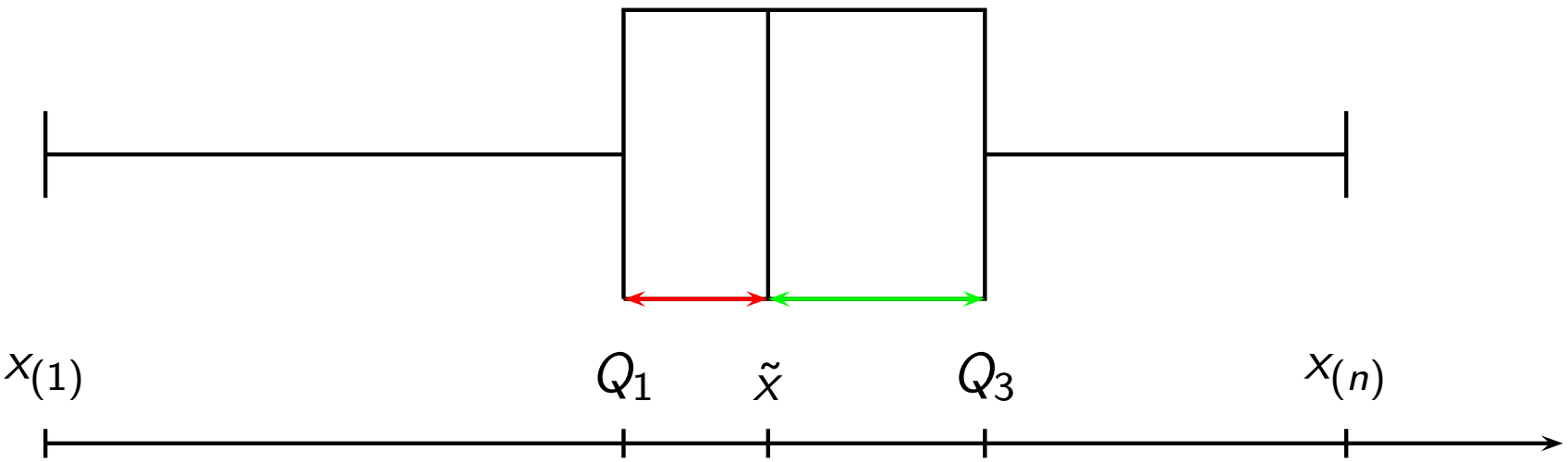
(2) Le coefficient empirique de dissymétrie de Pearson:

$$S_k = \frac{\bar{x} - x_M}{s}.$$

(3) Le coefficient empirique de dissymétrie de Yule et Kendall:

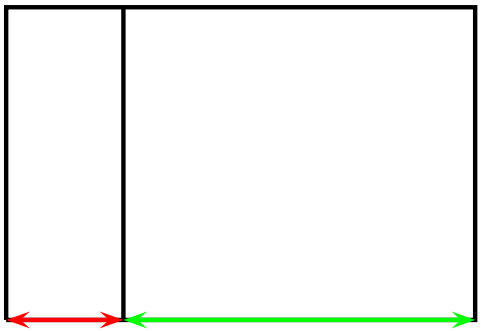
$$Y_k = \frac{Q_1 + Q_3 - 2\tilde{x}}{Q_3 - Q_1}$$

$$= \frac{Q_1 - \tilde{x} + Q_3 - \tilde{x}}{EIQ}$$



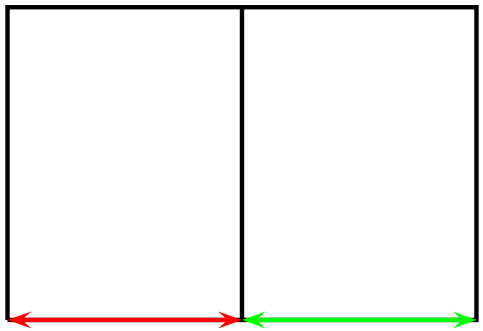
Signe du coefficient de Yule et Kendall

$$Y_k = \frac{Q_1 - \tilde{x} + Q_3 - \tilde{x}}{EIQ}$$



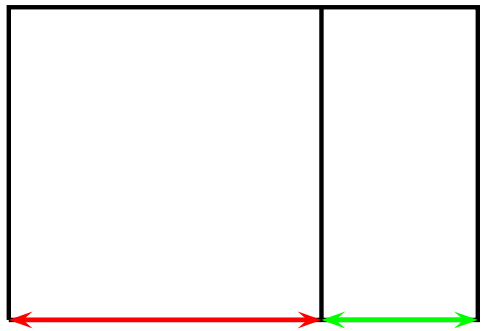
$Q_1 \quad \tilde{x} \quad Q_3$

$Y_k > 0$



$Q_1 \quad \tilde{x} \quad Q_3$

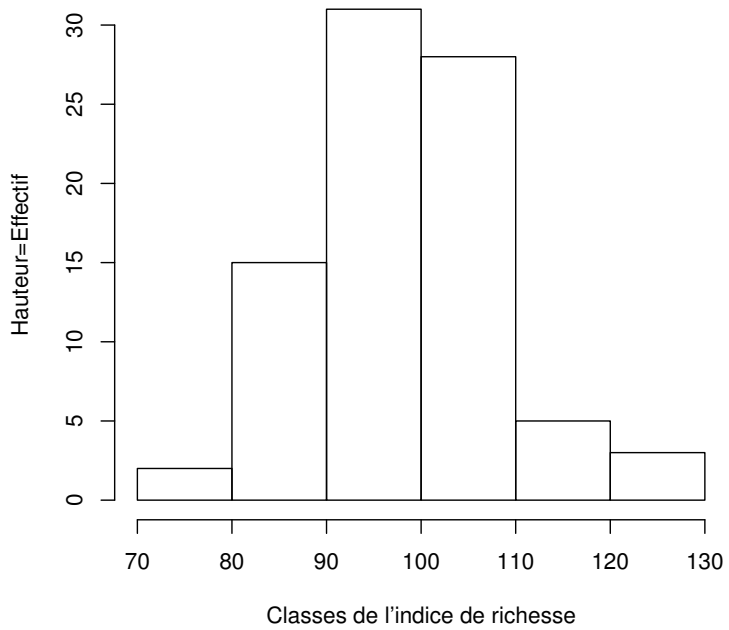
$Y_k \approx 0$

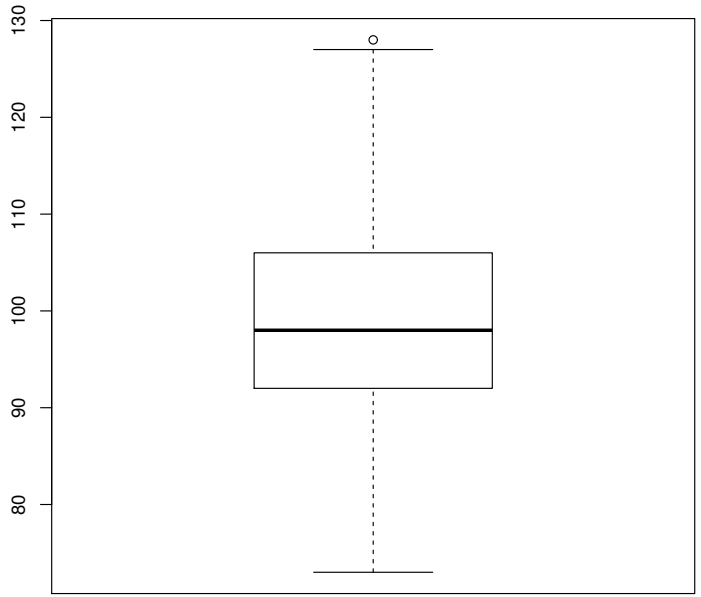


$Q_1 \quad \tilde{x} \quad Q_3$

$Y_k < 0$

Illustration sur la série des indices de richesse à Liège





$\gamma_1 = 0.32$ et $Y_k = 0.14$, $S_k = -0.66$

Les paramètres d'aplatissement

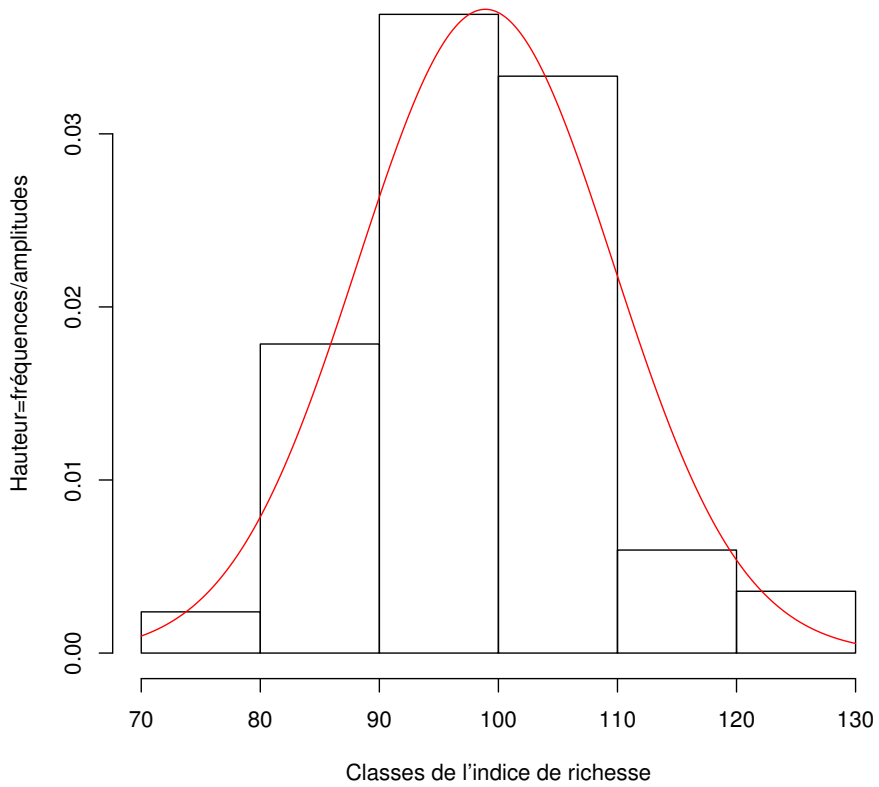
1. Le coefficient d'aplatissement de Pearson:

$$b_2 = \frac{m_4}{s^4}.$$

2. Le coefficient d'aplatissement de Fisher:

$$\gamma_2 = \frac{m_4}{s^4} - 3 = b_2 - 3.$$

Illustration



On obtient $b_2 = 3.23$.

Ce qu'il faut retenir de ce chapitre

Compétences pratiques:

- Définition, interprétation et exploitation à bon escient des paramètres statistiques

Position	Dispersion	Forme
Moyenne	Etendue	$\gamma_1 = m_3/s^3$
(et généralisations)		
Médiane	EIQ	$S_k = \frac{\bar{x}-x_M}{s}$
Quantiles	Var/Ecart-type	$Y_k = (Q_1 + Q_3 - 2\tilde{x})/EIQ$
Mode	Coef Variation	b_2 et γ_2

- Construction et exploitation des boîtes à moustaches;
- Calcul de fonctions d'influence;
- Exploitation de la formule de décomposition de la variance, de l'inégalité de Tchebychev.

LIÈGE université

Ce qu'il faut retenir de ce chapitre

Théorie: savoir citer et démontrer les propriétés suivantes:

- Effet d'un changement d'origine et d'échelle sur tous les paramètres.
- Définition et calcul de la moyenne de la série des valeurs centrées.
- Propriété d'optimalité de la moyenne (avec démonstration) et de la médiane (sans démonstration).
- Théorème de König-Huygens.
- Démonstration de la formule équivalente $s^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2$.
- Formule de décomposition de la variance pour une série décomposée en k sous-séries.
- Propriété de Tchebychev .

LIÈGE université