Statistique Descriptive
Chapitre 01 : Statistique Univariée

Il était une fois …

- Au départ, on a l'échantillon et une var X quantitative.
- n
  - l'effectif de l'échantillon
  - le nombre total d'objets, de sujets, de personnes,...dans l'échantillon
- On représente l'échantillon dans un tableau brut de données :

$$\{x_1, x_2, x_3, \ldots, x_n\}$$

Caractère quantitatif discret

nombre fini de valeurs

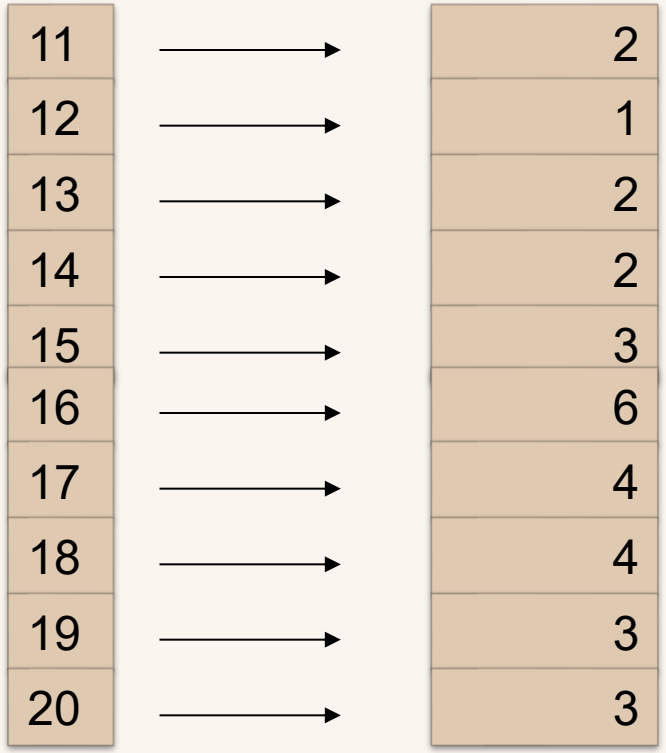
Exemple

- Une personne fait le relevé d'une température tous les jours à 8h30 pendant 30 jours et obtient les mesures suivantes :

12	20	13	11	15	16	16	14	17	13
11	18	16	17	17	15	16	16	15	19
18	16	17	14	18	19	19	20	20	18

Fréquence absolue n_i

- fréquence absolue d'une valeur x_i
  - n_i = nombre d'occurrences de la valeur x_i
  - effectif : $n = \sum n_i$



n = 30

Fréquence relative fr_i

- fréquence relative d'une valeur x_i
  - le rapport $\frac{n_i}{n}$

    effectif n = nombre total de valeur observées

$$f_i = \frac{n_i}{n} \text{ (en \%)}$$

n = 30

11	→	0,067	7 %
12	→	0,033	3 %
13	→	0,067	7 %
14	→	0,067	7 %
15	→	0,100	10 %
16	→	0,200	20 %
17	→	0,133	13 %
18	→	0,133	13 %
19	→	0,100	10 %
20	→	0,100	10 %

Fréquence cumulée d'une valeur x_i

- fréquence absolue cumulée d'une valeur x_i
  = somme des fréquences absolues de cette valeur et des valeurs inférieures.

$$nc_i = \sum_i n_i$$

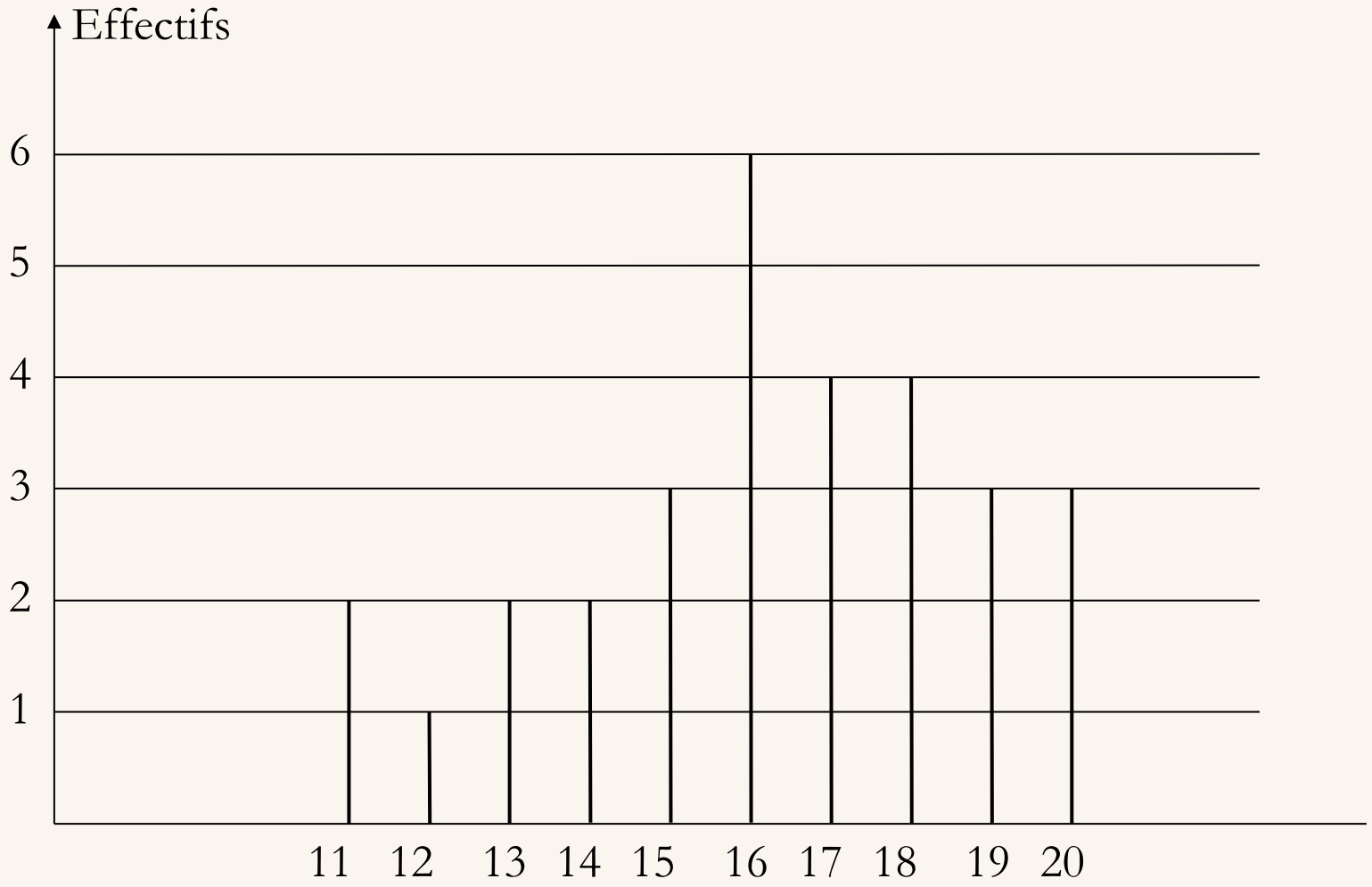
- fréquence relative cumulée d'une valeur x_i
  = somme des fréquences relatives de cette valeur et des valeurs inférieures.

$$fc_i = \sum_i f_i$$

Fréquence cumulée : exemple

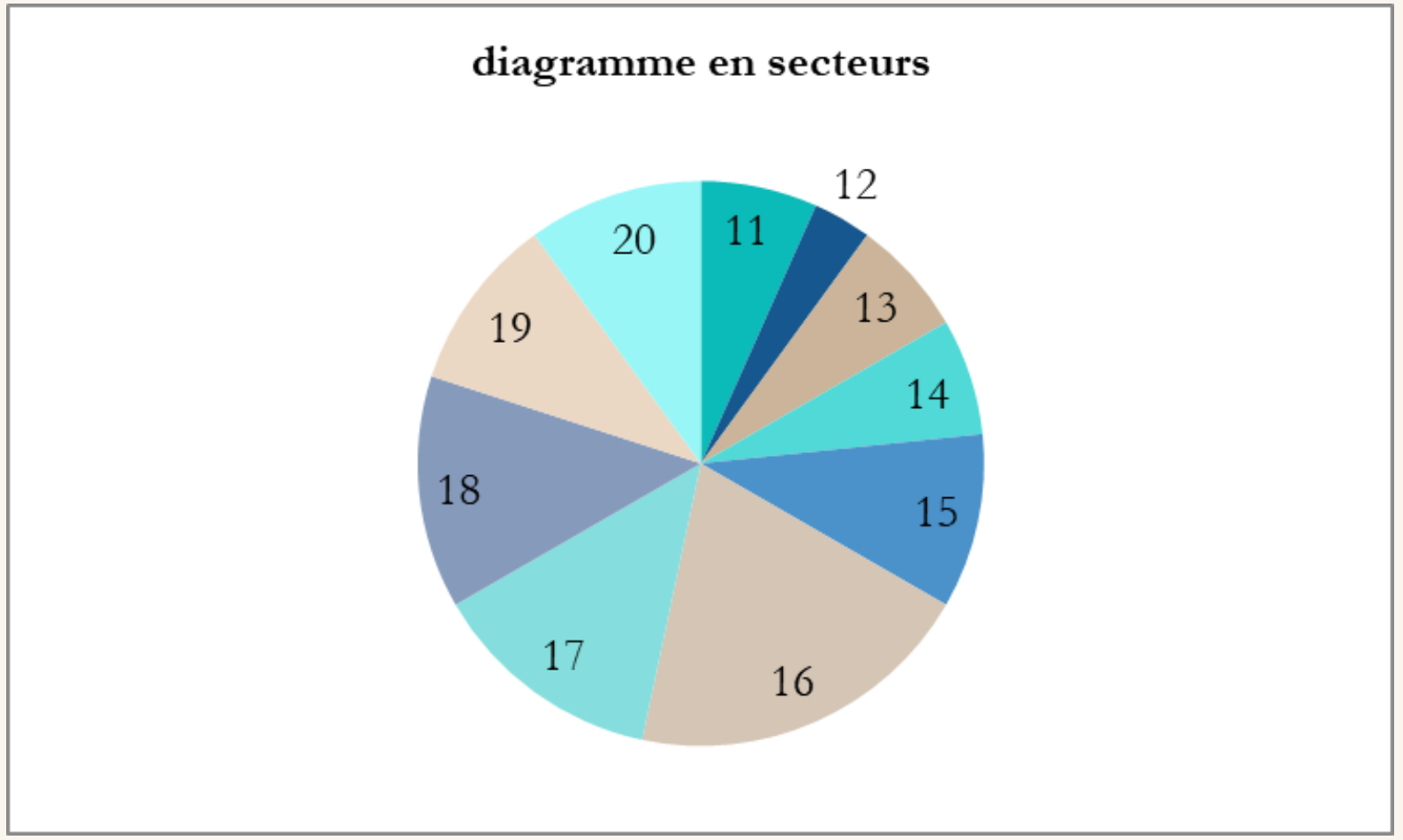
x_i	n_i	fr_i	fr_i (%)	nc_i	frc_i	frc_i (%)
11	2	0,067	6,7	2	0,067	6,7%
12	1	0,033	3,3	3	0,100	10,0%
13	2	0,067	6,7	5	0,167	16,7%
14	2	0,067	6,7	7	0,233	23,3%
15	3	0,100	10,0	10	0,333	33,3%
16	6	0,200	20,0	16	0,533	53,3%
17	4	0,133	13,3	20	0,667	66,7%
18	4	0,133	13,3	24	0,800	80,0%
19	3	0,100	10,0	27	0,900	90,0%
20	3	0,100	10,0	30	1,000	100%

Diagramme en bâtons

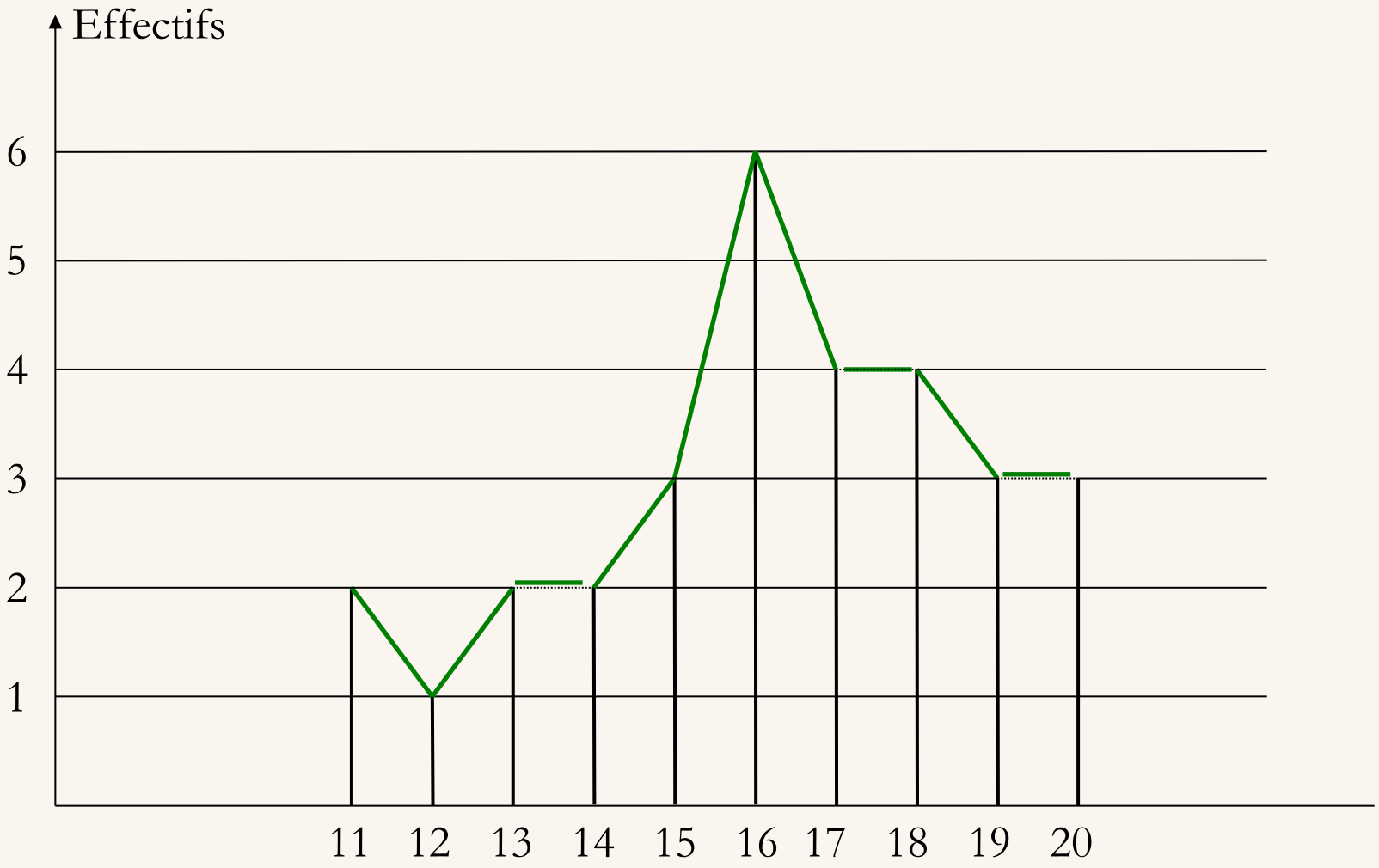


aussi avec fréquence relative

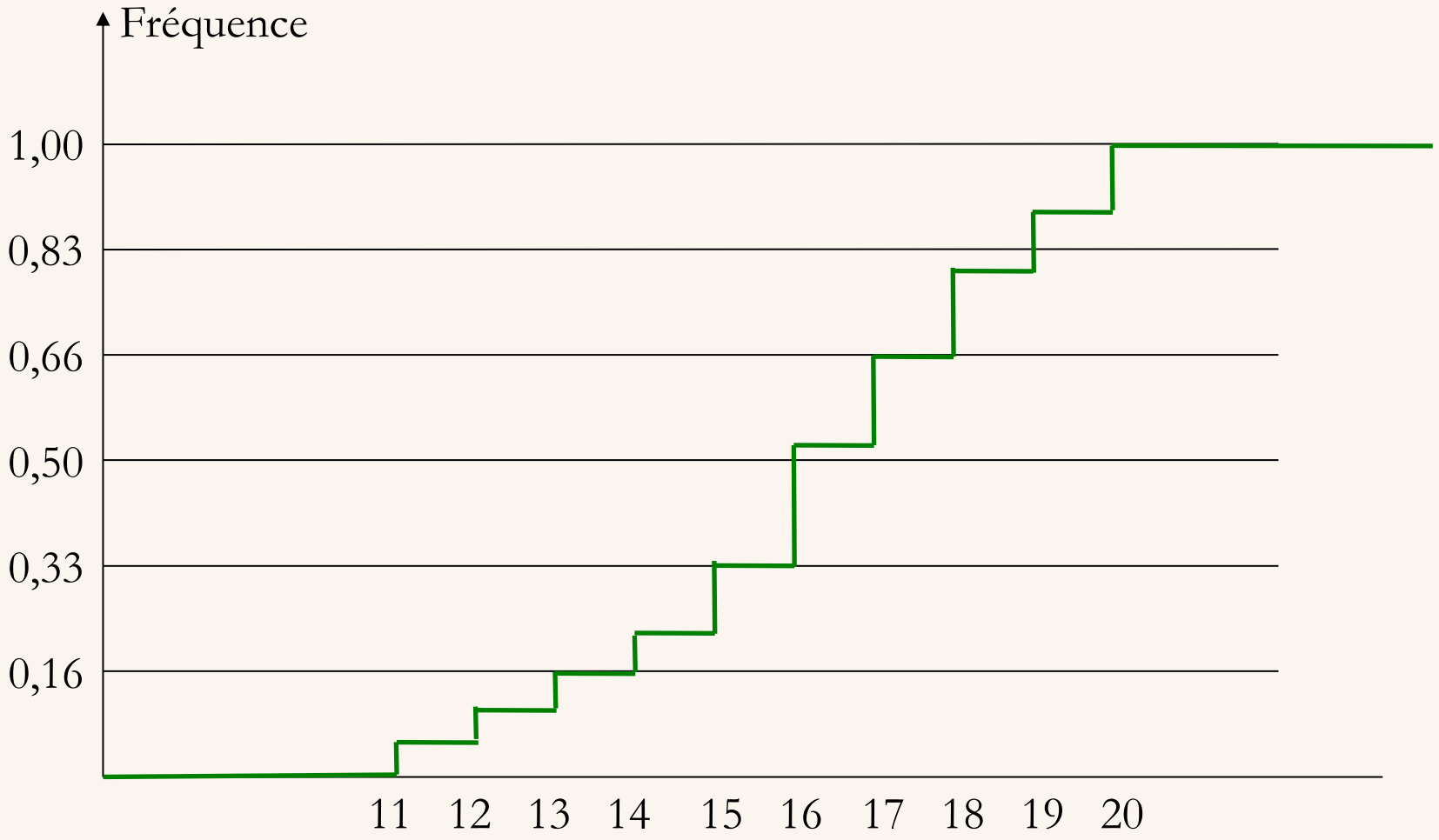
Diagramme en secteurs



Polygone des fréquences



Polygone des fréquences cumulées



Paramètres numériques de réduction de données

- Médiane & Quartiles
- Mode(s)
- Moyenne
- Variance & Écart-type
- Étendue & Étendue interquartile

Médiane

- médiane d'un ensemble de valeurs
  - la valeur prise dans l'ensemble trié
  - qui sépare celui-ci en deux sous-ensembles de même effectif.
- interprétation
  - tendance centrale de l'échantillon
  - peu de sensibilité aux valeurs extrêmes

Médiane

- n impair → médiane = valeur centrale de l'échantillon
- n pair → médiane = moyenne des deux valeurs centrales
- Exemples

$M = x_{\left(\frac{n+1}{2}\right)}$

échantillon : (28, 14, 11, 21, 13)

échantillon trié : (11, 13, 14, 21, 28)

$M = x_{\left(\frac{5+1}{2}\right)} = x_3 = 14$

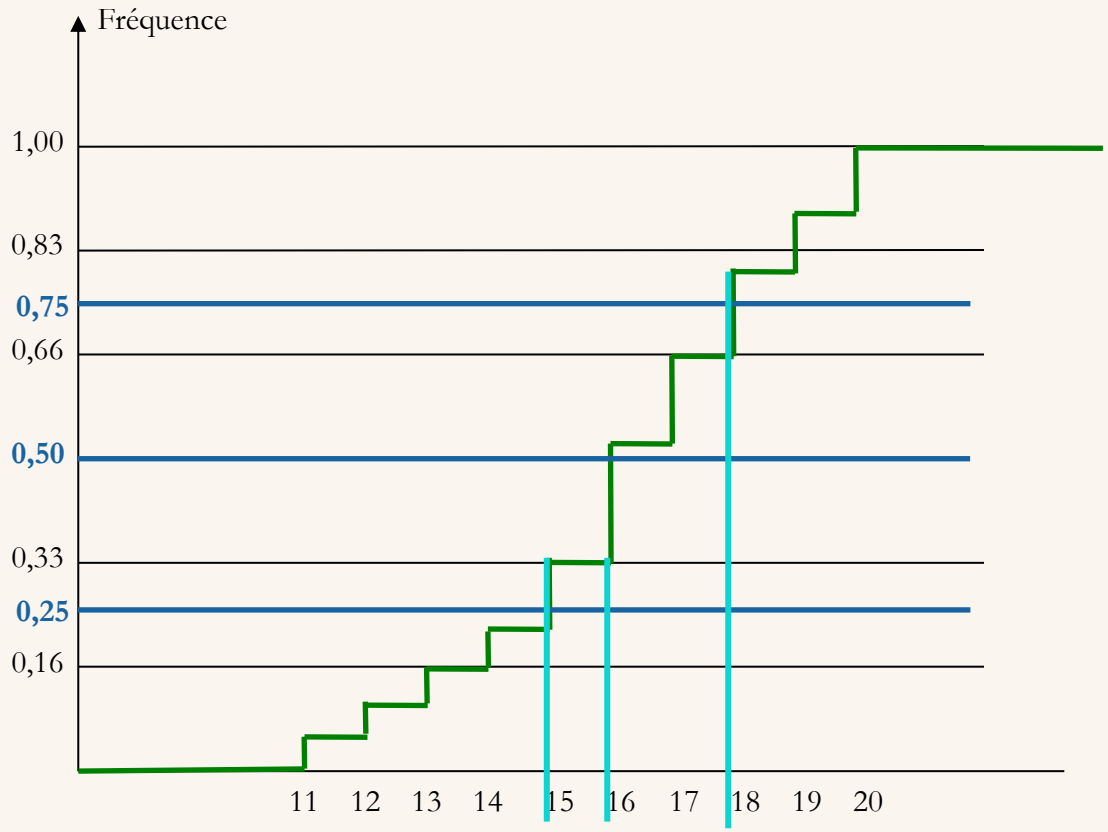
$$M = \frac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2}$$

échantillon trié : (11, 12, 13, 14, 23, 28)

$$M = \frac{13 + 14}{2} = 13{,}5$$

Quartiles

- quartiles d'un ensemble de valeurs
  - bornes divisant l'échantillon en quatre parties égales
    - extension de la médiane
  - 1er quartile (Q1)
    - médiane 1ère moitié des données
    - 25%
  - 2ème quartile (Q2)
    - médiane ensemble des données
    - 50%
  - 3ème quartile (Q3)
    - médiane 2ème moitié des données
    - 75%



quartiles : 15, 16, 18

Étendue (R) & Étendue interquartile (IQR)

- étendue
  - différence entre
    la valeur maximale et
    la valeur minimale
    de l'échantillon
- interprétation
  - une idée de la dispersion totale des données

$$R = x_n - x_1$$

- étendue interquartile
  - différence entre Q3 et Q1
- interprétation
  - mesure de la dispersion des valeurs centrales de l'échantillon

$$IQR = Q_3 - Q_1$$

Quartiles, étendue : exemple

x_i	n_i	fr_i (%)	nc_i	frc_i (%)
11	2	6,7	2	6,7%
12	1	3,3	3	10,0%
13	2	6,7	5	16,7%
14	2	6,7	7	23,3%
15	3	10,0	10	33,3%
16	6	20,0	16	53,3%
17	4	13,3	20	66,7%
18	4	13,3	24	80,0%
19	3	10,0	27	90,0%
20	3	10,0	30	100,0%

$R = 20 - 11 = 9$

$Q1 = 15$

$Q2 = 16\ (médiane)$

$Q3 = 18$

$IQR = 18 - 15 = 3$

Mode

- mode d'un ensemble de valeurs
  - la valeur la plus fréquente dans l'échantillon
- plusieurs modes possibles

Mode : exemple

x_i	n_i	fr_i	fr_i (%)	nc_i	frc_i	frc_i (%)
11	2	0,067	6,7	2	0,067	6,7%
12	1	0,033	3,3	3	0,100	10,0%
13	2	0,067	6,7	5	0,167	16,7%
14	2	0,067	6,7	7	0,233	23,3%
15	3	0,100	10,0	10	0,333	33,3%
16	**6**	**0,200**	**20,0**	**16**	**0,533**	**53,3%**
17	4	0,133	13,3	20	0,667	66,7%
18	4	0,133	13,3	24	0,800	80,0%
19	3	0,100	10,0	27	0,900	90,0%
20	3	0,100	10,0	30	1,000	100,0%

MODE : 16 (une seule valeur)

Moyenne

- moyenne d'un ensemble de valeurs
  - rapport entre
    - somme de toutes les valeurs et
    - nombre de valeurs de l'ensemble.
- interprétation
  - tendance centrale de l'échantillon
  - sensibilité aux valeurs extrêmes

$$\overline{x} = \frac{\sum_{i=1}^{n} n_i x_i}{n}$$

Variance

$$v^2 = \frac{\sum_{i=1}^{n} n_i x_i^2}{n} - m^2$$

- variance d'un ensemble de valeurs
  - somme des carrés des écarts à la moyenne pondérée par leur probabilité d'occurrence
- interprétation
  - mesure de la dispersion d'une variable par rapport à sa moyenne.
  - sensibilité aux valeurs extrêmes
- unité de la variance = carré de l'unité de l'individu
  - Par exemple, si on mesure le poids d'une population en **kg**, alors la variance s'exprime en **kg²**.

Ecart-type

- écart-type d'un ensemble de valeurs
  - racine carrée de la variance

$$s = \sqrt{v^2}$$

- interprétation
  - mesure de la dispersion d'une variable par rapport à sa moyenne.
  - sensibilité aux valeurs extrêmes
- unité de l'écart-type = unité de l'individu
  - Par exemple, si on mesure le poids d'une population en **kg**, alors l'écart-type s'exprime également en **kg**.

Exemple

x_i	n_i	$n_i x_i$	$n_i x^2_i$
11	2	22	242
12	1	12	144
13	2	26	338
14	2	28	392
15	3	45	675
16	6	96	1536
17	4	68	1156
18	4	72	1296
19	3	57	1083
20	3	60	1200
TOTAL	**30**	**486**	**8062**

$$m = 486 / 30 = 16,2$$

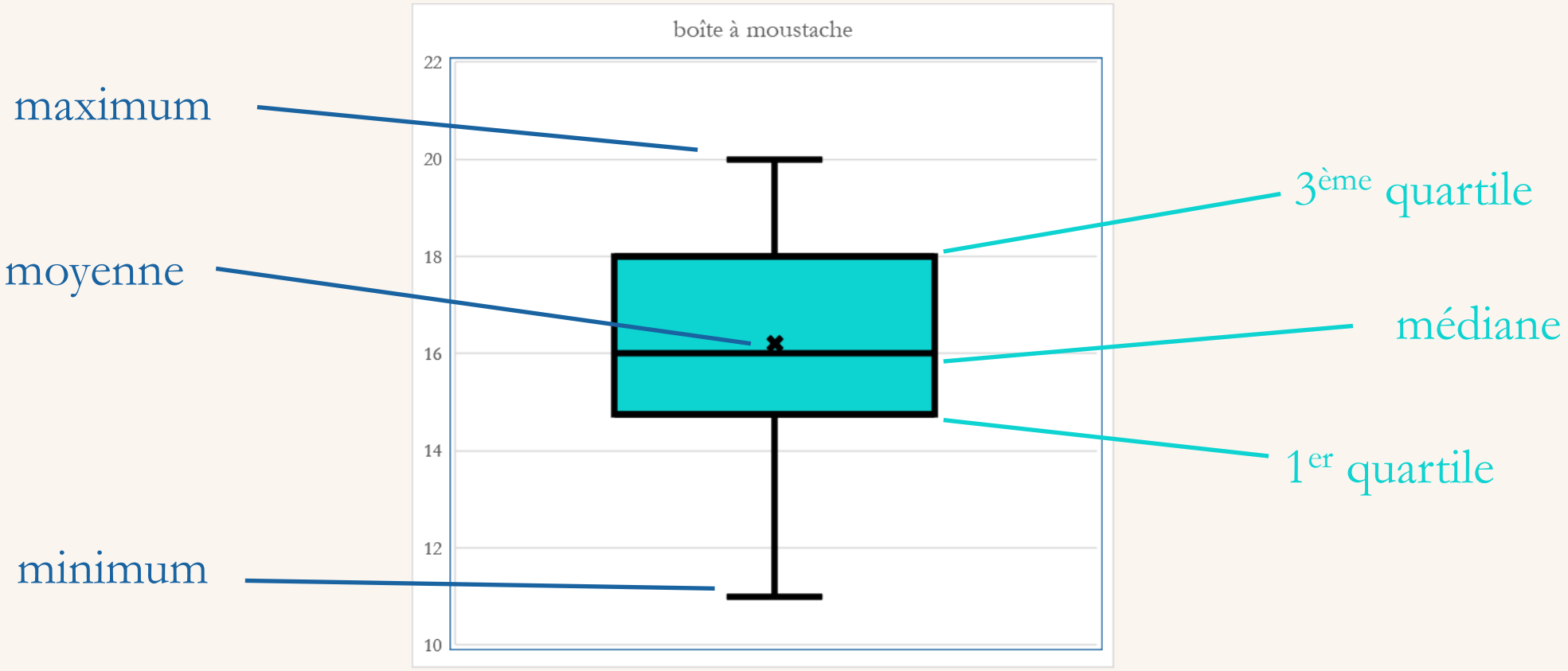
$$v^2 = (8062 / 30) - 16,2^2$$

$$= 268,73 - 262,44$$

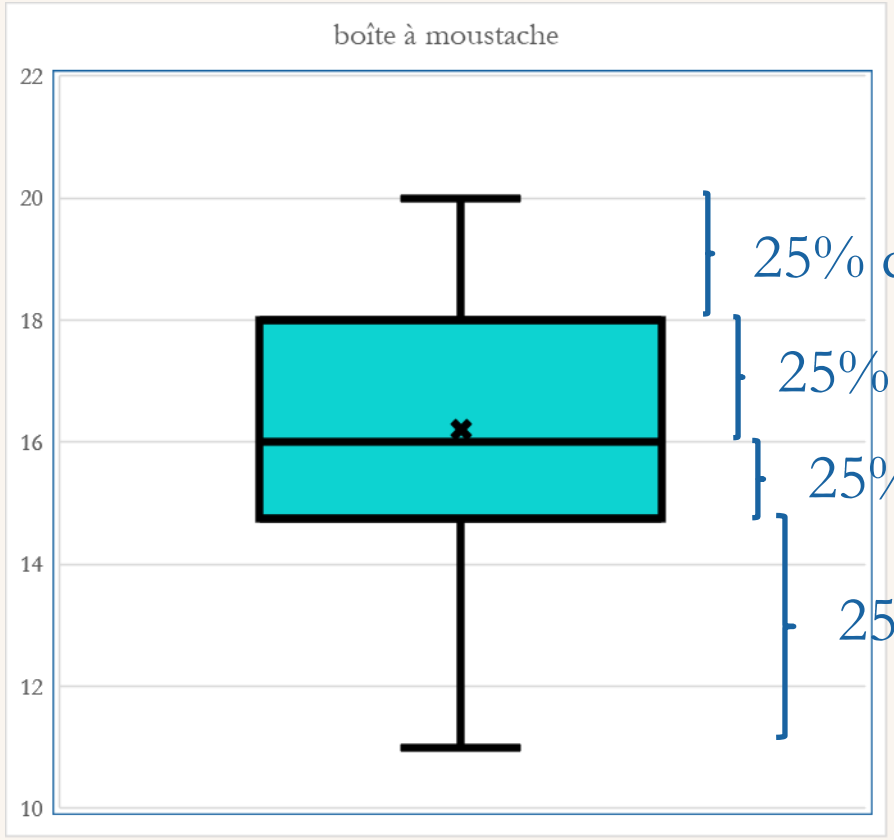
$$= 6,29$$

$$s = \sqrt{v^2} = \sqrt{6,29} = 2,51$$

Boîte à moustache



Boîte à moustache



Caractère quantitatif continu

toutes les valeurs sont possibles dans un intervalle

Exemple

- Un groupe de 50 personnes est décrit suivant le caractère quantitatif « salaire semestriel » (en millier d'euro) :

13,2	13,3	13,4	13,5	13,7	13,9	14,0	14,1	14,1	14,2
14,2	14,3	14,4	14,6	14,7	14,9	15,0	15,0	15,1	15,1
15,1	15,2	15,2	15,3	15,3	15,3	15,5	15,5	15,7	15,7
15,8	15,9	16,1	16,1	16,2	16,3	16,5	16,5	16,6	16,6
16,7	16,8	16,9	17,1	17,2	17,2	17,3	17,5	17,6	17,6

Remarque

nombre de données différentes trop élevé

impossible de présenter les données comme dans les exemples précédents

Solution : grouper les valeurs en classe

Classe	C_i	n_i
[13,14[	13,5	6
[14,15[	14,5	10
[15,16[	15,5	16
[16,17[	16,5	11
[17,18[	17,5	7

Combien de classes ? Règle de Sturges

- Quand on pense que les données sont réparties d'une manière symétrique et sans valeur aberrante, le règle de Sturges permet d'estimer le nombre de classes nécessaires.
- Notons k, le nombre de classes à calculer et N, le nombre de données de l'échantillon.
- Règle de Sturges

$$k = 1 + \frac{10}{3} \log_{10} N$$

- Exemple : Si nous disposons de 50 données, alors nous pourrions travailler avec (1 + 10 * log 50 / 3) = 5 ou 6 classes.

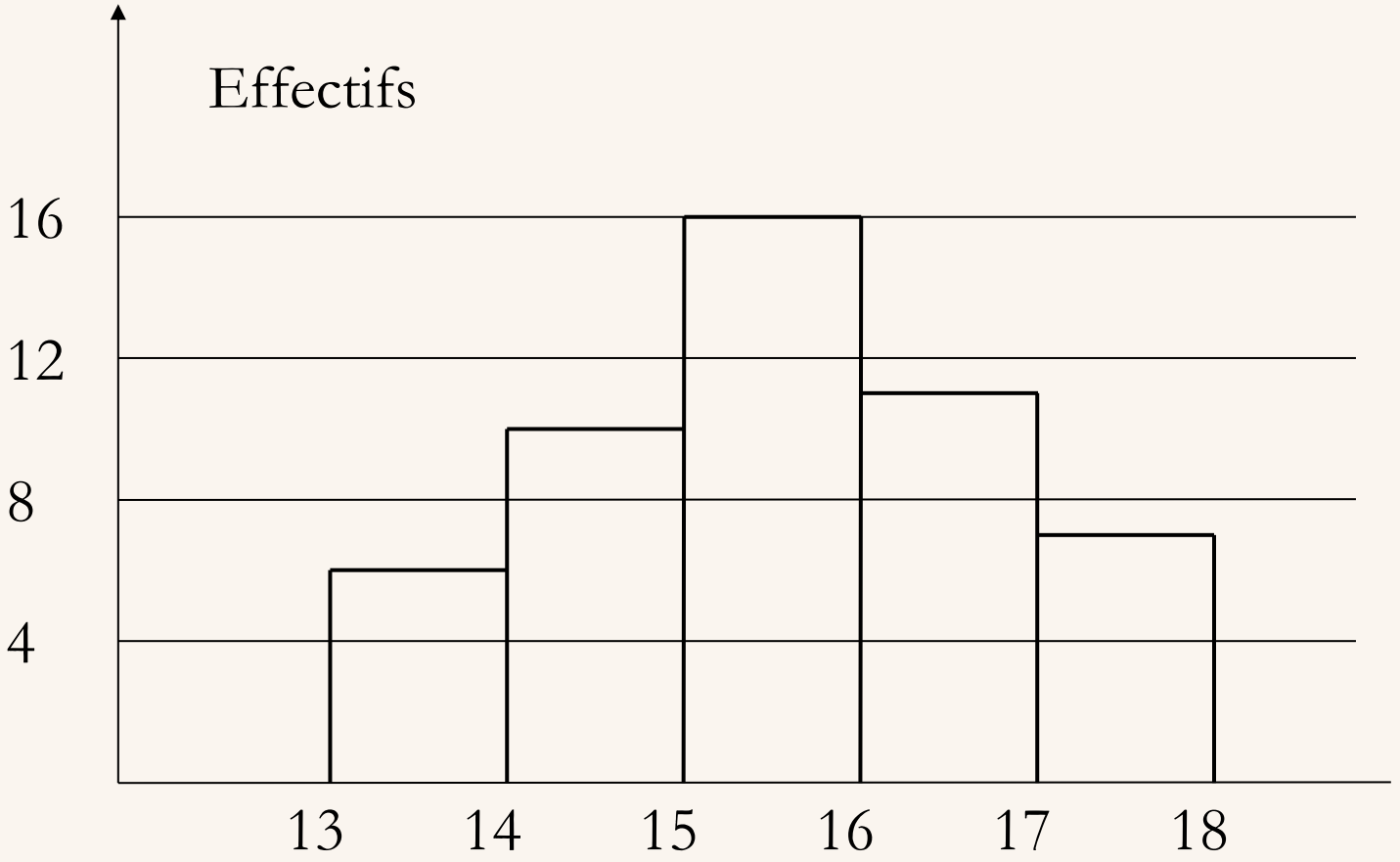
Définitions

- fréquence absolue d'une classe C_i
  = nombre n_i d'observations appartenant à la classe C_i
- fréquence relative d'une classe C_i
  = rapport n_i/n
  où n est le nombre de valeurs dans l'ensemble
- fréquence absolue cumulée d'une classe C_i
  = somme des fréquences absolues de cette classe et des classes inférieures.
- fréquence relative cumulée d'une classe C_i
  = somme des fréquences relatives de cette classe et des classes inférieures.

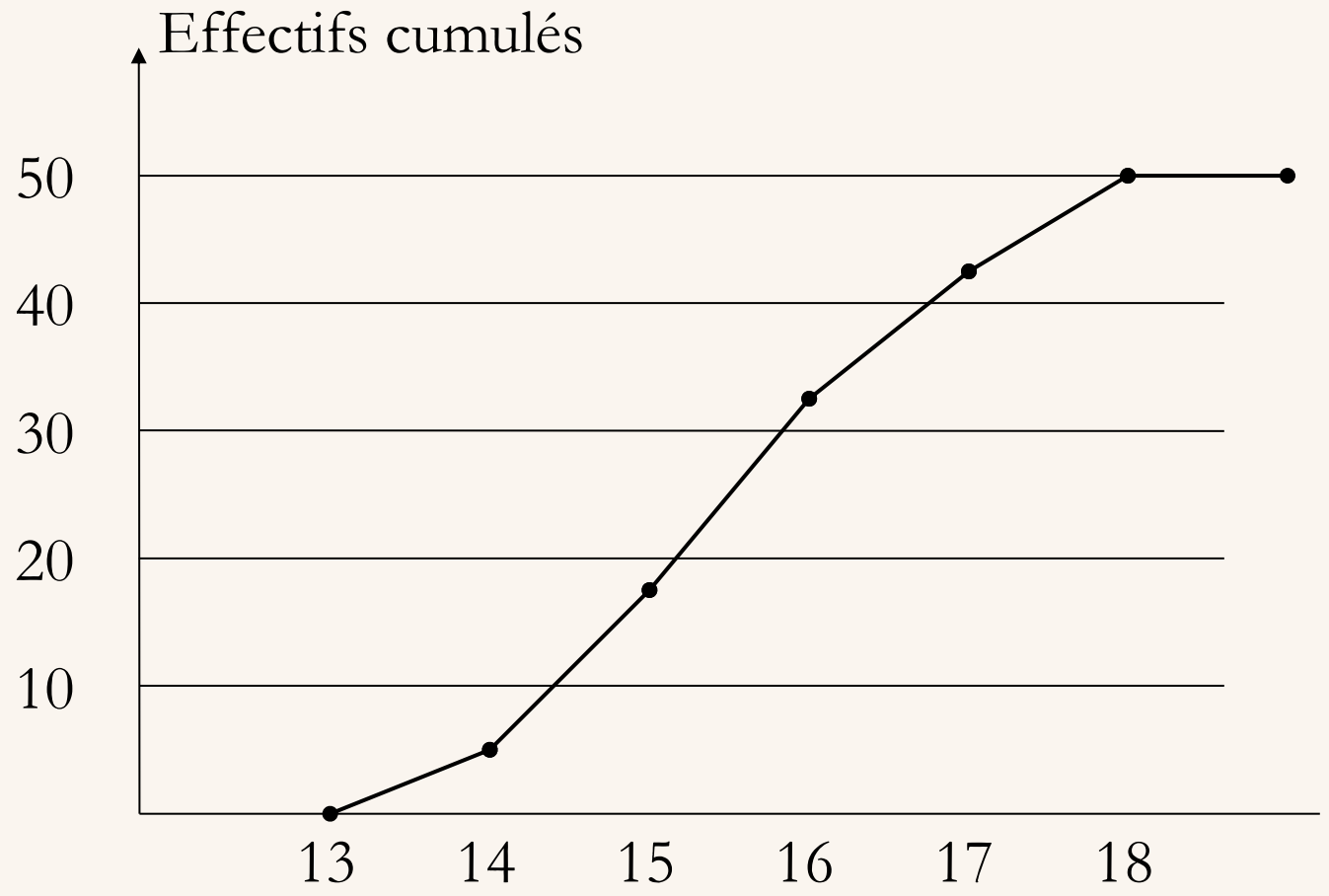
Exemple

Classe	C_i	n_i	fr_i	fr_i (%)	nc_i	frc_i	frc_i (%)
[13,14[	13,5	6	0,12	12	6	0,12	12%
[14,15[	14,5	10	0,20	20	16	0,32	32%
[15,16[	15,5	16	0,32	32	32	0,64	64%
[16,17[	16,5	11	0,22	22	43	0,86	86%
[17,18[	17,5	7	0,14	14	50	1,00	100%

Histogramme



Graphique des effectifs cumulés



Classe médiane

- médiane d'un ensemble de valeurs
  (réparties en classes et triées)
  = classe qui sépare l'ensemble
  en deux sous-ensemble de même effectif.

Étendue & Étendue interquartile

- étendue
  - différence entre la valeur maximale et la valeur minimale de l'échantillon
- étendue interquartile
  - différence entre Q3 et Q1

$$R = C_n - C_1$$

$$IQR = Q_3 - Q_1$$

Classe modale

- classe modale
  = la classe la plus représentée d'un ensemble de classes
- Il peut exister plusieurs classes modales.

Exemple

Classe	C_i	n_i	fr_i	fr_i (%)	nc_i	frc_i	frc_i (%)
[13,14[	13,5	6	0,12	12%	6	0,12	12%
[14,15[	14,5	10	0,20	20%	16	0,32	32%
[15,16[	15,5	**16**	**0,32**	**32%**	32	**0,64**	**64%**
[16,17[	16,5	11	0,22	22%	43	0,86	86%
[17,18[	17,5	7	0,14	14%	50	1,00	100%

CLASSE MODALE : [15,16[

CLASSE MEDIANE : [15,16[

$$R = C_n - C_1 = 17{,}5 - 13{,}5 = 4$$

$$IQR = Q_3 - Q_1 = 16{,}5 - 14{,}5 = 2$$

Moyenne

- Moyenne d'un ensemble de valeurs
  = rapport entre la somme de toutes les valeurs et le nombre de valeurs de l'ensemble.

$$m = \frac{\sum_{i=1}^{n} n_i C_i}{n}$$

Variance & Ecart-type

- variance d'un ensemble de valeurs
  - somme des carrés des écarts à la moyenne pondérée par leur probabilité d'occurrence

$$v^2 = \frac{\sum_{i=1}^{n} n_i C_i^2}{n} - m^2$$

- écart-type d'un ensemble de valeurs
  - racine carrée de la variance

$$s = \sqrt{v^2}$$

Exemple

Classe	C_i	n_i	n_iC_i	$n_iC^2_i$
[13,14[	13,5	6	81,0	1093,50
[14,15[	14,5	10	145,0	2102,50
[15,16[	15,5	16	248,0	3844,00
[16,17[	16,5	11	181,5	2994,75
[17,18[	17,5	7	122,5	2143,75
TOTAL		50	778,0	12178,50

$$m = 778 \,/\, 50 = 15{,}56$$

$$v^2 = (12178{,}5 \,/\, 50) - 15{,}56^2$$

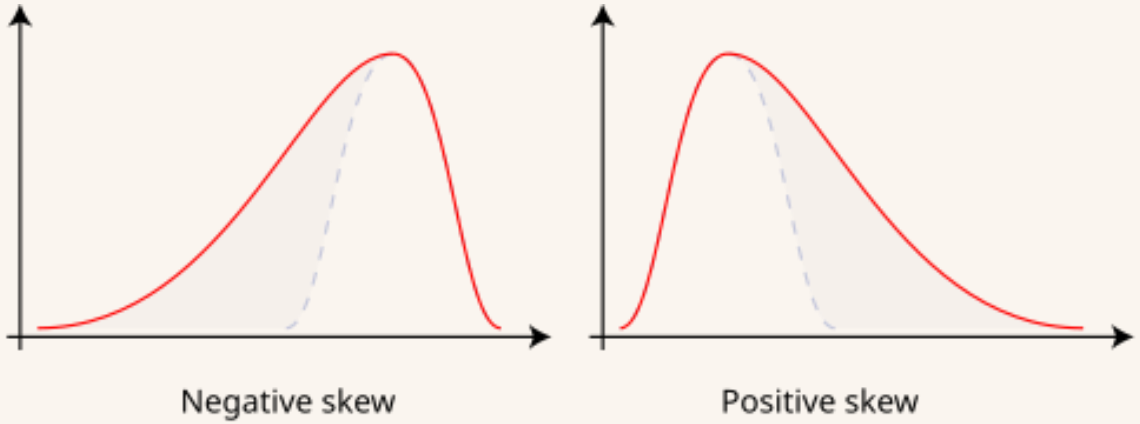
$$= 243{,}57 - 242{,}11$$

$$= 1{,}46$$

$$s = \sqrt{v^2} = \sqrt{1{,}46} = 1{,}208$$

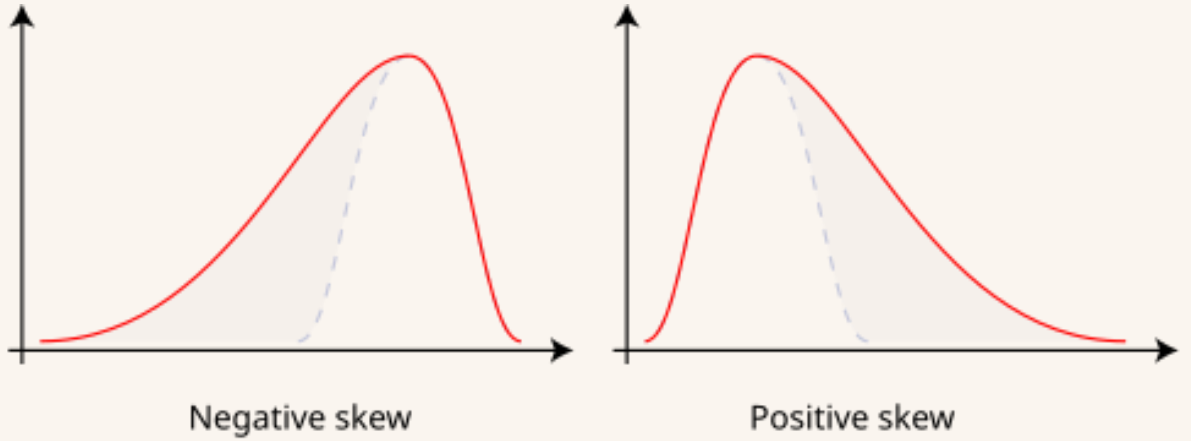
Asymétrie *(skewness)*

Asymétrie : tendance de la distribution à être décalée vers la gauche ou vers la droite par rapport à sa moyenne.



- distribution symétrique
  - moyenne, médiane et mode alignés
  - mode = médiane = moyenne
- distribution asymétrique à droite
  - "asymétrie positive"
  - queue longue vers la droite
  - mode < médiane < moyenne
- distribution asymétrique à gauche
  - "asymétrie négative"
  - queue longue vers la gauche
  - mode > médiane > moyenne

Asymétrie *(skewness)*



- Coefficient s'asymétrie

$$\gamma_1 = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{x_i - \mu}{\sigma}\right)^3$$

- Tableur : `COEFFICIENT.ASYMETRIE.P(plage de données)`
- D'autres coefficients existent
- Ces règles sont de très bons indicateurs mais ne sont pas absolues.

Inégalité & Indice de Gini

- L'indice de Gini mesure l'inégalité d'une distribution univariée.
- Intuitivement, c'est la différence moyenne entre deux valeurs tirées au hasard dans la population, rapportée à ce qu'elle serait au maximum.
- Indice de Gini :

$$G = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} |x_i - x_j|}{2n^2 \bar{x}}$$

Inégalité & Indice de Gini





Valeur de G

G vaut toujours entre 0 et 1 (%)

0 = Égalité parfaite : tout le monde a la même valeur

1 = Inégalité maximale : une personne concentre tout

Exemple

Deux personnes avec revenus de 0 et 100.

- Moyenne = 50.
- somme écarts absolus = 200
- G = 0,5

Deux personnes avec revenus de 50 et 50.

- Moyenne = 50
- somme écarts absolus = 0
- G = 0

Si l'on ajoute des individus à 0 : G monte vers 1.