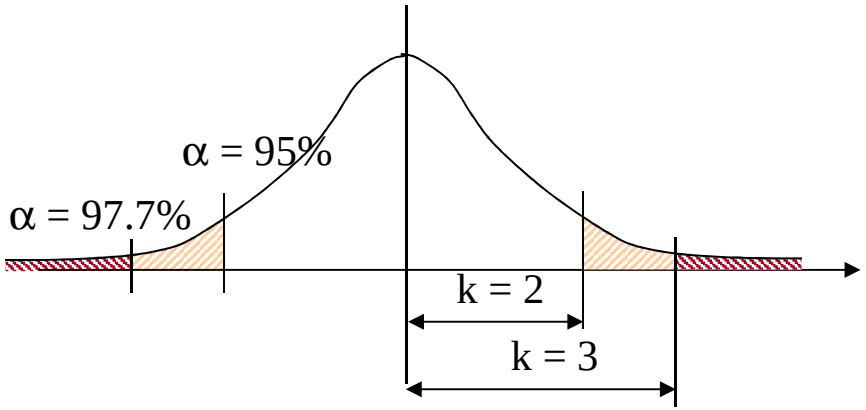
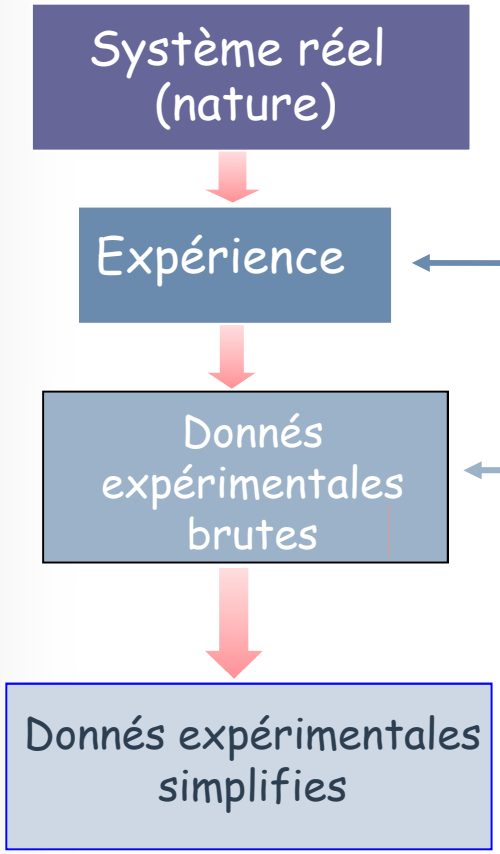

UNIVERSITE
MONTPELLIER

Cours no.4.
Statistique descriptive.



Démarche scientifique



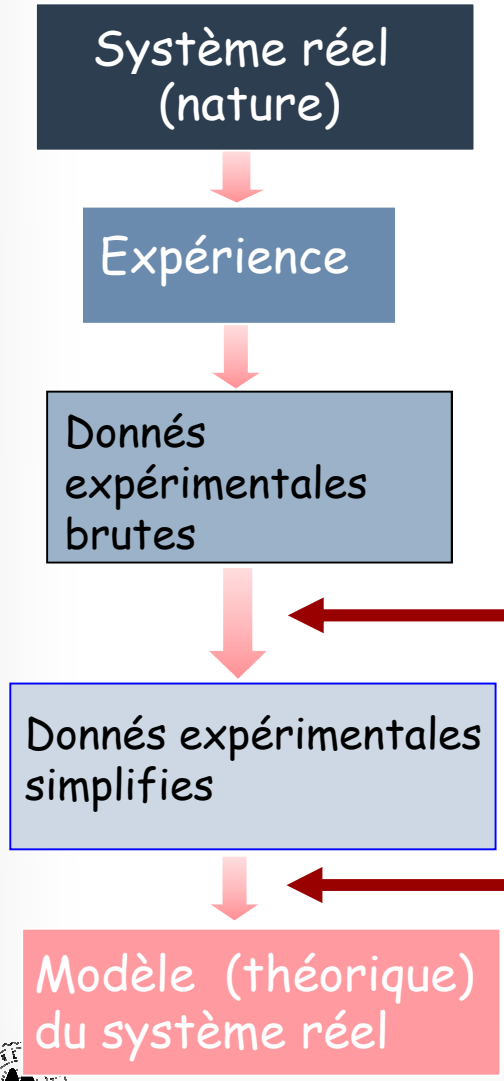
Nous avons à étudier un ensemble bien défini mais assez vaste d'individus ou d'observations.

Dans un premier temps, on mesure certaines caractéristiques de la population originale.

Les mesures sont classées et organisées (en tableaux ou en graphiques).

La seconde étape consiste, en fonction du problème posé ou d'une décision à prendre, à simplifier cette information à des diagrammes concis ou des coefficients simples et significatifs.

Statistiques et la démarche scientifique.



statistique descriptive

statistique inferentielle

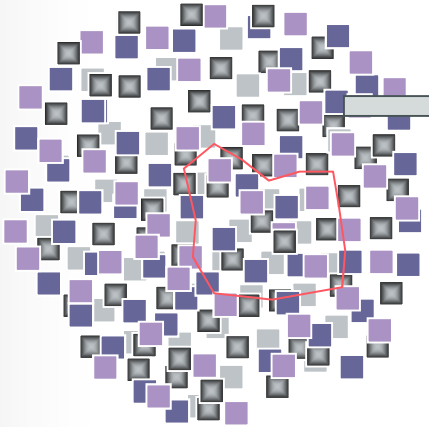
pour interpréter une masse de données

il faut accepter de perdre de l'information

au profit de la meilleure compréhension.

Il est alors très important de comprendre comment la simplification s'est opérée.

Vocabulaire général - rappels.



Population

un ensemble d'objets
soumis à une étude statistique:

- finie ou non-finie ni dénombrable
- réalité tangible ou hypothétique.

Échantillon

un nombre limité d'individus
d'une population

il doit être représentatif
d'une population pour qu'on puisse
extrapoler les résultats de son
étude vers la population dont
il est issue.

Individu

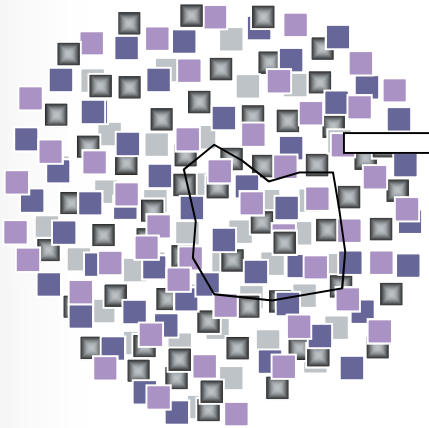
(ou unité statistique,
unité d'observation) –
un élément de la population.

Caractère

propriété (caractéristique):

- qualitatif (modalité)
  ou quantitatif (valeur);
- déterministe ou aléatoire;
- continu ou discret.

Vocabulaire général - rappels.



Population

un ensemble d'objets
soumis à une étude statistique:

?

Recensement

une interrogation exhaustive
de tous les individus
d'une population.

Échantillon

un nombre limité d'individus
d'une population

?

Sondage

une interrogation
de l'échantillon.

Individu

(ou unité statistique,
unité d'observation) –
un élément de la population.

?

Mesure

Série statistique.

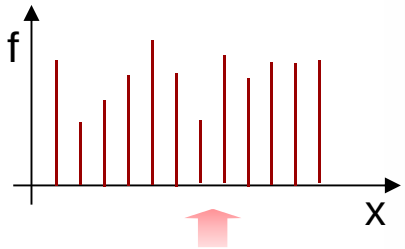
résultat d'un sondage concernant un caractère des individus d'une population

EXEMPLE : On a étudié, sur un échantillon de 1000 personnes, les habitudes alimentaires des clients de restauration rapide. 200 ont préféré y prendre leur petit déjeuner, 300 le déjeuner, 100 le dîner et 400 les snacks.

On étudie un caractère X d'une population. Dans un sondage d'un échantillon de **n** individus, $\mathbf{x_i}$ a été observée $\mathbf{n_i}$ fois.

- $\mathbf{n_i}$ - effectif de x_i
- $\mathbf{f_i}$ - fréquence de x_i :

$$f_i = \frac{n_i}{n} \Longrightarrow \begin{cases} \bullet \quad 0 \le f_i \le 1 \\ \bullet \quad \sum_{i=1}^{k} f_i = 1 \end{cases}$$



distribution de fréquences

Analyse de la fréquence facilite la comparaison des échantillons de tailles différentes

Distribution de fréquence.

diagramme en bâtons

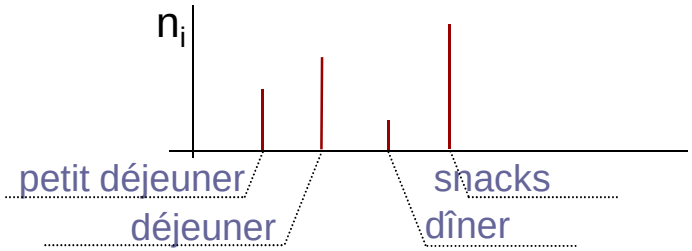


diagramme en barres

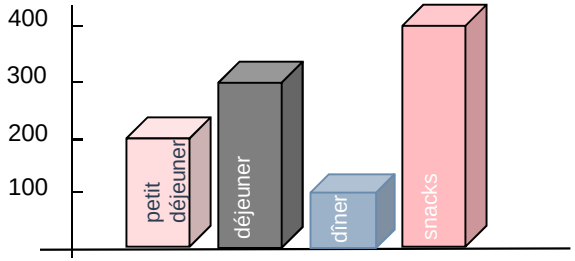
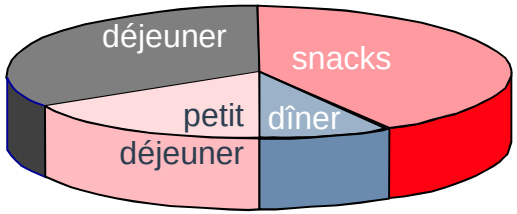


diagramme à secteurs

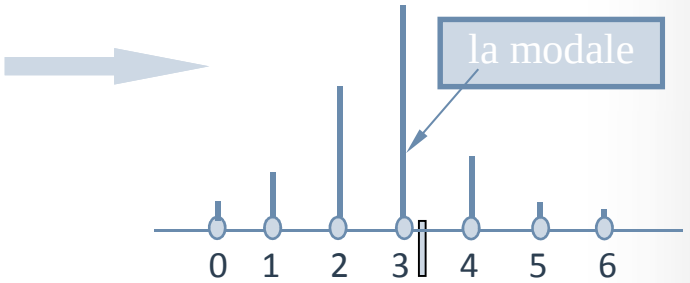


Analyse de la fréquence facilite la comparaison des échantillons de tailles différentes

Distribution de fréquence.

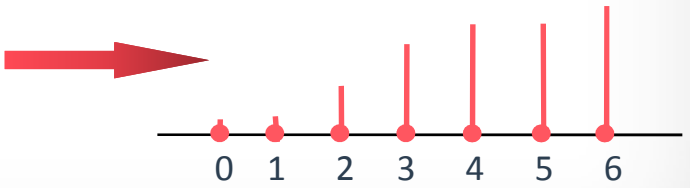
EXEMPLE : On a étudié, sur un échantillon de 1000 familles françaises, le nombre d'enfants à charge. Représenter d'une manière synthétique le résultat de cette étude.

x_i	n_i	f_i	F_i
0	50	.05	.05
1	100	.1	.15
2	250	.25	.4
3	440	.44	.84
4	137	.137	.977
5	20	.02	.997
6	3	.003	1



❍ F_i - **fréquence cumulée** de x_i :

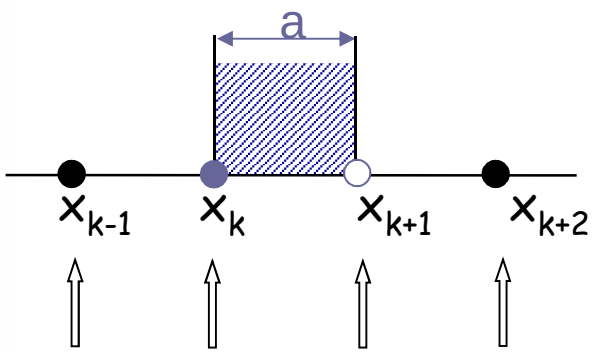
$$\forall (i = 1,2 \ldots k) \qquad F_i = \sum_{j=1}^{i} f_j = \sum_{j \le i} f_j$$



Distribution de fréquence.

Cas particulier : résultat d'un GRAND sondage

Si le nombre d'observations est fini, mais très grand,
on répartit les valeurs de x_i en **classes**:



- exhaustives et disjointes ;
- au moins 5, pas plus de 25;
- effectif de chaque classe $n_i \geq 5$;

Bornes des classes :

- **naturelles**
  (p.ex. les notes délimitant les mentions, tranches d'impôt sur revenus...)
- classes à **amplitudes égales** :

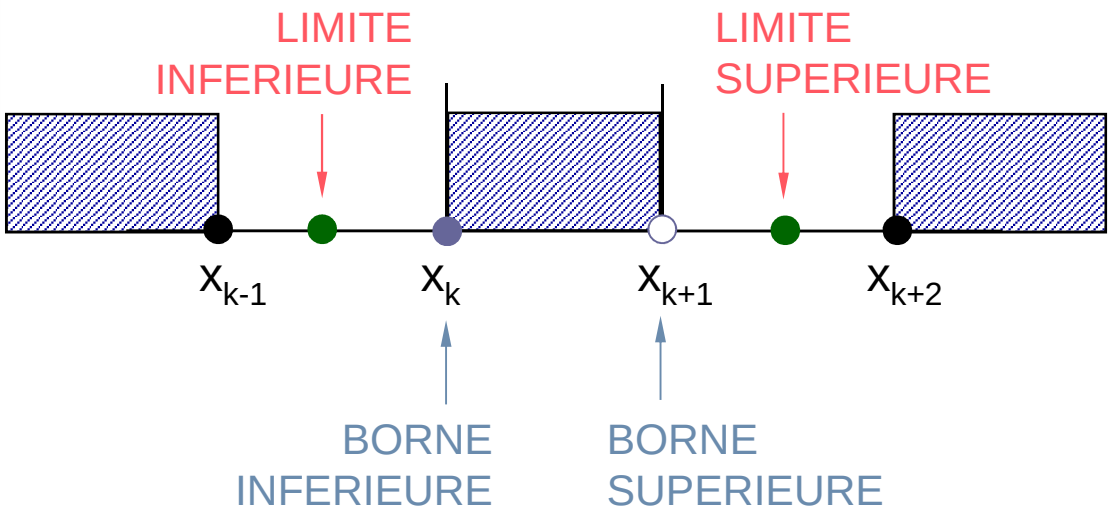
  amplitude de classe $a = \dfrac{x_{max} - x_{min}}{k}$
- classes à **fréquences égales** :

  $f_1 = f_2 = f_3 = ... = f_k$

Distribution de fréquence.

Cas particulier : résultat d'un GRAND sondage

Limites réelles d'une classe :



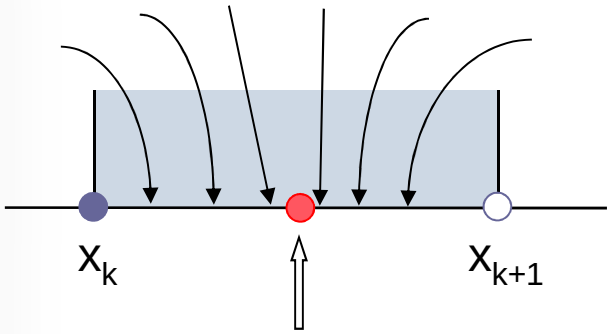
$$\text{limite inférieure} = \frac{x_k + x_{k-1}}{2}$$

$$\text{limite supérieure} = \frac{x_{k+2} + x_{k+1}}{2}$$

Distribution de fréquence.

Cas particulier: résultat d'un GRAND sondage

centre de classe :



centre de classe

- on ne distingue pas des valeurs inscrites dans la même classe ;
- à chaque valeur est attribuée une valeur unique, appelée 'centre de classe' :

$$x_{centre} = \frac{x_{k+1} + x_k}{2}$$

Autres paramètres :

- effectif de la classe $c_i \rightarrow \mathbf{n_i}$; (effectif total: $n = \sum_{i=1}^{k} n_i$)
- fréquence de la classe $\mathbf{f_i}$: $f_i = \frac{c_i}{n}$
- fréquence cumulée de la classe $\mathbf{F_i}$: $F_i = \sum_{j \le i} f_i$

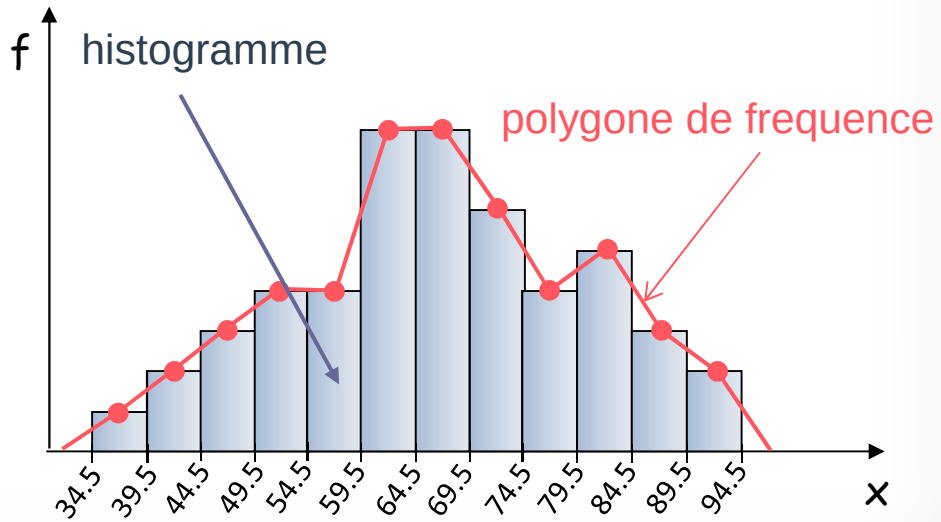
Distribution de fréquence.

Cas particulier: résultat d'un GRAND sondage

EXEMPLE: Les données ci-dessus représentent (en unités arbitraires) le niveau de pollution à l'ozone à Marseille durant 50 jours consécutifs. Représenter d'une manière synthétique le résultat de cette étude.

37	42	44	47	48	52	90	54	56	55	53	92	60	61	73	75	72	46	62	60	59	58	76	71	63
62	63	67	64	64	68	83	85	86	88	67	65	66	68	69	66	70	72	74	82	78	80	79	80	81

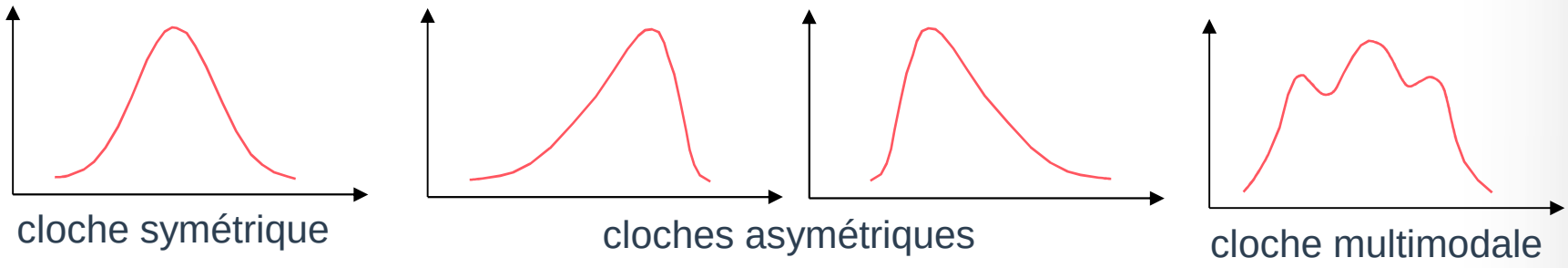
classe	x_{centre}	n_i	f_i
35-39	37	1	.02
40-44	42	2	.04
45-49	47	3	.06
50-54	52	4	.08
55-59	57	4	.08
60-64	62	8	.16
65-69	67	8	.16
70-74	72	6	.12
75-79	77	4	.08
80-84	82	5	.10
85-89	87	3	.06
90-94	92	2	.04



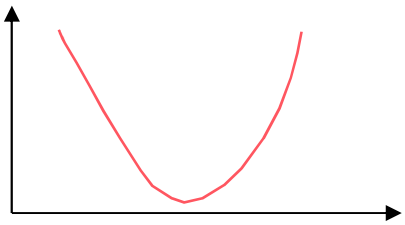
Distributions de fréquence.

Les courbes de fréquence réelles (observées lors d'une étude) appartiennent TOUJOURS à une de 3 catégories:

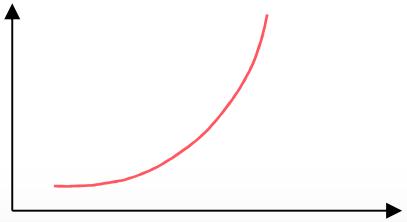
distributions en cloche



distribution en U

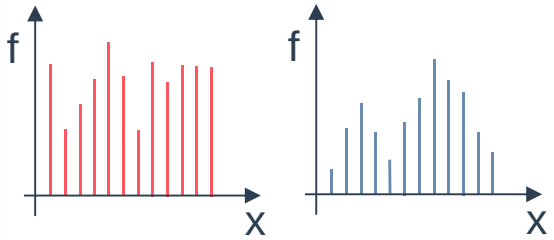


distribution en J



Description des distributions de fréquence.

Une fois un nombre suffisant de résultats expérimentaux récolté,
nous pouvons facilement les représenter sous forme de distribution de fréquence.



Les représentations graphiques
sont peu commodes
pour toute étude ultérieure.

Par contre, trois paramètres suffissent pour décrire totalement
l'allure de la courbe de distribution, sans avoir recours au graphe:

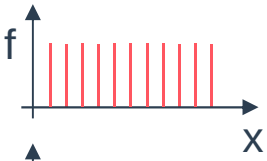
- paramètres de position (tendance centrale);
- paramètres de dispersion (de variabilité);
- paramètres de forme.

Aucune de ces grandeurs ne définit pas la distribution d'une manière exhaustive!

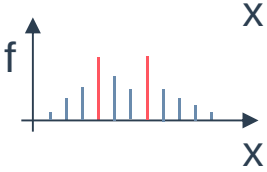
Paramètres de position (tendance centrale).

modale — la fréquence de son apparition est maximale

peut ne pas exister (valeurs équiprobables)



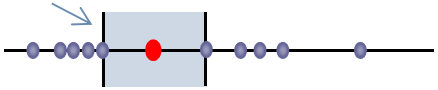
peut ne pas être unique (bi- ou plurimodalité)



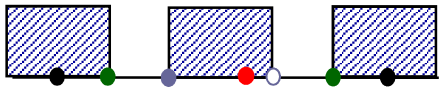
médiane — partage une série ordonnée en deux parties d'effectifs égaux.

si le nombre de mesures n est impair, la médiane est la valeur indexée n/2 + 0.5

intervalle médian



si n est pair, la médiane est la moyenne de deux valeurs du milieu.



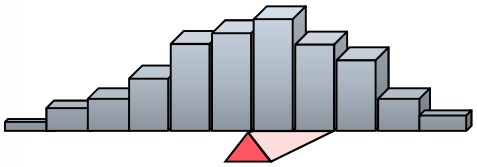
si les mesures sont groupées en classes, la médiane est:

$$Me = L + \frac{e_{median}}{f_{median}}\left(\frac{n}{2} - F_{median}\right)$$

Paramètres de position (tendance centrale).

moyenne (arithmétique)

- variable individuelle : …… $\bar{x} = E(x) = \mu = \frac{1}{n}\sum_{i=1}^{n} x_i$
- variable discrète : si **X** peut prendre **k** valeurs différentes (**k ≤ n**) et la fréquence de la valeur x_i est f_i, .. $\bar{x} = \sum_{i=1}^{k} f_i x_i$



- variable groupée en **k classes** : …… $\bar{x} = \sum_{i=1}^{k} f_i x_{(centre)i}$
- variable continue : si **X ∈]a, b[** , ……. $\bar{x} = \int_{a}^{b} x f(x) dx$

où f(x) – fonction de distribution de fréquence.

La moyenne arithmétique est fortement affectée par des valeurs extrêmes !!!

EXEMPLE : Quatre cadres dans une entreprise gagnent 2500, 3200, 3700 et 48000 euros par mois. Quel est le salaire moyen d'un cadre dans cette entreprise?

⇨ $\bar{x}$ **= 14 350 euro**

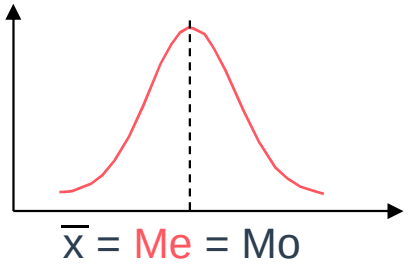
Paramètres de position.

Il existe une relation empirique entre les trois paramètres de position:

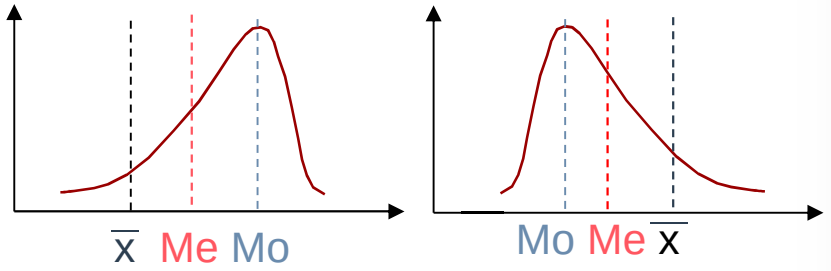
- la moyenne arithmétique $\overline{x}$,
- la médiane Me,
- la modale Mo:

$$\overline{x} - Mo = 3(\overline{x} - Me)$$

cloche symétrique



distribution en cloche asymétrique

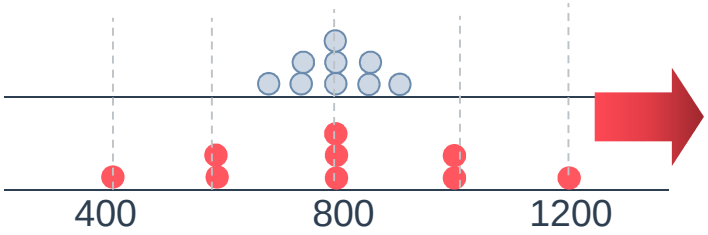


Paramètres de dispersion (variabilité).

On a mesure la durée de vie (en heures) de deux séries de 9 ampoules :

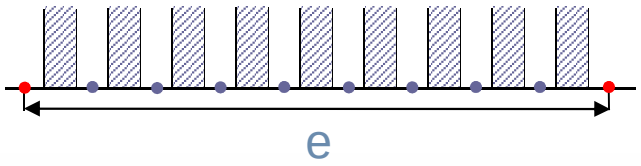
1ere série : 780 790 790 800 800 800 810 810 820

2nde série : 400 600 600 800 800 800 1000 1000 1200



Les deux séries ont même médiane, même modale, même moyenne, et pourtant elles sont très différentes.

étendue : différence entre la plus grande et la plus petite valeur prise par le caractère: $e = x_{max} - x_{min}$



Pour des données groupées en classes, étendue est égale à la différence entre la limite supérieure de la dernière classe et la limite inférieure de la première.

Paramètres de dispersion (variabilité).

quartile - valeur du caractère correspondant à :

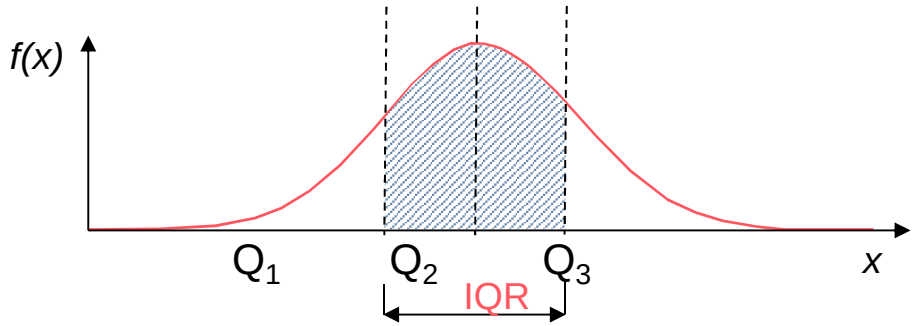
premier quartile Q_1 - 25% des effectifs cumulés croissants;

deuxième quartile Q_2 - 50% des effectifs cumulés croissants;

troisième quartile Q_3 - 75% des effectifs cumulés croissants;

distance interquartile (IQR, interquartile range) –

étendue de valeurs entre premier et troisième quartile, contenant 50% des valeurs



On défini aussi des déciles, centiles (ou percentiles, milliles....)

Paramètres de dispersion (variabilité).

variance — moyenne des carrés des écarts des x_i de la valeur moyenne :

$$V(x) = \overline{x^2} - \overline{x}^2$$

- variable individuelle : …… $V(x) = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2$
- variable discrète : si **X** peut prendre **k** valeurs différentes (**k ≤ n**) et la fréquence de la valeur x_i est f_i, .. $V(x) = \sum_{i=1}^{n} f_i\left(x_i - \overline{x}\right)^2$
- variable groupée en **k classes** : …… $V(x) = \sum_{i=1}^{k} f_i (x_{(centre)i} - \overline{x})^2$

  erreur de regroupement → correction d'Alan Sheppard:

  $V(x)_{vrai} = V(x)_{calcule} - a^2/12$

  a - amplitude des classes

- variable continue : si **X ∈]a, b[** , ……. $V(x) = \int_a^b (x - \overline{x})^2 f(x)dx$

  où f(x) – fonction de distribution de fréquence.

Paramètres de dispersion (variabilité).

Théorème de Koenig-Huyghens :

$$V(x) = \overline{x^2} - \bar{x}^2$$

écart type

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

ATTENTION !

Calculatrices scientifiques possèdent une touche « σ » ou « σ_{n-1} » qui correspond à l'estimation de s dans un échantillon de taille n :

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

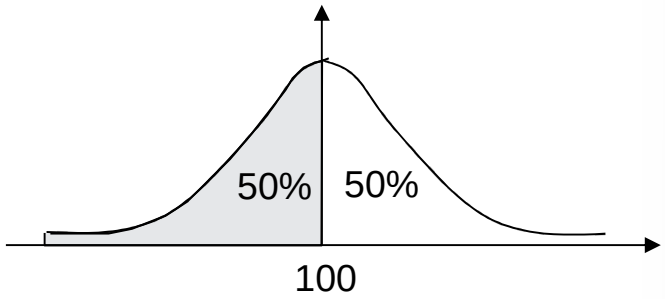
Si nombre de mesures n est grand, n ≈ n-1 et σ(x) ≈ s .

Paramètres de dispersion (variabilité).

EXEMPLE :
Supposons, que chaque pot de confiture doit contenir 100g de produit.
En prenant un échantillon à la sortie de la chaîne de remplissage,
le fabricant peut mesurer la dispersion de sa production: il trouve S = 2g.

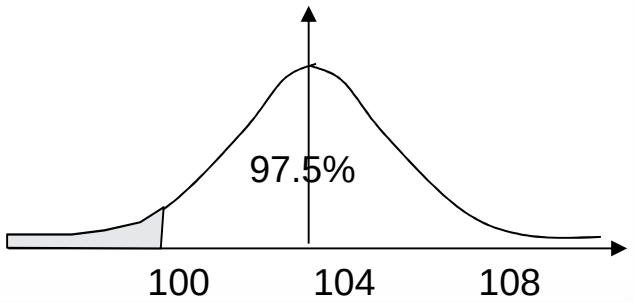
CONCLUSION 1:

il décide de régler ses machines afin que la production moyenne soit centrée sur 100 g.



CONCLUSION 2:

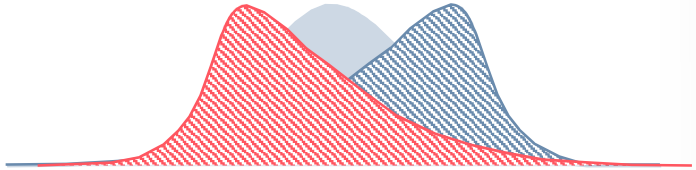
il décide de régler ses machines afin que la production moyenne soit centrée sur 104 g.



Paramètres de forme.

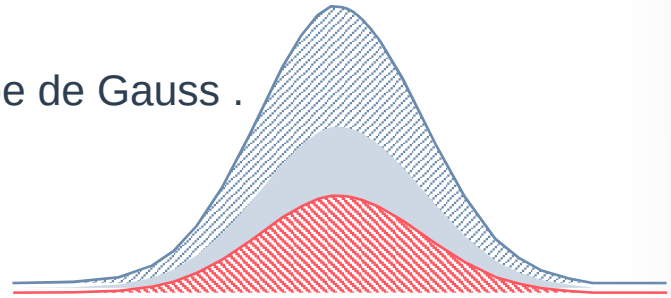
dissymétrie - coefficient γ_1 de Fischer : $\gamma_1 = \dfrac{\overline{x^3} - 3\bar{x}\overline{x^2} + 2\bar{x}^3}{\sigma^3}$

- $\gamma_1 = 0 \rightarrow$ distribution symétrique ;
- $\gamma_1 < 0 \rightarrow$ distribution étalée vers la gauche ;
- $\gamma_1 > 0 \rightarrow$ distribution étalée vers la droite.



aplatissement - coefficient γ_2 de Fischer : $\gamma_2 = \dfrac{\overline{x^4} - 4\bar{x}\overline{x^3} + 6\bar{x}^2\overline{x^2} - 3\bar{x}^4}{\sigma^4} - 3$

- $\Rightarrow$ $\gamma_2 = 0 \rightarrow$ distribution de forme gaussienne ;
- $\Rightarrow$ $\gamma_2 < 0 \rightarrow$ distribution plus aplatie que la courbe de Gauss .
- $\Rightarrow$ $\gamma_2 > 0 \rightarrow$ distribution moins aplatie



Annexe 1. Autres types de moyenne.

moyenne pondérée

On associe avec chaque valeur x_1, x_2, …, x_n un coefficient w_1, w_2, …, w_n. w_i décrit l'importance de la valeur x_i (son poids).
La moyenne pondérée de x_i est :

$$\bar{x} = \frac{x_1 w_1 + x_2 w_2 + \ldots + x_n w_n}{w_1 + w_2 + \ldots + w_n} = \frac{\sum_{i=1}^{n} x_i w_i}{\sum_{i=1}^{n} w_i}$$

EXEMPLE :
Un étudiant doit passer 2 tests, 1 partiel et 1 examen final en stats. Chaque épreuve est notée sur 100 points. Le partiel compte 3 fois plus que le test, et l'exam final 5 fois plus. L'étudiant obtient : 70 et 80 points en tests, 65 en partiel et 85 en examen final.
Quelle est sa note (sur20)?

$$\bar{x} = \frac{70 \cdot 1 + 80 \cdot 1 + 65 \cdot 3 + 85 \cdot 5}{1+1+3+5} = \frac{770}{10} = 77 \quad (\text{sur } 100)$$

$$\bar{x} = \frac{77}{5} = 15.4 \quad (\text{sur } 20)$$

Annexe 1. Autres types de moyenne.

moyenne géométrique

Soient des valeurs x_1, x_2, …x_n.
Leur moyenne géométrique est:

$$q = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n}$$

EXEMPLE : 1000 euros déposés dans une banque sont devenus 5000 euros au bout de 3 ans.
Quel est le taux d'intérêt t pratiqué par cette banque?

somme déposée : 1000

après 1 an : $1000 + 1000 \cdot \tau = (1 + \tau) \cdot 1000$

après 2ans : $(1 + \tau) \cdot 1000 + (1 + \tau) \cdot 1000 \cdot \tau = (1 + \tau)^2 \cdot 1000$

après 3 ans : $(1 + \tau)^2 \cdot 1000 + (1 + \tau)^2 \cdot 1000 \cdot \tau = (1 + \tau)^3 \cdot 1000 = 5000$

$$(1 + \tau)^3 = 5$$

$$\tau = \sqrt[3]{5} - 1 = 0.71 = 71\%$$

Annexe 1. Autres types de moyenne.

moyenne harmonique

Soient des valeurs x_1, x_2, …x_n. Leur moyenne harmonique est la valeur réciproque de la moyenne des valeurs des $1/x_i$:

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

EXEMPLE : Un cycliste a parcouru une distance de 30 km. Les 10 premiers kilomètres il a roulé avec la vitesse 35 km/h, les 10 suivants – avec la vitesse 48 km/h, et les 10 derniers – avec 40km/h. Quelle a été sa vitesse moyenne sur ce trajet?

$$\text{vitesse moyenne} = \bar{v} = \frac{\text{distance totale}}{\text{temps total}}$$

$$t_{0-10} = \frac{s_{0-10}}{v_{0-10}} = \frac{10}{35} = 0.286\,h$$

$$t_{0-10} = \frac{s_{10-20}}{v_{1-20}} = \frac{10}{48} = 0.208\,h$$

$$t_{0-10} = \frac{s_{20-30}}{v_{20-30}} = \frac{10}{40} = 0.25\,h$$

$$\bar{v} = \frac{30}{0.286+0.208+0.25} = \frac{30}{0.744} = 40.32\ \text{km/h}$$

Annexe 1. Autres types de moyenne.

moyenne quadratique

Soient des valeurs x_1, x_2, …x_n. Leur moyenne quadratique est la racine carréede la moyenne des carrés de valeurs x_i:

$$\sqrt{\overline{x^2}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i)^2}$$

COMPAREZ: Quelle est la moyenne de nombres 2, 3, 5 et 7 ?

moyenne arithmétique → $\bar{x} = \frac{2+3+5+7}{4} = \frac{17}{4} =$ **4.25**

moyenne géométrique → $\bar{x} = \sqrt[4]{2\cdot 3\cdot 5\cdot 7} = \sqrt[4]{210} =$ **3.807**

moyenne harmonique → $\bar{x} = \frac{4}{\frac{1}{2}+\frac{1}{3}+\frac{1}{5}+\frac{1}{7}} = \frac{4}{\frac{105+70+42+30}{210}} = \frac{840}{247} =$ **3.401**

moyenne quadratique → $\sqrt{\overline{x^2}} = \sqrt{\frac{2^2+3^2+5^2+7^2}{4}} = \frac{\sqrt{4+9+25+49}}{2} = \frac{\sqrt{87}}{2} =$ **4.664**