Statistique descriptive univariée

Anicet E. T. Ebou, ediman.ebou@inphb.ci

Objectifs pédagogique

- Connaître et savoir calculer les différents paramètres de position, de dispersion et les moments;
- Connaître et savoir calculer les différents paramètres de forme et d'aplatissement;
- Être capable de faire un diagramme en tiges et feuilles et les boîtes à moustaches.

01

Paramètres de position

Mode

- Le **mode** est la valeur distincte correspondant à l'effectif le plus élevé. Il est noté x_M;
- Le mode peut être calculé pour tous les types de variable, quantitative et qualitative;
- Le mode n'est pas nécessairement unique;
- Quand une variable continue est découpée en classes, on peut définir une classe modale (classe correspondant à l'effectif le plus élevé).

Mode

Quel est le mode de la variable ‘espece-arbre’?

X_j (Espèce d’arbre)	N_j (Effectif)	F_j (Fréquence)
AA	7	0.35
CP	5	0.25
TG	3	0.15
TI	5	0.25
n = 20		1

Moyenne arithmétique

La moyenne est la somme des valeurs observées divisée par leur nombre:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_i + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

La moyenne peut aussi être calculée à partir des valeurs distinctes et des effectifs:

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

Moyenne arithmétique

La moyenne ne peut être définie que sur une variable quantitative.

Par exemple, les nombres d'enfants de 8 familles sont les suivants 0, 0, 1, 1, 1, 2, 3, 4.

La moyenne est $\bar{x} = \frac{0+0+1+1+1+2+3+4}{8} = \frac{12}{8} = 1.5.$

Moyenne arithmétique

On peut aussi faire les calculs avec les valeurs distinctes et les effectifs. On considère le tableau

x_j	n_j
0	2
1	3
2	1
3	1
4	1
	8

$$\begin{aligned}\bar{x} &= \frac{2\times 0+3\times 1+1\times 2+1\times 3+1\times 4}{8}\\ &= \frac{3+2+3+4}{8}\\ &= 1.5.\end{aligned}$$

Moyenne géométrique

Si $x_i \geqslant 0$, on appelle moyenne géométrique la quantité:

$$G = \left(\prod_{i=1}^{n} x_i \right)^{1/n} = (x_1 \times x_2 \times \cdots \times x_n)^{1/n}$$

On peut écrire la moyenne géométrique comme l'exponentielle de la moyenne arithmétique des logarithmes des valeurs observées:

$$G = \exp \log G = \exp \log \left(\prod_{i=1}^{n} x_i \right)^{1/n} = \exp \frac{1}{n} \log \prod_{i=1}^{n} x_i = \exp \frac{1}{n} \sum_{i=1}^{n} \log x_i.$$

Moyenne géométrique

Exemple: Supposons que le taux de croissance d'un oranger est de 80%, 16 % et 42 % pour trois années successives. Que va-t-on obtenir après 3 ans si l'oranger produit 100 orange la première année ?

- Après 1 an on a, 100 x 1,8 = 180 oranges;
- Après 2 ans on a, 100 x 1,8 x 1,16 = 208.8 oranges;
- Après 3 ans on a, 100 x 1,8 x 1,16 x 1,42 = 296.496 oranges;

Moyenne géométrique

- Après 1 an on a, 100 x 1,8 = 180 oranges;
- Après 2 ans on a, 100 x 1,8 x 1,16 = 208.8 oranges;
- Après 3 ans on a, 100 x 1,8 x 1,16 x 1,42 = 296.496 oranges.

Si on calcule la moyenne arithmétique des taux de croissance on obtient:

$\bar{x}$ = (1.8 + 1.16 + 1.42) / 3 = 1.46

Moyenne géométrique

Si on calcule la moyenne géométrique des taux, on obtient:

$$G = (1.8 \times 1.16 \times 1.42)^{1/3} = 1.436612407$$

Le bon taux moyen est bien G et non $\bar{x}$, car si on applique 3 fois le taux moyen G aux 100 orange, on obtient

$$100 \times G^3 = 100 \times 1.436612407^3 = 249.95999$$

Moyenne harmonique

Si $x_i \geqslant 0$, on appelle harmonique la quantité: $H = \frac{n}{\sum_{i=1}^{n} 1/x_i}.$

Il est judicieux d'appliquer la moyenne harmonique sur des vitesses.

Moyenne harmonique

Exemple: Un guépard parcourt 4 terrains de 10 km. Les vitesses respectives pour ces étapes sont de 10 km/h, 30 km/h, 40 km/h, 20 km/h. Quelle a été sa vitesse moyenne ?

Moyenne harmonique

Un raisonnement simple nous dit qu'il a parcouru la première étape en 1 heure, la deuxième 20 minutes, la troisième en 15 minutes et la quatrième en 30 minutes. Elle a donc parcouru le total des 20 km en

1 h + 20 min + 15 min + 30 min = 2 h 05 minutes = 2,083h.

Sa vitesse moyenne est donc: 40 / 2,083 = 19,2 km/h.

Moyenne harmonique

Si on calcule la moyenne arithmétique des vitesses, on obtient:

$\bar{x}$ = (10 + 30 + 40 + 20) / 4 = 25 km/h

Si on calcule la moyenne harmonique des vitesses, on obtient:

H = 4 / (1/10 + 1/30 + 1/40 + 1/20) = 19,2 km/h

La moyenne harmonique est donc la manière appropriée de calculer la vitesse moyenne.

Moyenne pondérée

Dans certains cas, on n'accorde pas le même poids à toutes les observations.

Par exemple, si on calcule la moyenne des notes pour un programme d'étude, on peut pondérer les notes de l'étudiant par le nombre de crédits ou par le nombre d'heures de chaque cours.

Moyenne pondérée

Si $w_i > 0$, i = 1, . . . , n sont les poids associés à chaque observation, alors la moyenne pondérée par w_i est définie par :

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

Moyenne pondérée

Exemple: Supposons que les notes sont pondérées par le nombre de crédits, et que les notes de l'étudiant soient les suivantes:

Note	5	4	3	6	5
Crédits	6	3	4	3	4

La moyenne pondérée des notes par les crédits est alors

$$\bar{x}_w = \frac{6 \times 5 + 3 \times 4 + 4 \times 3 + 3 \times 6 + 4 \times 5}{6 + 3 + 4 + 3 + 4} = \frac{30 + 12 + 12 + 18 + 20}{20} = \frac{92}{20} = 4.6.$$

Médiane

La médiane, notée $x_{1/2}$, est une valeur centrale de la série statistique obtenue de la manière suivante:

- On trie la série statistique par ordre croissant des valeurs observées. Avec la série observée : 3 2 1 0 0 1 2,

  on obtient : 0 0 1 1 2 2 3

Médiane

- La médiane $x_{1/2}$, est la valeur qui se trouve au milieu de la série ordonnée

0 0 1 1 2 2 3

On note alors, $x_{1/2} = 1$.

Médiane: cas d'une série impaire

Examinons une manière simple de calculer la médiane. Deux cas doivent être distingués.

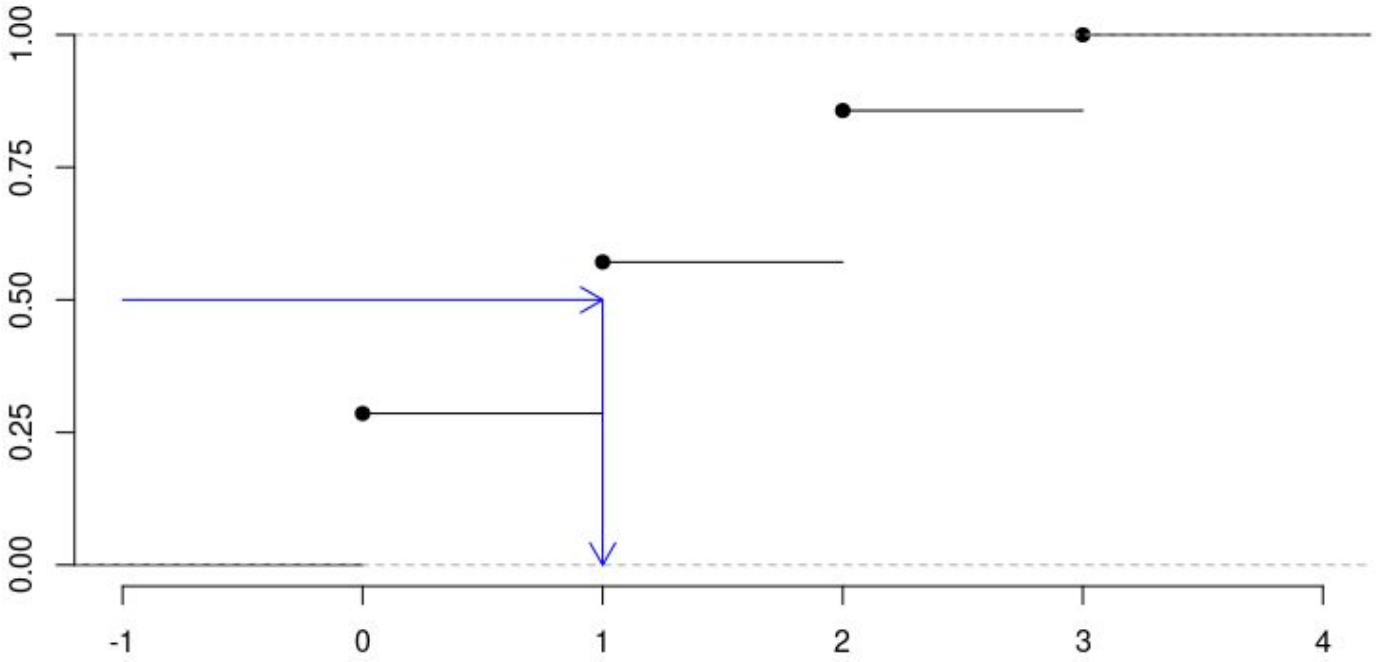
Si *n* est impair, il n'y a pas de problème (ici avec *n* = 7), alors $x_{1/2} = 1$:

0 0 1 1 2 2 3

Médiane: cas d'une série impaire

La médiane peut être définie comme l'inverse de la fonction de répartition pour la valeur 1/2: $x_{1/2} = F^{-1}(0.5)$.

Médiane: cas d’une série impaire



Médiane: cas d'une série paire

Si n est pair, deux valeurs se trouvent au milieu de la série (ici avec n = 8)

0 0 1 1 2 2 3 4

La médiane est alors la moyenne de ces deux valeurs:

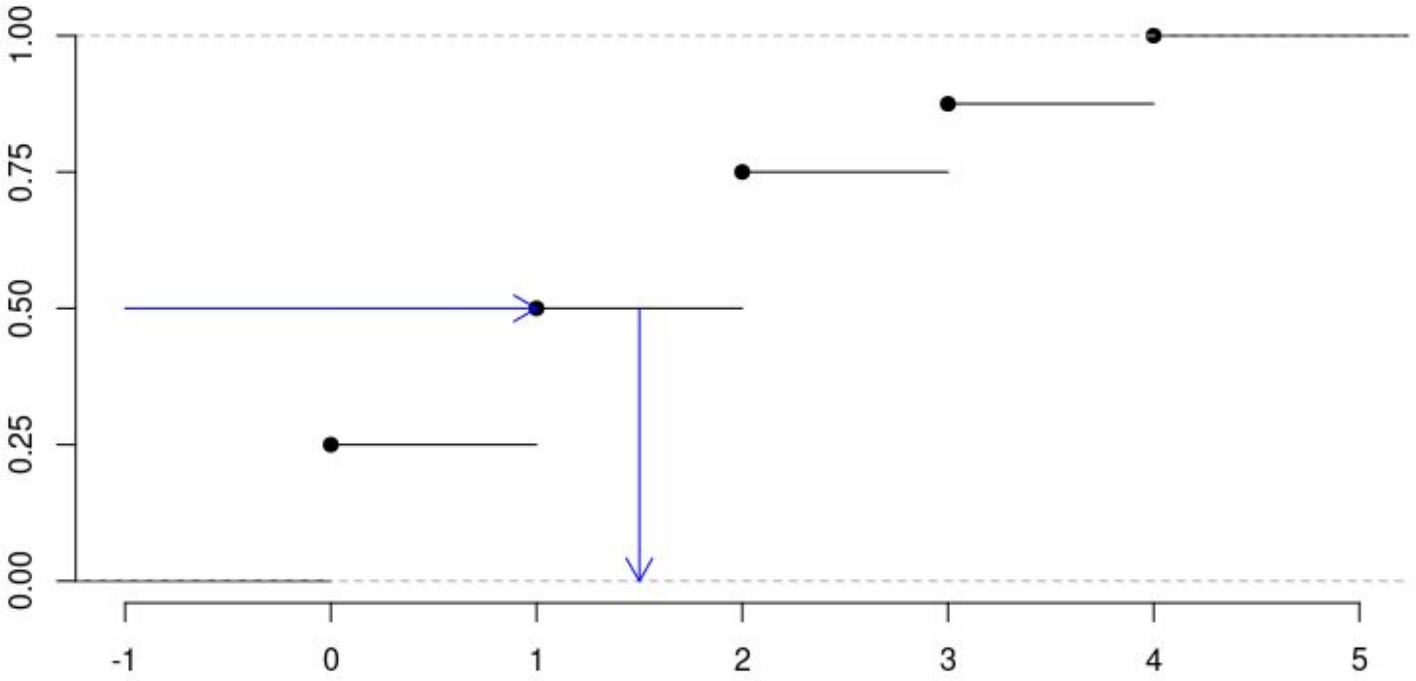
$$x_{1/2} = \frac{1+2}{2} = 1.5.$$

Médiane: cas d'une série paire

La médiane peut toujours être définie comme l'inverse de la fonction de répartition pour la valeur 1/2 :

$$x_{1/2} = F^{-1}(0.5)$$

Médiane: cas d'une série paire



Médiane: remarque

En général on note $x_{(1)}$, …, $x_{(i)}$, … $x_{(n)}$ la série ordonnée par ordre croissant. On appelle cette série ordonnée la statistique d'ordre. Cette notation, très usuelle en statistique, permet de définir la médiane de manière très synthétique:

- Si n est impair $x_{1/2} = x_{\left(\frac{n+1}{2}\right)}$
- Si n est pair, $x_{1/2} = \frac{1}{2}\left\{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right\}.$

Médiane: remarque

La médiane peut être calculée sur des variables quantitatives et sur des variables qualitatives ordinales.

Quantiles

Les quantiles sont les valeurs qui divisent un jeu de données en intervalles de même probabilité égale. La notion de quantile d'ordre *p* (où $0 < p < 1$) généralise la médiane.

Formellement un quantile est donné par l'inverse de la fonction de répartition: $x_p = F^{-1}(p)$

Quantiles

- Si *np* est un nombre entier, alors

$$x_p = \frac{1}{2}\left\{x_{(np)} + x_{(np+1)}\right\}$$

- Si np n'est pas un nombre entier, alors

$$x_p = x_{(\lceil np \rceil)}$$

où $\lceil np \rceil$ représente le plus petit nombre entier supérieur ou égal à *np*.

Quantiles

- La médiane est le quantile d'ordre p = 1/2.
- Si *F(x)* est la fonction de répartition, alors *F(xp)* ≥ *p.*
- On utilise souvent:
  - $x_{¼}$ le premier quartile;
  - $x_{3/4}$ le troisième quartile;
  - $x_{1/10}$ le premier décile.

Quantiles: exemple

Soit la série statistique 12, 13, 15, 16, 18, 19, 22, 24, 25, 27, 28, 34 contenant 12 observations (n = 12).

- Le premier quartile : Comme *np* = 0.25 × 12 = 3 est un nombre entier, on a $x_{1/4} = \frac{x_{(3)} + x_{(4)}}{2} = \frac{15+16}{2} = 15.5.$
- La médiane : Comme np = 0.5 × 12 = 6 est un nombre entier, on a

$$x_{1/2} = \frac{1}{2}\left\{x_{(6)} + x_{(7)}\right\} = (19+22)/2 = 20.5.$$

Quantiles: exemple

Le troisième quartile : Comme *np* = 0.75 × 12 = 9 est un nombre entier, on a

$$x_{3/4} = \frac{x_{(9)} + x_{(10)}}{2} = \frac{25 + 27}{2} = 26.$$

Quantiles en langage R

```
> x <- c(12, 13, 15, 16, 18, 19, 22, 24, 25, 27, 28, 34)

> quantile(x, type = 2)
```

02

Paramètres de dispersion

Etendue

L'*étendue* est simplement la différence entre la plus grande et la plus petite valeur observée:

$$\mathrm{E} = x_{(n)} - x_{(1)}.$$

Distance interquartile

La *distance interquartile* est la différence entre le troisième et le premier quartile:

$$IQ = x_{3/4} - x_{1/4}.$$

Variance

La *variance* est la somme des carrés des écarts à la moyenne divisée par le nombre d'observations :

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

Variance

La *variance* est la somme des carrés des écarts à la moyenne divisée par le nombre d'observations :

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2. \qquad \text{ou} \qquad s_x^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2.$$

La variance peut également être définie a partir des effectifs et des valeurs distinctes:

$$s_x^2 = \frac{1}{n}\sum_{j=1}^{J} n_j(x_j - \bar{x})^2. \qquad \text{ou} \qquad s_x^2 = \frac{1}{n}\sum_{j=1}^{J} n_j x_j^2 - \bar{x}^2.$$

Variance

Quand on veut estimer une variance d'une variable *X* à partir d'un échantillon (une partie de la population sélectionnée au hasard) de taille *n*, on utilise la variance "corrigée" divisée par *n* − 1.

$$S_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = s_x^2\frac{n}{n-1}$$

La plupart des logiciels statistiques calculent S^2_x et non s^2_x.

Ecart-type

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

$$s_x = \sqrt{s_x^2}.$$

Quand on veut estimer l'écart-type d'une variable *X* partir d'un échantillon de taille *n*, utilise la variance "corrigée" pour définir l'écart type

$$S_x = \sqrt{S_x^2} = s_x \sqrt{\frac{n}{n-1}}.$$

Ecart-type

La plupart des logiciels statistiques calculent S_x et non s_x.

L'écart-type: exercice

Soit la série statistique 2, 3, 4, 4, 5, 6, 7, 9 de taille 8. Calculer la variance et l'écart-type de cette série.

03

Moments

Moments

1. On appelle *moment à l'origine* d'ordre r $\in \mathbb{N}$ le paramètre

$$m'_r = \frac{1}{n}\sum_{i=1}^{n} x_i^r .$$

2. On appelle *moment centré* d'ordre r $\in \mathbb{N}$ le paramètre

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

Moments

Les moments généralisent la plupart des paramètres. On a en particulier:

- $m'_1 = \bar{x}$,
- $m_1 = 0$,
- $m'_2 = \frac{1}{n}\sum_i x_i^2 = s_x^2 + \bar{x}^2$
- $m_2 = s_x^2$.

04

Paramètres de forme

Coefficient d'asymétrie de Fisher

Le moment centré d'ordre trois est défini par: $m_3 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3.$

Il peut prendre des valeurs positives, négatives ou nulles. L'asymétrie se mesure au moyen du coefficient d'asymétrie de Fisher

$$g_1 = \frac{m_3}{s_x^3},$$

où s^3 x est le cube de l'écart-type.

Coefficient d’asymétrie de Yule

Le coefficient d’asymétrie de Yule est basé sur les positions des 3 quartiles (1er quartile, médiane et troisième quartile), et est normalisé par la distance interquartile :

$$A_Y = \frac{x_{3/4} + x_{1/4} - 2x_{1/2}}{x_{3/4} - x_{1/4}}.$$

Coefficient d’asymétrie de Pearson

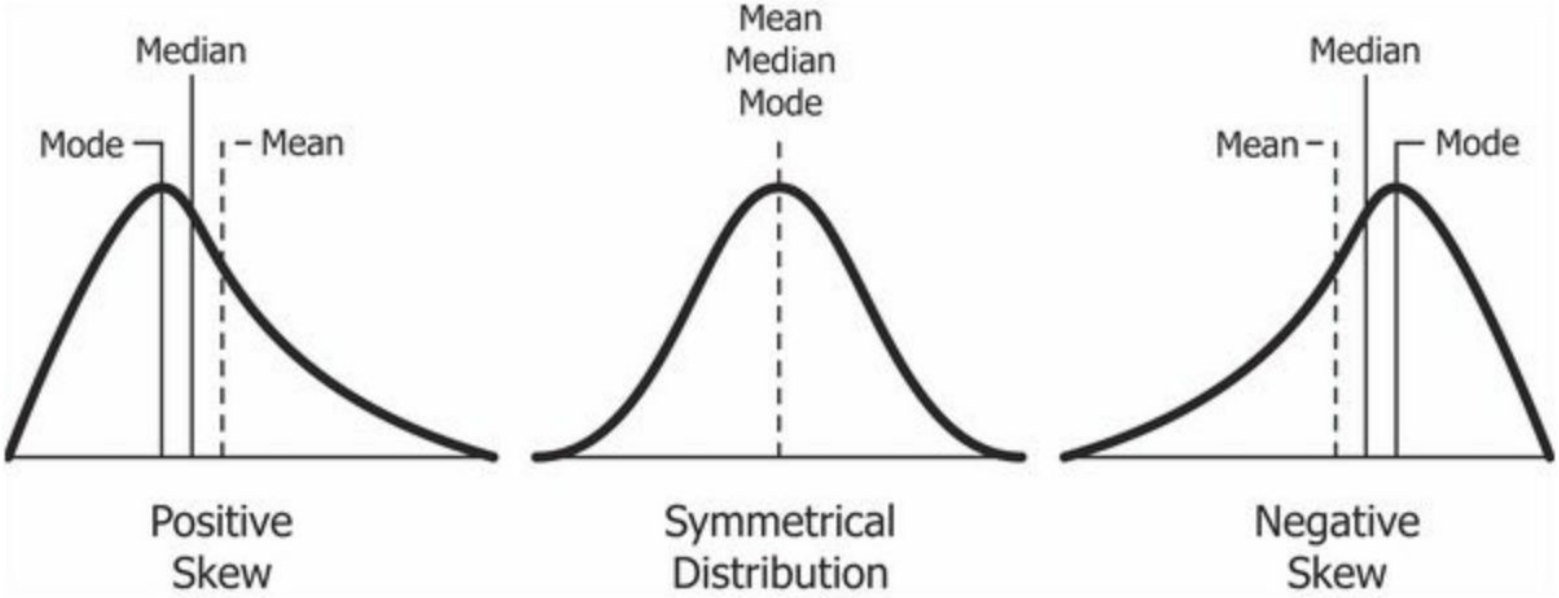
Le coefficient d’asymétrie de Pearson est basé sur une comparaison de la moyenne et du mode, et est standardisé par l’écart-type :

$$A_P = \frac{\bar{x} - x_M}{s_x}.$$

Propriétés des coefficients d’asymétrie

Tous les coefficients d’asymétrie ont les mêmes propriétés, ils sont nuls si la distribution est symétrique, négatifs si la distribution est allongée à gauche (left asymmetry), et positifs si la distribution est allongée à droite (right asymmetry).

Propriétés des coefficients d'asymétrie



Propriétés des coefficients d’asymétrie

Certaines variables sont toujours très asymétriques à droite, comme les revenus, les tailles des entreprises, ou des communes. Une méthode simple pour rendre une variable symétrique consiste alors à prendre le logarithme de cette variable.

05

Paramètre d’aplatissement

Paramètre d'aplatissement

L'aplatissement est mesuré par le coefficient d'aplatissement de Pearson:

$$\beta_2 = \frac{m_4}{s_x^4},$$

ou le coefficient d'aplatissement de Fisher:

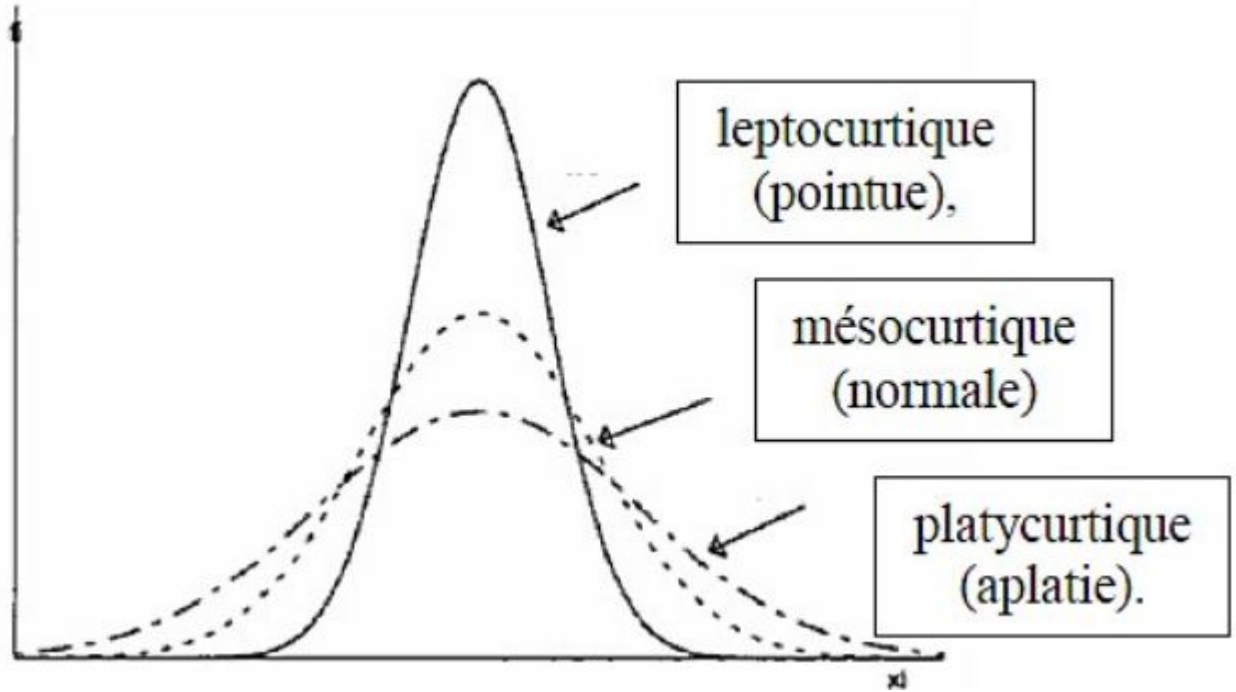
$$g_2 = \beta_2 - 3 = \frac{m_4}{s_x^4} - 3$$

où m^4 est le moment centré d'ordre 4, et s^4_x est le carré de la variance.

Paramètre d’aplatissement

- Une courbe mésokurtique si $g_2 \approx 0$.
- Une courbe leptokurtique si $g_2 > 0$. Elle est plus pointue et possède des queues plus longues.
- Une courbe platykurtique si $g_2 < 0$. Elle est plus arrondie et possède des queues plus courtes.

Paramètre d'aplatissement (kurtosis)



06

Diagramme en tiges et feuilles

Diagramme en tiges et feuilles

Le diagramme en tiges et feuilles ou *Stem and leaf diagram* est une manière rapide de présenter une variable quantitative. Par exemple, si l'on a la série statistique ordonnée suivante :

15, 15, 16, 17, 18, 20, 21, 22, 23, 23, 23, 24, 25, 25, 26,

26, 27, 28, 28, 29, 30, 30, 32, 34, 35, 36, 39, 40, 43, 44,

La tige du diagramme sera les dizaines et les feuilles seront les unités.

Diagramme en tiges et feuilles

A partir de la série précédente, on obtient le diagramme:

```
1 | 55678

2 | 012333455667889

3 | 0024569

4 | 034
```

Diagramme en tiges et feuilles en langage R

```
> X <- c(15,15,16,17,18,20,21,22,23,23,23,24,25,25,26,26,
27,28,28,29,30,30,32,34,35,36,39,40,43,44)

> stem(X, 0.5)
```

07

Boîte à moustache

Boîte à moustache

La boîte à moustaches, ou diagramme en boîte, ou encore boxplot en anglais, est un diagramme simple qui permet de représenter la distribution d'une variable. Ce diagramme est composé de

- Un rectangle qui s'étend du premier au troisième quartile. Le rectangle est divisé par une ligne correspondant à la médiane.
- Ce rectangle est complété par deux segments de droite.

Boîte à moustache

Pour dessiner les deux segments, on calcule d'abord les bornes:

$$b^- = x_{1/4} - 1.5IQ \quad \text{et} \quad b^+ = x_{3/4} + 1.5IQ$$

où IQ est la distance interquartile.

- On identifie ensuite la plus petite et la plus grande observation comprise entre ces bornes. Ces observations sont appelées "valeurs adjacentes".

Boîte à moustache

- On trace les segments de droite reliant ces observations au rectangle.
- Les valeurs qui ne sont pas comprises entre les valeurs adjacentes, sont représentées par des points et sont appelées “valeurs extrêmes”.

Boîte à moustache en langage R

```
> InsectSprays

> boxplot(count ~ spray, data = InsectSprays, col = "lightgray")
```

Boîte à moustache en langage R