

# Statistique descriptive

## Chapitre 4 : Le regroupement en classes



---

Jaouad Madkour

Faculté d'Économie et de Gestion de Tanger · 27 août 2026

Pourquoi des classes

Combien de classes, et lesquelles

Le prix des amplitudes inégales

Le centre de classe

Ce qu'il faut retenir

## Pourquoi des classes

---

## Le tableau du chapitre 3 est impossible ici

Le salaire prend **381** valeurs différentes sur **500** salariés. Un tableau à **381** lignes ne résume rien : il est plus long que la question qu'il prétend traiter.

Pour une variable continue, on n'a pas le choix : il faut **regrouper les valeurs en intervalles**.

## Classe

Une **classe** est un intervalle  $[e_{i-1} ; e_i[$ , fermé à gauche et ouvert à droite.

- $e_{i-1}$  et  $e_i$  sont les **bornes**;
- $a_i = e_i - e_{i-1}$  est l'**amplitude**;
- $c_i = \frac{e_{i-1} + e_i}{2}$  est le **centre de classe**.

### La convention d'écriture n'est pas décorative

$[4\,000 ; 6\,000[$  signifie : de 4 000 inclus à 6 000 **exclu**.

Un salarié à 6 000 dirhams appartient à la classe suivante, et à elle seule. C'est ce qui garantit que les modalités sont **incompatibles**, exigence posée au chapitre 2.

Un tableau écrit avec des classes  $[4\,000 ; 6\,000]$  et  $[6\,000 ; 8\,000]$  est faux.

Combien de classes, et lesquelles

---

### Deux règles empiriques

règle de Sturges :  $k \approx 1 + 3,3 \log_{10}(n)$

règle de la racine :  $k \approx \sqrt{n}$

### Sur nos données

Sturges :  $1 + 3,3 \times \log_{10}(500) = 1 + 3,3 \times 2,70 = 9,9$ , soit une dizaine de classes.

Racine :  $\sqrt{500} = 22,4$ .

Les deux règles ne s'accordent pas. Elles ne sont qu'indicatives.

### Le vrai critère est l'arbitrage

Trop peu de classes : on écrase la forme de la distribution, tout devient plat.

Trop de classes : le tableau redevient bruité, chaque classe contient trop peu d'individus pour être stable.

**Le bon nombre est celui qui laisse voir la forme sans laisser voir le hasard.** Il se juge sur l'histogramme, pas sur une formule — et il faut souvent essayer deux ou trois découpages.

### Trois principes

1. Des bornes **rondes** et lisibles : 4 000, 6 000, 8 000 — jamais 4 137.
2. Aucune classe **vide** à l'intérieur du tableau.
3. Des classes **plus larges là où les données sont rares** — c'est-à-dire dans les queues.

### Le découpage retenu pour les salaires

| Classe            | $n_i$      | $f_i$        | $a_i$ | $N_i$ | $F_i$ |
|-------------------|------------|--------------|-------|-------|-------|
| [2 000 ; 4 000[   | 31         | 0,062        | 2 000 | 31    | 0,062 |
| [4 000 ; 6 000[   | 157        | 0,314        | 2 000 | 188   | 0,376 |
| [6 000 ; 8 000[   | 166        | 0,332        | 2 000 | 354   | 0,708 |
| [8 000 ; 10 000[  | 79         | 0,158        | 2 000 | 433   | 0,866 |
| [10 000 ; 14 000[ | 57         | 0,114        | 4 000 | 490   | 0,980 |
| [14 000 ; 20 000[ | 10         | 0,020        | 6 000 | 500   | 1,000 |
| <b>Total</b>      | <b>500</b> | <b>1,000</b> |       |       |       |

### Pourquoi les deux dernières classes sont plus larges

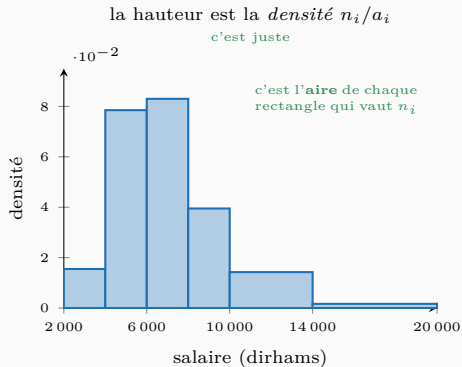
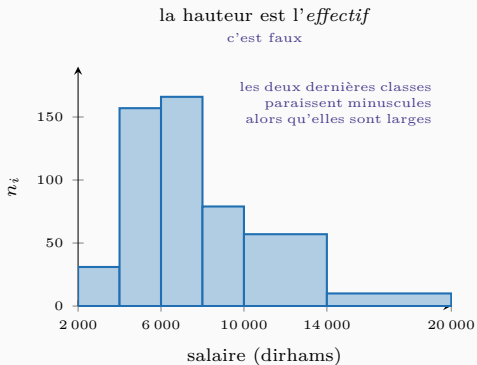
Au-delà de 10 000 dirhams, les salariés se raréfient. Découper de 2 000 en 2 000 jusqu'à 20 000 donnerait des classes à 3 ou 4 individus, illisibles et instables.

**Mais élargir les classes a un prix, et il faut le payer :** l'histogramme ne peut plus se lire en effectifs.

## Le prix des amplitudes inégales

---

# L'erreur qu'il faut voir une fois



classes d'amplitudes inégales : 2 000, 2 000, 2 000, 2 000, 4 000, 6 000 dirhams

## Densité de fréquence

$$d_i = \frac{n_i}{a_i} \quad \text{ou, en fréquences,} \quad \frac{f_i}{a_i}$$

Dans un histogramme, c'est la densité qui donne la hauteur du rectangle, jamais l'effectif.

## La règle, en une phrase

Dans un histogramme, c'est l'aire qui représente l'effectif, pas la hauteur.

Aire du rectangle = base  $\times$  hauteur =  $a_i \times \frac{n_i}{a_i} = n_i$ . L'amplitude se simplifie : chaque rectangle a une aire égale à son effectif, quelles que soient les largeurs choisies.

### Le calcul complet sur nos classes

| Classe            | $n_i$ | $a_i$ | $d_i = n_i/a_i$ |
|-------------------|-------|-------|-----------------|
| [2 000 ; 4 000[   | 31    | 2 000 | 0,0155          |
| [4 000 ; 6 000[   | 157   | 2 000 | 0,0785          |
| [6 000 ; 8 000[   | 166   | 2 000 | 0,0830          |
| [8 000 ; 10 000[  | 79    | 2 000 | 0,0395          |
| [10 000 ; 14 000[ | 57    | 4 000 | 0,0143          |
| [14 000 ; 20 000[ | 10    | 6 000 | 0,0017          |

### Ce que la correction change

Sur l'histogramme en effectifs, la classe  $[10\,000 ; 14\,000[$  paraît deux fois plus fournie que la première classe. En densité, elle est en réalité **au même niveau**.

Un lecteur non averti aurait conclu que les hauts salaires sont plus nombreux qu'ils ne le sont. **C'est le moyen le plus courant de mentir avec un histogramme, et il est presque toujours involontaire.**

### La classe ouverte

Un tableau publié se termine souvent par « 20 000 dirhams et plus » : la borne supérieure est inconnue.

On ne peut alors ni calculer la densité, ni le centre de classe, ni donc la moyenne.

### Les deux solutions

1. **Fermer arbitrairement** la classe, en lui donnant l'amplitude de la précédente. C'est la convention la plus répandue.
2. **Renoncer à la moyenne** et travailler avec la médiane, qui n'a pas besoin des valeurs extrêmes.

La seconde est plus honnête. C'est aussi pourquoi les revenus se publient en médiane : la classe la plus haute est toujours ouverte.

## Le centre de classe

---

## Le centre de classe

$$c_i = \frac{e_{i-1} + e_i}{2}$$

Pour tous les calculs qui suivent — moyenne, variance, moments — **on remplace chaque individu de la classe par le centre de sa classe.**

## L'hypothèse d'équirépartition

On suppose que les individus sont **uniformément répartis** à l'intérieur de chaque classe. C'est faux en général, mais les erreurs se compensent partiellement d'une classe à l'autre.

## Une hypothèse qu'on oublie (suite)

### Le coût de l'approximation, chiffré

Moyenne calculée sur les 500 valeurs exactes : 7 158 dirhams.

Moyenne calculée à partir du tableau en classes, avec les centres :

$$\bar{x} = \frac{3000(31) + 5000(157) + 7000(166) + 9000(79) + 12000(57) + 17000(10)}{500} = 7\,210$$

Écart : 52 dirhams, soit 0,7 %.

### La leçon

L'approximation est excellente **tant que les classes sont fines là où les données sont denses.**

Ici l'essentiel de l'écart vient des deux dernières classes, larges et asymétriquement remplies : à l'intérieur de [10 000 ; 14 000[, les salariées se pressent près de la borne basse, si bien que le centre 12 000 les surestime.

C'est un argument de plus pour ne pas élargir les classes du milieu de la distribution — et un rappel que la moyenne calculée sur données groupées reste une **approximation**, à ne pas donner au dirham près.

Ce qu'il faut retenir

---

## Six points

1. Une variable continue **doit** être regroupée en classes  $[e_{i-1} ; e_i[$ .
2. Sturges et la racine donnent un ordre de grandeur, pas une réponse : le nombre de classes se juge sur l'histogramme.
3. Bornes rondes, aucune classe vide, classes plus larges dans les queues.
4. **Densité**  $d_i = n_i/a_i$  : c'est elle qui donne la hauteur d'un histogramme, jamais l'effectif.
5. Dans un histogramme, **l'aire représente l'effectif**.
6. Le **centre de classe** remplace les individus dans tous les calculs : approximation excellente si les classes denses sont fines.

## La suite

Le tableau est construit. Chacune de ses colonnes appelle maintenant sa propre représentation graphique — et le choix du graphique dépend, une fois encore, du type de la variable.

C'est le chapitre 5.