

Statistique descriptive

Chapitre 3 : Le regroupement en classes

Jaouad Madkour

12 octobre 2026

Plan du chapitre

Pourquoi des classes	2
L'obligation	2
La convention d'écriture	2
Combien de classes	2
Deux règles empiriques	2
Trop peu, assez, trop	3
Le choix des bornes	3
Le découpage retenu pour les salaires	3
Amplitudes inégales, le prix	4
L'erreur qu'il faut voir une fois	4
La densité	4
Le calcul des densités	5
Ce que la correction change	5
La classe ouverte	5
Fermer une classe : ce que cela change	6
Le centre de classe	6
Une hypothèse qu'on oublie	6
Centre de classe contre moyenne réelle	6
Le coût de l'approximation, chiffré	7
Interpoler à l'intérieur d'une classe	7
S'entraîner	7
Exercice 1 : construire les classes	7
Exercice 1 : solution	8
Exercice 1 : solution, la lecture	8
Exercice 2 : interpoler dans les classes	8
Exercice 2 : solution	8
Ce qu'il faut retenir	9
Le résumé du chapitre	9

Pourquoi des classes

L'obligation

Le tableau du chapitre 2 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

Elle 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 1. Un tableau écrit avec des classes $[4\,000; 6\,000]$ et $[6\,000; 8\,000]$ est faux.

Bornes, amplitude, centre

Classe	e_{i-1}	e_i	a_i	c_i
$[4\,000; 6\,000[$	4 000	6 000	2 000	5 000
$[10\,000; 14\,000[$	10 000	14 000	4 000	12 000
$[14\,000; 20\,000[$	14 000	20 000	6 000	17 000

Combien de classes

Deux règles empiriques

Sturges et la racine

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

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

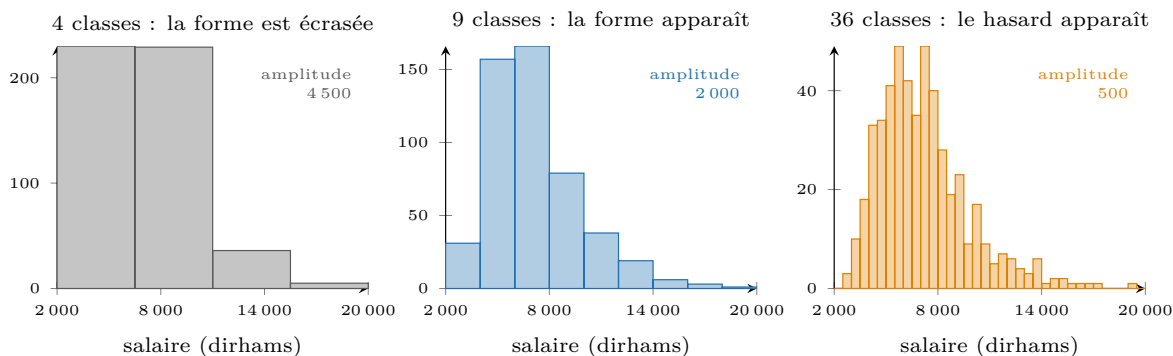
Sur nos données

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.**

Trop peu, assez, trop



Le vrai critère est l'arbitrage

Trop peu de classes écrase la forme ; trop de classes laisse voir le hasard, chaque classe contenant trop peu d'individus pour être stable. **Le bon nombre laisse voir la forme sans laisser voir le hasard** : il se juge sur l'histogramme, en essayant deux ou trois découpages.

Le choix des bornes

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 raisonnement, pas à pas, sur les salaires

1. Étendue : $19\,050 - 2\,810 = 16\,240$ dirhams.
2. Sturges propose $k \approx 10$, soit une amplitude d'environ $16\,240/10 \approx 1\,600$, arrondie à 2 000.
3. Bornes rondes de 2 000 à 20 000 : neuf classes de 2 000.
4. Au-delà de 10 000, les cinq dernières classes ne comptent que 38, 19, 6, 3 et 1 salariés : on les fusionne en $[10\,000 ; 14\,000[$ et $[14\,000 ; 20\,000[$.

Le découpage retenu pour les salaires

Six classes, dont deux plus larges

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

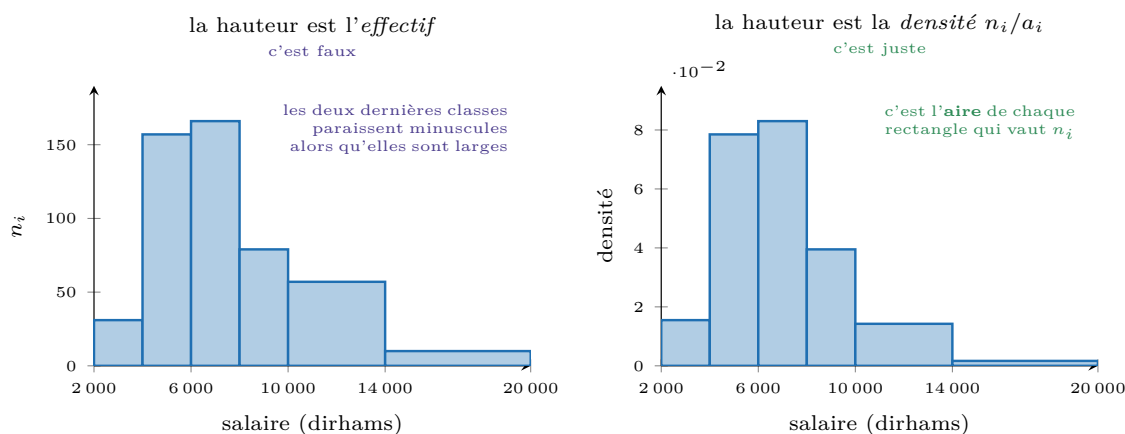
[14 000 ; 20 000[	10	0,020	6 000	500	1,000
Total	500	1,000			

Pourquoi les deux dernières classes sont plus larges

Découper de 2000 en 2000 jusqu'à 20 000 donnerait des classes à 1, 3 ou 6 individus, illisibles et instables. **Mais élargir les classes a un prix** : l'histogramme ne peut plus se lire en effectifs.

Amplitudes inégales, le prix

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

La densité

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 des densités

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

La densité se lit en salariés **par dirham** ; multipliée par 1 000, en salariés par tranche de 1 000 dirhams : 83 dans la classe modale.

Ce que la correction change

Une erreur presque toujours involontaire

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** : 0,0143 contre 0,0155.

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.**

Une vérification qui ne trompe pas

La somme des aires $a_i d_i$ redonne l'effectif total :

$$2\,000 \times 0,0155 + 4\,000 \times 0,0143 + 6\,000 \times 0,0017 = 433 + 57 + 10 = 500$$

La classe ouverte

Un cas particulier fréquent

Un tableau publié se termine souvent par « 14 000 dirhams et plus » : la borne supérieure est inconnue. On ne peut alors calculer ni 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.

Fermer une classe : ce que cela change

Deux fermetures, deux moyennes

Supposons la dernière classe publiée comme « 14 000 et plus », avec ses 10 salariés.

- Fermée à 20 000 : centre 17 000, moyenne groupée 7 210 dirhams.
- Fermée avec l'amplitude précédente, à 18 000 : centre 16 000, et

$$\bar{x} = 7\,210 - \frac{10 \times (17\,000 - 16\,000)}{500} = 7\,210 - 20 = 7\,190$$

Peu de salariés, mais un effet visible

Dix salariés sur 500 déplacent la moyenne de 20 dirhams selon une convention que personne ne peut vérifier. **La médiane, elle, ne bouge pas.**

Le centre de classe

Une hypothèse qu'on oublie

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.

Centre de classe contre moyenne réelle

Ce que gagnent vraiment les salariés de chaque classe

Classe	Centre c_i	Moyenne réelle	Écart
[2 000 ; 4 000[	3 000	3 528	−528
[4 000 ; 6 000[	5 000	5 078	−78
[6 000 ; 8 000[	7 000	6 995	+5
[8 000 ; 10 000[	9 000	8 838	+162
[10 000 ; 14 000[	12 000	11 470	+530
[14 000 ; 20 000[	17 000	15 927	+1 073

La lecture

Dans les classes centrales, le centre est presque exact. Aux deux bouts, il **pousse les salariés vers l'extérieur** : trop bas en bas, trop haut en haut. On s'en souviendra pour la variance, au chapitre 7.

Le coût de l'approximation, chiffré

La moyenne à partir du tableau

Moyenne sur les 500 valeurs exactes : 7 158 dirhams. À partir du tableau, chaque salarié est remplacé par son centre ; en milliers de dirhams :

$$\sum n_i c_i = 3 \times 31 + 5 \times 157 + 7 \times 166 + 9 \times 79 + 12 \times 57 + 17 \times 10 = 3\,605$$

$$\bar{x} = 3\,605\,000/500 = 7\,210 \text{ dirhams}$$

Écart : 52 dirhams, soit 0,7 %.

La leçon

L'approximation est excellente **tant que les classes sont fines là où les données sont denses**. L'essentiel de l'écart vient des deux dernières classes, larges et remplies surtout près de leur borne basse. La moyenne sur données groupées reste une **approximation**, à ne pas donner au dirham près.

Interpoler à l'intérieur d'une classe

Une proportion entre deux valeurs quelconques

Sous l'hypothèse d'équirépartition, la part d'une classe située sous une valeur x est proportionnelle à la longueur parcourue :

$$n_i \times \frac{x - e_{i-1}}{a_i}$$

Combien gagnent moins de 5 000 dirhams ?

Toute la première classe, 31 salariés, plus la moitié de [4 000 ; 6 000[:

$$31 + 157 \times \frac{5\,000 - 4\,000}{2\,000} = 31 + 78,5 = 109,5 \text{ salariés, soit } 21,9 \%$$

Le fichier exact en compte 98 : les salariés de cette classe sont un peu plus nombreux dans sa moitié haute. C'est le même calcul qui donnera la médiane au chapitre 6.

S'entraîner

Exercice 1 : construire les classes

Les temps de trajet de 24 salariés

Durée du trajet domicile-travail, en minutes, rangée par ordre croissant :

8, 12, 15, 18, 20, 22, 25, 25, 28, 30, 32, 35, 35, 38, 40, 42, 45, 48, 50, 55, 60, 70, 85, 110.

1. Calculer l'étendue et le nombre de classes proposé par Sturges.
2. On retient les classes $[0 ; 20[$, $[20 ; 40[$, $[40 ; 60[$, $[60 ; 80[$ et $[80 ; 120[$. Dresser le tableau : n_i , f_i , a_i , d_i , F_i .
3. Quelle classe a la plus forte densité ? Pourquoi la dernière classe est-elle plus large, et quelle hauteur lui donner dans l'histogramme ?

Exercice 1 : solution

Étendue, Sturges et tableau

1. Étendue $110 - 8 = 102$ minutes ; Sturges : $1 + 3,3 \times \log_{10}(24) = 1 + 3,3 \times 1,38 = 5,6$, soit cinq ou six classes.

Classe	n_i	f_i	a_i	d_i	F_i
$[0 ; 20[$	4	0,167	20	0,20	0,167
$[20 ; 40[$	10	0,417	20	0,50	0,583
$[40 ; 60[$	6	0,250	20	0,30	0,833
$[60 ; 80[$	2	0,083	20	0,10	0,917
$[80 ; 120[$	2	0,083	40	0,05	1,000

Exercice 1 : solution, la lecture

Densité et dernière classe

3. La plus forte densité est celle de $[20 ; 40[$: 0,5 salarié par minute, soit 10 salariés pour 20 minutes.

La dernière classe est élargie parce que les longs trajets sont rares : en $[80 ; 100[$ et $[100 ; 120[$, on n'aurait qu'un salarié chacune. Sa hauteur est sa **densité**, $2/40 = 0,05$: moitié de celle de $[60 ; 80[$, alors que les deux classes ont le même effectif.

À retenir

Deux classes de même effectif mais d'amplitudes différentes n'ont **pas** la même hauteur. C'est l'aire, $a_i \times d_i = 2$, qui est la même.

Exercice 2 : interpoler dans les classes

Les salaires entre 7 000 et 9 000 dirhams

On n'a que le tableau en six classes des 500 salaires (diapositive « Le découpage retenu »).

1. Sous l'hypothèse d'équirépartition, estimer le nombre de salariés dont le salaire est compris entre 7 000 et 9 000 dirhams.
2. Le fichier exact en compte 136. Calculer l'erreur relative de l'estimation.
3. Estimer de même la part des salariés gagnant au moins 12 000 dirhams.

Exercice 2 : solution

Deux moitiés de classes

1. La moitié haute de $[6\,000 ; 8\,000[$ et la moitié basse de $[8\,000 ; 10\,000[$:

$$166 \times \frac{8\,000 - 7\,000}{2\,000} + 79 \times \frac{9\,000 - 8\,000}{2\,000} = 83 + 39,5 = 122,5 \text{ salariés}$$

2. Erreur : $(122,5 - 136)/136 = -9,9\%$. L'estimation sous-évalue : dans ces deux classes, les salariés ne sont pas uniformément répartis.

3. La moitié haute de $[10\,000; 14\,000[$ plus toute la dernière classe :

$$57 \times \frac{14\,000 - 12\,000}{4\,000} + 10 = 28,5 + 10 = 38,5 \text{ salariés, soit } 7,7 \%$$

Ce qu'il faut retenir

Le résumé du chapitre

Sept points

1. Une variable continue **doit** être regroupée en classes $[e_{i-1}; e_i[$.
2. Sturges donne un ordre de grandeur ; le nombre de classes se juge sur l'histogramme.
3. Bornes rondes, pas de 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 ; **l'aire représente l'effectif**.
5. Une classe ouverte doit être fermée par convention — ou l'on renonce à la moyenne.
6. Le **centre de classe** remplace les individus dans les calculs : bonne approximation au centre, biaisée aux deux bouts.
7. Sous l'équirépartition, on **interpole** à l'intérieur d'une classe.

La suite

Chaque colonne du tableau appelle maintenant sa représentation graphique, qui dépend du type de la variable. C'est le chapitre 4.