

Statistique descriptive

Chapitre 11 : Distributions conjointe, marginales, conditionnelles

Jaouad Madkour

12 octobre 2026

Plan du chapitre

Trois familles de lois	2
Les trois définitions	2
Trois questions distinctes	2
Distributions conditionnelles	2
La définition	2
Les distributions conditionnelles du salaire	3
Sur nos données	3
Le tableau des conditionnelles	4
La moyenne totale	4
Le théorème de la moyenne totale	4
L'effet de structure	4
Décomposer la variance	5
Le second théorème	5
Le calcul, pas à pas	5
Le rapport de corrélation	6
Comment on commente ce nombre	6
Trois variables explicatives comparées	6
Ce que ρ^2 ne suppose pas	7
Marginales et indépendance	7
Conjointe, marginale, conditionnelle	7
Le cas de l'indépendance	7
Le tableau de synthèse	8
Les trois familles, récapitulées	8
Les deux théorèmes	8
S'entraîner	8
Exercice 1 : toutes les moyennes baissent	8
Exercice 1 : solution	9
Exercice 2 : deux entreprises fusionnent	9

Exercice 2 : solution	9
Ce qu'il faut retenir	10
Le résumé du chapitre	10

Trois familles de lois

Les trois définitions

Les trois familles

Distribution	Ce qu'elle décrit	Combien
conjointe	le couple $(x; y)$	une seule
marginale	une variable, l'autre oubliée	deux
conditionnelle	une variable, l'autre fixée	autant que de modalités

Où elles se lisent dans le tableau

- La **conjointe** est l'intérieur du tableau : les n_{ij} .
- Les **marginale**s sont la dernière colonne et la dernière ligne.
- Les **conditionnelles** sont les lignes ou les colonnes prises **une par une**, chacune ramenée à son propre total.

Trois questions distinctes

Une question par famille

- Conjointe : « Quelle part des salariés est Bac+5 et gagne plus de 10 000 dirhams ? »
- Marginale : « Quel est le salaire moyen de l'entreprise ? »
- Conditionnelle : « Quel est le salaire moyen **des Bac+5** ? »

La troisième est celle qui intéresse presque toujours le décideur. C'est elle qui permet de comparer des groupes.

Les trois réponses

Conjointe : $(32 + 8)/500 = 8\%$ des salariés sont Bac+5 et gagnent 10 000 dirhams ou plus.
Marginale : 7 158 dirhams. Conditionnelle : 9 534 dirhams.

Distributions conditionnelles

La définition

Distribution conditionnelle

La distribution de y **sachant** $x = x_i$ est la répartition de y **restreinte aux individus** pour lesquels x vaut x_i .

Ses fréquences sont les profils-lignes du chapitre 10 :

$$f_{j|i} = \frac{n_{ij}}{n_{i\bullet}} \quad \text{avec} \quad \sum_j f_{j|i} = 1$$

Le geste, en une phrase

Conditionner, c'est réduire la population, puis recalculer.

On ne travaille plus sur 500 salariés mais sur les 107 Bac+5. Tous les indicateurs des chapitres 5 à 9 s'appliquent alors à cette sous-population.

Les distributions conditionnelles du salaire

Le salaire en classes, sachant le diplôme (en pourcentage de chaque ligne)

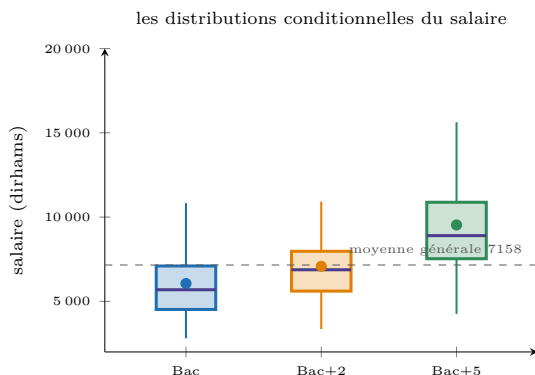
Diplôme	2 à 4	4 à 6	6 à 8	8 à 10	10 à 14	14 à 20	Total
Bac	12,8	43,4	29,7	10,0	3,7	0,5	100
Bac+2	1,7	31,0	42,5	14,4	9,8	0,6	100
Bac+5	0,0	7,5	25,2	29,9	29,9	7,5	100
Ensemble	6,2	31,4	33,2	15,8	11,4	2,0	100

Classes de salaire en milliers de dirhams.

La lecture

La classe la plus fréquente se déplace d'un cran à chaque diplôme. Chaque ligne somme à 100 ; la ligne « Ensemble » est la marginale du chapitre 3.

Sur nos données



Les trois moyennes conditionnelles

$$\bar{y}_{\text{Bac}} = 6\,059 \quad \bar{y}_{\text{Bac}+2} = 7\,081 \quad \bar{y}_{\text{Bac}+5} = 9\,534$$

Leur moyenne, *pondérée par les effectifs*, redonne exactement la moyenne générale 7 158 : c'est le **théorème de la moyenne totale**.

Et la variance se décompose

$$\underbrace{6\,750\,122}_{\text{totale}} = \underbrace{5\,010\,426}_{\text{intra, résiduelle}} + \underbrace{1\,739\,696}_{\text{inter, expliquée}}$$

$\eta^2 = 1\,739\,696 / 6\,750\,122 = 0,258$: le diplôme explique 25,8 % de la variance des salaires.

Le tableau des conditionnelles

Le salaire selon le diplôme

Diplôme	n_j	Moyenne	Écart-type	Médiane
Bac	219	6 059	2 021	5 690
Bac+2	174	7 081	2 045	6 875
Bac+5	107	9 534	2 872	8 900
Ensemble	500	7 158	2 598	6 770

La lecture, en trois constats

1. **Le niveau monte avec le diplôme** : +17 % du Bac au Bac+2, +35 % du Bac+2 au Bac+5. La prime au diplôme s'accélère.
2. **La dispersion aussi** : l'écart-type des Bac+5 dépasse celui des deux autres groupes.
3. **La moyenne d'ensemble n'est celle d'aucun groupe** : 7 158 dirhams ne décrit personne. Ce constat justifie à lui seul le conditionnement.

La moyenne totale

Le théorème de la moyenne totale

L'énoncé

$$\bar{y} = \frac{1}{n} \sum_j n_j \bar{y}_j = \sum_j f_j \bar{y}_j$$

La moyenne générale est la moyenne des moyennes conditionnelles, pondérée par les effectifs.

Vérification

$$\frac{219 \times 6\,059 + 174 \times 7\,081 + 107 \times 9\,534}{500} = \frac{3\,579\,153}{500} = 7\,158 \quad \checkmark$$

Moyenne **simple** des trois : $(6\,059 + 7\,081 + 9\,534)/3 = 7\,558$. **Faux de 400 dirhams.**

L'effet de structure

La moyenne bouge, aucun salaire ne bouge

Ce théorème explique un phénomène qui déroute les non-statisticiens : **la moyenne générale peut évoluer alors qu'aucune moyenne conditionnelle ne bouge**. Il suffit que les **poids** f_j changent.

C'est l'**effet de structure**, au cœur de toute analyse de l'évolution des salaires, des prix ou de la productivité. Une hausse du salaire moyen ne signifie pas que quelqu'un a été augmenté.

L'entreprise recrute 50 Bac+5

Payés à la moyenne de leur groupe, 9 534 dirhams :

$$\bar{y}' = \frac{3\,579\,153 + 50 \times 9\,534}{550} = \frac{4\,055\,853}{550} = 7\,374 \text{ dirhams}$$

Le salaire moyen gagne 216 dirhams, soit +3% — sans un dirham d'augmentation pour personne.

Décomposer la variance

Le second théorème

Décomposition de la variance

$$\underbrace{V(y)}_{\text{totale}} = \underbrace{\frac{1}{n} \sum_j n_j V_j(y)}_{\text{variance intra-groupe}} + \underbrace{\frac{1}{n} \sum_j n_j (\bar{y}_j - \bar{y})^2}_{\text{variance inter-groupes}}$$

Ce que chaque terme représente

- La variance **intra** est la moyenne des variances internes aux groupes : ce que le diplôme **n'explique pas**, la variance **résiduelle**.
- La variance **inter** mesure l'écartement des moyennes de groupe : ce que le diplôme **explique**, la variance **expliquée**.

Le calcul, pas à pas

Moyennes en milliers de dirhams, variances en millions de dh au carré

Diplôme	n_j	$\bar{y}_j$	V_j	$n_j V_j$	$(\bar{y}_j - 7,158)^2$	$n_j (\bar{y}_j - \bar{y})^2$
Bac	219	6,059	4,085	894,7	1,209	264,7
Bac+2	174	7,081	4,183	727,9	0,006	1,0
Bac+5	107	9,534	8,249	882,7	5,646	604,1
Total	500			2 505,3		869,8

Les trois variances

Intra = $2\,505,3/500 = 5,010$; inter = $869,8/500 = 1,740$; total = $6,750$ millions de dh^2 :

$$\underbrace{6\,750\,122}_{\text{totale}} = \underbrace{5\,010\,426}_{\text{intra}} + \underbrace{1\,739\,696}_{\text{inter}}$$

Le rapport de corrélation

Rapport de corrélation

$$\eta^2 = \frac{\text{variance inter}}{\text{variance totale}} \quad 0 \leq \eta^2 \leq 1$$

C'est la **part de la variance expliquée** par la variable qualitative.

Sur nos données

$$\eta^2 = \frac{1\,739\,696}{6\,750\,122} = 0,258$$

Le diplôme explique 25,8 % de la variance des salaires.

Comment on commente ce nombre

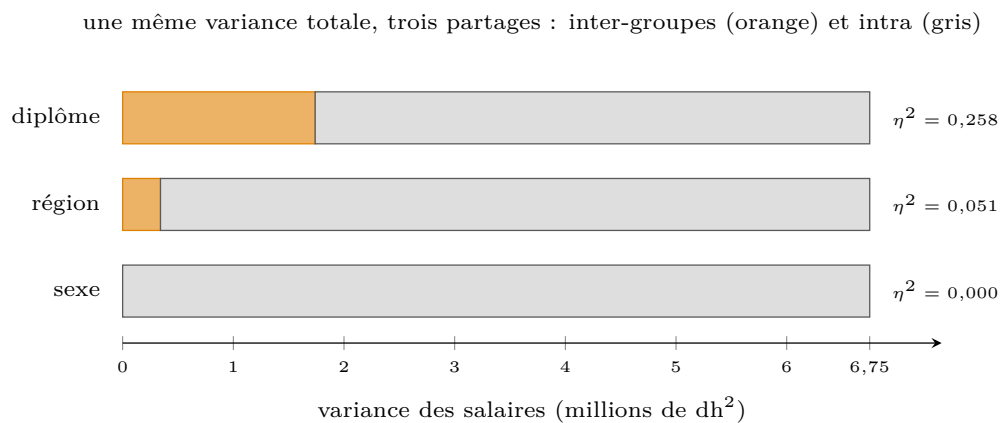
Un quart, et le reste

Un quart de la dispersion salariale s'explique par le seul diplôme. **Les trois quarts restants tiennent à autre chose** : l'ancienneté, le poste, la région, la performance individuelle — et une part irréductible d'arbitraire. À « le diplôme détermine-t-il le salaire ? », on répond « il pèse pour un quart » : c'est autrement plus utile qu'un oui ou un non.

Les deux cas extrêmes

- $\eta^2 = 0$: toutes les moyennes conditionnelles sont égales. La variable qualitative **n'apporte rien**.
- $\eta^2 = 1$: aucune variance à l'intérieur des groupes. Connaître le groupe donne **exactement** la valeur.

Trois variables explicatives comparées



Diplôme, région, sexe

Le diplôme explique 25,8 % de la variance, la région 5,1 %, le sexe **rien du tout** : les moyennes des femmes et des hommes sont égales. Pourtant leurs distributions diffèrent (chapitre 6) — η^2 ne voit que les **moyennes**.

Ce que η^2 ne suppose pas

Aucune forme de relation

η^2 fonctionne avec une variable explicative purement nominale, sans ordre. C'est ce qui le distingue du coefficient de corrélation du chapitre 13, qui n'existe qu'entre deux variables quantitatives et ne capte que la liaison **linéaire**.

Son usage

Classer des variables explicatives par leur pouvoir explicatif, avant toute modélisation : c'est exactement ce que fera l'**analyse de la variance** en statistique inférentielle, en se demandant si l'écart entre les moyennes de groupe peut tenir au hasard.

Marginales et indépendance

Conjointe, marginale, conditionnelle

Distribution marginale

Elle s'obtient en **oubliant** l'autre variable : ce sont les distributions étudiées dans toute la partie III.

$$f_{i\bullet} = \sum_j f_{ij} \qquad f_{\bullet j} = \sum_i f_{ij}$$

Le lien entre les trois familles

$$f_{ij} = f_{i\bullet} \times f_{j|i}$$

Conjointe = **marginale** $\times$ **conditionnelle**. Cette relation reviendra en probabilités sous la forme $P(A \cap B) = P(A)P(B | A)$: le même énoncé, dans un autre langage.

Le cas de l'indépendance

La caractérisation

x et y sont indépendantes si et seulement si

$$f_{ij} = f_{i\bullet} \times f_{\bullet j} \quad \text{pour toutes les cases}$$

Toutes les conditionnelles sont alors égales à la marginale : conditionner n'apporte **rien**.

Une vérification sur une case

Bac+5 et $[10\,000; 14\,000[$: $f_{ij} = 32/500 = 0,064$, alors que $f_{i\bullet} \times f_{\bullet j} = 0,214 \times 0,114 = 0,024$. Deux fois et demie plus : le diplôme et le salaire ne sont **pas** indépendants.

Le tableau de synthèse

Les trois familles, récapitulées

Ce qu'il faut savoir de chacune

	Conjointe	Marginale	Conditionnelle
Où	intérieur du tableau	les marges	une ligne, ramenée à son total
Fréquences	f_{ij}	$f_{i\bullet}, f_{\bullet j}$	$f_{j i}$
Somme	1 sur tout le tableau	1	1 par ligne
Combien	une	deux	k
Ce qu'elle donne	la liaison	le profil d'ensemble	la comparaison de groupes

Les deux théorèmes

Moyenne totale, variance totale

$$\bar{y} = \sum_j f_j \bar{y}_j \quad V(y) = \underbrace{\overline{V_j(y)}}_{\text{intra}} + \underbrace{V(\bar{y}_j)}_{\text{inter}}$$

La phrase à retenir

La moyenne des moyennes redonne la moyenne. La moyenne des variances ne redonne pas la variance — il manque la variance des moyennes. C'est la formulation la plus compacte du chapitre, et elle vaut d'être apprise telle quelle.

S'entraîner

Exercice 1 : toutes les moyennes baissent

Une entreprise sur deux années

	Effectif année 1	Salaire moyen année 1	Effectif année 2	Salaire moyen année 2
Ouvriers	80	4 000	50	3 900
Cadres	20	10 000	50	9 800

1. Calculer le salaire moyen de l'entreprise chaque année.
2. Comment évoluent les salaires moyens de chaque catégorie ? Et celui de l'entreprise ?
3. Expliquer ce paradoxe apparent.

Exercice 1 : solution

Deux moyennes pondérées

1. Année 1 : $(80 \times 4\,000 + 20 \times 10\,000)/100 = 5\,200$ dirhams. Année 2 : $(50 \times 3\,900 + 50 \times 9\,800)/100 = 6\,850$ dirhams.
2. Les ouvriers perdent 2,5 %, les cadres 2 % ; l'entreprise gagne 31,7 %.

L'effet de structure

3. La part des cadres passe de 20 % à 50 %. Le poids du groupe le mieux payé augmente si fort qu'il l'emporte sur la baisse de **chaque** moyenne. **Comparer deux moyennes générales sans regarder la structure est trompeur.**

Exercice 2 : deux entreprises fusionnent

La variance de l'entreprise fusionnée

Entreprise	Salariés	Salaire moyen	Écart-type
A	60	6 000	1 500
B	40	8 500	2 000

1. Calculer le salaire moyen de l'entreprise fusionnée.
2. Calculer sa variance intra, sa variance inter, puis sa variance totale et son écart-type.
3. Pourquoi l'écart-type de l'ensemble dépasse-t-il celui de chaque entreprise ? Calculer η^2 .

Exercice 2 : solution

Les deux théorèmes

1. $\bar{y} = (60 \times 6\,000 + 40 \times 8\,500)/100 = 7\,000$ dirhams.
2. Intra : $(60 \times 1\,500^2 + 40 \times 2\,000^2)/100 = (135 + 160)\,000\,000/100 = 2\,950\,000$. Inter : $(60 \times 1\,000^2 + 40 \times 1\,500^2)/100 = 1\,500\,000$. Totale : $4\,450\,000$, d'où $\sigma = 2\,110$ dirhams.

La lecture

3. À la dispersion interne de chaque entreprise s'ajoute l'**écart entre leurs moyennes** : 2 500 dirhams. $\eta^2 = 1,5/4,45 = 0,34$: l'entreprise d'origine explique un tiers de la variance des salaires après la fusion.

Ce qu'il faut retenir

Le résumé du chapitre

Sept points

1. Trois familles : **conjointe**, **marginale**, **conditionnelles**.
2. Conditionner, c'est **réduire la population puis recalculer**.
3. **Moyenne totale** : moyenne des moyennes conditionnelles, pondérée par les effectifs.
D'où l'**effet de structure**.
4. **Décomposition de la variance** : totale = intra + inter.
5. η^2 : part de variance expliquée par la variable qualitative. Ici 25,8 % pour le diplôme.
6. η^2 ne voit que les écarts entre **moyennes** de groupe.
7. $f_{ij} = f_{i\bullet} f_{\bullet j}$ caractérise l'**indépendance**.

La suite

Quand les **deux** variables sont quantitatives, on mesure directement dans quel sens et avec quelle force elles varient ensemble : c'est la **covariance**, dont la formule se déduit de celle de la variance. C'est le chapitre 12.