

Statistique descriptive

Chapitre 10 : Croiser deux variables

Jaouad Madkour

12 octobre 2026

Plan du chapitre

Le changement de question	2
Une variable de plus, une question nouvelle	2
Presque toute l'économie est bivariée	2
Deux formes de présentation	2
Selon le type des variables	2
Le tableau de contingence	3
L'anatomie	3
Les notations	3
Le diplôme et la région	3
Trois tableaux en un	4
Trois phrases, un seul chiffre de départ	4
Les profils-colonnes, en entier	4
Les profils, en image	5
Détecter une liaison	5
Observé contre théorique	5
La lecture des écarts	6
Le nuage de points	6
La représentation reine	6
Comment le lire	7
Un nuage résumé en tableau	7
Ce que l'épaisseur du nuage signifie	7
Le point moyen	8
Les pièges du nuage	8
S'entraîner	8
Exercice 1 : les marges ne suffisent pas	8
Exercice 1 : solution	9
Exercice 2 : du nuage au tableau croisé	9
Exercice 2 : solution	9
Ce qu'il faut retenir	10

Le changement de question

Une variable de plus, une question nouvelle

Ce que neuf chapitres n'ont pas su dire

Nous savons tout dire du salaire pris seul, et tout dire de l'ancienneté prise seule. Nous ne savons toujours pas si les salariés les plus anciens sont mieux payés. **Aucun indicateur univarié ne peut répondre à cette question** : il faut regarder les deux variables ensemble, individu par individu.

La donnée bivariée

On observe sur chaque individu un **couple** $(x_i; y_i)$. L'information essentielle n'est plus dans chaque colonne prise à part : elle est dans le **lien** entre les deux.

Presque toute l'économie est bivariée

Quatre questions, huit variables

Question	Les deux variables
La demande dépend-elle du prix ?	prix, quantité
La publicité fait-elle vendre ?	budget publicitaire, ventes
Le diplôme paie-t-il ?	diplôme, salaire
L'inflation suit-elle la masse monétaire ?	M , indice des prix

C'est le point de départ de toute l'économétrie : le cours d'introduction à l'économétrie commence exactement là où cette partie s'arrête.

Deux formes de présentation

Selon le type des variables

Trois croisements possibles

Croisement	Présentation	Graphique
qualitative $\times$ qualitative	tableau de contingence	barres empilées
quantitative $\times$ qualitative	indicateurs par groupe	boîtes côte à côte
quantitative $\times$ quantitative	tableau ou liste de couples	nuage de points

Ce chapitre traite les deux extrêmes

Le tableau de contingence pour le cas tout qualitatif, le nuage de points pour le cas tout quantitatif. Le cas mixte est celui du chapitre 11.

Le tableau de contingence

L'anatomie

le tableau de contingence : deux variables qualitatives croisées

n_{ij}	Tanger	Casa	Rabat	Autres	$n_{i\bullet}$
Bac	76	59	44	40	219
Bac+2	63	46	35	30	174
Bac+5	41	32	23	11	107
$n_{\bullet j}$	180	137	102	81	500

à l'intérieur les effectifs *conjoint* n_{ij} : combien de salariés ont à la fois ce diplôme et cette région.

dernière colonne la distribution *marginale* du diplôme : on a sommé sur les régions.

dernière ligne la distribution *marginale* de la région.

Une ligne prise seule, ramenée à son total, est une distribution *conditionnelle* : la région, sachant le diplôme.

Les marges se déduisent du tableau ; l'inverse est faux.

Les notations

Le vocabulaire

- n_{ij} : effectif **conjoint** — individus ayant la modalité i et la modalité j .
- $n_{i\bullet} = \sum_j n_{ij}$: total de la ligne i — **marge en ligne**.
- $n_{\bullet j} = \sum_i n_{ij}$: total de la colonne j — **marge en colonne**.
- $n = \sum_i \sum_j n_{ij}$: effectif total.

Le point qu'il faut avoir compris

Le tableau **détermine** les marges : il suffit de sommer. Les marges **ne déterminent pas** le tableau : connaître la répartition des diplômes et celle des régions ne dit rien de leur croisement.

Toute l'information sur la liaison est dans l'intérieur du tableau, jamais dans les marges.

Le diplôme et la région

Le tableau de contingence du fil rouge

	Tanger-Tétouan	Casablanca	Rabat	Autres	$n_{i\bullet}$
Bac	76	59	44	40	219
Bac+2	63	46	35	30	174
Bac+5	41	32	23	11	107
$n_{\bullet j}$	180	137	102	81	500

Deux contrôles

Chaque marge redonne la distribution univariée du chapitre 2 : 219, 174, 107 pour le diplôme ; 180, 137, 102, 81 pour la région. Et les deux marges totalisent 500.

Trois tableaux en un

Les trois lectures d'une même case

Fréquence	Formule	Question
conjointe	$f_{ij} = \frac{n_{ij}}{n}$	quelle part des salariés est Bac+5 et à Rabat ?
profil-ligne	$\frac{n_{ij}}{n_{i\bullet}}$	parmi les Bac+5, quelle part à Rabat ?
profil-colonne	$\frac{n_{ij}}{n_{\bullet j}}$	parmi ceux de Rabat, quelle part de Bac+5 ?

Sur la case Bac+5 et Rabat

$n_{ij} = 23$, $n_{i\bullet} = 107$, $n_{\bullet j} = 102$, $n = 500$: conjointe $23/500 = 4,6\%$; profil-ligne $23/107 = 21,5\%$; profil-colonne $23/102 = 22,5\%$.

Trois phrases, un seul chiffre de départ

Ne jamais confondre les lectures

- « 4,6 % des salariés sont des Bac+5 travaillant à Rabat. »
- « 21,5 % des Bac+5 travaillent à Rabat. »
- « 22,5 % des salariés de Rabat sont des Bac+5. »

Les trois sont exactes et disent des choses différentes. **La confusion entre profil-ligne et profil-colonne est l'erreur d'interprétation la plus fréquente en statistique appliquée** — c'est celle qui fait dire n'importe quoi aux sondages.

Où les deux profils divergent fortement

Case Bac+5 et « Autres » : $11/107 = 10,3\%$ des Bac+5 travaillent hors des trois villes, mais $11/81 = 13,6\%$ seulement des salariés de ces régions sont Bac+5. Même case, deux pourcentages, deux questions.

Les profils-colonnes, en entier

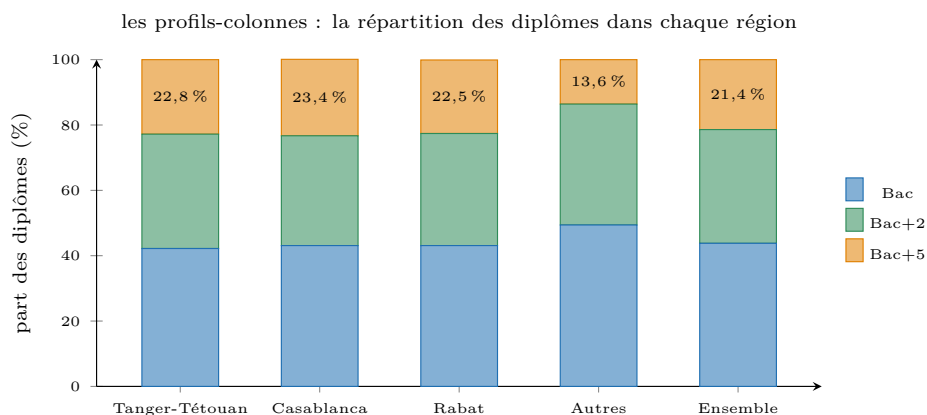
La répartition des diplômes dans chaque région (en pourcentage)

	Tanger-Tétouan	Casablanca	Rabat	Autres	Ensemble
Bac	42,2	43,1	43,1	49,4	43,8
Bac+2	35,0	33,6	34,3	37,0	34,8
Bac+5	22,8	23,4	22,5	13,6	21,4
Total	100	100	100	100	100

La colonne de référence

La colonne **Ensemble** est la marge, en pourcentage. Une région dont le profil s'en écarte se distingue de l'entreprise : c'est le cas de « Autres », et d'elle seule.

Les profils, en image



Le test visuel de l'indépendance

Si diplôme et région étaient indépendants, toutes les barres seraient **identiques** à celle de l'ensemble. Trois le sont presque ; la quatrième montre deux fois moins de Bac+5.

Détecter une liaison

Le principe

Deux variables sont **indépendantes** si les profils-lignes sont tous identiques — et alors égaux à la marge. Autrement dit : connaître le diplôme ne change rien à la répartition par région.

Les effectifs théoriques

Sous l'hypothèse d'indépendance, chaque case devrait valoir

$$n_{ij}^{\text{th}} = \frac{n_{i\bullet} \times n_{\bullet j}}{n} \quad \text{par exemple : } \frac{107 \times 81}{500} = 17,3$$

Observé contre théorique

Effectifs observés et théoriques

	Tanger-Tétouan	Casablanca	Rabat	Autres
Bac	76 / 78,8	59 / 60,0	44 / 44,7	40 / 35,5
Bac+2	63 / 62,6	46 / 47,7	35 / 35,5	30 / 28,2
Bac+5	41 / 38,5	32 / 29,3	23 / 21,8	11 / 17,3

Les écarts à l'indépendance, observé moins théorique

	Tanger-Tétouan	Casablanca	Rabat	Autres
Bac	-2,8	-1,0	-0,7	+4,5
Bac+2	+0,4	-1,7	-0,5	+1,8
Bac+5	+2,5	+2,7	+1,2	-6,3

Chaque ligne et chaque colonne d'écarts somme à zéro.

La lecture des écarts

Une seule case se détache

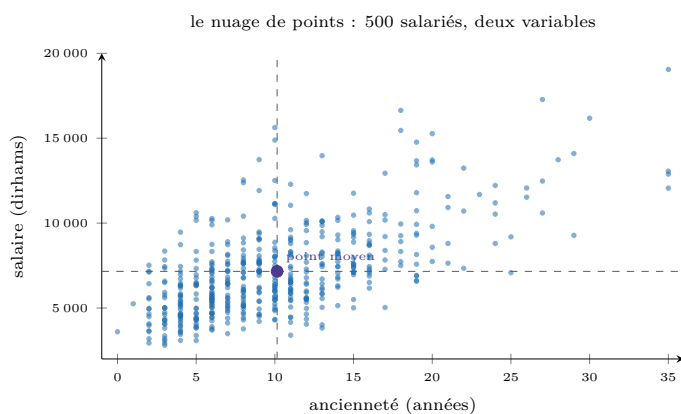
11 **Bac+5 dans les autres régions, contre 17,3 attendus.** Les diplômés du supérieur sont sous-représentés hors des trois grandes villes — ce qui correspond à une réalité économique connue, la concentration des emplois qualifiés dans les métropoles. Partout ailleurs, diplôme et région sont pratiquement indépendants.

La limite de ce chapitre

Nous venons de **constater** un écart. Nous ne pouvons pas dire s'il est significatif ou s'il tient au hasard de l'échantillon : cela demande le **test du khi-deux**, qui relève de la statistique inférentielle. La descriptive s'arrête au constat.

Le nuage de points

La représentation reine



chaque point est un salarié.

Le nuage **monte** : les deux variables varient dans le même sens.

Il est **épais** : la relation existe mais elle est loin d'être mécanique.

Le **point moyen** $(\bar{x}; \bar{y})$ appartient toujours à la droite d'ajustement.

Comment le lire

Quatre questions, dans cet ordre

1. **Sens** : le nuage monte-t-il, descend-il, ou n'a-t-il pas de direction ?
2. **Forme** : la tendance est-elle une droite, une courbe, une marche d'escalier ?
3. **Intensité** : le nuage est-il resserré autour de sa tendance, ou épais ?
4. **Anomalies** : y a-t-il des points isolés, des groupes distincts ?

Le nuage salaire contre ancienneté

1. **Sens** : il monte. Les deux variables varient dans le même sens.
2. **Forme** : la tendance paraît linéaire — un ajustement par une droite est légitime.
3. **Intensité** : le nuage est **épais**. Deux salariés de même ancienneté peuvent avoir des salaires très différents.
4. **Anomalies** : quelques points élevés à droite, aucun groupe séparé.

Un nuage résumé en tableau

Ancienneté et salaire, en classes

Ancienneté	2 à 4	4 à 6	6 à 8	8 à 10	10 à 14	14 à 20
0 à 4 ans	19	42	18	3	0	0
5 à 9 ans	10	72	68	19	11	0
10 à 14 ans	2	40	49	30	16	2
15 à 19 ans	0	3	27	21	12	3
20 ans et plus	0	0	4	6	18	5

En colonnes : classes de salaire, en milliers de dirhams.

La diagonale

Les gros effectifs suivent une **diagonale** descendante, du coin haut-gauche au coin bas-droit ; les deux autres coins sont vides. C'est la signature, dans un tableau, d'une liaison positive. Le chapitre 11 en tirera les distributions conditionnelles.

Ce que l'épaisseur du nuage signifie

L'ancienneté n'explique pas tout

Elle dit que le diplôme, la région, le poste occupé entrent aussi en jeu. C'est ce que le chapitre 13 chiffrera avec r^2 — et ce qui justifie de passer, plus tard, à la régression **multiple**, qui fait intervenir plusieurs variables explicatives à la fois.

Dans le tableau

Parmi les salariés de 5 à 9 ans d'ancienneté, 10 gagnent moins de 4 000 dirhams et 11 plus de 10 000 : à ancienneté voisine, un écart de un à trois.

Le point moyen

Point moyen

$$G = (\bar{x} ; \bar{y})$$

C'est le centre de gravité du nuage : (10,14 ans ; 7 158 dirhams) sur nos données.

Sa propriété

La droite d'ajustement du chapitre 13 passera **toujours** par le point moyen, quelle que soit la forme du nuage. C'est une conséquence du fait que la somme des écarts à la moyenne est nulle (chapitre 5), et c'est un **contrôle utile** : si votre droite ne passe pas par le point moyen, le calcul est faux.

Les pièges du nuage

Trois configurations à reconnaître

La liaison non linéaire. Un nuage en cloche ou en U a une structure évidente, et pourtant tous les indicateurs linéaires du chapitre 13 y seront nuls.

Le point aberrant. Un seul point très éloigné peut créer de toutes pièces une corrélation qui n'existe pas — ou en masquer une réelle.

Les groupes mélangés. Deux sous-populations, chacune sans liaison interne, peuvent former ensemble un nuage qui monte : c'est le **paradoxe de Simpson**.

La règle absolue

On regarde toujours le nuage avant de calculer un coefficient, et jamais l'inverse. Le chapitre 13 montrera quatre jeux de données ayant exactement le même coefficient de corrélation et quatre nuages sans rapport.

S'entraîner

Exercice 1 : les marges ne suffisent pas

200 clients, deux agences, deux produits

Une banque connaît seulement les marges de son tableau : 120 clients à Tanger et 80 à Tétouan ; 90 clients ont un produit d'épargne, 110 un crédit.

1. Construire le tableau qu'on obtiendrait si agence et produit étaient indépendants.
2. Le tableau réel est le suivant. Vérifier qu'il a les mêmes marges.

	Épargne	Crédit
Tanger	30	90
Tétouan	60	20

3. Calculer les profils-lignes des deux tableaux et conclure.

Exercice 1 : solution

Les deux tableaux

1. $120 \times 90/200 = 54$, $120 \times 110/200 = 66$, $80 \times 90/200 = 36$, $80 \times 110/200 = 44$.
2. Lignes : $30 + 90 = 120$ et $60 + 20 = 80$; colonnes : $30 + 60 = 90$ et $90 + 20 = 110$. Mêmes marges.

Les profils

3. Sous indépendance, les deux agences ont le même profil : $54/120 = 36/80 = 45\%$ d'épargne. Dans le tableau réel : $30/120 = 25\%$ d'épargne à Tanger, $60/80 = 75\%$ à Tétouan. **Mêmes marges, liaison très forte** : les marges ne disaient rien du croisement.

Exercice 2 : du nuage au tableau croisé

Dix salariés d'une agence

Ancienneté	2	3	4	5	6	8	10	12	15	20
Salaire	4,5	5,0	6,5	5,5	7,5	6,0	8,5	7,5	9,0	11,0

Ancienneté en années, salaire en milliers de dirhams.

1. Calculer le point moyen.
2. Classer chaque salarié selon que son ancienneté et son salaire sont au-dessus ou au-dessous de leur moyenne, et dresser le tableau croisé 2×2 .
3. Calculer les effectifs théoriques sous indépendance. Que conclure ?

Exercice 2 : solution

Point moyen et tableau croisé

1. $\bar{x} = 85/10 = 8,5$ ans ; $\bar{y} = 71/10 = 7,1$ milliers de dirhams.

	Salaire $\leq 7,1$	Salaire $> 7,1$	Total
Ancienneté $\leq 8,5$	5	1	6
Ancienneté $> 8,5$	0	4	4
Total	5	5	10

La comparaison

3. Théoriques : $6 \times 5/10 = 3$ dans chacune des deux premières cases, $4 \times 5/10 = 2$ dans les deux autres. On observe 5 et 4 sur la diagonale au lieu de 3 et 2 : **liaison positive nette**. Un seul salarié — 6 ans, 7 500 dh — est « mal placé ».

Ce qu'il faut retenir

Le résumé du chapitre

Sept points

1. La donnée bivariée est un **couple** $(x_i; y_i)$; l'information est dans le **lien**.
2. Le **tableau de contingence** détermine ses marges ; les marges ne déterminent pas le tableau.
3. Trois lectures d'une case : conjointe, **profil-ligne**, **profil-colonne** — ne jamais les confondre.
4. Sous **indépendance**, $n_{ij}^{\text{th}} = n_{i\bullet} n_{\bullet j} / n$; les écarts à ces effectifs signalent la liaison.
5. Indépendance : des profils tous égaux à la marge.
6. Le **nuage de points** se lit dans l'ordre : sens, forme, intensité, anomalies.
7. La droite d'ajustement passe toujours par le **point moyen**.

La suite

Un tableau croisé contient trois familles de distributions — conjointe, marginales, conditionnelles — reliées par deux théorèmes. C'est le chapitre 11.