

Statistique descriptive

Chapitre 11 : Croiser deux variables



Jaouad Madkour

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

Le changement de question

Les deux formes de présentation

Le tableau de contingence

Le nuage de points

Ce qu'il faut retenir

Le changement de question

Ce que dix 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 bivariable

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 ce chapitre s'arrête.

Les deux formes de présentation

Trois croisements possibles

Croisement	Présentation	Graphique
qualitative $\times$ qualitative	tableau de contingence	barres groupées
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 12.

Le tableau de contingence

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 *conjointes* 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.

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.

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

$n_{ij} = 23, n_{i\bullet} = 107, n_{\bullet j} = 102, n = 500.$

Lecture	Calcul	Résultat
Conjointe	23/500	4,6 %
Profil-ligne	23/107	21,5 %
Profil-colonne	23/102	22,5 %

Trois nombres, trois phrases, un seul chiffre de départ

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

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

$$n_{ij}^{\text{th}} = \frac{n_{i\bullet} \times n_{\bullet j}}{n}$$

Le test visuel sur nos données

	Tanger	Casa	Rabat	Autres
Bac observé / théorique	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

La lecture

Les écarts sont faibles partout, sauf une case : **11 Bac+5 dans les autres régions, contre 17,3 attendus.**

Diagnostic : les diplômés du supérieur sont légèrement 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.

(suite)

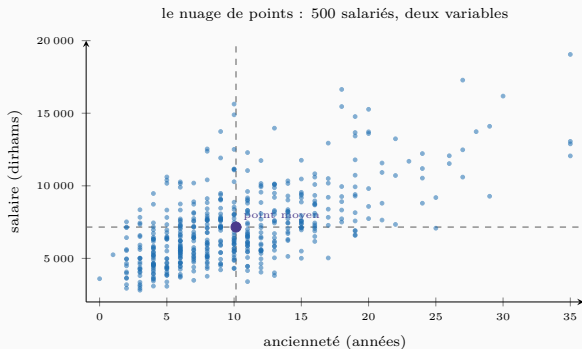
La limite de ce chapitre

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 — c'est exactement la frontière tracée au chapitre 1.

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.

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.

Ce que l'épaisseur du nuage signifie

Elle dit que l'ancienneté **n'explique pas tout**. Le diplôme, la région, le poste occupé entrent aussi en jeu.

C'est précisément ce que le chapitre 14 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.

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 14 passera **toujours** par le point moyen, quelle que soit la forme du nuage.

C'est une conséquence directe du fait que la somme des écarts à la moyenne est nulle (chapitre 6), et c'est un **contrôle utile** : si votre droite ne passe pas par le point moyen, le calcul est faux.

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 14 y seront nuls. Le nuage le voit; le coefficient ne le voit pas.

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**, et il est la cause de nombreuses conclusions fausses en sciences sociales.

La règle absolue

On regarde toujours le nuage avant de calculer un coefficient, et jamais l'inverse.

Le chapitre 14 montrera quatre jeux de données ayant exactement le même coefficient de corrélation et quatre nuages sans rapport. C'est la démonstration définitive de cette règle.

Ce qu'il faut retenir

Six 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 signalent la liaison.
5. Le **nuage de points** se lit dans l'ordre : sens, forme, intensité, anomalies.
6. La droite d'ajustement passe toujours par le **point moyen**.

La suite

Le tableau croisé contient en réalité **trois familles de distributions** : conjointe, marginales et conditionnelles. Chacune a sa moyenne et sa variance, et elles sont reliées entre elles par deux théorèmes remarquables.

C'est le chapitre 12.