

R2.08 - Statistique descriptive Cours 2 - Description bidimensionnelle

A. Ridard

A propos de ce document

- Pour naviguer dans le document, vous pouvez utiliser :
 - le menu (en haut à gauche)
 - l'icône en dessous du logo IUT
 - les différents liens
- Pour signaler une erreur, vous pouvez envoyer un message à l'adresse suivante :
anthony.ridard@univ-ubs.fr

Plan du cours

- 1 Liaison entre deux variables qualitatives
- 2 Liaison entre une variable quantitative et une variable qualitative
- 3 Liaison entre deux variables quantitatives
 - Interpréter le nuage de points
 - Construire le modèle
 - Mesurer la qualité du modèle

Après les descriptions unidimensionnelles, on s'intéresse aux liaisons entre les variables observées, c'est l'étude des corrélations.



- I Les indicateurs de « dépendance » varient selon la nature des variables.

- 1 Liaison entre deux variables qualitatives
- 2 Liaison entre une variable quantitative et une variable qualitative
- 3 Liaison entre deux variables quantitatives

On considère les deux variables qualitatives X et Y respectivement à r et s modalités.

Les données pour notre échantillon de taille n sont alors présentées sous la forme d'un tableau, appelé **tableau de contingence**, à r lignes et s colonnes renfermant les effectifs n_{ij} d'individus ayant pour X la valeur x_i et pour Y la valeur y_j :

	y_1	...	y_s
x_1	n_{11}	...	n_{1s}
\vdots			\vdots
x_r	n_{r1}	...	n_{rs}

Pour construire un tel tableau avec Python¹, on pourra utiliser la fonction **pd.crosstab()** ou la méthode **.pivot_table()** (cf. section 4.3.12 du livre)

1. On dit que l'on fait un « tri croisé » en opposition au « tri à plat » réalisé à une dimension.

Il est souvent complété par les **marges en lignes** $n_{i.}$ et les **marges en colonnes** $n_{.j}$:

	y_1	...	y_s	
x_1	n_{11}	...	n_{1s}	$n_{1.}$
\vdots				\vdots
x_r	n_{r1}	...	n_{rs}	$n_{r.}$
	$n_{.1}$...	$n_{.s}$	n

Avec évidemment $n_{i.} = \sum_{j=1}^s n_{ij}$ et $n_{.j} = \sum_{i=1}^r n_{ij}$.

On peut en déduire le tableau des **profils-lignes** :

	y_1	...	y_s
x_1	$\frac{n_{11}}{n_{1.}}$...	$\frac{n_{1s}}{n_{1.}}$
⋮			⋮
x_r	$\frac{n_{r1}}{n_{r.}}$...	$\frac{n_{rs}}{n_{r.}}$

Et celui des **profils-colonnes** :

	y_1	...	y_s
x_1	$\frac{n_{11}}{n_{.1}}$...	$\frac{n_{1s}}{n_{.s}}$
⋮			⋮
x_r	$\frac{n_{r1}}{n_{.1}}$...	$\frac{n_{rs}}{n_{.s}}$

Lorsque tous les profils-lignes sont identiques, on parle d'**indépendance empirique** entre X et Y puisque la connaissance de X ne change pas le comportement de Y . Dans ce cas, tous les profils-colonnes sont également identiques.

Cette condition d'indépendance empirique se traduit par :

$$\forall j, \frac{n_{1j}}{n_{1.}} = \frac{n_{2j}}{n_{2.}} = \dots = \frac{n_{rj}}{n_{r.}}$$

Ce qui entraîne :

$$\frac{n_{ij}}{n_{i.}} = \frac{n_{.j}}{n}$$

Ou encore :

$$n_{ij} = \frac{n_{i.} n_{.j}}{n}$$

On mesure alors l'écart à l'indépendance de la manière suivante :

$$d^2 = \sum_{i=1}^r \sum_{j=1}^s \frac{\left(n_{ij} - \frac{n_{i.} n_{.j}}{n} \right)^2}{\frac{n_{i.} n_{.j}}{n}}$$

- 1 Liaison entre deux variables qualitatives
- 2 Liaison entre une variable quantitative et une variable qualitative
- 3 Liaison entre deux variables quantitatives

On considère une variable qualitative X à k modalités (catégories) et l'on note :

- n_1, \dots, n_k les effectifs observés pour X
- $\bar{y}_1, \dots, \bar{y}_k$ les moyennes de Y pour chaque catégorie
- \bar{y} la moyenne (totale) de Y

Le comportement de Y peut-il s'expliquer à l'aide X ? Si oui, dans quelle mesure?

On peut décomposer la variance de Y sous la forme :

$$s_y^2 = \frac{1}{n} \sum_{j=1}^n (y_j - \bar{y})^2 = \frac{1}{n} \sum_{i=1}^k n_i (\bar{y}_i - \bar{y})^2 + \frac{1}{n} \sum_{i=1}^k n_i s_i^2$$

où les s_i^2 sont les variances de Y à l'intérieur de chaque catégorie.



Variance de Y = Variance **inter-catégories** + Variance **intra-catégories**

On calcule alors la part des variations de Y expliquées par X de la manière suivante :

$$e^2 = \frac{\frac{1}{n} \sum_{i=1}^k n_i (\bar{y}_i - \bar{y})^2}{s_y^2}$$

Il s'agit du **rapport de corrélation empirique** de Y sachant X .

- 1 Liaison entre deux variables qualitatives
- 2 Liaison entre une variable quantitative et une variable qualitative
- 3 Liaison entre deux variables quantitatives**

On considère deux variables quantitatives X et Y .

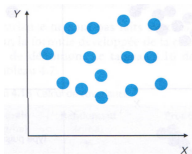
Le comportement de Y peut-il s'expliquer à l'aide de X ? Si oui, dans quelle mesure?
Plus précisément, peut-on trouver une fonction f telle que $Y \simeq f(X)$?

Cette modélisation est réalisée en 3 étapes :

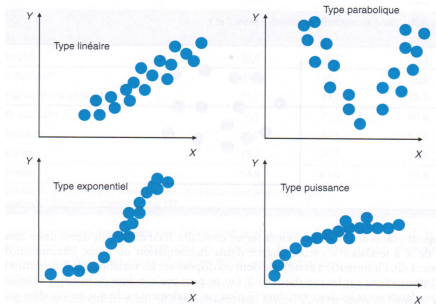
- On construit le nuage de points pour :
 - infirmer ou confirmer l'intuition de dépendance
 - déterminer la forme du modèle (linéaire, puissance, exponentielle, logistique)
- On construit le modèle :
 - s'il est linéaire, on utilise la méthode des moindres carrés ordinaires (MCO)
 - sinon, on effectue un changement de variables pour se ramener au cas linéaire
- On mesure la qualité du modèle

- 1 Liaison entre deux variables qualitatives
- 2 Liaison entre une variable quantitative et une variable qualitative
- 3 Liaison entre deux variables quantitatives**
 - **Interpréter le nuage de points**
 - Construire le modèle
 - Mesurer la qualité du modèle

Indépendance : absence de lien entre X et Y



Dépendance de formes différentes



On définit la **covariance empirique** de (X, Y) de la manière suivante :

$$Cov_{xy} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$



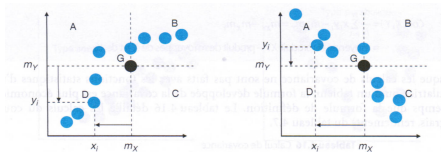
La covariance est un indicateur de monotonie

- si elle est positive, alors X et Y varient dans le même sens
- si elle est négative, alors X et Y varient dans le sens contraire
- Si la covariance est nulle ou presque nulle, alors il n'y a pas de tendance croissante ou décroissante et les variables sont dites non corrélées.

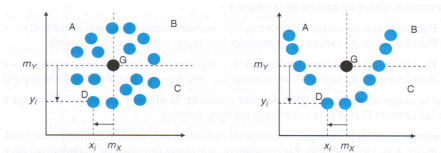


- La covariance n'est pas un indicateur d'indépendance
- Deux var. indépendantes sont non corrélées mais la réciproque est fausse

Covariance positive (resp. négative)



Covariance nulle



$m_X = \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ et $m_Y = \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$ désignent les moyennes empiriques

- 1 Liaison entre deux variables qualitatives
- 2 Liaison entre une variable quantitative et une variable qualitative
- 3 Liaison entre deux variables quantitatives**
 - Interpréter le nuage de points
 - Construire le modèle**
 - Mesurer la qualité du modèle

Lorsque le nuage de points suggère une relation linéaire entre X et Y , on cherche l'équation de la droite $y = ax + b$ qui ajuste « au mieux » le nuage de points.

Autrement dit, on cherche a et b qui minimisent :

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^N (y_i - (ax_i + b))^2 = f(a, b)$$

où $\hat{y}_i = ax_i + b$ est la valeur de Y estimée par le modèle à partir de x_i

Ce problème a pour solution² la droite définie par les deux conditions suivantes :

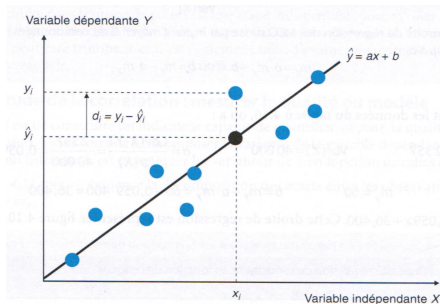
- elle passe par le point moyen (\bar{x}, \bar{y})
- son coefficient directeur est $a = \frac{Cov_{xy}}{s_x^2}$

La droite d'ajustement des moindres carrés ordinaires a donc pour équation :

$$y = \frac{Cov_{xy}}{s_x^2}(x - \bar{x}) + \bar{y}$$

2. On résout le système correspondant à l'annulation des deux dérivées partielles de f

En fait, on a pour tout i : $y_i = (ax_i + b) + d_i = \hat{y}_i + d_i$ avec $\frac{1}{n} \sum_{i=1}^n d_i = 0$



En moyenne, Y se comporte donc comme $aX + b$

De nombreux modèles non linéaires se ramènent au modèle linéaire :

- Le modèle **puissance** $Y = cX^d$ se ramène à

$$Y' = \ln c + dX' \text{ avec } Y' = \ln Y \text{ et } X' = \ln X$$

- Le modèle **exponentielle** $Y = ce^{dX}$ se ramène à

$$Y' = \ln c + dX \text{ avec } Y' = \ln Y$$

- Le modèle **logistique** $Y = \frac{e^{c+dX}}{1 + e^{c+dX}}$ se ramène à

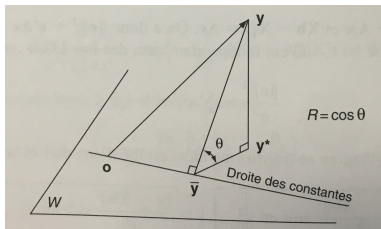
$$Y' = c + dX \text{ avec } Y' = \ln \frac{Y}{1 - Y}$$

- 1 Liaison entre deux variables qualitatives
- 2 Liaison entre une variable quantitative et une variable qualitative
- 3 Liaison entre deux variables quantitatives**
 - Interpréter le nuage de points
 - Construire le modèle
 - Mesurer la qualité du modèle**

On va décomposer, là encore, la variance de Y :

$$s_y^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2 = \frac{1}{n} \sum_{i=1}^n n_i (y_i - \hat{y}_i)^2 + \frac{1}{n} \sum_{i=1}^n n_i (\hat{y}_i - \bar{y})^2$$

Ce résultat s'obtient à partir du théorème de Pythagore, rappelez-vous :



Variance de Y = Variance **résiduelle** + Variance **expliquée**

On calcule alors la part des variations de Y expliquées par le modèle :

$$r^2 = \frac{\frac{1}{n} \sum_{i=1}^n n_i (\hat{y}_i - \bar{y})^2}{s_y^2}$$

Il s'agit du **coefficient de détermination**.



C'est aussi le carré du **coefficient de corrélation linéaire** défini par :

$$r = \frac{Cov_{xy}}{s_x s_y}$$

D'ailleurs, on peut exprimer la pente de la droite d'ajustement en fonction de r :

$$a = \frac{Cov_{xy}}{s_x^2} = \frac{r s_y}{s_x}$$



Le modèle qui ajuste au mieux les données (d'apprentissage) n'est pas forcément celui qui aura les meilleures prévisions. Il s'agit du phénomène de sur-apprentissage que vous verrez l'année prochaine en Machine Learning...