Analyse bivariée

Il est souvent important d'étudier une population en utilisant deux caractères. Dans ce cas, on parle de statistique bivariée.

L'objectif de l'analyse bivariée est d'étudier les éventuelles relations entre deux variables statistiques

1. Deux caractères quantitatifs:

Représentation graphique et/ou avec Tableau

Les couples de points

$$(x_1, y_1), \ldots, (x_i, y_i), \ldots, (x_n, y_n)$$

Peuvent être regroupés dans un tableau et/ou représentés comme un ensemble de points dans un plan,

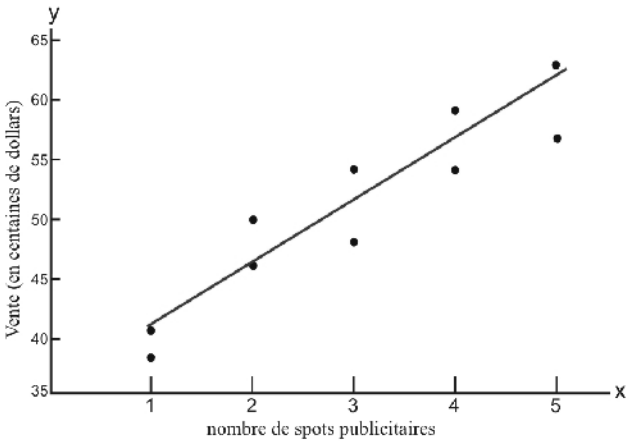
ce qui produit *un nuage de points*.

Exemple:

Afin d'étudier la relation entre la publicité et la vente pour un magasin d'équipement bureautique, le tableau T1 présente des données prises au cours de trois mois sur le nombre de spots publicitaires diffusés en fin de semaine et les ventes réalisés au cours de la semaine suivante.

Semaine	*NB spots X*	*Vol vente (centaines Dollars)*
1	2	50
2	5	57
3	1	41
4	3	54
5	4	54
6	1	38
7	5	63
8	3	48
9	4	59
10	2	46

La représentation graphique est donnée comme suit :



La droite en rouge fournit une approximation de la relation entre x et y est appelé la tendance.

<u>Etude de la liaison entre deux variables</u>

Même si lorsque le nombre d'observations est réduit, un simple graphique nous permet souvent de voir rapidement s'il y a ou non une relation entre les deux variables et de quel type elle est si jamais elle existe.

La représentation graphique ou le tableau ne permettons pas de façon sûre de décider si deux variables sont liées, dépendants ou totalement indépendants et elles sont parfois trempeuses

Paradoxe de Simpson :
C'est un paradoxe statistique dans lequel un phénomène observé sur deux ou plusieurs tableaux croisés semble s'inverser dans certains cas lorsque ces tableaux sont combinés ou agrégés. Par conséquent, il devient claire d'être prudent dans l'interprétation des relations entre deux variables que l'on pourrait faire à partir d'un tableau croisé agrégé.

Exemple (Paradoxe de Simpson):
les tableaux suivants présentent une comparaison entre deux juges selon le nombres de leur jugements (dans le Tribunal et la Cour) maintenus ou annulés à la Cour suprême durant trois années.

	Juge 1		
Jugement	**Tribunal**	**Cour**	**Total**
Maintenu	29 (91%)	100 (85%)	129
Annulé	3 (9%)	18 (15%)	21
Total (%)	32 (100%)	118 (100%)	150

	Juge 2		
Jugement	**Tribunal**	**Cour**	**Total**
Maintenu	90 (90%)	20 (80%)	110
Annulé	10 (10%)	5 (20 %)	15
Total (%)	100 (100%)	25 (100 %)	125

Juge 1
obtient un meilleur score

En revanche :

Si nous composons le tableau croisé agrégés qui suit :

	Juge 1		
Jugement	**Tribunal**	**Cour**	**Total**
Maintenu	29 (91%)	100 (85%)	129
Annulé	3 (9%)	18 (15%)	21
Total (%)	32 (100%)	118 (100%)	150

	Juge 2		
Jugement	**Tribunal**	**Cour**	**Total**
Maintenu	90 (90%)	20 (80%)	110
Annulé	10 (10%)	5 (20 %)	15
Total (%)	100 (100%)	25 (100 %)	125

Juge 1
obtient un meilleur score

	Juge		
Jugement	**Juge 1**	**Juge 2**	**Total**
Maintenu	129 (86%)	110 (88%)	239
Annulé	21 (14%)	15 (12 %)	36
Total (%)	150 (100%)	125 (100 %)	275

Juge 2
obtient un meilleur score

1. Notions de dépendance et d'indépendance

1. Variables liées

On dit que variables sont liées si les variations de l'une dépendent des variations de l'autre. On distingue deux cas :

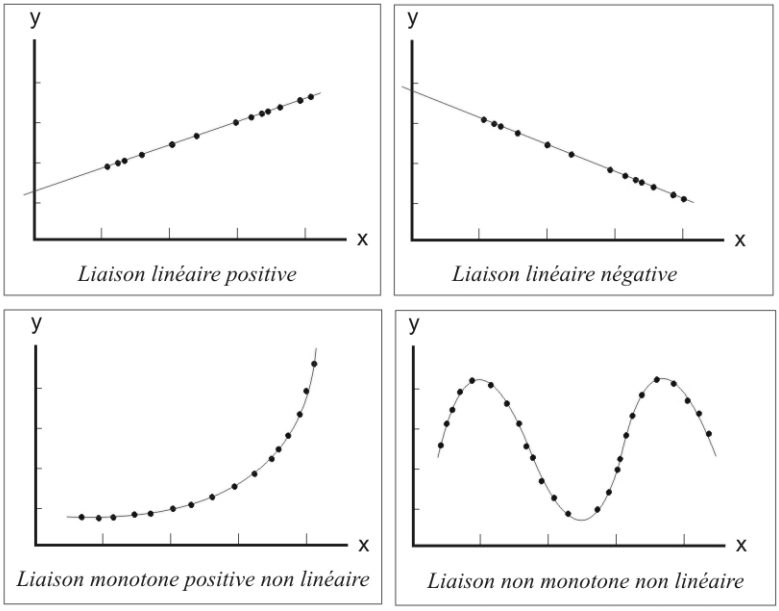
a). Liaison fonctionnelle

On dit qu'il existe une liaison fonctionnelle entre x et y si à chaque valeur de l'une des variables correspond une seule valeur de l'autre.

On note la relation : $y = f(x)$ où y est une fonction de x

Graphiquement, la liaison fonctionnelle peut prendre plusieurs formes de courbes, mais dans le cas linéaire, elle se traduit par l'alignement du nuage de points qui prend l'allure d'une ligne droite.

Dans la figure suivante, nous illustrons quelques types de liaisons fonctionnelles qui peuvent exister entre deux variables continues :



b). La dépendance

En sciences humaines, il est très rare d'identifier statistiquement une liaison fonctionnelle (lien de causalité) ou une indépendance totale. Le cas le plus fréquent est la dépendance, c-à-d, les deux variables entretiennent une relation plus ou moins forte selon les cas.

On note la relation :

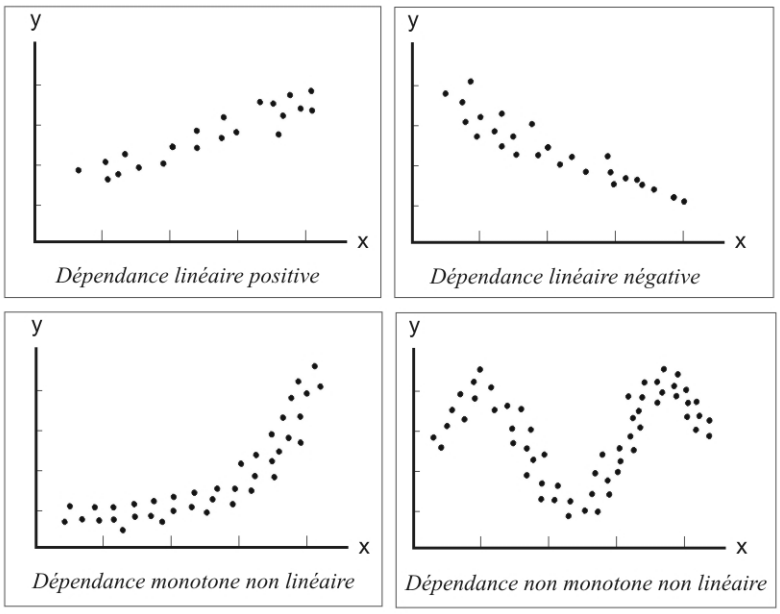
$$y = f(x) \pm \varepsilon$$

partie certaine — partie aléatoire

Cette incertitude peut être imputée à quatre types d'impact pouvant exister d'une manière isolée ou concomitante :

1. l'intervention d'autres variables.
2. la présence de facteurs aléatoires d'erreurs.
3. des erreurs d'échantillonnage.
4. des erreurs de mesure (relatives aux instruments, méthodes et aux techniques utilisés ou adoptés).

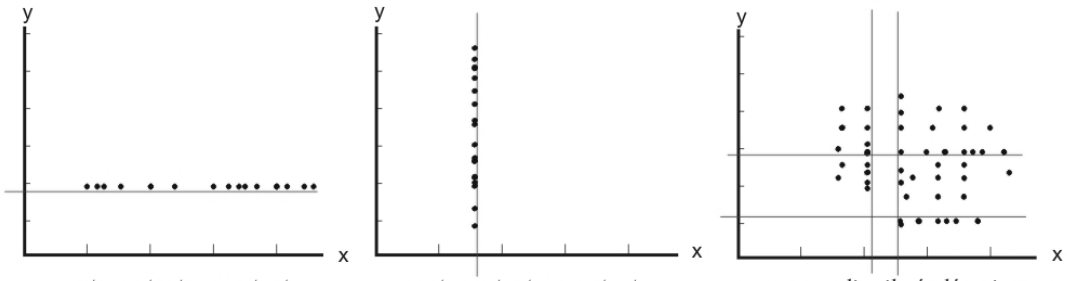
Dans la figure suivante, nous illustrons quelques types de dépendance :



2. Variables indépendantes

L'indépendance signifie l'absence totale de relation entre les deux variables x et y u'elles varient indépendamment l'une de l'autre.

1: la connaissance de la valeur de l'une des deux variables n'apporte aucune information sur la valeur de l'autre variable.
2: la répartition des points est similaire à celle produite par le hasard.



y est constante pour tout x

x est constante pour tout y

x et y sont distribué aléatoirement
Pour une valeur x, il y a plusieurs valeurs y et vice versa

2. Mesures de la relation entre deux variables quantitatives

a. Analyse graphique

L'intérêt du graphique est qu'il nous permet :
- de déterminer la forme globale des points
- voir s'il existe une forme de liaison ou de régularité dans le nuage de points;
- situer les proximités entre les individus;
- détecter visuellement les points qui s'écartent des autres;
- vérifier s'il n'y a pas de regroupement suspects, laissant entendre qu'il y a en réalité d'autres variables qui influencent le positionnement des points.

L'analyse graphique est une bonne manière de comprendre ou au moins pour prendre la première impression sur la relation entre les deux variables.

b. Analyse Numérique

1. La covariance

La covariance peut être utilisée dans la première étape de mesure pour confirmer ou infirmer l'existence d'une relation entre deux variables x et y

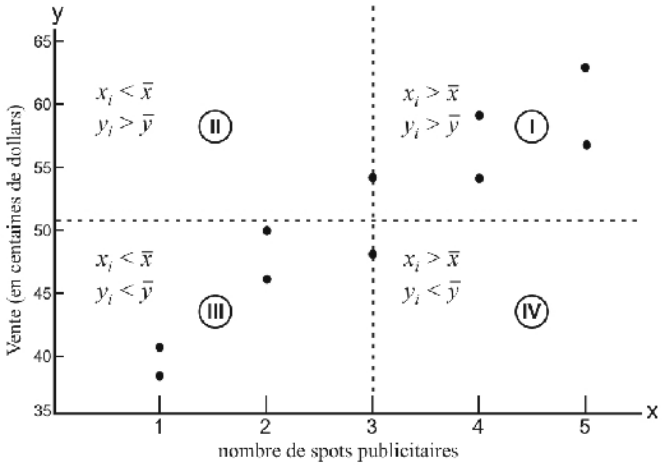
$$S_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \quad \text{Ou bien} \quad S_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} x_i y_i - \bar{x}\,\bar{y}$$

Reconsidérons l'exemple précédent du magasin d'équipement bureautique pour voir la relation entre le nombre de spots publicitaires et la vente .

x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$
2	50	-1	-1	1
5	57	2	6	12
1	41	-2	-10	20
3	54	0	3	0
4	54	1	3	3
1	38	-2	-13	26
5	63	2	12	24
3	48	0	-3	0
4	59	1	8	8
2	46	-1	-5	5
Total = 30	Total = 510	Total = 0	Total = 0	Total = 99

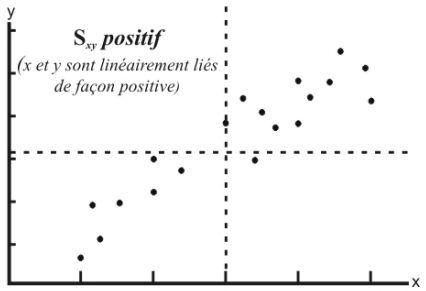
$$S_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = 99/(10-1) = 11$$

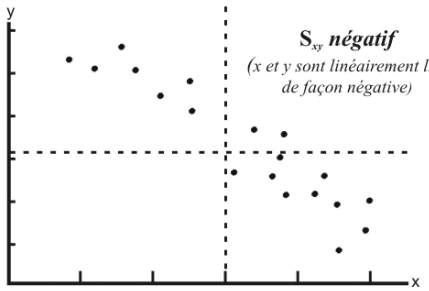
Interprétation de la covariance

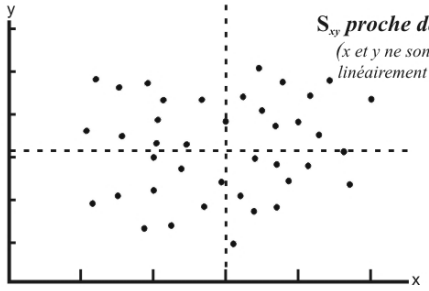


La covariance de deux variables indépendantes est nulle, bien que la réciproque ne soit pas toujours vraie.

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







L'inconvénient de la covariance comme mesure de l'association entre deux variables est qu'elle dépend des unités de mesures des variables x et y (centimètre / mètre comme unité de mesure de taille par exemple et kg / gr comme unité de mesure de poids).

2. Coefficient de corrélation de Pearson

Est définie comme suit :

$$r_{xy} = \frac{S_{xy}}{\delta_x \delta_y}$$

S_{xy} : la covariance de l'échantillon
δ_x : l'écart type de l'échantillon x
δ_y : l'écart type de l'échantillon y

$$r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}} = \frac{n\sum xy - \sum x \sum y}{\sqrt{n\sum x^2 - (\sum x)^2}\, n\sum y^2 - (\sum y)^2}$$

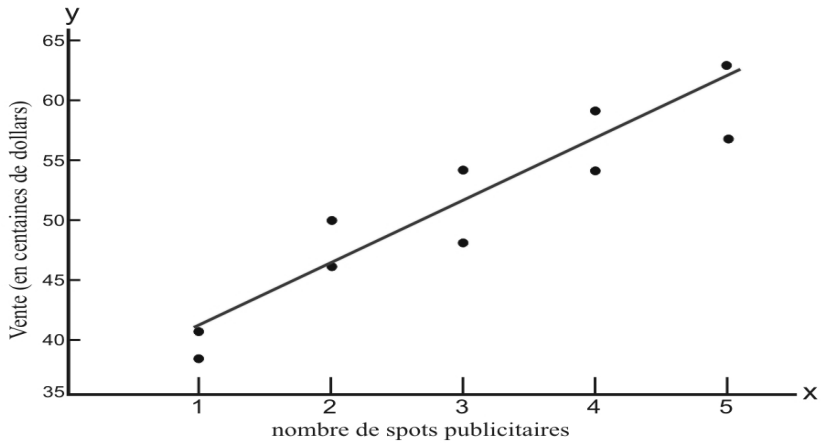
Exemple (Continuons sur le même exemple)

x_i	y_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$y_i - \bar{y}$	$(y_i - \bar{y})^2$
2	50	-1	1	-1	1
5	57	2	4	6	36
1	41	-2	4	-10	100
3	54	0	0	3	9
4	54	1	1	3	9
1	38	-2	4	-13	169
5	63	2	4	12	144
3	48	0	0	-3	9
4	59	1	1	8	64
2	46	-1	1	-5	25
Total = 30	Total = 510	Total = 0	Total = 20	Total = 0	566

$$\delta_x = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{20}{9}} = 1,49$$

$$\delta_y = \sqrt{\frac{\sum(y_i - \bar{y})^2}{n-1}} = \sqrt{\frac{566}{9}} = 7,93$$

$$r_{xy} = \frac{11}{1,49 * 7,93} = 0,93$$



Le cœfficient de corrélation varie entre - 1 et + 1.
Des valeurs proches de – 1 ou de + 1 révèlent une forte relation linéaire.
Plus le coefficient est proche de zéro, plus la relation est faible.

Le tableau suivant résume la nature de relation :

Valeur de coefficient	Explication
+1	relation positive complète
0.70 – 0.99	relation positive forte
0.50 – 069	relation positive moyenne
0.01 – 0.49	relation négative faible
0	pas de relation linéaire

Valeur de coefficient	Explication
-1	relation négative complète
-0.70 – -0.99	relation négative forte
-0.50 – -069	relation négative moyenne
-0.01 – -0.49	relation négative faible