L'ANALYSE DE LA RELATION ENTRE DEUX VARIABLES

A - Corrélation et régression linéaires simples

Souvent on est amené à s'interroger sur la nature de la relation qui peut exister entre deux variables. Y'a-t-il une relation entre deux variables données ?
De quelle intensité et de quelle forme est-elle ?

Pour le besoin de l'analyse, on se limitera ici à deux séries statistiques x et y. L'analyse porte ainsi sur **une distribution à deux caractères** (ou dimensions) ou **bivariée** et a pour objet l'étude **de l'intensité** de la relation (*la corrélation*) et **sa forme** entre les deux variables (*la régression*).

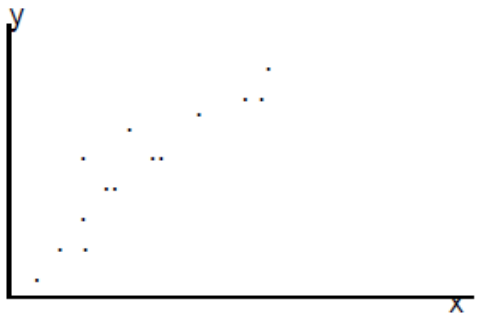
La distribution bivariée est une distribution où à chaque observation i correspond un couple de deux modalités xi, yi, elle se présente sous forme d'une liste (ou d'un tableau). Dans ce chapitre on s'intéressera à la relation linéaire entre deux variables quantitatives, sous la forme d'une liste de couples x et y. On étudiera dans les chapitres suivants la relation courbe entre deux variables ou dans un tableau.

1 - PRESENTATION : *Nuage de points et Table de contingence*

Comme pour le cas d'une distribution d'une variable (univariée), on a deux manières de présenter une distribution bivariée : le graphique et le tableau.

1.1 - Le graphique *: le nuage de points*

L'objectif étant de représenter la relation qui existe entre deux variables x et y, dans un système d'axes orthonormé x, y, on représente chaque observation i par un point à l'intersection de ses deux coordonnées xi, yi respectives, on obtient ainsi un ensemble de points qu'on appelle *nuage de points*.



Nuage de points

1.2 - Le tableau à double entrée

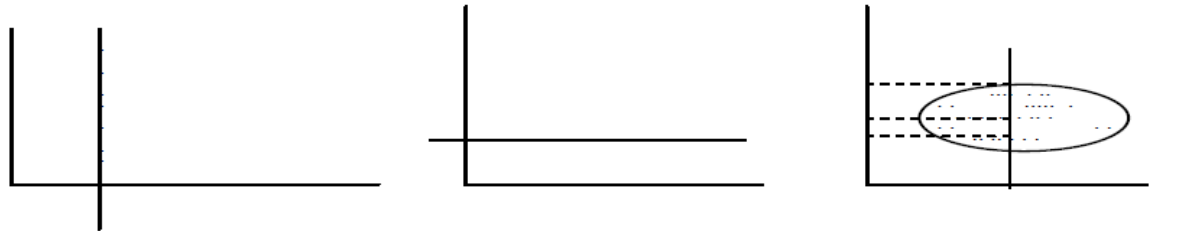
C'est un tableau à double entrée où on a en lignes les observations et en colonnes les deux variables x et y. A chaque observation i, on a deux valeurs xi et yi correspondant aux variables x et y.

2 - INDEPENDANCE, DEPENDANCE ET LIAISON FONCTIONNELLE

Le graphique ou le tableau est de nature à permettre de voir si les deux caractères sont liés, dépendants ou totalement indépendants. Ce premier test est à confirmer (ou infirmer) par un autre test plus puissant par la suite, il permet néanmoins une première sélection évitant la perte de temps et les calculs souvent inutiles. 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. Le calcul permet ensuite de confirmer ou d'infirmer cette première conclusion.

2.1 - L'indépendance

Elle exprime l'absence totale de relation entre les deux variables x et y. Elle se manifeste graphiquement par un nuage de points sous forme d'un alignement parallèle à l'un des axes (x ou y) ou d'une disposition circulaire ou elliptique. L'alignement parallèle exprime que l'une des variables est constante quelque soit la valeur de l'autre tandis que la disposition circulaire ou elliptique exprime une distribution aléatoire des observations les unes par rapport aux autres puisque chaque valeur de x (ou de y) correspond simultanément à plusieurs valeurs de la seconde variable.



x est constante pour tout y aléatoirement

y est constante pour tout x

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

Dans un tableau, l'indépendance s'exprime par des données égales ou proches, des colonnes ou des lignes proportionnelles pour l'une ou les deux variables. Dans ce cas, la connaissance de la valeur d'une variable ne permet guère celle du second caractère. Les deux variables sont indépendantes, le calcul ne doit être entamé.

2.2 - La liaison fonctionnelle

Il existe une liaison fonctionnelle entre x et y si à chaque valeur de l'une correspond une valeur donnée de y. La connaissance de x (ou de y) nous permet ainsi de déterminer y (ou x) de façon unique. On note la relation: $y = f(x)$ où y est fonction de x.
La liaison fonctionnelle linéaire se traduit sur le graphique par un alignement parfait du nuage de points qui prend l'allure générale d'une ligne droite. Elle se traduit dans un tableau par la concentration des valeurs sur la diagonale, il y a une seule valeur par ligne et par colonne.

2.3 - La dépendance

Il est souvent rare de trouver une liaison fonctionnelle ou une indépendance totale, le cas le plus fréquent est la dépendance: les deux variables entretiennent une relation plus ou moins forte selon les cas.
La dépendance s'exprime par une liaison plus complexe qui incorpore *une partie certaine* de y (qu'on peut déterminer avec certitude en connaissant la valeur de x: $y = f(x)$ et *une partie*

aléatoire, probable (ε) et réduite (on lui donne souvent le symbole d'epsilon pour exprimer le caractère négligeable dans le modèle). Le modèle général s'écrit alors sous la forme:

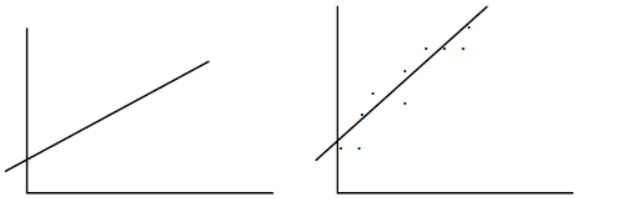
$y = f(x) \pm \varepsilon$

Partie certaine ± Partie aléatoire

Il existe une liaison entre x et y mais la connaissance de l'un ne permet de déterminer le second que dans une certaine probabilité, exprimée par ε qui est d'autant plus élevée que l'intervalle d'occurrence est étendu. Cette incertitude peut être imputée à quatre types d'effet qui peuvent exister d'une manière isolée ou concomitante :

1- l'intervention d'*autres variables que x* qui affectent la relation (entre y et x) et influent sur la valeur de y. Il est tout à fait évident que chaque élément (ou variable) se trouve très souvent lié et de là expliqué par plus d'une variable et la faiblesse de la relation exprime plutôt la présence de plus d'une variables en jeu qu'une totale indépendance. C'est ainsi que le niveau de scolarisation, par exemple, ne peut être imputé seulement à l'urbanisation, uniquement au revenu, essentiellement à l'appartenance socio-professionnelle, au genre ou à la tradition locale… Lorsqu'on ne retient qu'une seule de ces variables explicatives, c'est comme on ampute la réalité d'une partie plus ou moins importante selon les cas. A l'opposé, une forte corrélation ne signifie pas aussi que les autres variables n'interviennent pas.
2- la présence de *facteurs aléatoires d'erreurs*. En effet, en plus du facteur de base x, plus ou moins important, il y a toujours une multitude de facteurs à la fois inconnus et réduits qui interviennent dans le processus et finissent par dévier les différentes valeurs de y tantôt vers la hausse, tantôt vers la baisse ce qui explique la présence de résidus.
3- *des erreurs d'échantillonnage* peuvent être aussi imputés au choix des unités étudiées et des observations retenues pour l'analyse, un choix qui relève parfois du chercheur lui-même ou le dépasse souvent. Ce facteur d'erreur intervient systématiquement lorsque l'étude ne porte pas sur l'ensemble de la population, ce qui est souvent le cas dans la plupart des analyses.
4- des *erreurs de mesure* relatives aux instruments utilisés, aux méthodes adoptées et aux techniques sollicitées pour effectuer ces mesures. Il en est ainsi lorsqu'on embrasse un phénomène assez complexe qui correspond à plusieurs définitions comme l'urbanisation ou le chômage selon la définition retenue, on peut laisser passer une partie de la réalité.

La partie aléatoire est mesurée par la variance résiduelle, celle qui reste inexpliquée par x. Tout le problème consiste alors à minimiser cette partie aléatoire et la méthode adoptée s'appuie sur ce principe.



Liaison fonctionnelle — $y = f(x)$

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

Graphiquement, la dépendance s'exprime par *la présence d'une certaine tendance* dans le nuage de points sans que les points soient totalement alignés comme dans le cas de la liaison fonctionnelle ou totale. Le nuage de points prend *l'allure d'une ligne droite* dans le cas d'une relation linéaire mais il peut épouser plusieurs formes de courbes.

Dans un tableau, une relation linéaire de dépendance s'exprime par un rapport de proportionnalité entre les valeurs x, y. Les valeurs du tableau augmentent selon la diagonale (dans un sens ou dans l'autre) en diminuant tout autour avec des valeurs nulles de part et d'autre de cette diagonale principale.

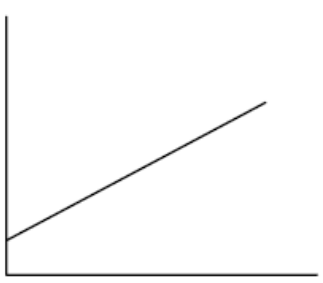
3 - LE MODELE LINEAIRE : *le rapport proportionnel*

Avant de procéder aux calculs, il faut bien s'assurer que le nuage de points a une allure générale linéaire et que les données présentent un rapport de proportionnalité quelconque vers la hausse ou la baisse.

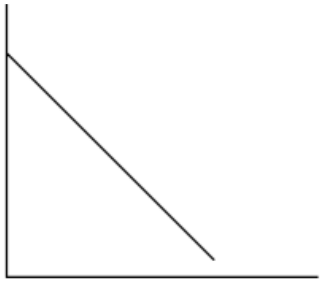
Le modèle linéaire exprime *un rapport de proportionnalité fixe et absolu* entre deux variables, à une variation absolue de la variable x d'une unité. La variation est ici prise dans le sens de hausse ou de baisse selon le type de relation qui lie les deux variables.

Le modèle linéaire a une équation de la forme $y = ax + b$. Le paramètre (a) représente *la pente*, qui exprime la quantité de variation de y lorsque x varie d'une unité tandis que (b) représente la valeur de y lorsque $x = 0$. Lorsque la valeur de a est positive, on a *une relation croissante ou directe* et les deux variables varient dans le même sens selon un rapport constant de 1 à a, c'est le cas par exemple de la scolarisation et de l'urbanisation. Quand elle est négative, on a *une relation décroissante ou inverse*, les variables x et y varient en sens opposé, lorsque l'une augmente l'autre diminue comme la pression et la température, l'offre et le prix d'un bien, l'urbanisation et le % de la population agricole...

Graphiquement l'équation est représentée par *une droite* dans un système d'axes orthonormé où a est *la pente angulaire de la droite* qui mesure la variation verticale (y) sur une distance de 1 unité de x (l'angle de la droite avec l'horizontale). Le paramètre b est *l'intersection* de la droite d'ajustement D avec l'axe des y. Lorsque a est positive, la droite est croissante (elle va dans le sens des axes x, y). Quand a est négative, la droite est décroissante : elle va dans le sens de l'axe x et en sens inverse de l'axe y.



Relation croissante ou directe ($a > 0$)



Relation décroissante ou inverse ($a < 0$)

On décide alors de faire une régression linéaire c'est-à-dire de déterminer par la méthode des moindres carrés ordinaires (MCO) l'équation de la droite qui représente le mieux cette liaison. Pour tracer la droite de régression, on doit utiliser l'équation $y = ax + b$. Le calcul des coefficients a et b est le suivant :

$$a = \frac{\sum_{i=1}^{n} x_i y_i - n \cdot \bar{x} \cdot \bar{y}}{\sum_{i=1}^{n} x_i^2 - n(\bar{x})^2}$$

$$b = \bar{y} - a\bar{x}$$

4 - LA CORRELATION LINEAIRE SIMPLE*: l'intensité de la relation*

La corrélation linéaire mesure le degré de liaison linéaire entre deux variables. Le coefficient le plus utilisé est le coefficient de corrélation linéaire ou coefficient de Bravais.

4.1 - Le coefficient de corrélation linéaire r

Le coefficient de corrélation linéaire (ou coefficient de Bravais Pearson) est égal au rapport entre *la covariance xy* et le produit des deux écarts-types de x et y (σx et σy), il est noté r :

r = Covariance (x, y)/σx.σy ou r = Cov xy/σx.σy

La covariance xy est la moyenne arithmétique du produit des écarts à la moyenne des deux variables x et y. Elle mesure la co-variation des deux variables l'une par rapport à l'autre et prend en compte leur variation commune. On peut l'écrire sous la forme:

Cov xy = Σ(xi - xmoy) (yj - ymoy)/n ou

Cov xy = Σ xi yj/n – xmoy*ymoy

La covariance est au plus égale au produit des deux écart-types: Cov xy <= σx.σy.
Son signe indique le sens de la relation (croissante ou décroissante), il est identique au signe de r et de a.

Le coefficient de corrélation linéaire varie de 0 en cas d'une indépendance totale à l'unité (1) en cas d'une liaison fonctionnelle ou totale y = f(x). *Le signe indique le sens de la relation*, elle est croissante ou directe si r > 0, décroissante ou inverse si r < 0:

Exemple: Fécondité et planning familial en Tunisie selon les régions dans les années 1980:

Région	Fécondité %° (F)	Planning Familial P % des femmes utilisant le P.F
Tunis	133.1	39.3
Sfax	137.4	33.1
Sahel	148.1	28.2
Nord-Est	137.3	32.6

L'examen du tableau et du nuage de points montre qu'on peut utiliser le modèle linéaire F = aPF + b. *La corrélation est élevée et négative, c'est-à-dire que la fécondité baisse avec le planning familial mais peut-on conclure qu'il y a une relation de cause à effet entre le planning familial et la niveau de la fécondité en Tunisie.*

	Moyenne	Variance	Ecart-type
Planning % x	27.02857143	84.39346939	9.186591827
Fécondité %° y	153	514.1657143	22.675222418
Modèle : y = ax + b Fécondité = a.Planning + b	Covariance : -201.9228578 Coefficient de corrélation r = -0.96934752439		

Seulement, il existe un seuil en delà duquel la valeur obtenue de r peut être imputée au hasard et n'exprime nullement la présence réelle d'une relation effective entre x et y. Il existe ainsi un seuil de signification séparant la dépendance de l'indépendance, c'est la table de Pearson qui va nous permettre de vérifier la signification de r obtenu.

Ce seuil est symétrique de part et d'autre de zéro (l et l') si bien que le signe n'intervient pas et la Table de Pearson ne donne que les valeurs positives. Le calcul du coefficient de corrélation n'est pas suffisant en soi, il faut tester la signification en utilisant *la Table de Pearson.*

4.2 - La signification de la corrélation linéaire*: la table de Pearson*

La signification du coefficient de corrélation est testée par la table de Pearson qui va nous permettre de dire si la valeur obtenue exprime bien ou la présence d'une relation réelle entre x et y ou elle est due totalement au hasard et n'a aucun sens.

La table de Pearson est une table qui indique la limite supérieure (l) des valeurs aléatoires de r susceptibles d'être imputées au hasard. Pour être significatif, le coefficient de corrélation linéaire calculé d*oit être supérieur ou égal au seuil indiqué par la table de Pearson*: *rcalc* ≥ *rthéo.*

Cette table donne les valeurs théoriques à atteindre ou à dépasser en fonction du nombre d'observations (n) ou du degré de liberté (v) en ligne, le risque d'erreur α ou le seuil de probabilité (1 - α) en colonne.

Le degré de liberté (v) est le nombre de fois qu'on peut choisir un élément dans un système donné. Très souvent, le dernier élément dans un système ne peut pas être choisi, en plus on a une autre contrainte exprimée par la relation liant y à x (en connaissant x, on peut déterminer y), d'où le nombre de degré de libertés v = n - p -1 où n: le nombre d'observations, p: le nombre de variables explicatives. Dans le cas de la corrélation simple (deux variables x, y) on a v = n – 1 - 1 = n - 2.
Le risque d'erreur α est la probabilité de se tromper en tirant une certaine conclusion. Un risque de 0.05 ou 5% correspond à un risque de se tromper 5 fois sur 100 observations. La probabilité correspondante de tirer une conclusion correcte est de (1- α), soit 95%.

Exemple: Si on reprend l'exemple de la fécondité et du planning familial on constate qu'on a obtenu un coefficient de -0.9693. On peut se demander dans quelle mesure cette valeur exprime bien la présence d'une relation entre le planning et la fécondité? Autrement, cette valeur peut-elle être due au hasard et n'a aucune signification ? En regardant la table de Pearson, on constate que à un degré de liberté égal à v = n - 2 = 7 - 2 = 5 et au seuil de signification de 99% (α = 0.01), on trouve la valeur 0.8329. La valeur observée étant supérieure à celle lue dans la table, on peut tirer la conclusion que la relation est hautement significative entre le planning familial et la fécondité. Autrement, la relation entre les deux phénomènes n'est pas due au hasard.

4.3- L'importance de la corrélation*: le coefficient de détermination et la variance expliquée*

La présence d'une corrélation significative est nécessaire mais pas suffisante pour continuer l'analyse. Il faut que la part de la variance expliquée de y par x soit assez élevée et dépasser la moitié (50%). C'est à partir de ce seuil de 50% qu'on peut dire que la relation entre x et y est importante et que x expliquerait y. Cette variance expliquée est exprimée par le carré du coefficient de corrélation r, appelé *coefficient de détermination* : r^2

4.4 - La variance résiduelle et l'erreur-type

Le coefficient de détermination prend en compte la partie de la variance de y liée à x, appelée *variance liée ou expliquée* laissant une partie qui ne peut pas être imputée à x ou variance résiduelle. Plus la corrélation est élevée et plus la variance résiduelle est réduite. *La variance résiduelle* (σ2y.x) est la variance qui n'est pas imputée à x et qui peut être due soit à d'autres variables, soit à des facteurs aléatoires. Elle est égale à la différence entre la variance totale de y et la variance imputée à x: $\sigma^2y.x = \sigma^2y\ (1 - r^2)$.

Variance Totale = Variance Expliquée + Variance Résiduelle

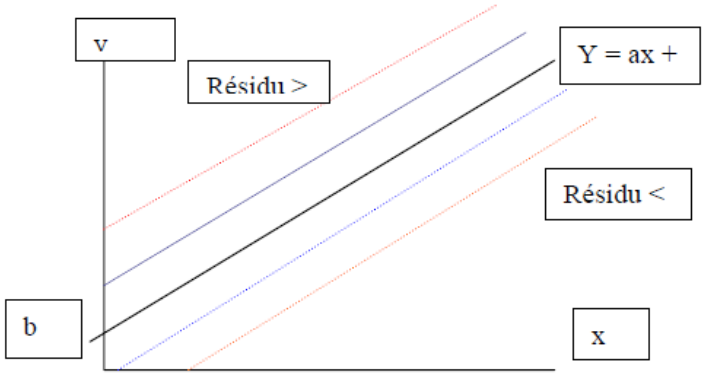
$1 =$	r^2	$+$	$1 - r^2$
$\sigma^2y =$	$\sigma^2y.\ r^2$	$+$	$\sigma^2y\ (1 - r^2)$

Elle mesure la variance des résidus par rapport à une droite d'ajustement qui passe le plus près possible de tous les points du nuage, appelée *droite de régression.* C'est la variance des distances des points à cette droite qui constitue en quelque sorte la moyenne. L'écart-type de ces résidus est appelé *erreur-type*: $e = \sigma y\ (1 - r^2)^{1/2}$

L'erreur-type constitue en quelque sorte l'erreur qu'on commet en remplaçant le nuage de points et le tableau par la droite et l'équation y = ax + b. En divisant les résidus par l'erreur-type on standardise les résidus, leur moyenne étant nulle (la droite) et l'écart-type est égal à l'erreur-type. Les résidus standardisés varient ainsi de -3 à +3.

Très souvent, les résidus se distribuent selon une loi normale autour de la droite de régression, les intervalles de demie amplitude un, deux et trois erreur-types contiennent 68.3%, 95.4% et 99.7% des observations. C'est ainsi que la moyenne des résidus est nulle et lorsqu'on divise les résidus sur l'erreur-type (e), on les standardise et on obtient ainsi des valeurs comprises entre -3 et +3.

Graphiquement, on détermine ces intervalles, en traçant de part et d'autre de la droite de régression des droites parallèles à partir des points situés sur l'axe y et de coordonnées : b +e, b+2e et b+3e d'un côté b-e, b-2e et b-3e. Les observations se trouvent d'un seul coup, classées sur le graphique selon une échelle de 6 classes en fonction de leur distance à la droite mesurée en erreur-type (e) allant de -3e et +3e. Si le graphique est bien tracé, il nous épargne des calculs fastidieux des résidus surtout lorsque le nombre des observations est élevé, il permet une classification directe sur le graphique. Tracer graphiquement les trois intervalles autour de la droite équivaut à la standardisation des résidus sur le tableau.



Droite d'ajustement et résidus

4.5 – La corrélation partielle

La corrélation partielle $R_{yx1,x2}$ est la corrélation entre les variables y et x1 lorsque la troisième variable x2 est réellement ou supposée constante. C'est le cas par exemple de la corrélation entre le niveau de dépense et la motorisation lorsque le taux d'urbanisation est constant. Autrement, dans les espaces où on a le même taux d'urbanisation, quelle est la corrélation entre la dépense et la motorisation.
La corrélation partielle Ryx1,x2 s'écrit comme suit:

$$Ryx1,x2 = (ryx1 - ryx2.rx1x2)/((1 - r^{2}_{yx2})(1- r^{2}_{x1x2}))^{1/2}.$$

Elle sert à mesurer l'effet d'une troisième variable exogène dans l'intensité de la corrélation des deux autres. On peut voir par exemple la relation entre la fécondité (y) et le planning familial (x1) à un niveau d'urbanisation donné (x2) ou de revenu (x3) ce qui revient à éliminer l'effet d'une variable donnée.