Statiqtique Bivariée

a) Test d'indépendance
   1. Variable aléatoire à caractère qualitatif :
      - Comparaison de plusieurs répartitions observées
   2. Variable aléatoire quantitative
      - Régression simple
      - Régression du second degré
      - Test de signification de corrélation
      - Précision de la corrélation

Statistique bivariée

V- Test d'indépendance :

5.1 Variable aléatoire à caractère qualitatif :

5.1.1 Comparaison de plusieurs Echantillons observées :

Dans les tests précédents le χ^2 consistait généralement à trouver si une distribution observée différait d'une distribution théorique connue.

Nous allons utiliser le test χ^2 pour comparer entre elles, des distributions relatives à plusieurs échantillons afin de déterminer si les différences observées sont significatives.

Dans ce cas les données figurent en général dans un tableau de contingence á double entrée. Ce tableau constitue une distribution d'effectifs associés à 2 variables.

Classes / Echantillons	Classe 1	Classe 2 ……..	Classe r	Total
Ech 1	O_{11}	O_{12}	O_{1r}	n_1
Ech 2	O_{21}	O_{22}	O_{2r}	n_2
.		.		
.		.		
.		.		
.		.		
.		.		
Ech l	O_{l1}	O_{l2}	O_{lr}	n_l
Total	n'	n'	n'	n

Le test d'indépendance se fait selon les étapes suivantes:

1- On pose H_0: « les variables qualitatives sont indépendantes. »

2- On calcule à partir du tableau de contingence, et pour chaque case , l'effectif théorique $\mathbf{C_{ij}}$ qui est le produit du total des effectifs observés de la ligne i correspondante (n_i) par le total des effectifs observés de la colonne j correspondante (n_j) devisé par l'effectif total n. Puis on calcule le χ^2 observé :

$$C_{ij} = \frac{n_i \times n_j}{n}$$

$$\chi^2 = \frac{\sum \left(O_{ij} - C_{ij}\right)^2}{C_{ij}}$$

$$ddl = (l-1)(r-1)$$

3- On lit le χ^2 théorique dans la table au seuil de signification choisi α et si:

$\chi^2 \geq \chi^2_\alpha \Rightarrow H_0$ est rejetée : il y a une dépendance entre les variables, et si

$\chi^2 < \chi^2_\alpha \Rightarrow H_0$ est acceptée : il y a une indépendance entre les variables.

Remarque : les effectifs théoriques par classe doivent être sont ≥ 5.

Exemple: Afin de déterminer s'il y a indépendance entre le groupe sanguin et le sexe au seuil α = 5 %, on a examiné 976 individus prélevés au hasard et on a trouvé les résultats suivants:

G.S / Sexe	AB	A	O	B	Total
Homme	25	215	200	60	500
Femme	15	207	194	60	476
Total	40	422	394	120	976

Solution:

H_0 : il existe une indépendance entre le groupe sanguin et le sexe.

On calcule les effectifs théoriques comme suit :

G.S / Sexe	AB	A	O	B	Total
Homme	20,49	216,18	201,84	61,47	500
Femme	19,50	205,8	192,15	58,52	476
Total	40	422	394	120	976

Par exemple: 500*40/976 = 20,49 …..etc

$$\chi^2 = \frac{(25-20{,}49)^2}{20{,}49} + \frac{(215-216{,}9)^2}{216{,}9} + \frac{(200-201{,}84)^2}{201{,}84} + \frac{(60-61{,}47)^2}{61{,}47} + \frac{(15-19{,}51)^2}{19{,}51} + \frac{(207-205{,}81)^2}{205{,}81} + \frac{(194-192{,}15)^2}{192{,}15} + \frac{(60-58{,}52)^2}{58{,}52}$$

$\chi^2 = 2{,}154$

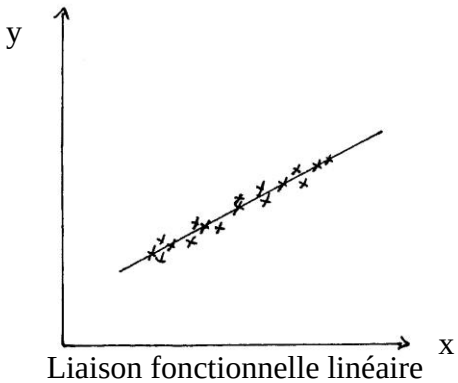
$\text{ddl} = (4-1)(2-1) = 3$

Puisque $\chi^2 < \chi^2_\alpha \Rightarrow H_0$ est acceptée entre les G.S et le sexe, il existe une indépendance.

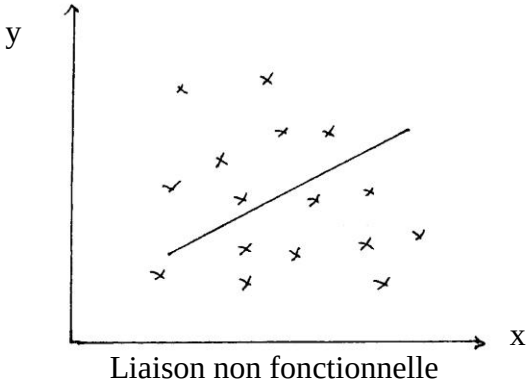
5.1.2 Variables Aléatoires quantitatives :

Dans ce cas l'étude statistique porte simultanément sur 2 ou plusieurs variables à caractère quantitatif, le problème est de déterminer s'il existe une liaison fonctionnelle (**corrélation**) entre les variables pour un même individu. L'objectif est donc d'étudier par les méthodes de régression cette corrélation indiquant une dépendance entre deux (régression simple) ou plusieurs variables (régression multiple).

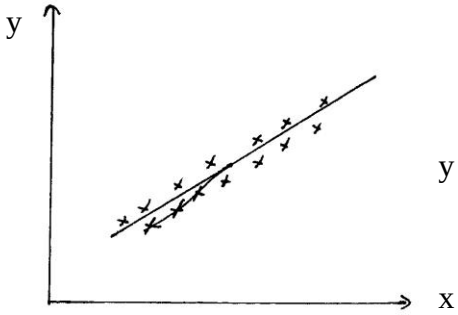
Les graphes suivants indiquent les 3 cas de dépendance qui peuvent se présenter entre 2 variables aléatoires x et y:



Liaison fonctionnelle linéaire
(dépendance idéale)



Liaison non fonctionnelle
(Indépendance)



y = ax + b

Liaison fonctionnelle
expérimentale

C'est cette dernière liaison fonctionnelle qui va faire l'objet d'étude dans ce qui suit:
Déterminer une corrélation entre 2 VA revient à caractériser leur degré de dépendance par un coefficient numérique.

Régression simple

On cherche á trouver une **liaison mathématique** entre les variables Y et X, qui peut être un **modèle déterministe** traitant des variable déterministes « prises » sans perturbation **ε** ou bien un **modèle stochastique** (probabiliste) traitant des variables aléatoires soumises á l'expérimentation affectés á un risque d'erreur donné.

$Y_i = a\ X_i + b$ $\longrightarrow$ modèle déterministe.

$Y_i = a\ X_i + b + \varepsilon_i$ $\longrightarrow$ modèle stochastique (probabiliste).

Avec

Notre objectif est de déterminer le model stochastique qui semble être plus proche á la réalité, $Y = a\ X + b + \varepsilon$ avec l'hypothèse que $\sum \varepsilon_i = 0$.

Déterminer ce modèle revient à déterminer ses constantes a et b par la méthode de régression simple qui se base sur la méthode des moindres carrés.

Remarque: La régression est dite simple parce que la variable Y est expliquée par une seule variable explicative X.

L'estimation des constantes de régression **a** et **b** sera faite sur la base de l'observation, c'est á dire des données expérimentales, avec :

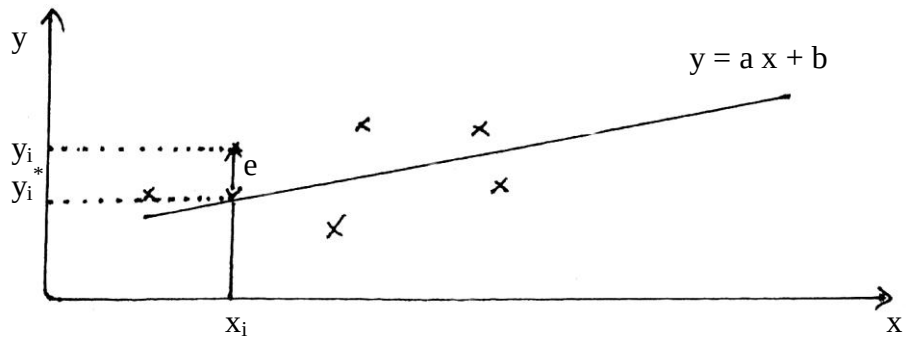
$X_i = \{x_1, x_2, \ldots, x_i\}$,

$Y_i = \{y_1, y_2, \ldots, y_i\}$

$\varepsilon_i = \{\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_i\}$

Cela veut dire que chaque couple de mesure (x_i , y_i) se réalise avec une erreur ou perturbation ε_i, et comme on est en expérimentation on doit supposer que la moyenne des erreurs doit tendre vers zéro, $E(\varepsilon_i) = 0$.

Notre Objectif est donc de chercher une fonction linéaire qui peut résumer le nuage de points (valeurs observées représentées sur le graphe) en une droite rectiligne. Autrement dit, trouver une droite passant le plus proche du nuage telle que la somme des carrés des écarts entre valeurs observées $\mathbf{y_i}$ et valeurs estimées par le modèle $\mathbf{y_i^*}$ soit le minimum possible, c'est ce qu'on appelle **la méthode des moindres carrés**.



$$e = \sum \left(y_i - y_i^*\right) = 0$$

$\sum \left(y_i - y_i^*\right)^2 \rightsquigarrow$ soit le minimum possible (méthode des moindres carrés).

$$\sum \left(y_i - y_i^*\right) = 0$$

$$\sum \left(y_i - ax_i - b\right) = 0$$

$$\sum y_i - \sum ax_i - \sum b = 0$$

$$\sum y_i - a\sum x_i - Nb = 0$$ N : taille de l'échantillon

$$\sum y_i - a\sum x = Nb \Rightarrow b = \frac{\sum y_i}{N} - a\frac{\sum x_i}{N}$$

$$\Rightarrow b = \bar{y} - a\overline{X}$$

$\overline{X}$: moyenne de la variable X

$\overline{Y}$: moyenne de la variable Y

$$\sum \left(y_i - y_i^*\right)^2 = \sum (y_i - axi - b)^2$$

$$= \sum (y_i - axi - \bar{y} + a\bar{x})^2$$

$$= \sum \left[(y_i - \bar{y}) - a(x_i - \bar{x})\right]^2$$

$$= \sum \left[(y_i - \bar{y})^2 - a^2(x_i - \bar{x})\right]^2 - 2a\sum \left[(y_i - \bar{y})(x_i - \bar{x})\right]$$

$\sum \left(y_i - y_i^*\right)^2 \rightsquigarrow$ soit le minimum implique que ,

$$\frac{\partial}{\partial a}\sum \left[(y_i - \bar{y})^2 - a^2(x_i - \bar{x})\right]^2 - 2a\sum \left[(y_i - \bar{y})(x_i - \bar{x})\right] = 0$$

$$a = \frac{\sum (y_i - \bar{y})(x_i - \bar{x})}{\sum (x_i - \bar{x})^2}$$

$$a = \frac{\frac{\sum (y_i - \bar{y})(x_i - \bar{x})}{N}}{\frac{\sum (x_i - \bar{x})^2}{N}} = \frac{Cov(x,y)}{Var(x)}$$

Exemple: soit un échantillon avec la distribution suivante :

x_i	2	5	7	10	12	14
y_i	7	11	14	18	20	23

Après calcul et en appliquant les formules ci-dessus on trouve le résultat suivant:

$(\bar{x} = 8{,}33)$ $\bar{y} = 15{,}50$

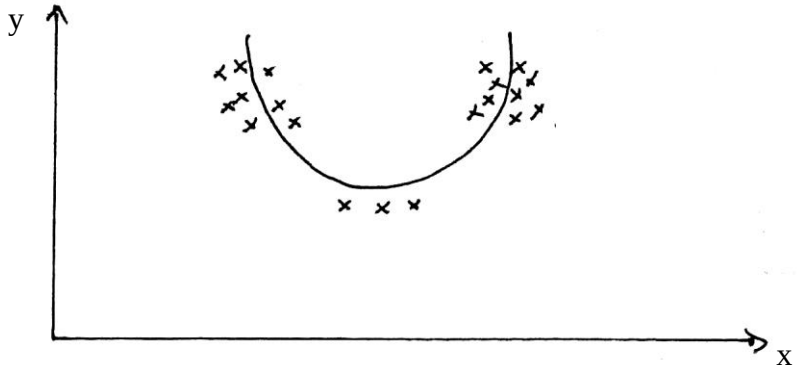
a=1,32 et b=4,50

le modèle de régression linéaire est donc de la forme y=1,32 x + 4,50

Régression de second degré :

Quand l'allure du nuage de point semble être une parabole, on pense automatiquement que le model de régression sera une équation du 2ème degré de la forme:

$y = ax^2 + bx + c$



En appliquant la méthode des moindres carrés, on détermine les constantes a, b et c.

$$C = \frac{\sum (x_i - \bar{x})^2 \sum (x_i - \bar{x})^2 (y_i - \bar{y})}{\left[\sum (x_i - \bar{x})^2\right]^2 - N \sum (x_i - \bar{x})^4}$$

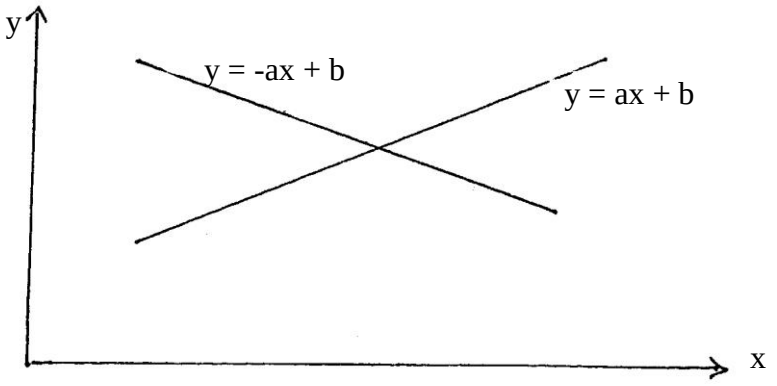
$$b = \frac{\sum (x_i - \bar{x})^2 (y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

$$a = \frac{-N.C}{\sum (x_i - \bar{x})^2}$$

Et de la même façon, on établit le model exponentiel, logarithmique…etc., à cet effet il est conseillé avant de procéder au calcul de la régression de faire une présentation graphique afin de choisir le model de régression convenable (parabolique, logarithmique……).

Coefficient de corrélation

L'intensité de la corrélation entre les variations de x et celles de y est mesurée par **le coefficient de corrélation de Pearson** qui est un terme sans dimension, il est exprimé en % et compris entre -1 et +1. Il vaut +1, ou -1 dans le cas d'une **liaison fonctionnelle**, il vaut 0 dans le cas où il y a une indépendance (**pas de corrélation)** entre X et Y.



Le cœfficient de corrélation prend le signe de la pente du modèle de régression (**a**), il est positif quand a est positif indiquant ainsi une corrélation positive et la droite de régression est donc croissante. Il est négatif quant le a est négatif indiquant une corrélation négative et la droite de corrélation est donc décroissante.

Le cœfficient de corrélation r est exprimé par la formule suivante :

$$r = \frac{Cor(x,y)}{\sqrt{Var(x).Var(y)}} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2.(y_i - \bar{y})^2}} = \frac{Cor(x,y)}{\sigma_x.\sigma_y}$$

Remarque : Dans la pratique on utilise parfois le r^2, qui est évidemment positif [0,1], et est appelé **coefficient de détermination.**

Exemple: Trouver pour l'exemple précédent :

$$r = \frac{Cor(x,y)}{\sqrt{Var(x).Var(y)}} \qquad r = 0{,}99$$

Cela veut dire que 99% de la variation de y est expliquée par la variation de X ou bien, on peut dire que 99% de la variance de Y est expliquée par le model de régression suivant Y= 1,32 x + 4,5.

Les 2 variables x, y sont très bien corrélées mais **attention** cette corrélation doit subir le **test de signification** pour qu'elle **soit acceptée**. Une bonne corrélation n'implique pas toujours un lien de causalité.

Test de signification de corrélation :

Ce test consiste á vérifier cette hypothèse $H_0 : r = 0$

Il s'agit donc de tester si le coefficient de corrélation **r** est significativement différent de zéro « 0 ».

Ce test se fait généralement par deux grandes méthodes, une méthode directe par le biais du T-test de Student ou l'utilisation directe de la table du Cœfficient de corrélation et une méthode indirecte utilisant la transformation Argument tangente hyperbolique de Fischer Z=Argth(r)

- **Méthode directe:**

  t-Test:

  L'hypothèse nulle $H_0 : r = 0$, sera testée par le calcul de l'expression:

$$t = \frac{r}{\sqrt{1-r^2}}\sqrt{n-2}$$

Puis comparer ce t ainsi avec le t_α de Student lu á partir de la table de Student au seuil α et avec un ddl = n- 2. H_0 sera rejetée indiquant une corrélation significative quant $t \geq t_\alpha$

Utilisation directe de la table du Coefficient de corrélation:

A l'aide de la formule $t = \frac{r}{\sqrt{1-r^2}}\sqrt{n-2}$, on a calculé la table de signification du coefficient de corrélation à différents seuil de signification α et ddl = n – 2.

Le principe est simple, On lit directement sur la table le r_α théorique en fonction du ddl correspondant et α choisie et on le compare au r observé, si $r \geq r_\alpha$ on rejette H_0.

Exemple1: A partir d'un échantillon de 37 individus on a calculé le r = 0,28

Est-ce que r est significativement différent de 0 au seuil de 5% ?

Solution :

$H_0 : r = 0$

ddl = 37 – 2 = 35

A partir de la table, pour $\alpha = 5\%$ et ddl = 35 on a lu $r_\alpha = 0{,}3246$

On remarque que $r < r_\alpha \Rightarrow (0{,}28 < 0{,}32) \Rightarrow H_0$ est acceptée $\Rightarrow$ r n'est pas significatif pour conclure une dépendance entre les 2 variables.

Les 2 variables sont complètement indépendantes et ne peuvent être corrélées même à α = 1% puisque $r_\alpha = 0{,}41$.

Exemple2: Pour un grand échantillon n=93 on trouvé r=0,24

La table nous livre pour $\alpha = 5\%$ et ddl= 93-2= 91 un r_α=0,22

On a $r \geq r_\alpha$, H_0 est donc rejetée indiquant une corrélation significative, le r cette fois-ci diffère significativement de zéro au seuil 5%.

Remarque :

- Pour avoir une bonne corrélation significative il faut que l'échantillon soit normal, généralement on a une loi normale quand **n est suffisamment grand.**
- On peut utiliser la table du cœfficient de corrélation pour déterminer la taille minimale **n** qui doit avoir un échantillon pour atteindre une corrélation significative au seuil donné α.