Chapitre 4 : Régression et Corrélation

1. Régression :

- Considérons deux variables aléatoires quantitatives X et Y.
- X est contrôlée (variable explicative) et Y est aléatoire (variable expliquée).
- Supposons un modèle linéaire (Y s'exprime linéairement en fonction de X) de la forme:

$$Y = \alpha X + \beta$$

Avec α: la pente et β: ordonnée à l'origine.

- Ces deux paramètres peuvent être estimés par a et b (comme vu au chapitre 2), et on aura la droite de régression au sens des moindres carrés à partir d'un échantillon de taille n :

$$Y = aX + b$$

- Pour mesurer la qualité de l'ajustement linéaire, on calcule le coefficient de détermination $R = r^2$ (où r est le coefficient de corrélation).
- R mesure la part de la variation totale de Y expliquée par le modèle de régression sur X.
- Si $R = 0$, le modèle n'explique rien, les variables X et Y ne sont pas corrélées linéairement.
- Si $R = 1$, les points sont alignés sur la droite, la relation linéaire explique toute la variation.
- Une valeur de R proche de 1 est nécessaire pour avoir un ajustement raisonnable mais en aucun cas suffisant.

1. 1. Test global de significativité de la régression :

Les hypothèses :

On teste la significativité globale du modèle, dans le cas de la régression linéaire simple, on teste :

H_0: $\alpha = 0$, contre :

H_1: $\alpha \neq 0$.

Statistique du test :

$$F_{calc} = (n-2)\frac{R^2}{1-R^2}$$

Valeurs critiques :

On utilise la table de Fisher à $(1, n-2)$ degré de liberté.

1. 2. Test sur les paramètres :

Est-ce-que le coefficient α est non nul, autrement dit la variable X a-t-elle réellement une influence sur Y ?

Est-ce-que le coefficient β est non nul, autrement dit faut-il une constante dans le modèle ?

a) Test sur la pente :

Pour pouvoir appliquer ce test, il faut rajouter des hypothèses (normalité et égalité des variances) sur les distributions de Y pour chaque valeur de X (distributions conditionnelles de Y).

Les hypothèses : (cas particulier valeur théorique égale à 0)

H_0: $\alpha = 0$, contre :

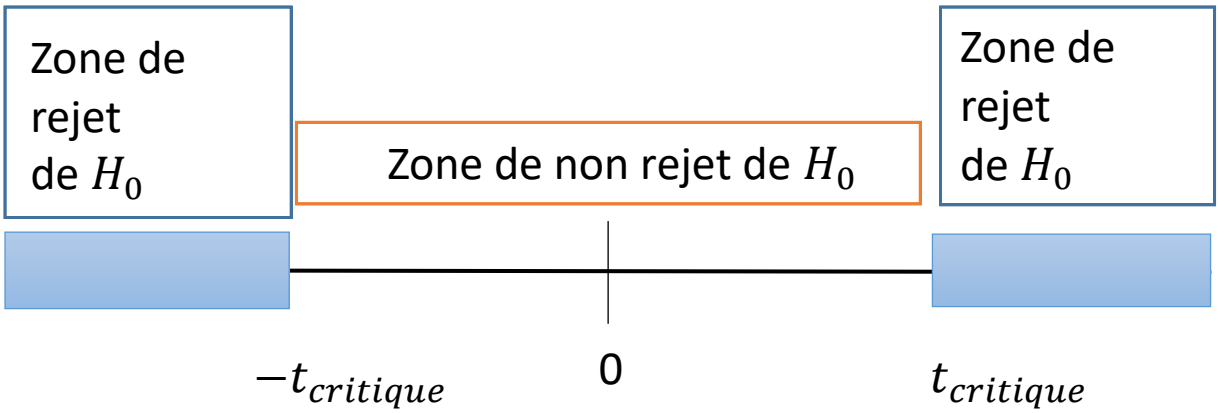
H_1: $\alpha \neq 0$.

Statistique du test :

$t_{calc} = \frac{a-0}{S_a}$ où $S_a^2 = \frac{\left(\frac{S^2_{ech\,Y}}{S^2_{ech\,X}}\right) - a^2}{n-2}$

Valeurs critiques :

On utilise la table de Student à $(n-2)$ degré de liberté.



b) Test sur l'ordonnée à l'origine :

- Il peut être intéressant de savoir si la droite estimée passe par l'origine, on teste donc l'ordonnée à l'origine par rapport à la valeur 0.
- Les hypothèses sous-jacentes à l'application de ce test sont identiques à celles du test précédent (normalité et égalité des variances des distributions conditionnelles de Y).

Les hypothèses :

H_0: $\beta = 0$, contre :

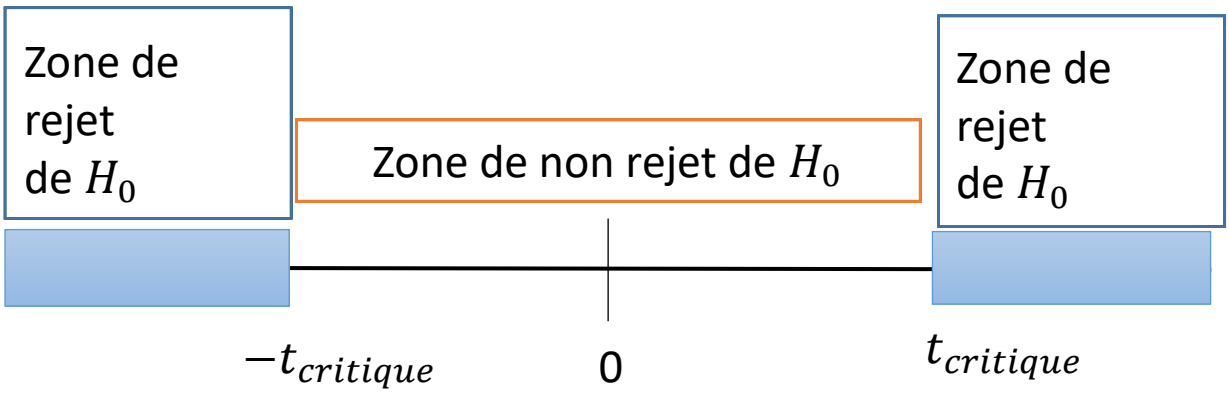
H_1: $\beta \neq 0$.

Statistique du test :

$$t_{calc} = \frac{b-0}{S_b} \text{ où } S_b^2 = \frac{S_a^2 \sum_{i=1}^{n} x_i^2}{n}$$

Valeurs critiques :

On utilise la table de Student à $(n-2)$ degré de liberté.



Exemple 1 :

Soit la série statistique double $(X,\ Y)$ où X est un temps de mesure en heures et Y un taux de précipitation au cours d'une réaction de synthèse chimique. Les données sont dans le tableau suivant :

X	0	0,5	1	1,5	2	2,5	3	3,5	4	4,5
Y	3	4,2	3,8	6,5	7,2	7,9	9,1	10	10,8	12,2

Au risque de 5%, effectuer un test sur la pente et sur l'ordonnée à l'origine.

Solution :

1. Test sur la pente :

H_0: $\alpha = 0$, contre :

H_1: $\alpha \neq 0$.

Calculs :

$$\bar{x} = 2{,}25 \ ; \ \bar{y} = 7{,}47 \ ; \quad S^2_{ech\,X} = 2{,}0625 \ ;$$

$$S^2_{ech\,Y} = 8{,}7861 \ ; \quad cov(X,Y) = 4{,}2125 \ ;$$

$$a = 2{,}0424 \quad ; \quad b = 2{,}8746$$

Donc l'équation de la droite de régression est

$$y = 2{,}0424\,x + 2{,}8746$$

$$S_a^2 = \frac{\left(\frac{S^2_{ech\,Y}}{S^2_{ech\,X}}\right) - a^2}{n-2} = \frac{\left(\frac{8{,}7861}{2{,}0625}\right) - (2{,}0424)^2}{8} = 0{,}01107$$

Donc $S_a = 0{,}1051$ et $t_{calc} = \frac{a-0}{S_a} = \frac{2{,}0424-0}{0{,}1051} = 19{,}43.$

$$t_{critique} = t(0{,}05;8) = 2{,}306$$

Décision :

$t_{calc} = 19{,}43 > t_{critique} = 2{,}306$, d'où le rejet de H_0.

Au risque de 5%, le temps de mesure a une influence significative sur le taux de précipitation au cours d'une réaction de synthèse chimique.

2. Test sur l'ordonnée à l'origine :

H_0: $\beta = 0$, contre :

H_1: $\beta \neq 0$.

Calculs :

$\sum_{i=1}^{10} x_i^2 = 71{,}25$;

$S_b^2 = \frac{S_a^2 \sum_{i=1}^{n} x_i^2}{n} = \frac{0{,}01107 * 71{,}25}{10} = 0{,}0789$;

$S_b = 0{,}2808$; donc

$t_{calc} = \frac{b-0}{S_b} = \frac{2{,}8746-0}{0{,}2808} = 10{,}24$ et

$t_{critique} = t(0{,}05; 8) = 2{,}306$

Décision :

$t_{calc} = 10{,}24 > t_{critique} = 2{,}306$, d'où le rejet de H_0.

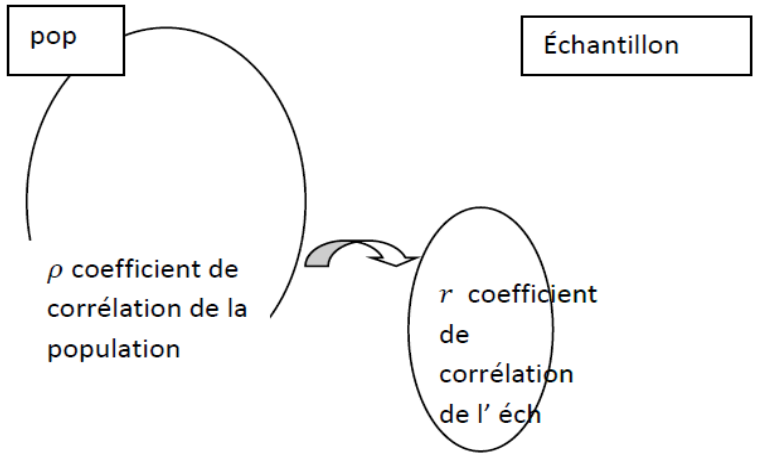
On peut dire que l'ordonnée à l'origine est très significativement différente de zéro.

2.Corrélation :

- Considérons deux variables aléatoires quantitatives X et Y.
- X et Y sont toutes aléatoires.
- Le coefficient de corrélation linéaire de l'échantillon r mesure l'intensité de l'association linéaire entre les valeurs de X et Y liées et issues d'un échantillon (formule de calcul déjà vue au chapitre 2).
- r explique la dispersion, conclure qu'il y a une corrélation linéaire significative entre X et Y implique qu'il est possible de trouver une équation linéaire qui exprime Y en fonction de X et que cette équation peut être utilisée pour prédire les valeurs de Y pour des valeurs de X données.
- Si $r = 0$, pas de liaison linéaire entre X et Y (il peut y avoir une autre forme de liaison: exponentielle…).
- Si $r < 0$, X et Y varient en sens opposés (si l'un croit, l'autre décroit).
- Si $r > 0$, X et Y varient dans le même sens (si l'un croit, l'autre croit).

2.1. Test sur le coefficient de corrélation ρ :

- ρ coefficient de corrélation de la population qui est estimé par r coefficient de corrélation de l'échantillon.



- Sous les hypothèses de normalité des distributions de X et Y, on peut tester l'existence d'une relation supposée linéaire entre X et Y .

Les hypothèses :

H_0:ρ $= 0$, contre:

H_1:ρ $\neq 0$, ρ > 0 ou ρ < 0.

Statistique du test :

$t_{calc} = \frac{r-0}{S_r}$ où $S_r^2 = \frac{1-r^2}{n-2}$.

Valeurs critiques :

On utilise la table de Student à $(n-2)$ degré de liberté.

Exemple 2:

Dans une maternité, on prélève un échantillon de 10 nouveau-nés et on mesure leurs poids et leurs tailles.

Poids xi (kg)	2,100	2,500	2,650	2,800	3,&00	3,200	3,350	3,650	3,750	4,250
Taille yj (cm)	40	45	43	50	48	51	49	51	53	55

Tester l'affirmation qu'il y a une corrélation linéaire significative entre la taille et le poids des nouveaux - nés.

Solution :

On suppose que les distributions du poids et de la taille sont normales.

Affirmer qu'il y a une corrélation linéaire significative est équivalent à dire que $\rho \neq 0$.

Nous testons les hypothèses suivantes :

H_0:ρ $= 0$, contre:

H_1:ρ $\neq 0$.

Calculs :

$$\bar{x} = 3{,}135 \ ; \ \bar{y} = 48{,}5 \ ; \quad S^2_{ech\,X} = 0{,}376 \ ;$$

$$S^2_{ech\,Y} = 19{,}25 \ ; \quad cov(X,Y) = 2,\ 4775 \ ;$$

$$r = 0{,}9208$$

$S_r^2 = \frac{1-r^2}{n-2} = \frac{1-(0{,}9208)^2}{10-2} = 0{,}0192$, donc $S_r = 0{,}1386$.

$$t_{calc} = \frac{r-0}{S_r} = \frac{0{,}9208-0}{0{,}1386} = 6{,}6436.$$

$$t_{critique} = t(0{,}05; 8) = 2{,}306.$$

Décision :

$t_{calc} = 6{,}6436 > t_{critique} = 2{,}306$, d'où le rejet de H_0.

Décision :

Il y a donc suffisamment de preuves pour confirmer l'hypothèse d'une corrélation linéaire entre la taille des bébés et leur poids.