

CORRELATION

La corrélation permet de retrouver et de quantifier un LIEN de dépendance entre 2 séries ou 2 facteurs.

Il faut que les séries suivent chacune une loi normale

1/Ranger les données selon un tableau à deux colonnes sous forme de couples

1/Exemple pratique :

Existe t-il une corrélation, entre le taux de cholestérol et le poids, chez 7 patients coronariens ?

| Sujets | Cholestérol X | Poids Y |
|--------|---------------|---------|
| 1 | 3 | 70 |
| 2 | 1.9 | 55 |
| 3 | 4 | 75 |
| 4 | 5 | 80 |
| 5 | 1.5 | 60 |
| 6 | 6 | 90 |
| 7 | 2 | 67 |

2/Calculer

ΣX : somme des X

ΣY : somme des Y

$\Sigma(XY)$: somme du produit XY

ΣX^2 : somme des X au carré

ΣY^2 : somme des Y au carré

2/Dans notre exemple :

| Sujets | X | X ² | Y | Y ² | XY |
|--------------|-------------------|----------------------|------------------|----------------------|-----------------------|
| 1 | 3 | 9 | 70 | 4900 | 210 |
| 2 | 1.9 | 3.61 | 55 | 3025 | 104.5 |
| 3 | 4 | 16 | 75 | 5625 | 300 |
| 4 | 5 | 25 | 80 | 6400 | 400 |
| 5 | 1.5 | 2.25 | 60 | 3600 | 90 |
| 6 | 6 | 36 | 90 | 8100 | 540 |
| 7 | 2 | 4 | 67 | 4489 | 134 |
| Somme | $\Sigma X = 23.4$ | $\Sigma X^2 = 95.86$ | $\Sigma Y = 497$ | $\Sigma Y^2 = 36139$ | $\Sigma(XY) = 1778.5$ |

3/Calculer le coefficient de corrélation r

$$r = \frac{N \times \sum(XY) - (\sum X)(\sum Y)}{\sqrt{N \times \sum X^2 - (\sum X)^2} \times \sqrt{N \times \sum Y^2 - (\sum Y)^2}}$$

N étant le nombre de couples

3/Dans notre exemple :

$$r = \frac{(7 \times 1778.5) - (23.4 \times 497)}{\sqrt{7 \times 95.86 - 23.4^2} \times \sqrt{7 \times 36139 - 497^2}}$$

r = 0.42

4/Pour qu'il y ait une corrélation entre deux distributions, il faut que la valeur absolue de r soit comprise entre 0 et 1.

4/Dans notre exemple

r = 0.42 il existe donc une corrélation 'moyenne' entre le poids et le taux de cholestérol chez nos 7 coronariens

5/Pour confirmer l'existence d'une corrélation, il faut appliquer le test t de conformité

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

Ce t est à comparer à la table de t (1- α /2) avec un degré de liberté $v = N-2$

Si t calculé est supérieur au t de la table donc r diffère de zéro, il existe donc une corrélation

Si t calculé est inférieur au t de la table donc r est égal à zéro, il n'existe pas de corrélation

5/Dans notre exemple :

$$t = \frac{0.42 \times \sqrt{7-2}}{\sqrt{1-0.42^2}}$$

$$t = 1.03$$

Degré de liberté $v = 7-2=5$

Sur la table $t = 1.28$

t calculé 1.03 est inférieur au t de la table donc il n'existe pas de corrélation entre le poids des 7 coronariens et le taux de cholestérol

Cet exemple résume bien la contradiction, qui peut exister entre un r calculé qui montre une corrélation moyenne et un test de conformité qui affirme le contraire. C'est finalement le test t de conformité qui a le plus de valeur statistique.

REGRESSION

La régression permet, non seulement de vérifier le lien entre deux distributions, mais aussi de le représenter graphiquement. Cette opération permet de poser une loi mathématique expliquant la relation d'une distribution par rapport à l'autre.

1/On représente la régression par une droite de la forme $y=ax + b$

$$\text{où } a = \frac{\sum XY - \frac{\sum X \sum Y}{N}}{\sum X^2 - \frac{(\sum X)^2}{N}}$$

et $b = \text{moyenne des } Y - a \times \text{moyenne des } X$

1/Reprenons l'exemple précédent :

| Sujets | X Cholestérol | X ² | Y Poids | Y ² | XY |
|--------------|-------------------|----------------------|------------------|----------------------|-----------------------|
| 1 | 3 | 9 | 70 | 4900 | 210 |
| 2 | 1.9 | 3.61 | 55 | 3025 | 104.5 |
| 3 | 4 | 16 | 75 | 5625 | 300 |
| 4 | 5 | 25 | 80 | 6400 | 400 |
| 5 | 1.5 | 2.25 | 60 | 3600 | 90 |
| 6 | 6 | 36 | 90 | 8100 | 540 |
| 7 | 2 | 4 | 67 | 4489 | 134 |
| Somme | $\Sigma X = 23.4$ | $\Sigma X^2 = 95.86$ | $\Sigma Y = 497$ | $\Sigma Y^2 = 36139$ | $\Sigma(XY) = 1778.5$ |

$$a = \frac{1778.5 - \frac{23.4 \times 497}{7}}{95.86 - \frac{23.4^2}{7}}$$

$a = 6.64$ correspond à la pente de la droite de régression

$b = 71 - (6.64 \times 3.34) = 48.83$ b correspond à l'ordonnée à l'origine

Donc la droite de régression s'écrit

$$Y = 6.64 X + 48.83$$

Dans le cas de notre exemple :

Poids = 6.64 Taux de cholestérol + 48.83

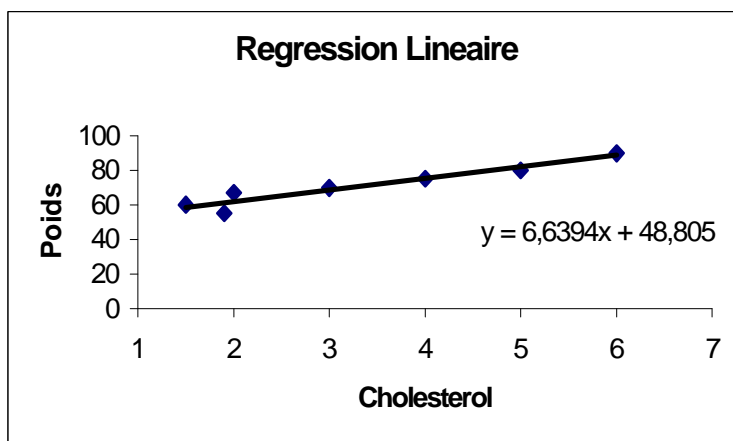


TABLE du t de STUDENT

| v Degré de liberté | P = 95 % | P = 99 % |
|--------------------|----------|----------|
| 1 | 12.706 | 63.657 |
| 2 | 4.303 | 9.925 |
| 3 | 3.182 | 5.841 |
| 4 | 2.776 | 4.604 |
| 5 | 2.571 | 4.032 |
| 6 | 2.447 | 3.707 |
| 7 | 2.365 | 3.499 |
| 8 | 2.306 | 3.355 |
| 9 | 2.262 | 3.250 |
| 10 | 2.228 | 3.169 |
| 11 | 2.201 | 3.106 |
| 12 | 2.179 | 3.055 |
| 13 | 2.160 | 3.012 |
| 14 | 2.145 | 2.977 |
| 15 | 2.131 | 2.947 |
| 16 | 2.120 | 2.921 |
| 17 | 2.110 | 2.898 |
| 18 | 2.101 | 2.878 |
| 19 | 2.093 | 2.861 |
| 20 | 2.086 | 2.845 |
| 21 | 2.080 | 2.831 |
| 22 | 2.074 | 2.819 |
| 23 | 2.069 | 2.807 |
| 24 | 2.064 | 2.797 |
| 25 | 2.060 | 2.787 |
| 26 | 2.056 | 2.779 |
| 27 | 2.052 | 2.771 |
| 28 | 2.048 | 2.763 |
| 29 | 2.045 | 2.756 |
| 30 | 2.042 | 2.750 |
| 40 | 2.021 | 2.704 |
| 60 | 2.000 | 2.660 |
| 120 | 1.980 | 2.617 |
| ∞ | 1.960 | 2.576 |