Statistiques Bivariées

Croisement Quanti - Quanti

. Statistiques Bivariées ≑ étudie de deux caractères (variables) simultanément chez les individus.

But: Etudier s'il y a un lien entre ces deux variables.

Exp: On prend deux variables:

X = taux d'alcoolisme

Y = taux de suicides

et on étudie s'il y a un lien entre ces 2 variables.

. **Type de croisement** : ces deux variables peuvent être :

1. quantitatif × quantitatif
2. qualitatif × quantitatif
3. qualitatif × qualitatif

Pour chaque type de croisement, on a une étude différente. On commence par le Quanti × Quanti.

1_ Nuage des points :

. X et Y deux variables quantitatives sur un échantillon de taille (n), les objectifs :

① Déterminer s'il y a un lien (ou une corrélation) entre les deux variables.

② Construire un modèle qui permet d'expliquer Y par X (ou l'inverse) s'il y a un lien.

2_ Droite de Regression :

* Soit $\{x_i\}_{i=1,\dots,n}$ les observations de X

et $\{y_i\}_{i=1,\dots,m}$ les observations de Y

objectif : trouver une fonction f tel que :
$y_i = f(x_i) + \varepsilon_i$ / ε = erreur.

On se restreint aux fcts affines : $f(x) = ax + b$

- On cherche les coefficients a et b qui minimisent l'erreur quadratique moyenne.

$$EQ(a,b) = \frac{1}{n} \sum_i \varepsilon_i^2$$
$$= \frac{1}{n} \sum_i (y_i - (ax_i + b)^2)$$

★ Droite de Regression de y en X est :

$$y = \hat{a}x + \hat{b}$$

tel que : $\hat{a} = \dfrac{C_{xy}}{S_x^2}$ $\hat{b} = \bar{y} - \hat{a}\bar{x}$

- $C_{xy} = \frac{1}{n} \sum_i (x_i - \bar{x})(y_i - \bar{y})$

c'est la covariance entre X et Y.

- S_x^2 = variance de X.

Ce droite traduit les variations de Y qui peuvent être expliquées par X.

★ Droite de Regression de X en Y.

$$x = \hat{a}y + \hat{b}$$

tel que : $\hat{a} = \dfrac{C_{xy}}{S_y^2}$ $\hat{b} = \bar{x} - \hat{a}\bar{y}$

Exp : lien $\overset{X}{\text{taille}}$ et $\overset{Y}{\text{poids}}$ chez les enfants de 6 ans.

- Equation de Y en x, $y = \hat{a}x + \hat{b}$

$$\begin{cases} \hat{a} = \dfrac{C_{xy}}{S_x^2} = \dfrac{16,27}{38,62} = 0,42 \\ \hat{b} = \bar{y} - \hat{a}\bar{x} = (20,30) - (0,42)(113,20) = -27,38 \end{cases}$$

donc $\boxed{y = 0,42x - 27,38}$

Exp : Si taille d'un enfant est $x = 115$, estimer le poids ?!!

$y = 0,42x - 27,38 = 0,42(115) - 27,38$
$= 20,92$ Kg.

Equation de La droite X en Y :

$$x = \hat{a}y + \hat{b}$$

$$\hat{a} = \frac{Cxy}{S_y^2} = \frac{16,27}{8,46} = 1,92$$

$$\hat{b} = \bar{x} - \hat{a}\bar{y} = 113,2 - (1,92)(20,30)$$
$$= +74,15$$

donc $\boxed{x = 1,92\,y + 74,15}$

taille   poids

Exp : Si le poids d'un enfant est 23 Kg. estimer son taille :

$$x = 1,92y + 74,15$$
$$= 1,92(23) + 74,15 = 118,31 \text{ cm}$$

- Covariance : C'est un indicateur numérique du lien entre X et Y : plus il est loin de 0, plus les variables sont liées. [Cxy = 0, pas de lien entre les 2 variables alors elles sont indépendantes].

⚠ La covariance n'est pas normé, donc il prendre n'importe quel valeur sur ℝ, pour cela pour étudier le lien entre X et Y, on définit le coefficient de Correlation.

. Coefficient de Corrélation (coefficient de Pearson) :

pour étudier le lien entre deux variables quantitative X et Y ; prend ces valeurs entre $[-1,1]$

$$\boxed{r_{xy} = \frac{C_{xy}}{s_x \cdot s_y}}$$

On a alors :

$$\hat{a} = \frac{C_{xy}}{S_x^2} \Rightarrow \hat{a} = \frac{r_{xy} \cdot S_y}{S_x}$$

Interpretation : ① $|r|$ proche de 1, alors X et Y sont très liés entre eux par une droite affine

② $r < 0$: X et Y varie en sens inverse. Si X augmente donc Y diminue.

Exp : $\begin{cases} X = \text{dépense} \\ Y = \text{Epargne} \end{cases}$

③ $r > 0$: X et Y varie de m̂ sens.

Exp : $\begin{cases} X = \text{taux nicotiné dans sang} \\ Y = \text{risque avoir maladie poumon} \end{cases}$

Si X augmente alors Y aussi

④ $|r| = 0$; On ne peut rien dire sur un lien eventuel entre X et Y.

Exple : pour poids et âge chez les enfants de 6 ans ;

$$r_{xy} = \frac{C_{xy}}{s_x \cdot s_y} = \frac{16,27}{\sqrt{38,62} \times \sqrt{8,46}} \simeq 0.9$$

$\begin{cases} \bullet\ r_{xy} \approx 1 \rightarrow \text{l'éq de droite est pleinement justifié.} \\ \bullet\ r_{xy} > 0 \rightarrow \text{plus la variable taille est grande plus le poids est important et vice vesa.} \end{cases}$

⑤

. **Prévisions :** on appele (prévisions). les valeurs donnés par la droite de regression.

Pour chaque x_i on peut calculer $\hat{y}_i$

$\hat{y}_i$ = c'est la valeur approché (estimé) de la vraie valeur y_i.

$$\boxed{\bar{\hat{y}} = \bar{y}}$$

mais

pas m̂ variance.

⇓

$$S^2_{\hat{y}} = r^2_{xy} \cdot S^2_y$$

Les $\hat{y}_i$ estimé ont m̂ moyenne que les vrai valeur y.

. **Résidus :** c'est l'ecart entre y_i et $\hat{y}_i$

$$\boxed{e_i = y_i - \hat{y}_i}$$

tel que : $\boxed{\hat{y}_i = \hat{a} x_i + \hat{b}}$

$$\boxed{\text{Erreur globale : } EQ(\hat{a}, \hat{b}) = \frac{1}{n} \sum_i e_i^2 = S^2_y \left(1 - r^2_{xy}\right)}$$

⚠ Variance $Y \neq$ Variance $\hat{Y}$.

$$\boxed{S^2_y = S^2_{\hat{y}} + S^2_e}$$

variance totale — variance expliqué — variance des résidus.

- Coefficient de détermination :

$$\boxed{R^2 = \frac{S^2_{\hat{y}}}{S^2_y} = r^2_{xy}}$$ (entre 0 et 1).

Donne proportion des observations expliqué par la droite de regression (plus il est élevée plus c mieux)

Exp : On donne :

$$\begin{cases} S^2_{\hat{y}} = 6{,}17 \\ S^2_e = 1{,}44 \\ S^2_y = 7{,}61 \end{cases}$$

donc $R^2 = \dfrac{6{,}17}{7{,}61} = 0{,}81$

$\Rightarrow$ 81% de la variation des poids observés est expliquée par la droite de régression.

$$\underbrace{S^2_y}_{\text{variance totale}} = \underbrace{S^2_{\hat{y}}}_{\text{variance expliquée}} + \underbrace{S^2_e}_{\text{variance résidu}}.$$

- $S^2_{\hat{y}} = r^2_{xy} \cdot S^2_y$
- $S^2_e = S^2_y (1 - r^2_{xy})$
- $S^2_y = \frac{1}{n} \sum_{k=1}^{n} (y_k - \bar{y})^2$

/ r_{xy} = coeff de corrélat°.

$r_{xy} = \frac{C_{xy}}{s_x \cdot s_y}$.

Outeliers et (valeur aberrantes) comment faire ??

+ Si on a une i (plusieurs) valeur aberrantes :

① On retire la i'ème observation

② On refait un nouveau modèle sans le ième

③ On calcule $\hat{y}_{(-i)}$ (la prevision) avec le nouveau modèle.
→ pour modèle sans (i)

④ On calcule le residu :

$$e_{(-i)} = y_i - \hat{y}_{(-i)}$$

pour le modèle sans (i).

⚠ Un residu a valeur importante signale une observation ab influente.

On : $\boxed{e_{(i)} = \frac{e_i}{1 - h_{ii}}}$ ou $\boxed{h_{ii} = \frac{1}{n} + \frac{1}{n-1} \frac{(x_i - \bar{x})^2}{S_x^2}}$

calcule avec modèle sans (i)

calcule avec modèle initial avec ièmeobs.

h_{ii} = Levier, prend ses valeurs entre : $\frac{1}{n} \leq h_{ii} \leq 1$

Si h_{ii} proche de 1 alors on a une observat° influente

- Les résidus standardisés doivent être toujours entre [± 2] [I.D.C à voir en ING 2].
[Ça c'est l'hypothèse de normalité des résidus]

$$\boxed{\delta_i = \frac{e_i}{\hat{\sigma} \cdot \sqrt{1 - h_{ii}}}}$$

Résidu standardisé