MOOC Statistique pour ingénieur
Thème 0 : statistique descriptive
Vidéo 2 : Statistiques à deux variables

F. Delacroix M. Lecomte

Institut Mines-Télécom
École Nationale Supérieure des Mines de Douai

Sommaire

Un exemple

- Test en compression d'éprouvettes de béton
- X=teneur en ciment (kg/m^3)
- Y=résistance à la compression (MPa)

Collecte des données

$n = 90$ mesures

$X \backslash Y$	60	80	100
300	15	4	1
350	10	20	10
400	5	10	15

$X \backslash Y$	y_1	$\dots$	y_j	$\dots$	y_c	Total
x_1			n_{1j}			
$\vdots$			$\vdots$			
x_i	n_{i1}	$\dots$	n_{ij}	$\dots$	n_{ic}	$n_{i.}$
$\vdots$			$\vdots$			
x_r			n_{rj}			
Total			$n_{.j}$			n

$n_{ij} =$ nombre d'observations avec $X = x_i$ et $Y = y_j$

Collecte des données

$n = 90$ mesures

$X \backslash Y$	60	80	100
300	15	4	1
350	10	20	10
400	5	10	15

$X \backslash Y$	y_1	$\dots$	y_j	$\dots$	y_c	Total
x_1			n_{1j}			
$\vdots$			$\vdots$			
x_i	n_{i1}	$\dots$	n_{ij}	$\dots$	n_{ic}	$n_{i.}$
$\vdots$			$\vdots$			
x_r			n_{rj}			
Total			$n_{.j}$			n

$n_{i.}$ = effectif marginal de la $i^{\text{ème}}$ ligne $= \sum_{j=1}^{c} n_{ij}$

Collecte des données

$n = 90$ mesures

$X \backslash Y$	60	80	100
300	15	4	1
350	10	20	10
400	5	10	15

$X \backslash Y$	y_1	$\dots$	y_j	$\dots$	y_c	Total
x_1			n_{1j}			
$\vdots$			$\vdots$			
x_i	n_{i1}	$\dots$	n_{ij}	$\dots$	n_{ic}	$n_{i.}$
$\vdots$			$\vdots$			
x_r			n_{rj}			
Total			$n_{.j}$			n

$$n_{.j} = \text{effectif marginal de la } j^{\text{ème}} \text{ colonne} = \sum_{i=1}^{r} n_{ij}$$

Collecte des données

$n = 90$ mesures

$X \backslash Y$	60	80	100
300	15	4	1
350	10	20	10
400	5	10	15

$X \backslash Y$	y_1	$\dots$	y_j	$\dots$	y_c	Total
x_1			n_{1j}			
$\vdots$			$\vdots$			
x_i	n_{i1}	$\dots$	n_{ij}	$\dots$	n_{ic}	$n_{i.}$
$\vdots$			$\vdots$			
x_r			n_{rj}			
Total			$n_{.j}$			n

$$n = \sum_{j=1}^{c} \sum_{i=1}^{r} n_{ij} = \sum_{i=1}^{r} n_{i.} = \sum_{j=1}^{c} n_{.j}$$

Distribution conjointe

Fréquence de la cellule C_{ij} : $f_{ij} = \frac{n_{ij}}{n}$

$X \backslash Y$	60	80	100	Total
300	15	4	1	$n_{1.} = 20$
350	10	20	10	$n_{2.} = 40$
400	5	10	15	$n_{3.} = 30$
Total	$n_{.1} = 30$	$n_{.2} = 34$	$n_{.3} = 26$	90

$X \backslash Y$	60	80	100	$f_{i.}$
300	16,7%	4,4%	1,1%	22,2%
350	11,1%	22,2%	11,1%	44,4%
400	5,6%	11,1%	16,7%	33,3%

$f_{i.} = \frac{n_{i.}}{n}$ distribution marginale en X

Distribution conjointe

Fréquence de la cellule C_{ij} : $f_{ij} = \dfrac{n_{ij}}{n}$

X \ Y	60	80	100	Total
300	15	4	1	$n_{1\cdot} = 20$
350	10	20	10	$n_{2\cdot} = 40$
400	5	10	15	$n_{3\cdot} = 30$
Total	$n_{\cdot 1} = 30$	$n_{\cdot 2} = 34$	$n_{\cdot 3} = 26$	90

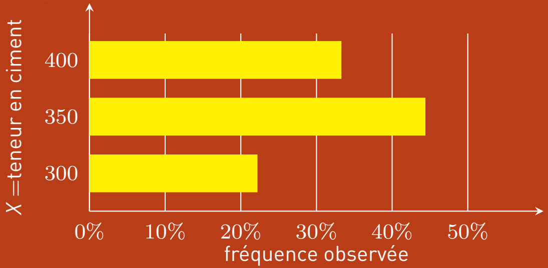
X \ Y	60	80	100	$f_{i\cdot}$
300	16,7%	4,4%	1,1%	22,2%
350	11,1%	22,2%	11,1%	44,4%
400	5,6%	11,1%	16,7%	33,3%
$f_{\cdot j}$	33,3%	37,8%	28,9%	

$f_{i\cdot} = \dfrac{n_{i\cdot}}{n}$ distribution marginale en X

$f_{\cdot j} = \dfrac{n_{\cdot j}}{n}$ distribution marginale en Y

Distributions marginales

- diagramme en barres



- teneur moyenne en ciment des éprouvettes

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{r} n_i.x_i = \sum_{i=1}^{r} f_i.x_i \simeq 355{,}5\, kg/m^3$$

Distribution conditionnelles

X \ Y	60	80	100	Total
300	15	4	1	20
350	10	20	10	40
400	5	10	15	30
Total	30	34	26	90

X \ Y	y_1	$\cdots$	y_j	$\cdots$	y_c	Total
y_j sachant que $X = x_i$	n_{i1}	$\cdots$	n_{ij}	$\cdots$	n_{ic}	$n_{i.}$

Distribution conditionnelles

$X \backslash Y$	60	80	100	Total
300	15	4	1	20
$f_{j/i}$	75%	20%	5%	

$X \backslash Y$	y_1	$\cdots$	y_j	$\cdots$	y_c	Total
y_j sachant que $X = x_i$	n_{i1}	$\cdots$	n_{ij}	$\cdots$	n_{ic}	$n_{i.}$

Fréquence de $Y = y_j$ sachant que $X = x_i$:

$$f_{j/i} = \frac{n_{ij}}{n_{i.}} = \frac{f_{ij}}{f_{i.}}$$

Indépendance

Définition

X et Y sont *indépendantes* si la distribution conditionnelle de Y sachant $X = x_i$ ne dépend pas de i :

$$\forall i,j, \quad f_{j/i} = f_{.j}$$
$$f_{ij} = f_{i.} \times f_{.j}$$

X \ Y	60	80	100	$f_{i.}$
300	16,7%	4,4%	1,1%	22,2%
350	11,1%	22,2%	11,1%	44,4%
400	5,6%	11,1%	16,7%	33,3%
$f_{.j}$	33,3%	37,8%	28,9%	

$$0{,}333 \times 0{,}222 \neq 0{,}167$$

X et Y ne sont pas indépendantes.

Sommaire

Covariance : un exemple

X engrais	20	24	28	22	32	28	32	36	41	41
Y rendement	16	18	23	24	28	29	26	31	32	34

$$\mathbb{C}\mathrm{ov}\,(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$$
$$= \frac{1}{n}\sum_{i=1}^{n}x_iy_i - \bar{x}\,\bar{y}$$

Covariance : un exemple

$$\bar{x} = \frac{1}{10}\sum_{i=1}^{10} x_i = 30{,}4$$

$$\bar{y} = \frac{1}{10}\sum_{i=1}^{10} y_i = 26{,}1$$

$$\frac{1}{n}\sum_{i=1}^{n} x_i y_i = 828{,}6$$

$$\begin{aligned}\mathbb{C}\mathrm{ov}\,(X,Y) &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \\ &= \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\,\bar{y}\end{aligned}$$

$$\mathbb{C}\mathrm{ov}\,(X,Y) = 828{,}6 - 30{,}4 \times 26{,}1 = 35{,}16$$

Propriétés de la covariance

Proposition

- *Symétrie :* $\mathbb{C}\mathrm{ov}\,(X,Y) = \mathbb{C}\mathrm{ov}\,(Y,X)$
- *lien avec la variance :* $\mathbb{C}\mathrm{ov}\,(X,X) = \mathbb{V}\,(X)$
- *transformation affine :* $\mathbb{C}\mathrm{ov}\,(aX+b, cY+d) = a\,c\,\mathbb{C}\mathrm{ov}\,(X,Y)$
- *Si X et Y sont indépendantes alors* $\mathbb{C}\mathrm{ov}\,(X,Y) = 0.$

⚠ La réciproque est fausse !

$\mathbb{C}\mathrm{ov}\,(X,Y)=0$ n'entraîne pas que X et Y sont indépendantes

Variance d'une somme

Théorème

$$\mathbb{V}(X+Y) = \mathbb{V}(X) + 2\mathbb{C}\mathrm{ov}(X,Y) + \mathbb{V}(Y)$$

Cas de variables décorrélées : $\mathbb{V}(X+Y) = \mathbb{V}(X) + \mathbb{V}(Y)$.

Inégalité de Cauchy-Schwarz

Pour $t \in \mathbb{R}$:

$$0 \leqslant \mathbb{V}\left(X + tY\right) = \mathbb{V}\left(X\right) + 2\,t\,\mathbb{C}\mathrm{ov}\left(X,Y\right) + t^2\mathbb{V}\left(Y\right)$$

$$\Delta = \left[2\mathbb{C}\mathrm{ov}\left(X,Y\right)\right]^2 - 4\mathbb{V}\left(X\right)\mathbb{V}\left(Y\right) = 4\left[\mathbb{C}\mathrm{ov}\left(X,Y\right)^2 - \mathbb{V}\left(X\right)\mathbb{V}\left(Y\right)\right] \leqslant 0$$

Théorème (Inégalité de Cauchy-Schwarz)

$$\left|\mathbb{C}\mathrm{ov}\left(X,Y\right)\right| \leqslant \sqrt{\mathbb{V}\left(X\right)\mathbb{V}\left(Y\right)} = \sigma(X)\,\sigma(Y)$$

Sommaire

Coefficient de corrélation linéaire

Définition

$$r(X,Y) = \frac{\mathbb{C}\mathrm{ov}\,(X,Y)}{\sigma(X)\,\sigma(Y)}$$

On a

$$-1 \leqslant r(X,Y) \leqslant 1$$

$r(X,Y) \simeq 0 \Rightarrow$ absence de relation linéaire (décorrélation)

$\not\Rightarrow$ indépendance

Exemple

Corrélation entre rendement et quantité d'engrais d'une parcelle de blé

X engrais	20	24	28	22	32	28	32	36	41	41
Y rendement	16	18	23	24	28	29	26	31	32	34

$$\sigma(X) \simeq 7{,}40 \qquad \sigma(Y) \simeq 5{,}91 \qquad \mathbb{C}\mathrm{ov}\,(X,Y) \simeq 35{,}16$$

$$r(X,Y) \simeq \frac{35{,}16}{7{,}40 \times 5{,}91} \simeq 0{,}89$$

Il y a corrélation linéaire forte.

INSTITUT
Mines-Télécom