Université 7 Novembre à Carthage
Ecole Supérieure de la Statistique et de l'Analyse de l'Information à Tunis

Année Universitaire 2010-2011
Première Année

Examen de Statistique Descriptive

Session Principale : 11 Janvier 2011

Cours de Mr Ghazouani et Mme Ouaili-Mallek — Durée 1h30

(2 pages)

Exercice 1 *On considère un couple de variables statistiques (T,P) où T désigne la taille en cm et P le poids en kg. L'information recueillie auprès d'un échantillon de 100 jeunes collégiens a permis de construire le tableau suivant :*

P \ T	[150 ; 155[	[155 ; 160[	[160 ; 165[	[165 ; 170[
[40 ; 45[	20	2	0	0
[45 ; 50[	9	18	5	1
[50 ; 55[	1	4	12	7
[55 ; 60[	0	1	6	14

On dispose en outre des informations suivantes :

$$\sum_{i=1}^{4} n_{i.}\ c_i = 4970 \qquad \sum_{j=1}^{4} n_{.j}\ c_j = 15935 \qquad \sum_{i=1}^{4} n_{i.}\ c_i^2 = 249775$$

$$\sum_{j=1}^{4} n_{.j}\ c_j^2 = 2542425 \qquad \sum_{i=1}^{4}\sum_{j=1}^{4} n_{ij} c_i c_j = 794387.6 \qquad \sum_{i=1}^{4}\sum_{j=1}^{4} \frac{n_{ij}^2}{n_{i.}n_{.j}} = 2$$

où les c indiquent les milieux des classes et les n les effectifs.

1. *Déduire de ce tableau les distributions marginales respectives des variables T et P.*

2. *Donner la distribution conditionnelle de la taille sachant que le poids de l'individu est compris entre 45 et 50 kg.*

3. *Les deux variables statistiques T et P sont-elles indépendantes? (justifier)*
4. *Quel aurait été l'effectif du pavé* $[50\ ;\ 55[\ X\ [160\ ;\ 165[$ *si les variables statistiques étaient indépendantes ?*
5. *Calculer la covariance des deux variables T et P.*
6. *Calculer la distance du χ^2. Interpréter le résultat.*
7. *On s'intéresse maintenant à la série de couples $(c_j, e_j)_{1\leq j\leq 4}$, où c_j désigne le centre de la j-ème classe de taille et e_j désigne l'effectif marginal qui lui est associé $(n_{.j})$. Après avoir donné le tableau associé à cette série, écrire l'équation de la droite de régression de E sur C.*
8. *Calculer le coefficient de corrélation linéaire entre ces deux variables. Commenter.*

Corrigé de l'exercice 1 *:*

1. *Distribution marginales de* T *:*

$Taille$	$[150\ ;\ 155[$	$[155\ ;\ 160[$	$[160\ ;\ 165[$	$[165\ ;\ 170[$	$Total$
$Effectif$	30	25	23	22	100

Distribution marginales de P *:*

$Poids$	$[40\ ;\ 45[$	$[45\ ;\ 50[$	$[50\ ;\ 55[$	$[55\ ;\ 60[$	$Total$
$Effectif$	22	33	24	21	100

2. *Distribution conditionnelle de la taille sachant que le poids de l'individu est compris entre 45 et 50 kg :*

$Taille$	$[150\ ;\ 155[$	$[155\ ;\ 160[$	$[160\ ;\ 165[$	$[165\ ;\ 170[$	$Total$
$Fréquence\ \left(f_j^2\right)$	0.27	0.55	0.15	0.03	1

3. T *et* P *ne sont pas indépendantes car* $n_{14} = 0$ *et* $n_{1.} * n_{.4} = 0.22 * 0.22 = 0.0484 \neq 0$.

4. *Si les variables statistiques* T *et* P *étaient indépendantes, la fréquence croisée associée au pavé* $[50\ ;\ 55[\ X\ [160\ ;\ 165[$ *aurait été* $f_{33} = 0.24 * 0.23 = 0.0552$*, soit un effectif croisé* $n_{33} = 5.52$

5.
$$s_{PT} = \frac{1}{100} \sum_{i=1}^{4} \sum_{j=1}^{4} n_{ij} c_i c_j - \overline{p}\ \overline{t}$$

$$\overline{p} = \frac{1}{100} \sum_{i=1}^{4} n_{i.}\ c_i = 49.7 \qquad \overline{t} = \frac{1}{100} \sum_{i=1}^{4} n_{.j}\ c_j = 159.35$$

Donc $s_{PT} = 7943.876 - 49.7 * 159.35 = 24.181$

6. *On a*
$$D^2 = n\left(\sum_{i=1}^{4} \sum_{j=1}^{4} \frac{n_{ij}^2}{n_{i.} n_{.j}} - 1 \right)$$

$D^2 = 100\,(2-1) = 100 >> 0$. *On peut donc conclure avec peu d'incertitude que les variables* T *et* P *ne sont pas indépendantes. Par ailleurs, la borne sup de* D^2 *vaut* $100 * 3 = 300$. *On ne peut donc pas envisager de liaison fonctionnelle entre les deux variables. Ceci était prévisible puisque pour tout* i *et pour tout* j, $n_{ij} \neq n_{i.}$ *et* $n_{ij} \neq n_{.j}$.

7. *Tableau de la répartion de* E *selon* C :

C	152.5	157.5	162.5	167.5
E	30	25	23	22

La droite de régression de E *sur* C *a pour équation :*

$$\widehat{e_i} = \widehat{\alpha} c_i + \widehat{\beta} \quad avec\ \ \widehat{\alpha} = \frac{s_{CE}}{s_C^2} \quad et\ \ \widehat{\beta} = \overline{e} - \widehat{\alpha}\overline{c}$$

$s_{CE} = \dfrac{1}{4}\sum_{i=1}^{4} c_i * e_i - \overline{c} * \overline{e}$

$\overline{c} = \dfrac{152.5 + 157.5 + 162.5 + 167.5}{4} = 160 \qquad \overline{e} = \dfrac{100}{4} = 25$

$\sum_{i=1}^{4} c_i * e_i = 152.5 * 30 + 157.5 * 25 + 162.5 * 23 + 167.5 * 22 = 15935$

D'où $s_{CE} = \dfrac{15935}{4} - 160 * 25 = -16.25$

$s_C^2 = \dfrac{1}{4}\sum_{i=1}^{4} c_i^2 - \overline{c}^2 = \dfrac{152.5^2 + 157.5^2 + 162.5^2 + 167.5^2}{4} - 160^2 = 31.25$

On a finalement : $\widehat{\alpha} = -\dfrac{16.25}{31.25} = -0.52$ *et* $\widehat{\beta} = 25 + -0.52 * 160 = -58.2$

et l'équation s'écrit :

$$\widehat{e_i} = -0.52 c_i - 58.2$$

8. *Coefficient de corrélation*

$$r_{CE} = \frac{s_{CE}}{s_C * s_E}$$

$s_E = \left[\dfrac{1}{4}\sum_{i=1}^{4} e_i^2 - \overline{e}^2\right]^{\frac{1}{2}} = \left[\dfrac{30^2 + 25^2 + 23^2 + 22^2}{4} - 25^2\right]^{\frac{1}{2}} = \sqrt{9.5} = 3.0822$

$s_C = \sqrt{31.25} = 5.5902$

D'où $r_{CE} = -\dfrac{16.25}{5.5902 * 3.0822} = -0.94$

Le coefficient de corrélation exprime une forte liaison linéaire entre les deux variables. Il s'agit néanmoins d'une relation négative.