



## Université de Poitiers Département de Mathématiques

Statistique descriptive, 1er semestre, année univ. 2009-2010

Fiche 4

### Statistiques descriptives bivariées

#### Exercice 1

##### Régression linéaire, approche analytique

On désigne par  $x = (x_1, x_2, \dots, x_n)$  un échantillon pondéré par les poids  $p = (p_1, \dots, p_n)$ . On suppose que les  $p_i$  sont strictement positifs et de somme 1. On suppose que la variance empirique  $s_x^2$  est strictement positive.

On munit  $\mathbb{R}^n$  du produit scalaire  $\langle x; y \rangle = \sum_{i=1}^n p_i x_i y_i$ . On définit le vecteur  $\underline{1} = (1; \dots; 1)$ .

1) Vérifier que  $(a; b) \mapsto \sum_i p_i (y_i - (ax_i + b))^2$  s'optimise comme la fonction

$$J(a, b) = a^2 \overline{x^2} + 2ab\bar{x} + b^2 - 2a\bar{xy} - 2b\bar{y}.$$

Montrer que  $J(a; b)$  admet un minimum. En déduire les équations de la droite de régression.

2) Vérifier que le coefficient de détermination  $r$  de la droite de régression  $\hat{y} = \hat{a}x + \hat{b}$  satisfait  $r^2 = \rho_{y, \hat{y}}^2$ .

3) Montrer que les variables résidu  $e = y - \hat{y}$  et  $x$  sont non-corrélées.

4) Montrer que

$$s_e^2 = s_y^2 - \frac{s_{x,y}^2}{s_x^2}.$$

En déduire que  $|\rho_{x,y}| = 1$  est équivalent à l'existence de constantes  $(a, b)$  telles que  $y = ax + b$ .

#### Exercice 2

##### Choix de modèle en régression

Six couples de données concernant deux variables  $X$  et  $Y$  sont reportés dans le tableau suivant :

$x_i$	1	2	3	4	5	6
$y_i$	7	10	10	12	17	16

1) Tracer le nuage de point correspondant aux données.

2) Déterminer la droite de régression  $y = ax + b$  et la valeur du coefficient de détermination.

3) Tracer la droite de régression, calculer la valeur prédite pour  $\bar{x}$ , puis construire le tableau donnant pour chaque valeur  $x_i$  la valeur du résidu correspondant.

4) On considère le modèle suivant :

$$y = b + ax^2 + e.$$

Expliquer comment on peut se ramener au cadre usuel du modèle linéaire (on pourra poser  $z = x^2$ ). Déterminer la droite de régression  $y = az + b$  et la valeur du coefficient de détermination.

5) Tracer la courbe de régression, puis la valeur prédite pour  $\bar{x}$ , puis construire le tableau donnant pour chaque valeur  $x_i$  la valeur du résidu correspondant.

6) Quel modèle vous semble préférable ?