

E.I.S.T.I. – Département de Mathématiques
1^{ère} année d'ingénieurs

STATISTIQUES DESCRIPTIVES
TD N°4 : statistiques bivariées
Quantitatif-Qualitatif / Quantitatif-Quantitatif

Quantitatif-Qualitatif

Dans une population Ω de taille n , on observe deux caractéristiques :

- une qualitative, $x = \{x_k\}_{k=1, \dots, n}$, à p modalités notées, $m_1, \dots, m_l, \dots, m_p$
- une quantitative continue $y = \{y_k\}_{k=1, \dots, n}$ de moyenne \bar{y} et de variance s_y^2 .

On suppose que les modalités de la série x définissent des sous-populations

$$\Omega = \Omega_1 \cup \dots \cup \Omega_p \quad \text{où} \quad \Omega_l \cap \Omega_{l'} = \emptyset,$$

On peut alors considérer les restrictions de la caractéristique y sur chacune des sous-populations et calculer les indicateurs numériques usuels pour chaque modalités de x ,

- moyennes : $\bar{y}_l, l=1, \dots, p$
- variances : $s_l^2, l=1, \dots, p$

....

Exercice 1 : Formules de décomposition

Montrer que

$$\bar{y} = \frac{1}{n} \sum_{l=1}^p n_l \bar{y}_l$$

$$s_y^2 = \frac{1}{n} \sum_{l=1}^p n_l (\bar{y}_l - \bar{y})^2 + \frac{1}{n} \sum_{l=1}^p n_l s_l^2 = s_E^2 + s_R^2$$

A quoi correspondent les termes s_E^2 et s_R^2 ?

On définit un indice de liaison entre les deux caractéristiques x et y par le rapport de corrélation

$$s_{y/x} = \sqrt{\frac{s_E^2}{s_y^2}}$$

Donner un encadrement de $S_{y/x}$. A quoi correspondent les cas $S_{y/x}=0$ et $S_{y/x}=1$?

Exercice 2

Dans une entreprise on étudie le lien entre le salaire mensuel, l'âge et le nombre d'enfants des cadres supérieurs.

On effectue une première étude entre le nombre d'enfants et le salaire.

- 1) Sur un même graphique, tracer les boîtes à moustaches. Commenter.
- 2) Calculer le rapport de corrélation. Qu'en pensez-vous ?

individus	Age	Salaire (k€)	nb enfants
1	junior	35	1
2	senior	51	3
3	expert	39	3
4	expert	40	1
5	senior	50	3
6	expert	37	3
7	junior	33	0
8	junior	36	0
9	expert	42	2
10	senior	44	0
11	expert	46	1
12	expert	36	2
13	expert	41	2
14	senior	49	1
15	junior	36	0
16	junior	34	2
17	junior	36	1
18	expert	36	3
19	junior	32	2
20	junior	35	2

salaires (k€)

nb enfants			
0	1	2	3
33	35	41	51
36	40	34	39
44	46	32	50
36	49	42	37
	36	36	36
		35	

Effectif	4	5	6	5
Moyenne	37,25	41,20	36,67	42,60
Mediane	36,00	40,00	35,50	39,00

Variance	22,25	37,70	15,87	53,30
Ecart-type	4,72	6,14	3,98	7,30
CV	0,13	0,15	0,11	0,17
Q1	35,25	36,00	34,25	37,00
Q3	38,00	46,00	39,75	50,00
IIQ	2,75	10,00	5,50	13,00
m	31,13	21,00	26,00	17,50
M	42,13	61,00	48,00	69,50

Moyenne	39,4
Variance	33,73

Quantitatif-Quantitatif

Dans une population Ω de taille n , on observe deux caractéristiques quantitatives continues, $x=\{x_k\}_{k=1,\dots,n}$, et $y=\{y_k\}_{k=1,\dots,n}$, de moyennes \bar{x} et \bar{y} et de variances s_x^2 et s_y^2 .

Exercice 3 : covariance et coefficient de corrélation linéaire et droite de régression

La *covariance* entre les deux caractéristiques est définie par

$$c_{xy} = \frac{1}{n} \sum_{k=1}^n (x_k - \bar{x})(y_k - \bar{y}) = \frac{1}{n} \sum_{k=1}^n x_k y_k - \bar{x} \bar{y}$$

C'est une forme bilinéaire symétrique, à valeurs réelles, telle que

$$c_{xx} = s_x^2$$

$$s_{x+y}^2 = s_x^2 + s_y^2 + 2c_{xy}$$

$$c_{xy}^2 \leq s_x^2 s_y^2 \text{ (inégalité de Cauchy-Schwartz)}$$

On définit alors le *coefficient de corrélation linéaire* par

$$r_{xy} = \frac{c_{xy}}{s_x s_y}.$$

- 1) Montrer que le coefficient de corrélation linéaire est symétrique, à valeurs dans $[-1,1]$ et correspond à la covariance des observations réduites et centrées. A quoi correspondent les valeurs -1 et $+1$?

La droite de régression de y sur x est construite en minimisant,

$$S(a, b) = \frac{1}{n} \sum_{i=1}^n (y_i - (ax_i + b))^2,$$

C'est-à-dire en minimisant la moyenne des écarts au carré entre l'observation y_i et la valeur de la droite au point x_i . On obtient alors

$$\hat{a} = \frac{c_{xy}}{s_x^2} = r_{xy} \frac{s_y}{s_x} \quad \text{et} \quad \hat{b} = \bar{y} - \hat{a}\bar{x}.$$

La série des valeurs prédites par la droite de régression est donnée par

$$\hat{y}_i = \hat{a}x_i + \hat{b},$$

et les *résidus* par

$$\hat{e}_i = y_i - (\hat{a}x_i + \hat{b}).$$

- 2) Montrer que la moyenne des valeurs prédites est égale à la moyenne de la série observée.
 3) Montrer que les résidus sont de moyenne nulle et sont non corrélés avec la série X . Qu'est-ce que cela signifie ?
 4) Montrer la formule de décomposition de la variance

$$s_y^2 = s_E^2 + s_R^2$$

où s_E^2 est la *variance expliquée* par la droite de régression, et s_R^2 est la *variance résiduelle*.

On peut alors montrer que le *coefficient de détermination*

$$R^2 = \frac{s_E^2}{s_y^2},$$

qui donne le taux de variance expliquée par la droite de régression, est égale au coefficient de corrélation linéaire au carré, $R^2 = r_{xy}^2$.

Exercice 4

On donne pour les six premiers mois de l'année 1982 les nombres d'offres d'emploi (concernant des emplois durables à temps plein) et de demandes d'emploi (déposées par des personnes sans emploi, immédiatement disponibles, à la recherche d'un emploi durable à plein temps). Les nombres sont exprimés en milliers.

Offres (x_i)	61	66,7	75,8	78,6	82,8	87,2
Demandes (y_i)	2034	2003,8	1964,5	1928,2	1885,3	1867,1

On a les résultats suivants

$$\bar{x} = 75,35 \quad \bar{y} = 1947,15 \quad s_x^2 = 97,49 \quad s_y^2 = 4329,14 \quad c_{xy} = -639,90$$

- 1) Tracer le nuage de points des demandes en fonction des offres.
- 2) Calculer le coefficient de corrélation linéaire. Conclusion
- 3) Déterminer la droite de régression. Tracer la droite sur le graphique.
- 4) Calculer les prévisions et les résidus. Vérifier les hypothèses sur les résidus.

Exercice 5 (PY Bernard, exercices corrigés de statistique descriptive, ed. economica)

Une étude a été menée auprès d'entreprises afin d'établir le lien entre les quantités commandées d'un bien, Y, et son prix, X et on obtient les observations suivantes.

Prix de vente (€)	Quantités commandées	log(Prix)	log(Quantités)
95	104	4,55	4,64
130	58	4,87	4,06
148	37	5,00	3,61
210	22	5,35	3,09
250	12	5,52	2,48
330	9	5,80	2,20
Moyenne		5,18	3,35
Variance		0,21	0,88
Covariance		-0,43	

- 1) Tracer le nuage de points. Conclusion.
- 2) On pose $u = \log(x)$ et $v = \log(y)$. Quelle est la relation entre u et v ?
- 3) Calculer le coefficient de corrélation linéaire entre u et v.
- 4) Trouver la droite de régression de v sur u.
- 5) En déduire la quantité qui serait commandée si le prix était fixé à 75€.