

Quelques exercices de statistique descriptive

1. Un supermarché reçoit 1000 caissettes comprenant chacune 9 pêches. La distribution du nombre de pêches abîmées par caissette est décrite dans le Tableau suivant:

x_j	0	1	2	3	4	5	6	7	8	9
n_j	126	307	317	173	55	17	2	2	0	1

- (a) En moyenne, combien y a-t-il de pêches abîmées par caissette?
 - (b) Combien de pêches abîmées un client trouvera-t-il le plus fréquemment s'il achète, sans être suffisamment attentif, une caissette de pêches dans ce supermarché?
 - (c) Si le supermarché décidait de ne conserver que la moitié des caissettes (évidemment celles contenant le moins possible de pêches abîmées), quel est le nombre maximum de pêches abîmées contenues dans ces caissettes privilégiées?
 - (d) Si que les employés du supermarché recevaient en cadeau 10% des caissettes (choisies parmi les meilleures), tandis que les 10% des caissettes les plus abîmées seraient gardées pour confectionner des confitures, déterminer le nombre moyen de pêches abîmées dans les caissettes restantes.
2. Rechercher graphiquement les quartiles des 4 séries suivantes :
- $S_1 = \{1, \dots, 8\}$ $S_2 = \{1, \dots, 9\}$
 $S_3 = \{1, \dots, 10\}$ $S_4 = \{1, \dots, 11\}$
3. Le tableau 3 de l'avant-dernière page donne la répartition selon l'éducation de la population âgée de 25 à 64 ans pour 13 pays industrialisés, ainsi que les taux de chômage (en %) observés dans ces pays parmi les groupes constitués des différents niveaux d'études.

La variable $X = \text{"Education"}$ est qualitative ordinale avec les quatre modalités:

- m_1 : Dernier diplôme obtenu: Enseignement primaire
- m_2 : Dernier diplôme obtenu: Enseignement secondaire
- m_3 : Dernier diplôme obtenu: Ens. sup. non-universitaire
- m_4 : Dernier diplôme obtenu: Enseignement universitaire

Par contre, la variable "Taux de chômage" est une variable quantitative continue (ici étudiée sur 4 sous-populations).

Considérer chacune des modalités m_1, \dots, m_4 de la variable X comme une variable univariée (quantitative continue) observée sur la population constituée des 13 pays.

Comparer les taux de chômage dans les quatre niveaux d'études à l'aide de boîtes à moustaches modifiées. Commenter.

4. Le tableau 4 de l'avant-dernière page donne, pour différentes régions de France, la valeur du PIB par habitant en 2004 (en Euros) ainsi que le taux de chômage pour la même période.

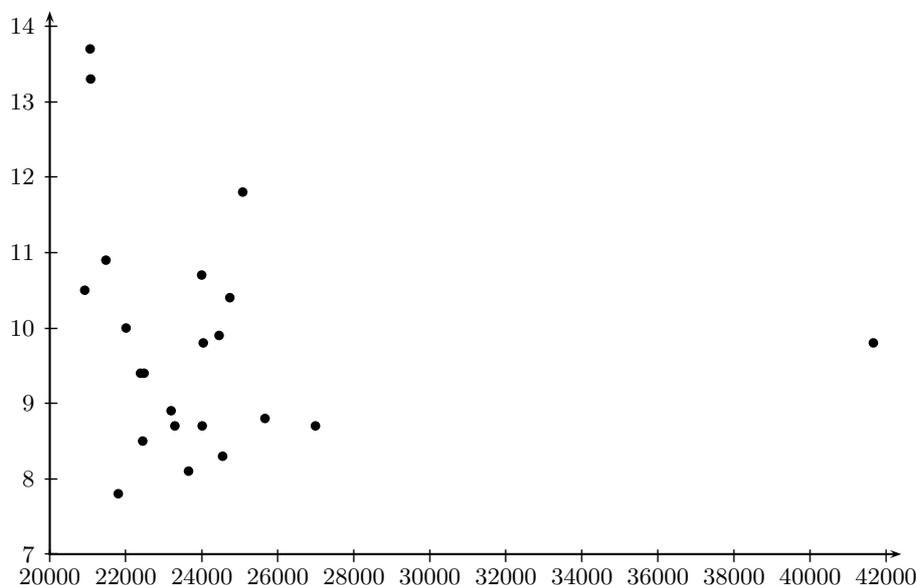
La figure ci-dessous représente le nuage de points de la série bivariée "PIB par habitant-Taux de chômage".

- (a) Calculer le coefficient de corrélation de cette série double sachant que

$$\begin{aligned}
 - \sum_{i=1}^{22} d_i &= 530\,938 \\
 - \sum_{i=1}^{22} e_i &= 216 \\
 - \sum_{i=1}^{22} d_i \cdot e_i &= 5\,194\,787,5 \\
 - \sum_{i=1}^{22} d_i^2 &= 13\,189\,365\,324
 \end{aligned}$$

$$- \sum_{i=1}^{22} e_i^2 = 2174$$

- (b) Un ajustement par une droite de régression est-il indiqué? Justifier.
 (c) Dans l'affirmative, calculer les coefficients d'une droite de régression et la représenter sur le graphique.



5. On interroge vingt personnes sur le nombre de TV au rez-de-chaussée de leur habitation et le nombre d'adultes y habitant. On obtient le tableau suivant :

TV	Adultes	
	1	2
0	2	0
1	4	8
2	1	5

Calculer la covariance de cette série et interpréter la valeur obtenue.

À l'aide de l'outil informatique

6. À partir du fichier « Données », représenter les graphiques relatifs aux caractères suivants:
- Gravité des faits reprochés aux dirigeants de la Carolorégienne
 - Nombre de frères et sœurs de 197 étudiants
 - Taille de 197 étudiants
7. À partir des données des deux dernière séries,
- calculer le nombre de frères et sœurs moyen ainsi que le nombre de frères et sœurs médian, les comparer et interpréter ;
 - calculer la taille moyenne et la taille médiane des 197 étudiants ;
 - calculer la taille moyenne et la taille médiane des 197 étudiants à partir de la répartition en classes, comparer avec les résultats du point précédent ;
 - calculer la variance et l'écart-type des deux séries.
8. Le Tableau de contingence 1 décrit la répartition d'une population constituée de 535 ménages selon les deux variables suivantes : X représente le nombre de pièces de l'habitation et Y correspond au nombre d'enfants du ménage.
- (a) Que représentent les effectifs n_{23} et n_{54} ? Calculer les fréquences correspondantes.

TAB. 1 – *Tableau de contingence pour la série double “Nombre de pièces - Nombre d’enfants”.*

Pièces	Enfants				
	0	1	2	3	4
1	7	3	2	1	0
2	24	32	21	2	1
3	16	35	54	26	4
4	9	28	74	55	12
5	4	12	46	13	12
6	2	6	16	11	7

- (b) Déterminer les distributions marginales des variables X et Y . Calculer les moyennes, médianes et variances marginales ainsi que les modes marginaux.
- (c) Déterminer la distribution conditionnelle de la variable Y sachant que le nombre de pièces du logement est égal à 4. Calculer la moyenne et la variance de cette distribution conditionnelle.
- (d) Calculer la covariance de la série double.
9. Le responsable des ressources humaines d’une entreprise s’intéresse au bien-être de ses employés. A cet effet, il interroge ceux-ci et souhaite, notamment, savoir si les dépenses mensuelles pour les loisirs sont corrélées avec les revenus des travailleurs. Il recueille les données suivantes (exprimées en une certaine unité monétaire) :

X	Y	X	Y	X	Y
752	85	492	81	679	80
855	83	569	81	902	226
871	162	462	80	918	260
734	79	907	243	828	82
610	81	643	84	875	186
582	83	862	84	809	77
921	281	524	82	894	223

- (a) Représenter graphiquement le nuage de points correspondant aux variables X et Y relatives, respectivement, aux revenus mensuels nets et aux dépenses mensuelles en loisirs.
- (b) Déterminer la valeur du coefficient de corrélation et interpréter cette valeur en tenant compte du contexte.
- (c) Déterminer l’équation de la droite de régression de Y en X obtenue par la méthode des moindres carrés et la représenter sur le graphique.
- (d) Déterminer la valeur du coefficient de détermination et l’interpréter.
- (e) Quelle est la pertinence de l’ajustement linéaire réalisé? Trouver une meilleure stratégie pour analyser cet ensemble de données.

Exercices supplémentaires

1. Répondant à une offre d’emploi, une personne s’interroge sur le montant de ses rémunérations futures en cas d’embauche. Le directeur de l’entreprise de vente à domicile lui répond que le salaire moyen de la firme est supérieur à 1600 euros par mois, mais que, pendant la période de formation, l’employé ne gagnera que 500 euros chaque mois, puis sera augmenté dans la suite.

Avant de signer le contrat d’embauche, la personne a mené son enquête et a obtenu les renseignements suivants:

- Le directeur gagne 12500 euros par mois.

- Le sous-directeur gagne 6000 euros par mois.
- Chacun des 4 chefs de secteur gagne 1200 euros par mois.
- Chacun des 5 techniciens gagne 950 euros par mois.
- Chacun des 10 démarcheurs gagne 600 euros par mois.

- (a) Le directeur a-t-il dit la vérité au candidat?
- (b) Quelle question le candidat aurait-il dû poser au patron pour avoir une estimation plus réaliste de son salaire futur?
2. Les poids des 22 étudiantes de première candidature Ingénieur de gestion sont donnés par la série ordonnée suivante:

$$S = \{47,48,49,50,53,55,55,55,56,56, \\ 58,59,61,62,62,63,63,63,64,65,65,66\}.$$

- (a) Calculer la moyenne arithmétique, la médiane et le mode de cette série et interpréter ces paramètres.
- (b) Calculer les quartiles Q_1 et Q_3 . Sachant que la série des poids des étudiants masculins de 1ère candidature IG correspond aux paramètres suivants:

$$Q_1 = 65; \tilde{x} = 69,5; Q_3 = 75,$$

tandis que les observations individuelles sont reprises ci-dessous, comparer les deux distributions de poids à l'aide de boîtes à moustaches.

68	70	67	75	72	71	67	65	60	60	65	65
77	95	85	70	70	72	66	75	90	65	62	70
52	60	59	65	68	71	97	65	57	75	77	75
85	56	77	67	62	52	67	72	79	60	72	69
58	55	75	75	78	65	95	65	90	72	72	60

- (c) Dans une usine, trois machines A , B et C fabriquent des tiges métalliques. En principe, les tiges devraient mesurer 250 centimètres. Avec le temps, les performances des trois machines se sont détériorées et le directeur de l'usine désire se renseigner sur la qualité de ses produits finis. Pour cela, il mesure toutes les tiges produites sur ces machines lors d'une semaine bien précise. Il obtient les renseignements suivants :
- La machine A a produit 1248 tiges. La longueur minimale obtenue est 215 cm et la longueur maximale 285 cm. La distribution des longueurs (groupées en 5 classes d'amplitude 15 cm) est donnée dans le tableau 2 :

Classes des longueurs	Effectifs
[215,230[204
[230,245[324
[245,260[288
[260,275[360
[275,290[72

TAB. 2 – Longueurs des tiges produites par la machine A

- A partir de la machine B , plus ancienne et donc plus lente, seules 17 tiges ont été fabriquées. Les longueurs ordonnées sont les suivantes :

215, 235, 237, 249, 251, 252, 252, 254, 255, 256, 257, 259, 260, 260, 261, 275, 285.

- A partir de la machine C , 1036 tiges ont été produites. Les caractéristiques de la distribution des longueurs sont les suivantes :

longueur minimale : 230 cm; longueur maximale : 270 cm; médiane : 250 cm; $Q_1 = 245$ cm et $Q_3 = 257$ cm.

- Pour les machines A et B , déterminer la médiane et les quartiles Q_1 et Q_3 .
- Représenter les unes à côté des autres les boîtes à moustaches (de base) pour les distributions des longueurs des trois machines et commenter.
- Commenter les affirmations suivantes :
 - Comme la longueur "normale" est 250 cm, seule la machine B fonctionne correctement.
 - Si on produit une tige sur la machine A , il y a une chance sur deux qu'elle ne fasse pas les 250 cm attendus.
 - Si on produit une tige sur la machine C , il y a une chance sur deux qu'elle ne fasse pas les 250 cm attendus.
 - La machine qui a la plus petite dispersion dans ses longueurs est la machine C .
 - A partir de la machine B , les trois quarts des tiges font au moins la longueur réglementaire.
 - Il vaut mieux produire les tiges sur les machines A ou B , car la longueur maximale relevée est 285 cm (contre 270 cm pour C).
 - La longueur moyenne des tiges produites par la machine C est 257 cm.

3. Soit S la série double suivante, due à Anscombe (1973) :

x_i	10	8	13	9	11	14	6	4	12	7	5
y_i	8,04	6,95	7,58	8,81	8,33	9,96	7,24	4,26	10,84	4,82	5,68

- Représenter le nuage de points de cette série bivariable.
- Rechercher l'équation de la droite des moindres carrés. Calculer les résidus.
- Anscombe a construit trois autres séries assez proches de S , à savoir :

S_1		S_2		S_3	
x	y	x	y	x	y
10	9,14	10	7,46	8	6,58
8	8,14	8	6,77	8	5,76
13	8,74	13	12,74	8	7,71
9	8,77	9	7,11	8	8,84
11	9,26	11	7,81	8	8,47
14	8,1	14	8,84	8	7,04
6	6,13	6	6,08	8	5,25
4	3,1	4	5,39	19	12,5
12	9,13	12	8,15	8	5,56
7	7,26	7	6,42	8	7,91
5	4,74	5	5,73	8	6,89

Pour chacune de ces séries, construire le nuage de points, déterminer l'équation de la droite des moindres carrés, calculer le coefficient de corrélation linéaire et comparer les résultats obtenus. Construire ensuite des diagrammes pour visualiser les résidus.

- (d) A la lumière de ces exemples, émettre des conjectures sur l'opportunité d'ajuster un nuage de points à l'aide de la droite des moindres carrés.
4. Le tableau 5 (Source : Rousseeuw - Leroy) donne les poids du corps en kilo et les poids du cerveau en grammes pour 28 espèces.
- Les valeurs des deux variables variant fortement, définir les nouvelles variables X et Y à l'aide d'une transformation logarithmique (prendre par exemple le logarithme népérien). Soit X la variable "Ln du poids du corps" et Y la variable "Ln du poids du cerveau".
- (a) Calculer les moyennes et variances marginales des deux variables.
- (b) Représenter un diagramme de dispersion pour la série double en mettant les valeurs de X en abscisses et les valeurs de Y en ordonnées.
- En utilisant les mêmes critères que lors de la construction d'une boîte à moustaches, déterminer s'il y a des observations extrêmes dans les séries marginales.
 - A partir du nuage de points de la série double, la relation entre les deux variables semble-t-elle croissante ou décroissante?
 - Vérifier cette impression en calculant la covariance. Interpréter le signe de la covariance en décomposant le plan en 4 parties à l'aide des droites $x = \bar{x}$ et $y = \bar{y}$.
- (c) Calculer la corrélation pour la série double.
- (d) Déterminer les droites de régression linéaire (par la méthode des moindres carrés) en prenant d'abord X comme variable explicative et Y comme variable dépendante et ensuite en permutant les rôles des deux variables. Représenter les deux droites de régression sur le nuage de points. Calculer le coefficient de détermination pour chacune des deux droites. Commenter.
- (e) En observant attentivement le nuage de points, les observations 6, 16 et 25 semblent se détacher de la tendance linéaire indiquée par les autres points. Cette remarque s'applique aussi (dans une moindre mesure) aux observations 14 et 17. Exprimer en mots les caractéristiques de ces deux groupes de données.
- (f) Recalculer la corrélation de la série double en laissant de côté les observations 6, 16 et 25. Commenter. En prenant la variable X comme variable explicative et la variable Y comme variable dépendante, déterminer de même la droite de régression de Y en X en laissant de côté ces trois observations.
- Représenter sur le même graphique les deux droites de régression de Y en X (avec le nuage de points).
 - Comparer les coefficients de détermination.

TAB. 3 – Répartition (%) de la population âgée de 25 à 64 ans selon l'éducation et taux de chômage (%) par niveaux d'éducation (population de 25 à 64 ans). Ces données correspondent à l'année 1995 (Pallage S. - Zimmerman CH., "Assurance chômage et sociétés" dans Finances publiques; Finances privées, Editions de l'Université de Liège.).

Pays	Niveaux d'études				Taux de chômage selon l'éducation			
	m_1	m_2	m_3	m_4	m_1	m_2	m_3	m_4
Allemagne	16	61	10	13	13,3	7,9	5,2	4,7
Belgique	47	29	14	11	13,4	7,5	3,5	3,6
Canada	25	28	30	17	13	8,6	7,5	4,6
Danemark	38	42	6	14	14,6	8,3	5,3	4,3
Espagne	72	12	4	12	20,6	18,5	16,6	13,8
Etats-Unis	14	53	8	25	10	5	3,6	2,5
France	32	50	8	11	14	8,9	5,9	7
Grèce	57	25	6	11	6,3	9	10,1	7,1
Irlande	53	27	10	10	16,4	7,6	5	3,4
Norvège	19	53	11	18	6,5	4	3,4	1,7
Royaume Uni	24	54	9	12	12,2	7,4	4,1	3,5
Suède	25	46	14	14	10,1	8,7	4,8	4,2
Suisse	18	61	12	9	5,8	2,8	1,5	2,6

TAB. 4 – PIB (en euros) et taux de chômage (en %) pour différentes régions de France

	PIB (d_i)	Chômage en % (e_i)
Corse	20 918	10.5
Languedoc-Roussillon	21 060	13.7
Nord-Pas-de-Calais	21 076	13.3
Picardie	21 477	10.9
Limousin	21 799	7.8
Lorraine	22 005	10
Basse-Normandie	22 385	9.4
Auvergne	22 445	8.5
Poitou-Charentes	22 477	9.4
Franche-Comté	23 190	8.9
Bourgogne	23 291	8.7
Bretagne	23 653	8.1
Haute-Normandie	23 994	10.7
Centre	24 010	8.7
Midi-Pyrénées	24 037	9.8
Aquitaine	24 452	9.9
Pays de la Loire	24 547	8.3
Champagne-Ardenne	24 738	10.4
Provence-Alpes-Côte d'Azur	25 073	11.8
Alsace	25 661	8.8
Rhône-Alpes	26 988	8.7
Ile-de-France	41 662	9.8

TAB. 5 – Poids du corps et du cerveau de 28 animaux

i	Nom	Pds du corps (Kg)	Pds du cerveau (g)
1	Castor	1.35	8.1
2	Vache	465	423
3	Loup gris	36.33	119.5
4	Chèvre	27.66	115
5	Cobaye	1.04	5.5
6	Diplodocus	11700	50
7	Éléphant d'Asie	2547	4603
8	Âne	187.1	419
9	Cheval	521	655
10	Babouin	10	115
11	Chat	3.3	25.6
12	Girafe	529	680
13	Gorille	207	406
14	Humain	62	1320
15	Éléphant d'Afrique	6654	5712
16	Triceratops	9400	70
17	Singe Rhesus	6.8	179
18	Kangourou	35	56
19	Hamster	0.12	1
20	Souris	0.023	0.4
21	Lapin	2.5	12.1
22	Mouton	55.5	175
23	Jaguar	100	157
24	Chimpanzé	52.16	440
25	Brachiosaurus	87000	154.5
26	Rat	0.280	1.9
27	Taupe	0.122	3
28	Cochon	192	180