EXAMEN DE STATISTIQUES

Documents autorisés : l'utilisation d'une feuille A4 recto-verso est autorisée.

Le tableau suivant résume avec une série statistique double une campagne de mesures de deux caractères notés x et y, campagne opérée dans une entreprise :

Indice i	1	2	3	4	5	6	7	8	9	10	11	12
x_i	-2.00	-1.42	-1.32	0.05	-0.03	0.94	-0.88	0.38	1.23	-0.75	1.74	1.56
y_i	-0.62	-0.11	-1.46	2.22	-0.01	0.61	-2.20	0.42	0.37	-0.25	1.33	0.92

1 Statistiques descriptives

1. Construire le nuage de points $P_i = (x_i, y_i)$, $i \in [1, 12]$, correspondant à cette série statistique double. Choisir et justifier les unités utilisées pour le graphique.
2. Calculer les coordonnées du *centre de gravité* $G = (\overline{x}, \overline{y})$ du nuage et le placer sur le graphique.
3. Calculer le coefficient de corrélation linéaire r du nuage.
4. Déterminer l'équation de la droite de régression D du nuage. Tracer D sur le graphique (marquer les points utilisés pour tracer D). Jugez-vous que (D) approche bien le nuage de points ?
5. Sur un graphique soigné, représenter l'histogramme illustrant la distribution (marginale) du caractère y, indiquer sur l'histogramme, la moyenne, la médiane des mesures y_i ainsi que leur écart-type.
6. On considère une série statistique 1D déduite de la précédente en multipliant les caractères pour tous les indices $z_i = x_i * y_i$. Sur un graphique soigné, représenter l'histogramme correspondant à la distribution du nouveau caractère z, représenter sur l'histogramme, la moyenne la médiane et l'écart-type des observations z_i.
7. Puisque tous les $z_i = x_i * y_i$ sont positifs, pouvions-nous, ou pas, anticiper la présence d'une dépendance entre les caractères x et y ? pourquoi ?

2 Estimation au maximum de vraisemblance

Un expert nous indique que le caractère y suit une loi normale dont on ignore les paramètres m et σ. On considère que les mesures y_i à notre disposition sont la réalisation d'un échantillon de 12 variables aléatoires i.i.d. On rappelle que la vraisemblance de y_i, en fonction de m et σ, est alors $L_{m,\sigma^2}(y_i) = \frac{1}{\sqrt{2\pi}\sigma} \exp - \frac{(y_i - m)^2}{2\sigma^2}$.

1. Expliquer pourquoi la vraisemblance $L_{m,\sigma^2}(y_1 \dots y_{12})$ des 12 mesures y_i s'écrit alors
$$L_{m,\sigma^2}(y_1 \dots y_{12}) = L_{m,\sigma^2}(y_1) \times L_{m,\sigma^2}(y_2) \times \dots \times L_{m,\sigma^2}(y_i) \times \dots \times L_{m,\sigma^2}(y_{12})$$
2. On fait 3 hypothèses pour rechercher parmi elles les paramètres les plus vraisemblables : $(m; \sigma) = (0; 1)$, $(m; \sigma) = (0.1; 1)$, $(m; \sigma) = (0.1; 1.2)$. Quelle est la méthode utilisée pour comparer les 3 hypothèses ? Que valent $L_{0;1}(y_1 \dots y_{12})$, $L_{0.1;1}(y_1 \dots y_{12})$, $L_{0.1,(1.2)^2}(y_1 \dots y_{12})$? Quels paramètres sont les plus vraisemblables ?
3. D'après les calculs faits dans la partie statistique descriptives, peut-on trouver de meilleures valeurs pour m et σ ?