Université de Pau et des Pays de l'Adour **Année 2019-2020**
Département de Mathématiques
M1-Génie Pétrolier - Analyse de données

TD-2 : Série bivariée (indépendance, corrélation, régression linéaire)

Exercice 1. L'étude palynologique d'un échantillon du Mésozoïque d'Aquitaine a permis de construire la table de contingence suivante :

Couleur / Richesse	Gris foncé	Grisâtre	Rougeâtre
Non fossilifère	26	19	18
Fossilifère	34	9	4

1. Préciser la population étudiée et la nature des deux variables.
2. Calculer les effectifs marginaux des deux distributions. Calculer l'effectif total.
3. Déterminer la distribution de la richesse des roches conditionnellement au fait que la couleur des roches est gris foncé.
4. Déterminer la distribution de la couleur des roches conditionnellement au fait qu'elles ne sont pas fossilifères.
5. Déterminer le mode de la distribution bivariée.
6. Peut-on considérer que les deux variables étudiées sont indépendantes ?

Exercice 2. Le tableau ci-dessous indique pour chaque couple de modalités de deux caractères non numériques, la fréquence d'apparition de ce couple dans une population donnée. On suppose que les modalités ont été codées : $E = \{5, 15, 25\}$ (pour la variable x) et $F = \{10, 20, 30\}$ (pour la variable y). Certaines de ces données ont été remplacées par les paramètres a, b et c.

x / y	10	20	30
5	0.09	a	0.06
15	0.15	0.25	b
25	c	0.10	0.04

1. Déterminer les réels a, b et c sachant que les deux variables x et y sont indépendantes.
2. Donner la distribution conditionnelle de y sachant $x = 5$.

Exercice 3. On donne pour les six premiers mois de l'année 1982 les nombres d'offres d'emploi (concernant les emplois durables et à plein temps) et le nombre des demandes d'emploi (déposées par des personnes sans emploi, immédiatement disponibles et à la recherche d'un emploi durable et à plein temps). Les données sont exprimées en milliers d'individus.

Offres (x)	61	66.7	75.8	78.6	82.8	87.2
Demandes (y)	2034	2003.8	1964.5	1928.2	1885.3	1867.1

1. Représenter le nuage de points. Le nuage de points vous semble-t-il aligné le long d'une droite ?
2. Trouver la droite de régression des demandes d'emploi en fonction des offres d'emploi et la tracer sur le graphique précédent.
3. Calculer le coefficient de corrélation entre x et y. Commenter.

Exercice 4. On donne pour les années 1975 à 1983, le prix du ticket de métro à Paris acheté à l'unité en août (x) et le prix moyen annuel du kilogramme de bananes dans la région parisienne (y).

Prix ticket (x)	2.2.	2.5	2.7	3	3.6	4.5	5	5.5	6
Prix Kg de bananes (y)	3.72	3.95	4.27	4.52	5.01	5.41	6.17	7.03	8.42

Répondre aux mêmes questions qu'à l'exercice précédent. Commenter.

Exercice 5. Le tableau ci-dessous donne les valeurs expérimentales du volume V (en cm^3) et de la pression P (en Kg par cm^3) d'un gaz. D'après les lois de la thermodynamique de Laplace pour un gaz parfait, on a la relation $PV^\gamma = C$ où γ et C sont des constantes.

Volume (v)	620	890	1013	1186	1454	1944	2313	3179
Pression (p)	6.7	4.3	3.48	2.644	1.997	1.35	1.1	0.71

1. Transformer le modèle pour obtenir un modèle linéaire.
2. Déterminer la droite de régression linéaire pour le modèle transformé. En déduire une estimation $\hat{\gamma}$ et $\hat{C}$ des constantes γ et C.
3. Estimer P lorsque le volume $v = 1000$ cm^3.

Exercice 6. Le taux d'équipement des ménages en matériel informatique est une variable y_t où t représente l'année de l'observation. On fait l'hypothèse d'un modèle logistique :

$$y_t = \frac{1}{1 + ae^{-bt}} ,$$

avec a une constante positive.

1. Identifier les individus de cette étude.
2. Par un changement de variable approprié, montrer que le modèle logistique peut être transformé en un modèle linéaire que l'on précisera.
3. Appliquer la méthode des moindres carrés sur le modèle transformé en utilisant les données ci-dessous :

Années	1988	1989	1990	1991	1992	1993	1994
t	1	2	3	4	5	6	7
y	0.45	0.57	0.69	0.78	0.86	0.91	0.93

4. Tracer la droite de régression du modèle transformé ainsi que le nuage de points correspondant.

5. En déduire une estimation des paramètres a et b. Sur un autre graphique, tracer la courbe estimée et le nuage de points des données initiales. Le modèle logistique vous semble-t-il bien spécifié ?

6. Prévoir le taux d'équipement en 1996. En quelle année le taux d'équipement sera-t-il de 99% ?

Exercice 7. (Première session - Mai 2019) Corrélation et moindres carrés.

Les données suivantes ont étées obtenues en étudiant l'évolution de la population d'une bactérie "Pseudomonas" en fonction du temps. L'unité permettant de dénombrer les bactéries vivantes est l'Unité Formant Colonie (CFU) par millilitre (ml).

Heure h_i	6	7	8	9	9.5	10
CFU/ml y_i	1.0×10^5	8.5×10^5	6.2×10^6	8.0×10^7	2.5×10^8	8.8×10^8

1. Tracer le nuage de points.

2. Á la vue du nuage de points, semble-t-il y avoir une corrélation entre ces deux variables ?

3. Calculer le coefficient de corrélation linéaire de Pearson.

4. Calculer le coefficient de corrélation non linéaire de Spearman. Conclure.

   On souhaite maintenant tester un modèle non linéaire selon lequel la le nombre de bactérie par millilitre évolue de manière exponentielle en fonction du nombre d'heures : $y = c\exp(ah)$ où c et a sont des constantes.

5. Transformer le modèle pour obtenir un modèle linéaire.

6. Représenter le nuage de points transformés.

7. Déterminer la droite de régression linéaire pour le modèle linéaire. En déduire une estimation $\hat{c}$ et $\hat{a}$ des constantes c et a.

8. Estimer y pour une heure $h = 8.5$.

Exercice 8. (Deuxième session - Juin 2019) Corrélation et régression linéaire.

Sur 12 ouvriers d'une entreprise, on a observé en 2015 l'ancienneté (X en années) et le salaire mensuel (Y en Euros).

x_i	7	15	15	16	5	12	2	20	14	9	15	8
y_i	8100	10200	8400	11400	6900	9600	6300	10500	10800	8100	9300	7500

1. Construire le nuage de points.

2. Semble-t-il y avoir une corrélation entre ces deux variables ? Justifier.

3. Semble-t-il y avoir une corrélation linéaire entre ces deux variables ? Justifier

4. Commenter la dispersion des données pour X et pour Y.

5. Calculer le coefficient de corrélation linéaire. Commenter.
6. Réaliser la régression linéaire de Y sur X. Tracer la droite de régression sur le graphique de la question 1.
7. Déterminer les valeurs ajustées, c'est-à-dire les valeurs
$$\widehat{y}_i = \widehat{a}x_i + \widehat{b}; i = 1, \ldots, 12,$$
où $\widehat{a}$ et $\widehat{b}$ sont les coefficients de la droite de régression (respectivement la pente et la valeur à l'origine).
8. Déterminer les résidus $y_i - \widehat{y}_i$. Calculer la moyenne des résidus : que constate-t-on ?
9. Deux ouvriers de l'entreprise ont respectivement 4 et 18 ans d'ancienneté ; donner une estimation de leur salaire à l'aide de la droite de régression.

Exercice 9. (Première session - Mai 2018) Corrélation et régression linéaire.
On considère une automobile roulant sur une route sèche à une certaine vitesse v. Cette automobile freine brusquement pour simuler un freinage d'urgence. On note d la distance de freinage. Les données sont les suivantes :

Vitesse v (en km/h)	30	50	70	90	110	130
Distance d (en m)	7	16	31	52	78	123

1. Tracer le nuage de points.
2. À la vue du nuage de points, semble-t-il y avoir une corrélation entre ces deux variables ?
3. Calculer le coefficient de corrélation linéaire de Pearson.
4. Calculer le coefficient de corrélation non linéaire de Spearman. Conclure.

On souhaite maintenant tester un modèle non linéaire selon lequel la distance de freinage dépendrait du carré de la vitesse. On pose alors $t = v^2$.

5. Calculer le coefficient de corrélation linéaire entre d et t.
6. Effectuer la régression linéaire. Quelle est l'équation de la courbe obtenue.
7. Quelle serait une estimation de la distance de freinage pour une vitesse de 95 km/h ? 150 km/h ?

Exercice 10. (Seconde session - Juin 2018) Corrélation et régression.
La société Métalex moule des pièces dans un four. L'ingénieur se demande s'il existe un lien entre la température (en degré Celsius $°C$) à laquelle les pièces sont moulées et leur résistance (en kg/cm^2). Il dispose des données suivantes transmises par l'atelier :

Temp. t (en $°C$)	100	120	140	160	180	200	220	240	260	280	300
Rés. r (en kg/cm^2)	46	48	49	51	52	53	54	55	56	56	56

1. Tracer le nuage de points.

2. À la vue du nuage de points, semble-t-il y avoir une corrélation entre ces deux variables ?

3. Calculer le coefficient de corrélation linéaire de Pearson.

4. Calculer le coefficient de corrélation non linéaire de Spearman. Conclure.

   On souhaite maintenant tester un modèle non linéaire selon lequel la résistance dépendrait du logarithme de la température. On pose alors $v = \ln(t)$.

5. Calculer le coefficient de corrélation linéaire entre v et r.

6. Effectuer la régression linéaire. Quelle est l'équation de la courbe obtenue.

7. Quelle serait une estimation de la résistance pour une température de $210°C$? $350°C$?