BUT Science des Données (SD)

R2-05 (STAT) – STATISTIQUE DESCRIPTIVE 2

- Statistique bivariée – Ajustement -

Auteur(s) : L. Laval

DEVOIR MAISON

Préambule

— Pour vous éviter l'usage d'un éditeur d'équations et la perte de temps que cela entraîne, les calculs pourront être présentés - de manière bien lisible - sur une feuille de papier qui sera photographiée ou scannée pour être incorporée dans votre rapport.
— Les calculs devront être programmés de préférence en langage **R**. L'utilisation du langage Python est néanmoins autorisée.
— Le rapport devra impérativement être au format PDF et déposé dans Moodle avec le(s) programme(s) **R** ou Python.

Un épidémiologiste réalise une étude statistique sur l'efficacité d'un sérum auprès de N individus infectés par la bactérie Zomb-2025 (bactérie bien connue car elle transforme les êtres humains en zombies). Chaque individu reçoit ainsi une dose x_i (en mL) de sérum et l'on mesure la quantité de bactéries y_i (en Unité Formant Colonie$/mL$) encore présente dans le corps 24 heures après l'injection. Le tableau ci-dessous résume les dix premiers résultats expériences

Dosage X (en mL)	0.00	0.20	0.41	0.61	0.82	1.02	1.22	1.43	1.63
Quantité de bactéries Y (en UFC/mL)	1.02	0.96	0.96	0.97	0.84	0.81	0.87	0.80	0.71

Nota : L'ensemble complet de données est fourni dans le fichier **donnees_r.csv** disponible dans Moodle.

Exercice n°1

L'objectif de cet exercice est d'analyser l'éventuelle corrélation entre le dosage de sérum et la quantité de bactéries Zomb-2024 mesurée après injection ; la finalité étant, si cette corrélation est avérée, de procéder à un ajustement (linéaire si pertinent) afin de disposer d'un modèle mathématique permettant d'effectuer des prédictions.

Travail demandé

1. Préciser, parmi X et Y, qu'elle est la variable explicative et la variable à expliquer ;
2. Représenter le nuage de points associé à la série statistique bivariée des couples $(x_i; y_i)$;
3. Effectuer une description de la variable statistique relative à la quantité de bactéries. Pour cela :
   (a) Calculer sa moyenne arithmétique et sa variance ;
   (b) Représenter la Box-Plot (boîte à moustaches) ;
   (c) Analyser les résultats obtenus.
4. Calculer et analyser la covariance de X et Y ;

Dans un premier temps, on s'attachera à étudier l'existence d'une dépendance linéaire entre les deux variables statistiques X et Y.

5. En perspective de l'utilisation du coefficient de corrélation de Pearson, étudier l'hypothèse relative la normalité de la distribution associée à chacune des deux variables X et Y (autrement dit, nous cherchons à vérifier si chaque variable suit une distribution normale). Pour cela, vous devez procéder à :
   (a) Une étude graphique à partir d'un Quantile-Quantile Plot (Q-Q Plot) ;
   (b) Une étude numérique basée sur le test de Shapiro-Wilk.

6. Quel que soit les résultats obtenus à la question précédente sur la nature des distributions, calculer et analyser le coefficient de corrélation de Pearson de X et Y.

   Nota : En langage **R**, il est demandé d'utiliser la fonction **cor.test()** et non pas simplement **cor()**, en expliquant/commentant les informations obtenues.

En perspective d'opérer des prédictions sur le niveau d'infection résiduel en fonction de la quantité de sérum injectée, on demande :

7. Déterminer, par la méthode des Moindres Carrés Ordinaires, un ajustement linéaire simple (c.-à-d. un ajustement matérialisé par une droite affine de la forme $y = ax + b$) et afficher la droite de régression en superposition du nuage de points établi à la question 2 ;

   Remarque : Pour tracer la droite en langage **R**, vous pouvez utiliser **abline()** ou **ggplot()**.

8. A partir de la figure précédente (question 7) et sans aucun calcul complémentaire, discuter de la pertinence a priori de cet ajustement linéaire ;
9. Expliquer et analyser les résultats de l'ajustement obtenus via la commande **summary()**

   Exemple :

   LMS <- lm(Y $\sim$ X)
   summary(LMS)

10. Déterminer la SCR (Somme des Carrés Résiduels ou Somme des Carrés des Résidus) et expliquer ce qu'elle représente comme information ;

Rappelons qu'à la question 8, un jugement subjectif (car non fondé sur des éléments quantitatifs calculés) a été opéré pour qualifier a priori la qualité de l'ajustement. A présent, l'objectif est d'évaluer, sur des bases quantitatives et objectives, la qualité de la régression linéaire établie à la question 7. Pour cela, il convient d'analyser les résidus (ou erreurs résiduelles).

Pour rappel : Un ajustement linéaire est valide si les résidus sont :

— indépendants ;
— distribués selon une loi Normale de moyenne nulle ;
— distribués de façon homogène (c.-à-d. avec une variance constante)

On vous demande ainsi :

11. Analyser les résidus, à savoir :
    (a) Évaluer l'indépendance des résidus ;
        i. A partir du tracé d'une estimation de l'autocorrélation des résidus (en exploitant, par exemple, la fonction **acf()** en langage **R**)
        ii. A partir du test de Dubin-Watson (en exploitant, par exemple, la fonction **dwtest()** du package **lmstest** en langage **R**).
    (b) Tester la normalité des résidus ;

        Nota : La normalité devra être étudiée de deux manières différentes, à partir :
        — du test de Shapiro-Wilk
        — d'un diagramme Quantile-Quantile (Q-Q plot)

    Remarque : Il n'est pas demandé ici de tester l'homogénéité.

12. Déterminer, par la méthode des Moindres Carrés Ordinaires, un ajustement exponentiel de la forme $y = ba^x$ et afficher la courbe de régression en superposition de la figure établie à la question 7 (contenant le nuage de points et la droite de régression de l'ajustement linéaire).

    Nota : Dans votre rapport, Vous devrez impérativement décrire les calculs mis en oeuvre pour déterminer les coefficients a et b de l'ajustement exponentiel. Une fois encore rappelons que les calculs pourront être présentés sur une feuille de papier qui sera photographiée ou scannée pour être incorporée au rapport.

13. Calculer, pour chacun des deux modèles d'ajustement (c.-à-d. le modèle linéaire $y = ax + b$ et le modèle exponentiel $y = ba^x$), le coefficient de détermination R^2, puis effectuer une analyse comparative des résultats obtenus.
14. Calculer, à partir de chacun des deux modèles, une prévision de la quantité de bactérie encore présente dans le corps 24 heures après l'injection d'une dose de 8 mL de sérum. Analyser les résultats et conclure sur la pertinence des deux modèles.

Bonus : Dire au bout de combien de temps Rick de la série "The Walking dead" va intervenir pour faire décroître la densité de zombies.