CORRÉLATIONS & REGRESSIONS LINÉAIRES
REVISION - MODELE DE SUJETS - ÉPREUVE PRATIQUE

Mihaela Iancu, Daniel Leucuța

Objectifs éducationnels

Après avoir terminé ce travail pratique, vous devriez être capables de :

- Représenter graphiquement la relation linéaire entre DEUX VARIABLES QUANTITATIVES
- Quantifier/calculer l'intensité de la corrélation linaire par le coefficient de corrélation de Pearson en utilisant la fonction CORREL
- Représenter la droite de régression linaire et déterminer la forme de dépendance linéaire
- Interpréter la droite de régression (la pente de la régression) et le coefficient de détermination

Scenario

Une étude a été réalisée pour évaluer :

i) La **relation linéaire significative** entre la circonférence abdominale et l'épaisseur intima-média de la carotide EIM (mm) **chez les sujets adultes souffrant d'hypertension artérielle.**

ii) **La dépendance linéaire** entre la circonférence abdominale et l'épaisseur intima-média de la carotide EIM (mm) **chez les sujets adultes souffrant d'hypertension artérielle.**

L'échantillon de l'étude était composé de 100 patients hypertendus sélectionnés au hasard, âgés entre 30 et 80 ans, qui sont traitées au département de médecine interne d'une clinique universitaire de Cluj Napoca (période du temps : décembre 2016 à décembre 2018). Les données ont été collectées et entrées dans le fichier **BD_TP13MGFR.**

Demandes

1. Créez un nouveau dossier **TP13NP ou N = votre nom, P = votre prénom** sur le bureau Bureau (Desktop) de Windows.
2. Enregistrez le fichier **BDTP13_MGFR** dans le dossier **TP13NP**.

Pour tester les objectifs de l'étude, nous suivrons les étapes suivantes :

REALISER UN BON GRAPHIQUE POUR EXAMINER LA RELATION LINAIRE

3. Dans la page nommée **Corrélation**, copier les deux variables spécifiées dans le Scénario.
4. Faites un diagramme (graphique) par nuage de points pour montrer le lien entre la Circonférence abdominale (cm) et la variable EIM. **(voir Conseils, pages 1-2)**

ETABLIR SI LES DEUX VARIABLES QUANTITATIVES SUIVENT LA LOI NORMALE

5. Dans la même page, en utilisant les statistiques descriptives que vous les calculerez via l'option l'option *DATA ANALYSIS*, déterminer si les variables d'intérêt **Circonférence abdominale** et **EIM** suivent la distribution normale de probabilité (lois Gaussienne). **(voir Conseils, page 3)**
6. Surligner (en rouge) les statistiques descriptives nécessaires à l'établissement de la loi Normale. **(voir Conseils, page 3)**

DETERMINATION DU COEFFICIENT DE CORRELATION (de PEARSON) DANS l'EXCEL

7. Dans la page nommée Corrélation, calculer le coefficient de corrélation de Pearson entre les deux variables à l'aide de la fonction **CORREL** **(voir Conseils, page 4)**

DETERMINATION DE LA SIGNIFICATIVITE DU COEFFICIENT DE CORRELATION

8. Dans la page nommée Corrélation, remplir le **Tableau 1**. **(voir Conseils, page 5)**
9. Interpréter la significativité du coefficient de corrélation en remplissant le **Tableau 2**. **(voir Conseils, page 6)**

ETABLIR LA FORME DE DEPENDENCE LINAIRE A L'AIDE DE LA DROITE DE REGRESSION

10. Copier le graphique (Scatter) fait au point 4 dans la nouvelle feuille nommée **Régression**
11. Ajoutez sur le graphique la droite de régression, l'équation de régression et le coefficient de détermination. **(voir Conseils, page 7)**
12. Au-dessous du graphique, écrire les interprétations pour:
    - le nouage de points **(voir Conseils, page 8)**
    - la pente de la droite de régression **(voir Conseils, page 8)**
    - le coefficient de détermination **(voir Conseils, page 8)**

Rappelez-vous que....

1. Etablir l'existence de la distribution de probabilité normale pour une variable quantitative peut se faire par plusieurs méthodes, parmi lesquelles on citera : 1) le calcul de statistiques descriptives (moyenne, médiane, mode, coefficient d'asymétrie, coefficient d'asymétrie) et 2) les méthodes graphiques : histogramme ou graphique en boîte à moustaches
2. Le bon graphique pour mettre en évidence le lien possible entre DEUX VARIABLES QUANTITATIVES est le graphique SCATTER (nuage de points).
3. La quantification de l'intensité (magnitude) du lien/corrélation linéaire entre DEUX VARIABLES QUANTITATIVES avec une distribution normale de probabilité sera obtenue par le coefficient de corrélation de Pearson
4. Le calcul du coefficient de corrélation linéaire (Pearson) sera effectué dans Excel à l'aide de la fonction CORREL
5. L'interprétation de la valeur du coefficient de corrélation de Pearson se fera à l'aide des règles empiriques de Colton
6. La dépendance linéaire entre deux variables QUANTITATIVES peut être montrer par la droite de régression et le coefficient de détermination

INFORMATIONS GENERALES SUR L'EPREUVE PRATIQUE:

Les absences des activités pratiques : devraient être motivées et récupérées jusqu'au jeudi 12.01.2023.

En général, **les sujets éliminatoires** pour l'épreuve pratique peuvent être de la forme suivante:

- ✓ Comment enregistrer, où renommer un fichier avec un nom identique à celui de la demande ;
- ✓ Application des formules données par utilisateur
- ✓ Création d'un graphique approprié pour les données
- ✓ Calcul des statistiques descriptives
- ✓ Interprétation des résultats d'un test statistique

En général, l'épreuve pratique peut avoir 2 questions (QI, QII) ayant le suivant contenu possible :

Question 1 (QI) :

Faire un dossier ayant un nom prédéfini et enregistrer tout le fichier dans une place précisée

Question 2 (QII) : analyse statistique sur Microsoft Excel:

- Application d'une formule donnée
- Application de la fonction IF
- Trier les données en fonction d'un critère donnée
- Calcul des statistiques descriptives comme (la moyenne arithmétique, médiane, mode, amplitude, écart type, erreur standard, coefficient de variation, minimum, 1er quartile, 2ème quartile (médiane), 3ème quartile, Maximum, skewness (asymétrie), excès de kurtosis (coef. d'aplatissement)
- Création d'une table de fréquence
- Création d'un Tableau de contingence
- Graphique en secteurs ou « diagramme en camembert »
- Graphique par colonnes ou barres
- Graphique box-plot
- Histogramme
- Diagramme par nuage de points (Scatter) avec la droite de régression
- Calcul des différentes types de probabilités
- Calcul de l'intervalle de confiance d'une moyenne sur Excel (avec Data analysis)
- Calcul de l'intervalle de confiance d'une fréquence par sa formule
- Interprétation d'un intervalle de confiance
- Faire/realiser un test statistique (les tests t de Student pour des échantillons independantes avec des variances égales ou inégales, le test de Student pour des échantillons apparies, test F de Fisher, le test de Khi deux)
- Interprétation des tests statistiques
- Calcul et l'interprétation du coefficient de corrélation de Pearson
- Calcul et l'interprétation du coefficient de détermination.

MODELE DE SUJETS - ÉPREUVE PRATIQUE- S1

Temps de travail (y compris la lecture) : 25 minutes

Les exigences écrites en gras sont des ELIMINATOIRES !

QI. **(0.25 pt) Créez sur le bureau un dossier nommé ExamenNP (ou N & P sont votre nom et prénom).**

QII. On a fait une étude pour évaluer les relations entre les différentes caractéristiques cliniques et le stade de la maladie sur un échantillon de patients avec l'insuffisance rénale chronique. Les informations ont été recueillies et présentées dans le fichier BD_Sim.xlsx.

1. **(0.25 pt) Enregistrez le fichier BD_Sim.xlsx et renommez-le ExcelNP.xlsx dans le dossier ExamenNP.**
2. (2 pt) Définir une nouvelle variable nommée Clearance de la créatinine finale (CC) et calculer ses valeurs pour chaque patient en utilisant la formule suivante :

   CC = 1.25 × Poids (en kg) × (140 − Age en année) / [72 × Créatinine finale (mg/dl)]

3. (1 pt) **Déterminer la médiane et la déviation standard de la Créatinine chez les hommes.**
4. (3 pt) **Réalisez le graphique approprié pour la répartition des patients en fonction du Stade de la maladie et Genre.**
5. (3,5 pt) Au seuil de signification 5%, on peut affirmer qu'il y a une différence statistiquement significative entre les moyennes d'urée sanguine chez les patients ayant un stade avancé de l'insuffisance rénale chronique et ceux qui souffrent d'un stade précoce ?? On admet que la variable Urée sanguine est normalement distribuée.
   a. (1,5 pt) pour répondre à la question réaliser un test statistique approprié sur Excel
   b. **(2 pt) dans la même feuille de calcul, interpréter le résultat du test en écrivant :**
      -formulation des hypothèses (H0 et H1)
      - interprétation de la p-valeur

MODELE DE SUJETS - ÉPREUVE PRATIQUE- S2

Temps de travail (y compris la lecture) : 25 minutes

Les exigences écrites en gras sont des ELIMINATOIRES !

QI. **(0.25 pt) Créez sur le bureau un dossier nommé ExamenNP (ou N & P sont votre nom et prénom).**

QII. On a fait une étude pour évaluer les relations entre les différentes caractéristiques cliniques et le stade de la maladie sur un échantillon de patients avec l'insuffisance rénale chronique. Les informations ont été recueillies et présentées dans le fichier BD_Sim.xlsx.

1. **(0.25 pt) Enregistrez le fichier BD_Sim.xlsx et renommez-le ExcelNP.xlsx dans le dossier ExamenNP.**
2. (2 pt) Définir une nouvelle variable dichotomique nommée Urémie. Utilisez la définition suivante à afficher pour chaque sujet la valeur de cette variable :

   Si Urée > 44 mg/dl affichage Oui, sinon afficher Non
3. (2 pt) **Déterminer le premier quartile et l'erreur standard de la Créatinine dans l'échantillon d'étude.**
4. (2 pt) **Réalisez le graphique approprié pour la distribution de l'Age dans l'échantillon.**
5. (3,5 pt) Au seuil de signification 5%, on peut affirmer qu'il y a une association statistiquement significative entre le Genre et le Stade de l'insuffisance rénale chronique ??
   a. (1,5 pt) pour répondre à la question réaliser un test statistique approprié sur Excel
   b. (2 pt) dans la même feuille de calcul, interpréter le résultat du test en écrivant :
   - **Formulation des hypothèses (H0 et H1)**
   - **Interprétation de la p-valeur**

EXERCICES de REVISION : - facultatifs

1. Réaliser un graphique approprié pour la distribution du Genre.
2. Réaliser un graphique approprié pour la distribution de la Créatinine finale.
3. Réaliser un graphique pour la répartition des patients en fonction de la fatigue et le stade de l'insuffisance rénale.
4. On considère les événements : A= {être femme} et B= {avoir un stade précoce}. Déterminer la probabilité Pr(A/B).
5. Quelle est la fréquence relative de l'insuffisance rénale à un stade avancé chez les patientes ?
6. Calculer 2 mesures descriptives de dispersions de la variable Créatinine finale à l'aide des fonctions prédéfinies de l'Excel.
7. Déterminer s'il existe une différence statistiquement significative entre la créatinine initiale et finale (en utilisant un test statistique approprié). On sait que les valeurs de la créatinine sont normalement distribuées.
8. Déterminer, à l'aide d'un test statistique, s'il existe une association statistiquement significative entre le Stade de l'insuffisance rénale et la Fatigue. Réaliser le test approprié (calculer sa p-valeur).
9. Y a-t-il une corrélation linéaire entre l'âge et l'urée sanguine dans l'échantillon d'étude ?
10. Soit m=7,12 mg/dl la moyenne de la créatinine initiale et S=2,83 mg/dl l'écart-type. Déterminer l'intervalle de confiance à 95% : $[m - 1{,}96 \cdot \frac{S}{\sqrt{n}};\ m - 1{,}96 \cdot \frac{S}{\sqrt{n}}]$.