Probabilités discrètes et statistiques descriptives

L2 Université Lyon 1 2021-2022

TP noté numéro 86435 à rendre avant le vendredi 17 décembre 2021 à 23H59

Il faut charger sur Tomuss une feuille de calcul R et une présentation des réponses aux questions de cette feuille dans les deux colonnes Tomuss prévues à cet effet.

Instructions pour la synthèse de présentation des résultats du TP

Le projet est à rédiger avec Word ou LibreOffice (ou LaTeX si vous savez déjà l'utiliser). Mais vous devrez convertir votre fichier au format pdf. Attention ! Quand on exporte un fichier en pdf, il manque parfois les formules mathématiques ou des graphes. Il faut soigneusement vérifier le fichier pdf et si nécessaire, essayer d'autres convertisseurs pdf (il en existe de nombreux gratuits sur le web, comme cutepdf).

Le nom des fichiers devra contenir le nom de famille des deux membres du groupe (sans accents) sous la forme Fisher_Pearson.pdf et Fisher_Pearson.R.

Rédigez soigneusement. Commentez à chaque fois vos graphes et vos résultats.

La première page devra contenir : noms, prénoms, formation, année universitaire, nom de l'UE, date, numéro du sujet.

- Ajoutez des numéros de page, s'ils n'y sont pas.
- Vérifiez les graphiques : il faut qu'ils soient centrés, que leurs titres soient en français.
- Harmonisez la mise en forme des titres. Le style des paragraphes doit être justifié, pas aligné à gauche.
- Relisez pour l'orthographe (au moins deux fois), relisez pour la ponctuation.
- Écrivez le code R dans un fichier à part (colonne à part sur Tomuss). Si vous copiez du code R dans votre présentation, utilisez une police adéquate (avec un interlettrage fixe, comme `Lucida Console`).

  Dans tous les cas, nettoyez le code des lignes inutiles, commentez-le de manière raisonnable.

• **Exercice 1. Statistiques**

Charger le jeu de données :

```
Air<-read.delim2("http://tinyurl.com/y39an7ef/DATA86435.csv",na.strings="-")
```

Ce jeu de données contient des informations issues de relevés de la pollution dans différentes stations proches de Lyon. Il contient les relevés horaires sur le mois de décembre 2018 de trois types de polluants principaux mesurés en microgrammes par mètre cube d'air :

a. Le dioxyde d'azote est relevé aux stations suivantes : sur l'A7 au sud de Lyon, sur la place Grand-Clément à Villeurbanne, à l'aéroport Saint-Exupéry, à la station Lyon-centre dans le 3^{e} arrondissement, à la sortie du tunnel de la Croix-Rousse, sur le périphérique de Lyon et sur la place Jean Jaurès dans Lyon 7^{e}.

b. Les particules de taille 10 micromètres (PM10) sont relevées aux mêmes stations.

c. L'ozone est seulement relevé à Lyon-centre et à l'aéroport Saint-Exupéry.

On parlera de station d'observation pour désigner un des lieux où une mesure de pollution a été relevée (Villeurbanne, la Croix-Rousse, l'aéroport Saint-Exupéry, Lyon périphérique, Lyon A7, Place Jaurès ou Lyon-centre). On parle de mesure (ou relevé) de pollution pour parler du chiffre relevé pour un polluant donné dans une station d'observation donnée. Certaines mesures ont en commun une station d'observation et d'autres mesures ont en commun le type de polluant observé.

Air est un **data.frame** dont les individus sont les jours d'observations et les caractères sont les 16 mesures de pollutions.

Les valeurs indiquées "-" dans le fichier ou NA par R sont les valeurs non disponibles (correspondant à une heure de non-observation du polluant donné à la station donnée).

On donnera les réponses numériques arrondies avec TROIS décimales.

A. Informations générales

1. Quel est le type de variable statistique de chacune des variables (nominale, ordinale, quantitative discrète ou continue) ?
2. Quel est le nombre d'heures d'observation de l'échantillon ? Quel est le nombre d'heures où les particules ont été mesurées dans toutes les stations ? (Indication : vous pouvez utiliser la fonction **is.na**).
3. Dans cette question, on va extraire des informations sur le nombre de jours d'observations de certains polluants en créant deux nouvelles variables booléennes PM10Obs et AzoteObs. On rappelle que ces variables vont associer à chaque heure d'observation un booléen (en formant un vecteur de booléens). Créer la variable booléenne PM10Obs ayant, une heure donnée, pour valeur « VRAI » si les particules PM10 ont été observées dans toutes les stations d'observations. De même, créer la variable booléenne AzoteObs ayant, une heure donnée, pour valeur « VRAI » si le dioxyde d'azote a été observé dans toutes les stations d'observations. Donnez la table de contingence de ces deux variables nominales. Quel est le nombre d'heures où l'azote et les particules ont été observés dans toutes les stations ?
4. En ignorant les heures de non-observation de l'ozone à Lyon-centre (on pourra utiliser la fonction **na.omit**), trouvez la moyenne empirique, variance empirique, variance empirique non-biaisée et le quartile à 25 % de la variable ozone mesurée à Lyon-centre. Est-ce que l'émission d'Ozone a dépassé le seuil d'information de $180\mu g/m^3$ d'ozone dans l'air ?
5. Dans la suite on va s'intéresser aux moyennes sur des périodes de 6H (mesures de la nuit : 0H à 6H, du matin : 7h à 11H, de l'après-midi : 12H à 17H, du soir : 18H à 23H) pour

limiter l'impact des heures de non observation et des erreurs de mesures. Pour chaque mesure, calculer le nombre d'heures de non-observation dans chacune des tranches de 6 heures. Trouver les 2 mesures qui ont encore des tranches de 6H (regroupées comme ci-dessus) sans aucune observation (on les exclura par la suite du data.frame dft).

Créer un data.frame dft avec, pour individus les tranches de 6H d'observations et comme variables, pour chaque mesure ayant des observations pour toutes les tranches, les moyennes des mesures de pollution que vous venez de calculer.

Dans la suite, on travaille sur les données du data.frame dft, créé à la dernière question.

B. Corrélation et Régression Linéaire

1. Calculez les covariances non-biaisées et les corrélations des 10 relevés de dft.
2. Soit y l'échantillon des mesures de moyenne par tranches de 6H du Dioxyde d'Azote à la Croix-Rousse et x l'échantillon des mesures de moyenne par tranches de 6H du Dioxyde d'Azote sur l'A7. On cherche à savoir si elles sont corrélées. Quelle hypothèse doit-on faire pour effectuer un test de Pearson. Effectuez un test de corrélation. Que concluez-vous ?
3. Sous la même hypothèse, trouvez (avec la commande **lm**) la droite de régression de y en fonction de x. Discutez la significativité statistique du résultat (commentez les résultats de la commande **summary**). Tracer le nuage de points correspondant et la droite de régression en couleur.
4. Si un jour donné de décembre 2018, on avait mesuré la valeur 200 mg/m^3 de Dioxyde d'Azote (la valeur du seuil d'information pour un pic de pollution pour ce polluant) sur l'A7, donner un intervalle de prédiction de niveau de confiance 95% pour la mesure de Dioxyde d'Azote à la sortie du tunnel de la Croix-Rousse.
5. A l'aide de la fonction gofTest, faites un test du χ^2 de normalité pour les échantillons y et x Qu'en concluez-vous ? Même question avec les variables z donnant la moyenne par tranches de 6H des mesures de Particules PM10 sur l'A7, et son logarithme népérien $z' = \ln(z)$ (obtenu dans R par la fonction **log**).

Dans la suite de cette partie, à la place de la description linéaire $x = az + b$, on va trouver une meilleure description sous la forme $x = cz' + d$, ou, encore $x = c\ln(z) + d$.

6. Trouvez (avec la commande **lm**) la droite de régression de x en fonction de z'. Discutez la significativité statistique du résultat (en commentant les résultats de la commande **summary**). Tracer le nuage de points correspondant et la droite de régression en couleur, puis deux droites donnant les bornes de l'intervalle de prédiction au niveau 80% (d'une autre couleur).
7. Tracer les nuages de points des mesures de Dioxyde d'Azote et de Particules PM10 sur l'A7 avec la courbe de régression (en supposant qu'il suffit de prendre l'image par un logarithme de la régression linéaire de x, z', ce qui n'est qu'une première approximation) et les bornes des intervalles de prédiction au niveau 80% pour ces mesures. Calculer les intervalles de prédiction (au niveau 95%) pour la mesure de Dioxyde d'Azote si la mesure en Particules est au niveau de l'un des deux seuils d'alerte réglementaires d=50, ou d=80. Est-ce que, selon vous, la mesure des Particules peut servir de référence pour la mesure de pollution au Dioxyde d'Azote ?

• **Exercice 2.**

On rappelle que la fonction **runif** permet de simuler une loi uniforme continue en obtenant un vecteur dont les nombres sont uniformément répartis dans l'intervalle spécifié.

1. Simulez dans un vecteur U de taille N, un grand entier, N réalisations de la loi uniforme sur $[-5, 2]$. Construisez un vecteur V dont l'élément V_i est le i-ème terme positif de U. Donnez les commandes qui vous permettent d'obtenir V.

   En comparant l'histogramme associé à V avec la densité d'une loi uniforme bien choisie, faites une hypothèse sur la loi dont V est un échantillon. Quel N vous a permis de conclure ? (Cette méthode de simulation d'une variable uniforme est dite simulation par rejet.)

2. Tester votre hypothèse en utilisant un test du χ^2 d'adéquation à une loi continue. Si vous répétez la simulation aléatoire 100 fois (avec une boucle), compter combien de fois vous conservez l'hypothèse nulle (au niveau de risque 5%) ? Qu'en concluez-vous sur votre hypothèse concernant la simulation ?

3. On donne les commandes suivantes :

   ```
   U <- runif(10000,min=-1,max=1); V <- runif(10000,min=0,max=1)
   S <- (1>abs(U)+V)
   U0 <- U[S]; V0 <- V[S];
   ```

   Tracez un nuage de points avec les points de l'échantillon (U, V) (vous pouvez utiliser l'option « pch='.' » pour améliorer la lisibilité). Trouvez sur quel ensemble de $\mathbb{R}^2$ l'échantillon du couple (U, V) est uniformément distribué ? (Vous pouvez si nécessaire simuler d'autres échantillons de tailles différentes.)

4. Tracez un nuage de points avec les points de l'échantillon $(U0, V0)$. Trouvez sur quel ensemble de $\mathbb{R}^2$ l'échantillon du couple $(U0, V0)$ est uniformément distribué ? (Vous pouvez si nécessaire simuler d'autres échantillons de tailles différentes et indiquer quelle taille N vous permet de conclure.)

5. En comparant un histogramme à une densité bien choisie, (on pourra utiliser la fonction **lines** pour tracer la courbe par dessus l'histogramme et utiliser l'option freq de la fonction traçant l'histogramme), trouvez graphiquement la densité de la loi dont $U0$ est un échantillon.

   En utilisant la notion de simulation par rejet, expliquez ce que fait le programme ci-dessus pour obtenir $U0$.