université PARIS-SACLAY

Travaux pratiques en langage R
M1 MIAGE FI

guillaume.postic@univ-evry.fr

Le langage de programmation R a été développé pour les calculs statistiques et l'analyse de données. Ces travaux pratiques ont pour objectif de vous familiariser à son utilisation. Il s'agira donc d'écrire des scripts en R, qui permettront de répondre à une série de questions. Celles-ci portent sur des données présentes dans le répertoire `TP/` localisé sur le dépôt https://github.com/guipostic/L2_databases.

Afin de réaliser ce TP, il est nécessaire d'avoir accès à un interpréteur R pour pouvoir exécuter les différents scripts. Nous utiliserons, pour cela, l'environnement de développement intégré (IDE) `RStudio`. Celui-ci est accessible en ligne (par le biais de Binder) à l'adresse suivante : http://mybinder.org/v2/gh/binder-examples/r/master?urlpath=rstudio.

Alternativement, vous pouvez bien sûr utiliser une installation locale de `RStudio` (ou de n'importe quel autre IDE) sur votre machine personnelle. Il est également possible de réaliser le TP avec un *notebook* Google Colab : https://colab.to/r.

Vos réponses aux différentes questions devront être rendues sous la forme d'un script, c'est-à-dire un fichier au format texte portant l'extension `.r`. Plusieurs scripts pourront être rendus si vous jugez cela utile. Ce format implique que la partie « rédactionnelle » des questions devra être rédigée sous la forme de commentaires de code (en langage R, les commentaires commencent par un `#`).

Problème 1

Le fichier `psychology.csv` contient les résultats d'une étude en psychologie du travail visant à établir un lien entre productivité et bien-être au travail. Pour cette étude, les chercheurs ont conçu deux mesures :
- un « indice de bien-être » (IBE), dont les valeurs sont des entiers compris entre 0 et 100,
- un « indice de productivité » (IP), dont les valeurs sont des nombres avec trois chiffres après la virgule et compris entre 0 et 10.

Les valeurs ont été obtenues auprès de 150 salariés, tous issus de secteurs d'activité professionnelle différents.

Question 1

Calculez des statistiques descriptives des distributions de l'IBE et de l'IP : moyenne, médiane, écart-type, IQR.

Question 2

Affichez une estimation de densité par histogramme pour chacune des deux variables : IBE et IP.

Question 3

Affichez les mêmes histogrammes que ceux de la question précédente, mais avec des largeurs d'intervalles divisées par deux.

Question 4

Affichez une estimation par noyau de la densité (en anglais : *kernel density estimation*) pour chacune des deux variables : IBE et IP.

Question 5

Refaites la Question 4
- en utilisant un noyau triangulaire
- et en changeant la méthode du choix de paramètre de lissage (*bandwidth selector*) par une validation croisée biaisée

Question 6

Afficher le nuage de points entre les deux variables IBE et IBP. Au regard de ce graphique, comment décririez-vous la relation entre ces deux variables ?

Question 7

Trouvez une transformation des données permettant de linéariser la relation entre les deux variables et affichez le nuage de points.

Question 8

À partir de ces données transformées, calculez le coefficient de corrélation de Pearson entre les deux variables. Calculez également les coefficients de corrélation de Spearman et Kendall.

Question 9

En reprenant le graphique de la question précédente, ajoutez

- une ligne diagonale (pente = 1), de couleur rouge,
- les noms des abscisse et ordonnée (c'est-à-dire, IBE et IP).

Question 10

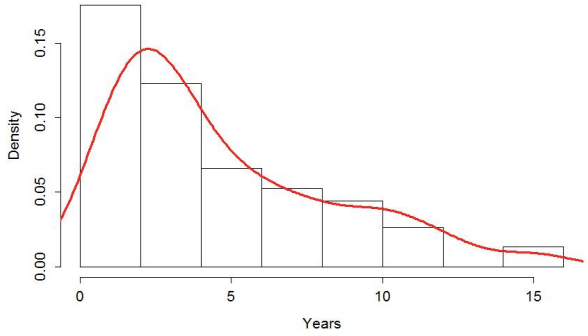
Par défaut, les graphiques sont enregistrés dans des fichiers au format PDF. Reprenez la réponse de la question précédente, en faisant en sorte de produire en sortie un fichier au format PNG.

Problème 2

Une entreprise a changé d'équipe dirigeante tous les 5 ans, à partir de 2007.
Les performances de 584 employés, présents de 2007 à 2022, ont été évaluées à l'issue des cinq années faisant suite à chaque changement de direction.
Les résultats ont été enregistrés dans les fichiers : `company2012.dat`, `company2017.dat` et `company2022.dat`.

Question 1

Affichez la distribution des performances pour les 3 années. Vous superposerez une estimation par noyau de densité au-dessus de chacun des 3 histogrammes, comme dans l'exemple ci-dessous.



Question 2

Visuellement, les performances semblent suivre une loi de probabilité normale. Effectuez un test de Shapiro-Wilk afin de vérifier cette hypothèse.

Question 3

Les actionnaires de l'entreprise se demandent si ces changements d'équipe dirigeante ont eu un effet significatif sur les performances des employés (en espérant que cet effet soit bon !). Répondez à leurs interrogations, en trouvant et calculant un test statistique approprié. Pour choisir le bon test, vous vous poserez (entre autres) la question de « l'appariement » des données.

Question 4

Trouvez et calculez un équivalent non-paramétrique du test de la question précédente.