Applications Shiny pour l'exploration visuelle interactive des données et l'enseignement des statistiques en SHS

Nicolas Audibert*

Résumé

La visualisation des données est incontournable en statistiques, toutefois la prise en main des fonctionnalités de R peut être complexe pour les non-spécialistes. Les trois applications Shiny présentées, dont le code est distribué sous licence GPL, permettent aux utilisateurs d'importer leurs données et offrent chacune un mode spécifique d'exploration des données via ggplot2. L'application iHist permet l'exploration interactive de distribution via un histogramme. Quant à elle, iScatter permet l'exploration interactive de nuages de points, avec la possibilité d'adapter l'affichage en fonction d'une ou deux variables qualitatives. Ces deux applications intègrent l'affichage d'informations complémentaires (éléments graphiques superposés, statistiques descriptives). Enfin iPlotDesigner, destinée à la comparaison de distributions entre modalités d'une à trois variables qualitatives, permet de naviguer entre représentations graphiques des mêmes données et d'ajuster de façon interactive le paramétrage de ces représentations, ainsi que de générer le code R commenté pour reproduire hors-ligne les mêmes figures. Outre l'objectif premier de didactique des statistiques appliquées à l'analyse de données linguistiques, ces applications peuvent apporter un gain de temps aux chercheurs plus expérimentés et être étendues à d'autres domaines scientifiques.

Mots-clefs (3 à 5) : Statistique – Data – Enseignement – Shiny

Développement

Bien que parfois négligée avec la généralisation de l'usage de modèles statistiques avancés, la visualisation des données constitue une étape indispensable de l'analyse quantitative, qu'il s'agisse d'exploration des distributions ou de présentation des résultats. Par manque de culture technique, certains acteurs des Sciences Humaines et Sociales et notamment les étudiants peuvent toutefois éprouver des difficultés à prendre en main les outils existants, parmi lesquels R occupe une place privilégiée. Ce constat, ainsi que celui d'une tendance croissante à n'interpréter que les prédictions des modèles statistiques de régression sans explorer au préalable la distribution des données brutes, a motivé le développement de trois applications Shiny dédiées à la visualisation interactive des données via les fonctions de ggplot2 et pouvant être utilisées dans le cadre de l'enseignement des statistiques descriptives et/ou inférentielles à un public de non-spécialistes. Ces trois applications open source partagent la fonctionnalité d'import des données dans différents formats courants, et en option l'application de filtres pour sélectionner un sous-ensemble. Elles sont localisées, permettant leur adaptation multilingue via la simple traduction d'un fichier de localisation.

* Laboratoire de Phonétique et Phonologie, UMR7018, CNRS, Univ. Sorbonne Nouvelle, Paris, France, nicolas.audibert@sorbonne-nouvelle.fr

La première, iHist (liens : version en français, GitHub) est destinée à l'exploration interactive de la distribution d'une variable quantitative via un histogramme paramétrable dans lequel les barres correspondant aux différentes plages de valeur peuvent être sélectionnées pour afficher le sous-ensemble de données correspondant. L'histogramme peut être complété de façon optionnelle par l'affichage graphique de quantiles et/ou des bornes de l'intervalle de confiance, ainsi que par des courbes représentant la densité de distribution et la distribution normale de même moyenne et écart-type à des fins de comparaison visuelle. La seconde, iScatter (liens : version en français, GitHub) qui reprend les mêmes principes généraux, est dédiée à l'exploration interactive de nuages de points. Elle permet d'intégrer de façon optionnelle l'adaptation des couleurs et/ou formes des points en fonction d'une ou deux variables qualitatives, et l'affichage des droites de régression.

Le mode d'exploration interactive des données intégré dans iHist et iScatter relève d'une approche différente de celle de Plotly (Plotly Tech. Inc., 2015), l'objectif étant ici de permettre la sélection graphique par l'utilisateur de groupes d'observations pour une interprétation en contexte. Ces deux applications proposent en outre l'affichage de statistiques descriptives (et des coefficients de corrélation dans le cas de iScatter), principalement à des fins didactiques. Elles ont fait l'objet d'une première publication (Audibert, 2024) dans un contexte d'exploration de valeurs aberrantes issues de l'analyse acoustique de données de parole.

La troisième application, iPlotDesigner (liens : version en français, GitHub) développée plus récemment, est consacrée à la comparaison de distributions entre modalités d'une à trois variables qualitatives. Elle permet de naviguer entre différentes représentations graphiques des mêmes données afin de mieux appréhender les avantages et limites de ces représentations. A des fins de publication ou de didactique de l'utilisation de ggplot2, elle génère le code R commenté permettant de reproduire hors-ligne les mêmes figures, le paramétrage de diverses options graphiques pouvant être ajusté de façon interactive.

Si les fonctionnalités proposées par iPlotDesigner se rapprochent d'une partie de celles proposées par l'outil libre jamovi, la prise en main de l'application iPlotDesigner par des utilisateurs non-spécialistes est plus immédiate et les options de paramétrage des figures son plus étendues. En outre, tandis que le code R équivalent généré par jamovi exploite des fonctions de haut niveau du package jmv avec des possibilités de paramétrage limitées, celui généré par iPlotDesigner s'appuie très majoritairement sur les fonctions de tidyverse. Il est de plus largement commenté dans une optique de didactique de l'utilisation de tidyverse et plus spécifiquement de ggplot2 et afin de faciliter d'éventuelles adaptations.

Le code source des trois applications est distribué sous licence libre GPL v3.

Références

N. Audibert: iHist et iScatter, outils en ligne d'exploration interactive de données : application aux valeurs aberrantes de f0 et de formants. *Actes des 35èmes Journées d'Études sur la Parole (JEP 2024)*, Toulouse, France, pp.598-607, 2024.

Plotly Technologies Inc: Collaborative data science. Montréal, QC. https://plot.ly. 2015

The jamovi project: jamovi (version 2.6.45), https://www.jamovi.org, 2026.