PairedData 0.9 : Un package R en S4 pour analyser les données numériques appariées

Stéphane Champely

PairedData 0.9 : Un package R en S4 pour analyser les données numériques appariées

S. Champely

Sciences et Techniques des Activités Physiques et Sportives
Centre de Recherche et d'Innovation sur le Sport
27-29, Boulevard du 11 Novembre 1918, 69622 Villerubanne cedex.
champely@univ-lyon1.fr

Mots clefs : Données appariées, Graphiques, Robustesse, Sport, S4.

Le dispositif apparié est l'un des plus utilisés en sciences du sport afin d'augmenter la puissance de comparaisons du type : avant *vs.* après entraînement, main droite *vs.* main gauche, temps réels *vs.* temps imaginés (*i.e.* un facteur intra à deux niveaux). Il est également classique de réaliser ces comparaisons sur plusieurs groupes (*i.e.* un facteur inter, le plus souvent à deux niveaux également : groupe traité *vs.* témoin).
L'analyse statistique employée est immanquablement un ou une combinaison de tests de Student appariés voire une analyse de variance à mesures répétées. Cependant, cette analyse élémentaire cache parfois diverses difficultés [8] : points extrêmes, différence de dispersion, multimodalité, hétéroscadasticité. Le problème principal est de repérer ces situations et, pour ce faire, la visualisation est un outil privilégié [9]. S'il existe dans la littérature plusieurs techniques graphiques dévolues aux données appariées ([10], [6], [3], [7]), elles sont rarement présentes dans les logiciels statistiques. Le package PairedData 0.5, en employant les outils du package ggplot2 [11], réunit ces propositions (cf. Figure 1) et les étend en autorisant la prise en compte de groupes.
Une fois les structures repérées graphiquement, le calcul de statistiques descriptives en permet une discussion plus précise. La tendance dans les publications scientifiques dans certaines disciplines (Médecine, Psychologie) des sciences du sport est de seconder les résumés habituels (moyenne, écart-type) par des tailles d'effet standardisées. Or, ces statistiques étant plus ou moins sensibles à la normalité des données, des versions robustes sont préférables. Les propositions d'Algina *et al.* [1] sont intégrées dans le package PairedData 0.5. Au delà de ces descriptions, les tests d'hypothèses permettant de comparer la centralité, mais aussi la dispersion, accompagnés par les intervalles de confiance associés, sont également peu disponibles. PairedData réunit les versions classiques (Student, Pitman, Morgan) et des alternatives robustes moins connues ([13], [12], [5], [2]).
L'un des objectifs du package est de servir d'outil pédagogique. Aussi, plusieurs jeux de données typiques des sciences du sport (biomécanique, psychologie, neurosciences) sont proposés, ainsi qu'un exemple illustrant les pièges du test de Student apparié (voir également Figure 1). Une version (RcmdrPlugin.PairedData), pour l'instant confidentielle, intégrée au package Rcmdr [4] permet aux utilisateurs moins avertis d'utiliser ces outils d'analyse de données appariées à l'aide d'une interface graphique en menus déroulants.
Enfin, une nouvelle version (0.9) en S4 du package PairedData est en construction. Basée sur la création d'un objet "paired", de type "dataframe", mais contraint à deux colonnes et possédant des "class" identiques (et même des "levels" identiques dans le cas de données catégorielles), cette version permet d'employer les génériques classiques (`show`, `summary`, `plot`) et de construire des méthodes adaptées à la fonction `t.test`, ainsi qu'à `var.test`.

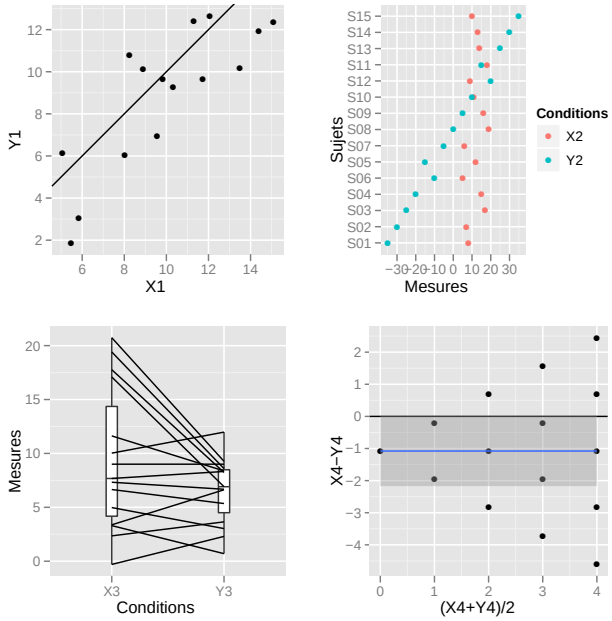

Figure 1: Divers graphiques adaptés au cas de données numériques appariées. Les données reflètent successivement diverses situations problématiques du test de Student apparié

Références

[1] Algina, J., Keselman, H.J., Penfield, R.D. (2005). Effect Sizes and their Intervals: the Two-Level Repeated Measures Case. *Educational and Psychological Measurement*, **65**, 241−258.

[2] Bonett, D.G., Seier, E. (2003). Statistical Inference for a Ratio of Dispersions using Paired Samples. *Journal of Educational and Behavioral Statistics*, **28**, 21−30.

[3] Cox, N. (2004). Speaking Stata: Graphing Agreemeent and Disagreement. *The Stata Journal*, **4**, 329−349.

[4] Fox, J. (2005). The R Commander: A Basic Statistics Graphical User Interface to R. *Journal of Statistical Software*, **14** (9), 1−42.

[5] Grambsch, P.M. (1994). Simple Robust Tests for Scale Differences in Paired Data. *Biometrika*, **81**, 359−372.

[6] McNeil, D.R. (1992). On Graphing Paired Data. *The American Statistician*, **46**, 307−310.

[7] Meek, D.M. (2007). Two Macros for Producing Graphs to Assess Agreement Between Two Variables. In Proceedings of Midwest SAS Users Group Annual Meeting.

[8] Preece, D.A. (1982). t is for Trouble (and Textbooks): a Critique of some Examples of the Paired-Samples t-Test. *The Statistician*, **31**, 169−195.

[9] Pruzek, R.M., Helmreich, J.E. (2009). Enhancing Dependent Sample Analyses with Graphics. *Journal of Statistics Education*, **17**.

[10] Rosenbaum, P.R. (1989). Exploratory Plot for Paired Data. *The American Statistician*, **43**, 108−110.

[11] Wickham, H. (2009). *ggplot2: elegant graphics for data analysis*. New York: Springer.

[12] Wilcox, R.R. (2005). *Introduction to Robust Estimation and Hypothesis Testing*. San Diego: Academic Press.

[13] Yuen, K.K. (1974). The Two-Sample Trimmed t for Unequal Population Variances. *Biometrika*, **61**, 165−170.