Interface Homme Machine : Export automatique des données

C Genolini[a]

[a]Inserm U1027
cgenolin@u-paris10.fr

Mots clefs : Statistique, Interface Homme Machine, GUI, export des données

1 L'export des données

1.1 Popeye

La première apparition publique de Popeye date du 19 décembre 1919. Marin borgne et tatoué, Popeye est régulièrement opposé à un adversaire patibulaire du nom de Brutus. Il doit ses nombreuses victoires (et sa popularité) à sa forte consommation d'épinards qui lui donne une force prodigieuse -sorte de potion magique à l'Américaine-.
Cette caractéristique des épinards serait liée à leur haute teneur en fer. Or, il est maintenant assez bien établi que les épinards ne sont pas spécialement riches en fer. La méprise vient d'une erreur de recopie : en 1870, le nutritionniste E. von Wolf mesura la teneur en fer des épinards. Il trouva 2,7 mg de fer pour 100g d'épinard, mais nota 27 mg pour 100 g... La légende des épinards-donneurs-de-force était née.

1.2 Retour en 2016

Au pays des statisticiens. De nos jours, la science a beaucoup progressé. Différents scientifiques ont établi des règles qui permettent d'éviter les biais, comme par exemple la démarche expérimentale hypothético-déductive de Claude Barnard. La liste des biais dont il faut se méfier lorsque l'on réalise une expérience n'en finit pas de s'allonger.
Du point de vue logiciel, différentes solutions proposent des aides au data management ou utilisent des solutions conservatives afin d'éviter au débutant de faire des erreurs (comme le `na.rm=FALSE` par défaut dans de très nombreuses fonction R). t.test).
Par contre, l'export des données reste le parent pauvre de l'analyse statistique.
En effet, a ce jour, il n'existe pas de méthode simple permettant d'exporter des données sans risque. Sous R, les résultats sont généralement livrés avec de nombreux commentaires. Ces commentaires, indispensables par ailleurs, empêche un export global des données. Par exemple, un t de Student est accompagné

```
        Welch Two Sample t-test

data:  x and y
t = 0.11067, df = 17.887, p-value = 0.9131
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.2830970  0.3145644
sample estimates:
mean of x mean of y
0.5237206 0.5079869
```

Les informations qui nous intéressent sont probablement le t, avec 2 ou 3 décimales, et le p. La solution la plus simple pour récupérer ces informations est simplement de faire un copier-coller ou une recopie manuelle des valeurs pertinentes. Dans le cadre d'analyses plus complexes, comme par exemple une régression linéaire, le nombre de valeurs pertinentes est plus important :

```
Call:
lm(formula = y ~ x1 + x2 + x3 + x4 + x5)

Residuals:
       1        2        3        4        5        6        7      ...
 0.38909 -0.10473 -0.03548  0.09858 -0.01355 -0.34607  0.17661      ...

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 14.19451   11.49901   1.234    0.285
x1          -1.83505    1.57979  -1.162    0.310
x2          -0.96425    1.18114  -0.816    0.460
x3          -0.01599    0.32191  -0.050    0.963
x4          -0.75019    0.60872  -1.232    0.285
x5           1.58241    1.19128   1.328    0.255

Residual standard error: 0.3005 on 4 degrees of freedom
Multiple R-squared:   0.44,     Adjusted R-squared:  -0.26
F-statistic: 0.6286 on 5 and 4 DF,  p-value: 0.6918
```

Certaines solutions existent, plus ou moins compliquées à mettre en œuvre. On peut par exemple extraire les coefficients, puis supprimer les colonnes qui ne nous intéressent pas et enfin exporter le tout dans un fichier. Mais plus les analyses se complexifient, plus ces solutions sont difficiles à mettre en œuvre. Au final, c'est le copier-coller ou la recopie manuelle qui sont le plus souvent utilisé, avec tous les risques qu'il comporte. En termes d'export des données, nous sommes encore très proches du XIX° siècle...

2 Solution alternative : export semi-automatique des données

2.1 R++ the Next Step

R++, the Next Step est un projet de développement d'une nouvelle implémentation de R. Il a pour vocation d'être compilable, d'intégrer en natif la gestion du parallélisme et de permettre l'exploitation des bases de données de grande dimension. Mais surtout, R++ est intégré dans une interface homme machine moderne et conviviale, spécifiquement conçue pour les analyses statistiques.

2.2 Interaction Homme Machine

L'Interaction Homme Machine est une science ayant pour objectif d'étudier la manière dont les humains interagissent avec les ordinateurs afin d'ensuite concevoir des outils plus ergonomiques. Pour cela, des séances de brainstorming réunissant utilisateurs (dans notre cas les statisticiens) et informaticiens sont organisées. Dans un premier temps, l'objectif est de définir les tâches qui sont particulièrement ardues, pénibles à réaliser ou a fort risque d'erreur. Ensuite, des solutions sont collectivement imaginées. Enfin, un prototype vidéo, illustration par l'exemple du problème et de sa solution, est élaboré.
Dans le cas présent, la "tâche ardue" identifiée a été l'export des données.

2.3 L'export semi-automatique

Dans cette présentation, nous vous proposons un nouvel axe de l'Interaction Homme Machine dédié spécifiquement à l'export des données. Nous présenterons le problème tel qu'il a été percu par les utilisateurs, puis différents prototypes vidéos avant de passer à la démonstration finale des différentes solutions. Afin que le copier-coller manuel des résultats disparaisse de nos habitudes...