Generating datasets, beta-error study

Emmanuelle Becker

Method

What's the problem ? Imagine you were teaching statistics to Bachelor students. Your aim is that they learn :

— the χ^2 test ;

— the tests used to compare two means (`t-test` or `Mann-Whitney-Wilcoxon` test).

You want them to practice these tests with `R` by giving them a small project to solve. You propose them a simulated dataset.

Dataset description The dataset studies the link between the use of IT devices by parents (parental use in hours per day of smartphone, television, computer) and the use of IT devices by young children under 3 y.o. (child use in minutes per day of smartphone, television, computer). You also register other values such as :

— the `Gender` of the parent (M,F) and the one of the child (M,F)

— the `age` of the parent and the age of the child (in years and months, respectively)

— the `body mass index` of the parent

— the size of the city the family lives within (small, medium, big, mega).

Question 1 List all the variables of the dataset. What variables are qualitative, which one are quantitative ?

Question 2 By using the `random` functions of `R`, such as `rnorm(), runif, sample ...`, write a script that generates a dataset that has the following properties.

— the number of couples (parent, child) studied is randomly comprised between 800 and 900 ;

— the values are realistic ;

— for parents, women have a greater use of smartphone that mens (very light effect)

— people living in big and mega cities have a smaller body mass index than people living in small and medium cities (very light effect) ;

— the older a child, the more he/she uses computer and smartphone (strong effect) ;

— the more the parent is watching television, the more the child is watching television (strong effect).

When you are satisfied of your dataset, save it in a file by using the `write.table()` function.

Question 3 Test that your dataset has the properties we asked for with appropriate tests (the linked variables are linked, the variables not linked are not linked). If it is not the case, explain why.

Question 4 Now use you script to generate 100 such datasets, and test the different properties inside these 100 datasets. Are the properties always respected ? If it is not the case, explain why.

Good luck !