Teaching Statistical Inference Using Many Samples from a Real Large Dataset

Jean-Hugues Chauchat
Université Lyon-2, Département Informatique et Statistique
5 avenue Pierre Mendès-France
69676, Lyon-Bron CEDEX, France
chauchat@univ-lyon2.fr

1.1 Before computer age in class room

Using large data bases allows us to renew statistical inference education, by asking the students to produce and analyse simulated samples resulting from real populations.

Formerly the statistics teacher used to ask the students to apply the theory on an experimental situation which resulted in a single sample of measurements. The student had to calculate a confidence interval, or to give the conclusion of a test, and then to interpret the result.

But, as each experimental situation (or each statement of exercise) produced only one sample, it was difficult for the student to understand the concepts of risk. And a great part of time was used for tiresome numerical calculations.

From the 80's, the use of the computers improved teaching of statistical methods. I will distinguish three stages :

1. numerical calculation,
2. random numerical simulation,
3. and, now, sample simulation from real populations.

1.2 Using computers for numerical calculation or digital simulation

At the beginning, the computer was only used to accelerate numerical calculations; that made it possible to get more time for modelling the problem and interpreting the results.

Now, digital simulations are used to illustrate the randomness of statistical conclusions

Multiple simulation softwares were built on the same model:

a) choice of a theoretical distribution for the random variable (or vector),
b) choice of the parameters (for exemple: means, variances, covariances, of a vector multinormal),
c) choice of a sample size
d) generation of several random samples
e) calculations on these samples: estimates, intervals, decisions of test.

(See, for exemples :

http://www.stanford.edu/~savage/software.htm http://www.ruf.rice.edu/~lane/stat_sim/index.html etc. le finlandais)

This type of simulations has many advantages:

- better understanding of the concept of risk in the statistical decision process: the confidence interval is random and varies from one sample to another; some time it does not contain the "true value" of the parameter; when the null assumption (Ho) is true, the procedure of test rejects sometimes this assumption, and, when (Ho) is a little bit false, the procedure very often accepts it;

- sensitivity of the results to the variations of the parameters: it is easy to compare the amplitudes of the confidence intervals, or the power of the tests, for various sizes of sample, various value of standard deviation, etc.

But it has some disadvantages:

- the random samples generation remains very abstract for the students; they do not see clearly the link between a long and expensive real experiment, and the samples automatically generated while pressing on a button;

- the distribution used to generate the samples are "perfect" (normal, multi-normals, exponential, etc.); that made confuse the ideal model (useful to understand a reality) and reality itself, which is never "normal" and even less "multi-normal"...

1.3 Teaching Statistical Inference Using Many Samples from a Real Large Dataset

The access to large real data bases allows a new use of simulation for the teaching of the statistics. A data base is selected as population of reference within which one can extract as many samples as necessary, according to given sampling process. Working on these sets of samples makes it possible to preserve a part of the advantages of the purely simulated data, while working on realistic samples.

For statistical education, this way of using simulation has new advantages:

- Using a real and finite population (all customers of a company, or all objects in a stock, or all the accidents on a territory during one year), the concept sample is more understandable; the student understands that each individual of the population (a person, an object, an accident) can, or not, being randomly present in the sample.

- One can understand the meaning of each variable (age, sex, income, profession, weight, value, hour of the day, place,)

- The meaning of “the parameter true value” or “the true hypothesis” is clear because we are able to calculate the population parameter.

- Variables are not drawing from theoretical distribution; they are dissymetric, multimodal, and so on, as they are in real life; their correlations are complex. Looking at his sample, each student has to choose one or more models. It becomes aware that no modelling is exact, but that some are adapted more than of others to solve a particular real problem; and also that certain modelling suggested are too related to the sample available and lead to “over-fitting”.

So that the students understand well the concept of random sample resulting from the base of data, my experiment shows that it is necessary to start with random samples from a population which relates to them personally; a good manner would be to draw random samples from the list of the students, using random numbers produced in situ with a pocket calculator, and then to write in the blackboard the results of the estimates and/or tests. After that, one can automate the sampling process.

RÉSUMÉ

L’accès à de grandes bases de données réelles permet un nouvel usage de la simulation pour l’enseignement de l’inférence statistique. Une base de données est choisie comme population de référence et on peut en extraire autant d’échantillons que nécessaire, selon une méthode d’échantillonnage définie. Pour les étudiants, travailler sur ces ensembles d’échantillons permet de mieux comprendre les propriétés des méthodes d’estimation et de tests d’hypothèses. Cette approche conserve les avantages des expérimentations utilisant des données simulées selon un modèle théorique, mais y ajoute les avantages du travail sur des données réelles, qui suivent rarement les lois théoriques.