DEMONSTRATIONS IN PROBABILITY AND STATISTICS USING EXCEL

Giuseppe Cicchitelli
University of Perugia, Italy
giuseppe.cicchitelli@stat.unipg.it

The list of technologies available in the teaching of statistics has been expanded in the last decade. It includes free internet-based analysis tools, multimedia textbooks, Java applets and software designed to help students learn and understand statistics. This paper presents some learning objects, developed in Excel, dealing with the following topics:

1. *Student's T as the ratio of two appropriate independent random variables.*
2. *Empirical comparison of the efficiency of different estimators of the population mean.*
3. *Graphical presentation of the idea of consistency.*

Proper use of these tools and advantages of the Excel-based software are discussed.

INTRODUCTION

The role of the computer in teaching Statistics has been widely discussed in recent years. A very large number of contributions has been offered on the potentialities and effectiveness of computer simulations in the presentation of topics, such as the central limit theorem, confidence intervals, and hypothesis testing. An extensive review of these contributions is contained in Mills (2002). Today this type of software is available also on the Internet, especially in the form of JAVA applets.

With this wide availability of software, some authors point out the necessity of conducting experiments aimed at ascertaining the real impact of these tools on students' conceptual learning and at identifying the best ways of using them (delMas, 1997).

Bartolucci *et al.* (2005) developed *Excel*-based software covering the principal topics of an introductory statistics course.

This paper presents three learning tools concerning probability and inference topics along with reflections on their proper use.

COMPUTER AIDS IN THE TEACHING OF PROBABILITY

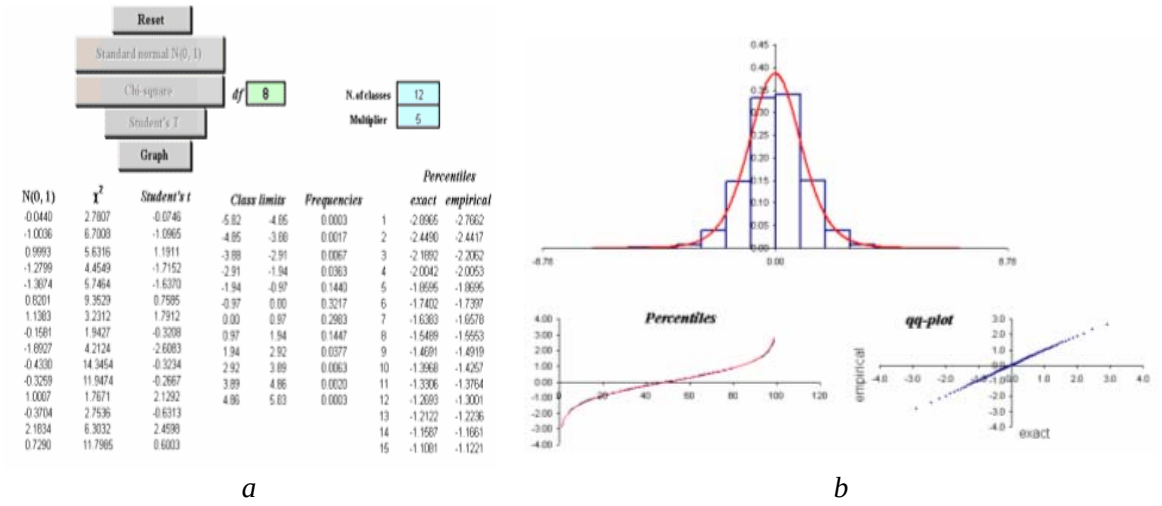
The exercise I will present here regards the demonstration that the ratio between a standard normal random variable (r.v.) divided by the square root of a chi-square (independent from the normal) divided by its degrees of freedom has a Student's *T* distribution. This aid belongs to the set of tools, contained in Bartolucci *et al*. (2005), devoted to the empirical demonstration of theorems concerning the transformation of multiple random variables, such as linear combinations of independent normal r.v., sum of independent chi-square r.v., etc.

The method used is that of Monte Carlo simulation: a large number of pairs of observations is generated (1,000 in the case being examined), the first drawn from the standard normal distribution, the second from a chi-square r.v. with a certain number of degrees of freedom; for each pair, the ratio indicated previously is calculated and, after that, the empirical frequency distribution is constructed together with exact and empirical percentiles. The result of the application of the software is shown in Figure 1a.

When the program is being started, one may choose - working in the cells in the upper part of the spreadsheet in Figure 1a - the degrees of freedom, the number of classes for the construction of the frequency distribution of the *T* observed values, and the multiplicative constant (named multiplier) used to determine the lower and the upper limits of the extreme classes of this distribution.

By clicking on the "Graphs" button one can obtain the histogram of the empirical frequency distribution of *T* observed values, together with the theoretical distribution, as well as two graphs devoted to the comparison between empirical and exact percentiles (see Figure 1b).

With this aid one shows how a certain sequence of operations produces the result predicted by the theory. This exercise may improve learning, linking concepts to a concrete and observable process.



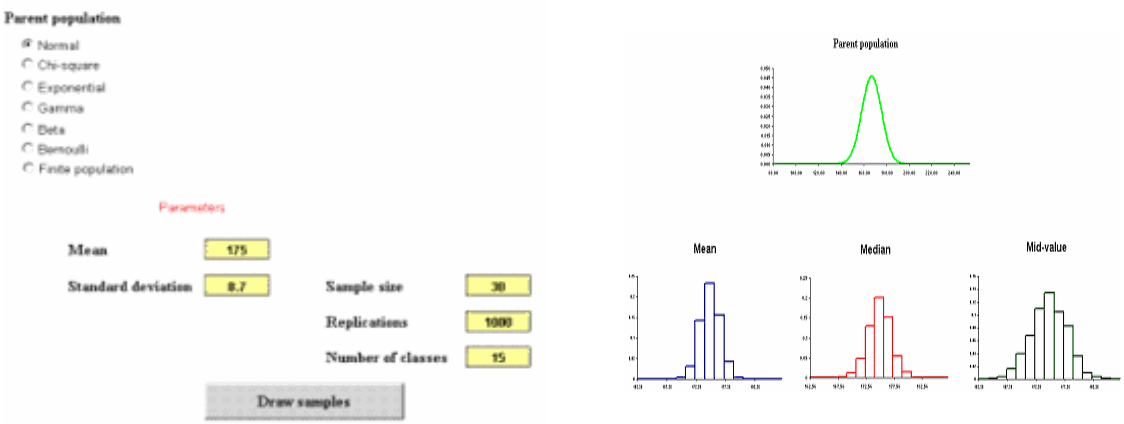
a *b*

Figure 1: Student's *T* distribution

POINT ESTIMATION

This demonstration illustrates the sampling behaviour of three estimators of the population mean: the sample mean, the sample median and the sample mid-value. The first step consists of choosing the parent population. The models considered are: the Bernoulli, normal, exponential, gamma and beta distributions; it is also possible to assume any finite population, preparing a file consisting of a column of numbers. Figure 2a shows the screen of the initial set-up: population and population parameters, sample size, number of replications and number of classes for the construction of the empirical distribution of the individual estimator.

After these preliminary operations, the "Draw samples" command starts the drawing of random samples that are shown as rows of numbers. At the end, for each sample, one obtains the values of the three estimators considered; finally, one can obtain the graphs of the empirical frequency distributions of the estimators. These representations make it possible to compare the estimators in terms of efficiency: the more efficient estimator is that whose graph shows a stronger concentration around the population mean (see Figure 2b).



a *b*

Figure 2: Efficiency comparison of three estimators

The software aims at showing concretely the phenomenon of sampling variability. Another goal is that of making it understood how the behavior of the estimators is linked to the nature of the parent population. For example, when repeating the procedure described above for nonsymmetrical populations, students can discover that the median and the mid-value do not behave like unbiased estimators.

CONSISTENCY

The initial setting of this tool is similar to that illustrated in the preceding paragraph. In addition, three sample sizes must be entered, in ascending order, in the cells marked with the symbols n_1, n_2 and n_3. By clicking on the "Draw samples" command, the desired number of samples are drawn, first those of size n_1, then those of sizes n_2 and n_3. Then, for each sample, the mean is calculated, and, finally, a graphical representation of the empirical distributions of the estimator is produced (see Figure 3), where the "trajectories" of the time series of the estimates for sample sizes n_1, n_2 and n_3 appear on the left, and the corresponding histograms are drawn on the right to point out how as *n* increases the distribution becomes more and more concentrated around the population mean.

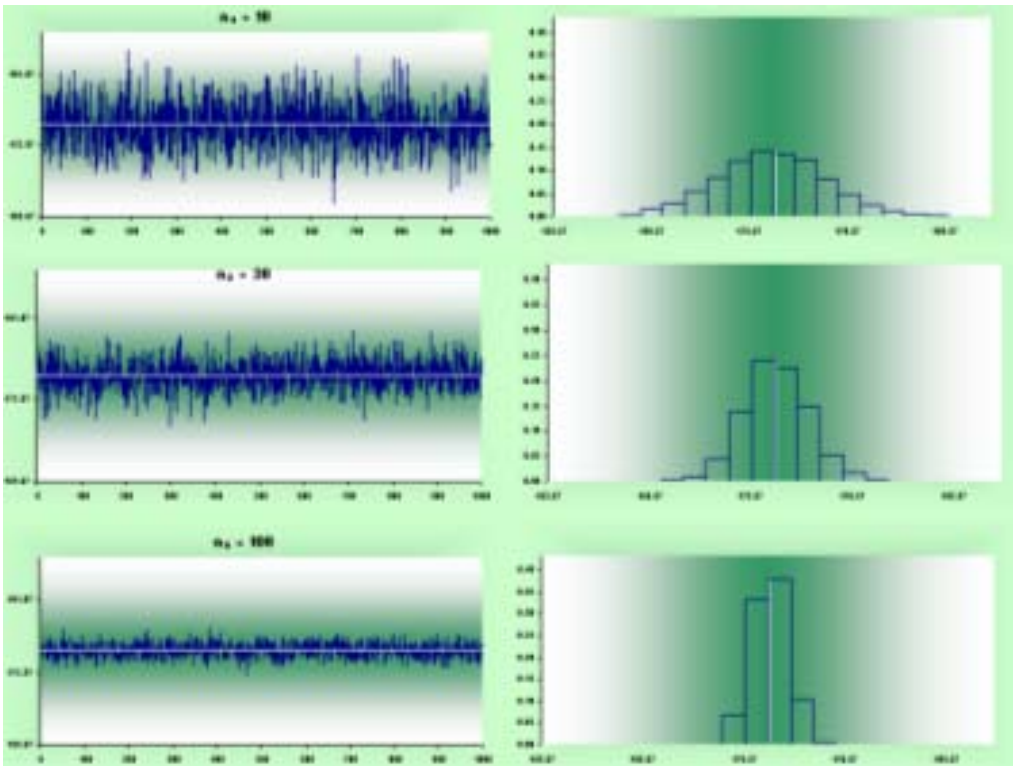


Figure 3: Sampling distribution of the mean for increasing sample size

The graphical presentation above may improve the conceptual learning of consistency, an idea difficult to understand relying only on its mathematical formulation. Perhaps, the graphs on the left are particularly useful: they allow students to grasp how the estimation error decreases as the sample size increases.

DISCUSSION

There is agreement on the importance of computer simulation methods in facilitating the understanding of difficult topics, such as the central limit theorem, the sampling distribution of a statistic, confidence intervals, hypothesis testing, etc. In fact, a computer simulation program is a sort of "laboratory" for experiments aimed at improving the learning about randomness, probability, and inference. Of course, computer technology does not promote understanding in and of itself; it should be constructed following some fundamental principals such as those suggested by Nickerson (1995): "a) View learning as a constructive process where the task is to provide guidance to facilitate exploration and discussion. b) Use simulation to draw students' attention to aspects of situation or problem that can easily dismissed or not observed under normal conditions. c) Provide a supportive environment that is rich in resources, aids exploration,…"

It seems to me that the program developed in Excel and described in Bartolucci *et al*. (2005) - of which three modules were presented in this note - matches almost entirely the principals indicated above. The Excel spreadsheet seems one of the most suitable systems to implement teaching tools: it is popular, immediate and easy to use; furthermore, students can "see how it is done" by looking at the cells; for example, they can reconstruct step by step the results produced by the software.

The software can be used either for demonstrations during a class or as lab activity. In the latter case, students are asked to investigate the topic and to answer a set of questions designed to guide their exploration of the concepts. The answers they give can be used for the evaluation of the software for future improvements.

REFERENCES

Bartolucci, F, Cicchitelli, G. and Manni, D. (2005). *La Statistica con Excel*. Rapporto tecnico, Dipartimento di Economia, Finanza e Statistica, Università di Perugia.

delMas, R. (1997). A framework for the development of software for teaching statistical concepts. In J. B. Garfield and G. Burril (Eds.), *Research on the Role of Technology in Teaching and Learning Statistics: Proceedings of the 1996 International Association of Statistics Education (IASE) Round Table Conference*, (pp. 85-99). University of Granada, Spain. Voorburg, The Netherlands: International Statistical Institute.

Mills, J. D. (2002). Using computer simulation methods to teach statistics: A review of the literature. *Journal of Statistics Education*, 10(1).

Nickerson, R. S. (1995). Can technology help teach for understanding? In D. N. Perkins (Ed.), *Software Goes to School: Teaching for Understanding with New Technologies*. New York: Oxford University Press.