A WEB-BASED EXERCISE TOOL USING RANDOMIZED DATASETS FOR STATISTICAL EDUCATION

Peter Pipelers, Olivier Thas, Dries De Vleeschauwer and Jean - Pierre Ottoy
Ghent University, Belgium
Peter.Pipelers@UGent.be

A learning environment for statistical education aims at providing on-line course material for distance learning. It typically also includes practical exercise material, providing the student the ability to test his/her statistical knowledge in a real-life situation. A desirable aspect of a self-test module is that students cannot cheat and copy the answers from their colleagues. The Elestat *project maintains this scenario with the construction of a web-based self-test module that contains many realistic datasets of which the observations are randomized before they are presented to a student, obliging the student to analyze a different dataset each time an exercise is started. The student is guided through the exercises in a step-by-step manner. The open source statistical software package* R *takes care of calculating the correct solution in real-time and offering immediate feedback. The web-technology is based on a collaboration between Java and the* Rserve-*interface. The* Elestat *project is accessible via the website http://www.Elestat.be.*

INTRODUCTION

A growing number of electronic learning environments have begun their march through the field of education. Many e-learning environments for statistical education aim at providing on-line course material for distance learning (Graham *et al*., 2000, and Stephenson, 2001). Some examples of this material consist of theory content, example exercises, applets and self-tests. Rather than only designing static learning content, the student must be provided with a tool to test his knowledge on the subject. It is hard to work out a self-test module which bans the opportunity to interchange correct answers between students and thus resists cheating and avoids repetition of questions. Moreover, a typical self-test consists of a database containing many multiple choice questions from which a random sample is taken and presented to each individual student (McLeod *et al*., 2003). If the generated database is sufficiently large, this structure will indeed solve the problem of students interchanging correct answers. But what if an enthusiastic student wants to make use of the self-test very frequently? The probability that this student gets the same question more than once is surely present.

In statistics, however, exercises with only multiple choice questions are not always appropriate. A very important topic in the field of applied statistics is the analysis of datasets. Each analysis includes the calculation of data-dependent quantities (e.g., test statistic, p-values, means, confidence intervals). Apart from multiple choice questions, where the correct computed value is listed among the selectable choices, open questions where the student needs to fill out his calculated values, are highly desirable. Although this type of questions may be very good, they still suffer from the drawback that they are not of interest anymore to the student when presented a second time.

In this paper, we introduce a web based self-test for statistical data analysis, which generates for every student a randomized dataset. The most important advantage of this approach is that even when a student is provided twice the same problem with the same questions, the data analysis will be different each time. As a consequence, the correct conclusions are possibly different too. This self-test tool is a part of a larger e-learning environment (www.Elestat.be). At Ghent University, the exercise tool has been used in a basic statistics course for students 3rd Bachelor in bio-engineering sciences.

The first section contains a more detailed discussion about the setup of such an electronic exercise environment. In the next section, some technical specifications of the random generator and navigation tools are given. Finally, the results of a survey are discussed in the final conclusion.

THE EXERCISE ENVIRONMENT

Each exercise is designed with a clear structure. First, the randomized dataset is provided to the student together with a description of the problem and details on the design of the study. Subsequently, the student is guided through the exercise with a series of questions. After each answer, the student gets brief feedback containing the correct answer and an explanation as to how the solution is obtained.

The real novelty of the environment lies in the use of randomized datasets. Each time a user visits an exercise, a new dataset is constructed with the use of a random generator. Due to this individual character, each time a student makes the same exercise, the answers and the conclusion may be different. Even when the same questions are phrased, the student still has to redo the statistical analysis because of the different observations in the datasets. The exercise starts with a description of a biological problem and the key research question. The latter is the original question as it was raised by the biologist. The dataset looks very realistic. In this way, the student performs a real-life analysis which is related to his future field of profession. A consequence of these generated datasets is that all necessary calculations need to be redone each time an exercise is loaded. Each exercise therefore pretends to be individualized to the student.

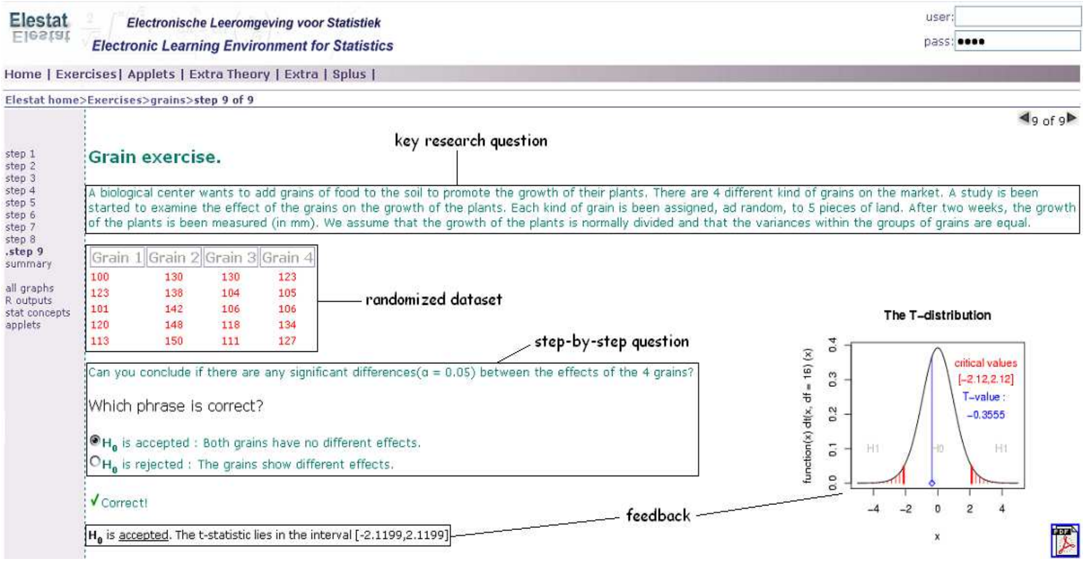
Each exercise in the environment consists of a sequence of step-by-step questions. This sequence serves as a kind of guideline through the exercise, so that the self-test is particularly useful in a basic course in statistics. The student is guided through the exercise until a final conclusion is to be made. When the first question appears on the screen, the environment waits for the student's answer. Since illustrations are convenient tools to support a conclusion, the exercise environment generates graphs too. They are also computed in real-time for each individualized dataset. Not only multiple choice questions are included, but, more importantly, open questions for which numerical results (e.g., p-values, means, parameter estimates, …) have to be entered. Once the answer is submitted, the correct answer is compared with this answer and feedback is given. Since the answers may involve data-dependent calculations, and since each dataset is different, the exercise environment must include a computation engine which computes all necessary quantities in real-time. More technical details are given in the next section. This process of question-answer is repeated until a final conclusion is reached. Finally, the student's score is reported.

For each exercise, the environment also provides additional information about the statistical concepts related to the exercise (e.g., *t*-tests, analysis of variance). This information can be accessed through the link in the left menu of the screen. For these concepts more theoretical background is provided in separate html pages. A more interactive tool to illustrate the theory behind the exercises is found in the use of a series of about 40 applets that have been developed in previous projects (Darius *et al*., 2000, 2002). The links in the left menu navigate the student directly to the relevant applets.

The *Elestat* exercise environment encourages students to test their knowledge of basic statistics. In order to integrate the environment into a course in basic statistics, exercises related to the following topics are considered: goodness-of-fit tests, Pearson test of independence, *t*-tests, analysis of variance, regression analysis and some non-parametric tests. In total 25 exercises are available already, but many more are in preparation.

Figure 1 shows a screenshot of an exercise on analysis of variance. The key research question remains always visible in the center of the screen along with the randomized dataset. The step-by-step questions appear directly under the individualized dataset. After answering the multiple choice question, feedback is immediately given. This feedback reproduces the problem-solving activity to follow and visualizes the interpretation with an on-line computed graph. The left menu contains hyperlinks to the statistical concept and to the applet that illustrates the concept of analysis of variance.

Finally, some questions related to the theoretical concepts are included. These tests are made up of five randomly selected questions.



Grain exercise.

A biological center wants to add grains of food to the soil to promote the growth of their plants. There are 4 different kind of grains on the market. A study is been started to examine the effect of the grains on the growth of the plants. Each kind of grain is been assigned, ad random, to 5 pieces of land. After two weeks, the growth of the plants is been measured (in mm). We assume that the growth of the plants is normally divided and that the variances within the groups of grains are equal.

Grain 1	Grain 2	Grain 3	Grain 4
100	130	130	123
123	138	104	105
101	142	106	106
120	148	118	134
113	150	111	127

Can you conclude if there are any significant differences($\alpha = 0.05$) between the effects of the 4 grains?

Which phrase is correct?

● H_0 is accepted : Both grains have no different effects.
○ H_0 is rejected : The grains show different effects.

✓ Correct!

H_0 is accepted. The t-statistic lies in the interval [-2.1199,2.1199]

Figure 1: Example of an *Elestat* exercise

TECHNICAL DETAILS

The setup of the environment requires that, for every individualized dataset, all statistical computations have to be completed before the start of the step-by-step questions. Therefore, issues such as calculation speed and the necessary user-friendly environment have influenced the implementation method. In our environment, the major parts are implemented with the Java programming language. Many websites nowadays use JavaServer Pages (JSP) and other Java-based technologies (Di Ciaccio, 1998). A major advantage of Java is its platform independency. With the construction of JavaServer Pages and the implementation of JavaBeans, the layout and navigation of the environment is easy to maintain. JavaBeans are software components which can be manipulated from several tools. Since online calculations must be made at runtime, there is need of a powerful calculation engine for doing statistical analysis. The solution was found with *Rserve* (Urbanek, 2003), a TCP/IP server which allows the programming language to use facilities of the freeware statistical software package *R* (Ihaka *et al.*, 1996). From Java, *R* syntax commands are send via a simple connection with this virtual server. The necessary computations are subsequently made on this server and the results are send back. The Java applications store these *R* results, collect the user answers and compare both. This generic style of programming avoids the initialization of *R* and provides the necessary computations immediately when the exercise is loaded. It can take up to ten seconds to load an exercise due to these calculations. During the exercise, the navigation system experiences no delay since all necessary information is locally stored. Figure 2 visually summarizes the procedure.

The powerful graphical capability of *R* is certainly an additional advantage. With simple *R* commands, it is easy to construct statistical plots which are relevant for the exercise and informative for the feedback. These graphs are saved to files on the server. Java provides elementary file handling to interrelate these graphs within the environment. The generation of the graphs is also finished before the exercise is started.

EVALUATION

By monitoring the web traffic, we can evaluate the popularity and difficulty of all 25 exercises. This enables us to further optimize the set of exercises.

The present environment has been set open for a test public of 150 students 3rd bachelor in bio-engineering sciences at Ghent University. They were encouraged to use the material for preparing the exam of a basic course in statistics. In order to assess the personal opinions of the students, we held a survey after ten weeks. 84% of the students admitted to have frequently used the exercise environment. To the statement that the online randomized exercises are useful when it comes to processing the material, almost 75% of the students agreed. This also explains why a common received remark was that more exercises are desirable.

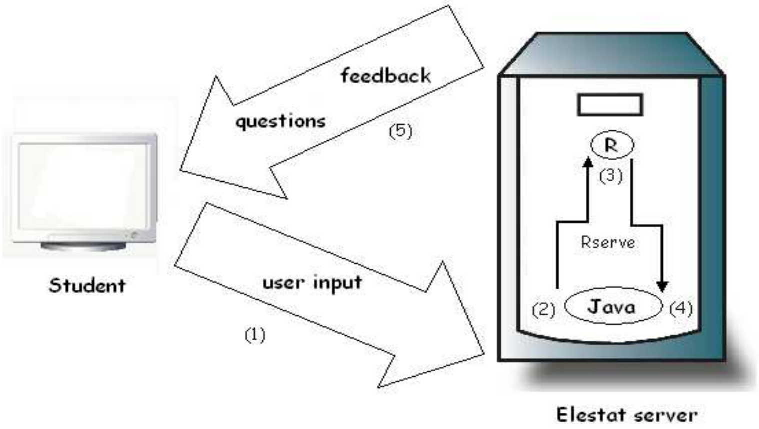


Figure 2: Modus operandi of the *Elestat* exercise environment:
(1) The student picks an exercise. (2) Java invokes the corresponding *R* calculations via the *Rserve* connection. (3) The *R* calculations are performed on the virtual server and send back. (4) Java stores the necessary information. (5) The questions are displayed with JSP-pages. What follows is a loop between the user input (1), Java comparing the answers (4) and the generated output and feedback (5).

REFERENCES

Darius, P., Ottoy, J. P., Solomin, A., Thas, O. Raeymaekers, B. and Michiels, S. (2000). A collection of applets for visualising statistical concepts. In *Proceedings of the 14th Conference of the International Association for Statistical Computing (COMPSTAT).*

Darius, P., Ottoy, J. P., Michiels, S., Thas, O. and Raeymaekers, B. (2002). Applets for experimenting with statistical concepts. In B. Phillips (Ed.), *Proceedings of the Sixth International Conference on Teaching of Statistics*, Cape Town. Voorburg, The Netherlands: International Statistical Institute.

Di Ciaccio, A. (1998). Hypermedia and WWW for the teaching of statistics. In L. Pereira Mendoza, L. Seu Kea, T. Wee Kee, amd W. Wong (Eds.), *Statistical Education - Expanding the Network: Proceedings of the Fifth International Conference on Teaching Statistics*, Singapore, Vol. 3, (pp. 953-960). Voorburg, The Netherlands: International Statistical Institute.

Graham C., Cagiltay K., Craner J., Lim B. and Duffy T. M. (2000). Teaching in a web based distance learning environment, an evaluation based on four courses. Center for Research on Learning and Technology, Technical Report No. 13-00.

Ihaka, R. and Gentleman, R. (1996). *R*: A language for data analysis and graphics. *Journal of Computational and Graphical Statistic*, 5(3), 299-314.

McLeod, I., Zhang, Y. and Yu, H. (2003) Multiple choice randomization. *Journal of Statistics Education*, Vol. 11(1).

Stephenson, W. R. (2001). Statistics at a distance. *Journal of Statistics Education*, 9(3).

Urbanek, S. (2003). *Rserve* - A fast way to provide *R* functionality to applications. In K. Hornik, F. Leisch and A. Zeileis (Eds.), *Proceedings of the 3rd International Workshop on Distributed Statistical Computing* (DSC 2003).