Assessment in Statistics Using the Personal Computer

Giuseppe Cicchitelli - Francesco Bartolucci - Antonio Forcina

Dipartimento di Scienze Statistiche
Via A. Pascoli
Perugia, Italy
pino@stat.unipg.it

1. Introduction

In Italian universities, students' performance is usually assessed by oral examinations. Although this system offers some advantages, it also has a number of weaknesses regarding both objectivity and efficiency, especially for crowded courses.

One method of assessment which can be considered objective is true/false testing. This method is well theorised in psychometrics. Responses to items can be thought of as observed values of random variables which are indicators of a latent parameter, student ability. Various models in Item Response Theory (see Goldstein and Lewis, 1996, and Hambleton and Swaminathan, 1985) relate the probability of a correct response to this parameter. These models assume that:

- Student ability is a unidimensional latent trait;
- The probability of a correct response depends both on student ability and on item-specific parameters. This probability is modelled by the so-called Item Characteristic Curve, which usually has a logistic form;
- A student's responses to different items in a test are independent, and the responses of different students are also independent.

One of the most commonly used models is given by the following formula

$$(1) \qquad p_{ij} = e^{a_j(\theta_i - b_j)} \Big/ [1 + e^{a_j(\theta_i - b_j)}],$$

where p_{ij} is the probability that student *i* responds correctly to item *j*, θ_i is the ability level of student *i*, and a_j and b_j are, respectively, the discrimination index and the difficulty level of item *j*.

The computer is an important resource in preparing banks of items, in administering the test and in processing the results.

We present an integrated computer-based approach to all three of these stages of the process, and the results of using it with university students in a course on descriptive statistics.

2. The procedure

The procedure consists of three modules. The first is used to create test items using texts in ASCII format, formulas, tables, graphs and objects such as input boxes, double check boxes and buttons. The second one is used to administer the test, i.e. to select the items from the archive (according to the parameters established by the examiner), present them on the screen, and record the student's responses. In the current configuration, the items are presented in groups of four, as shown in Figure 1. The test items can be identified beforehand, or they can be selected randomly from the archive, arranged by topic. When the test is completed, a file is produced containing all the information on the student's performance.

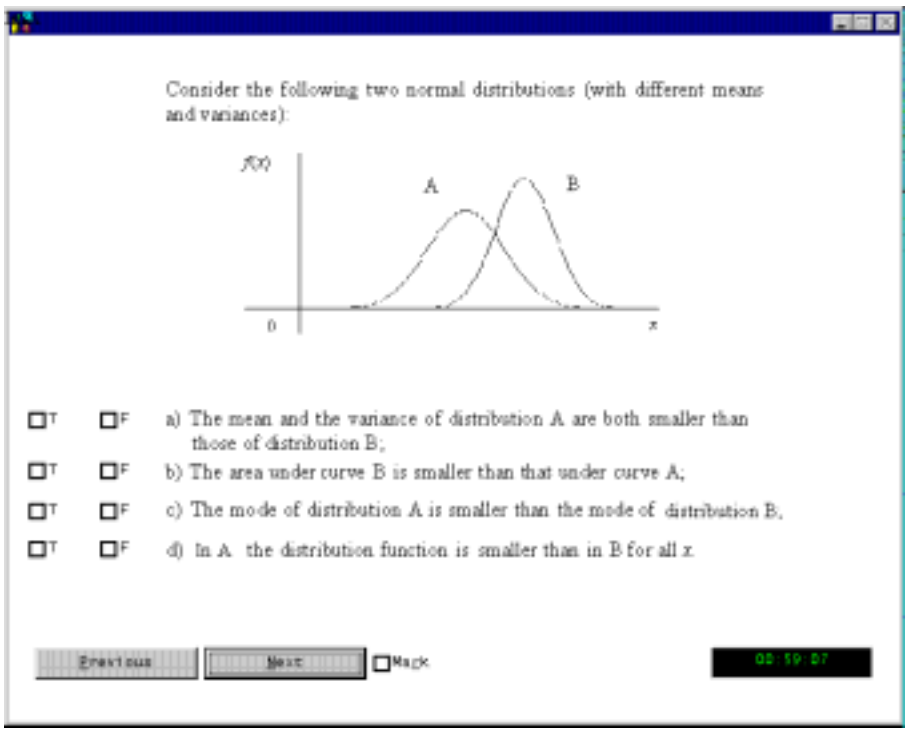


Figure 1

The third module is devoted to the estimation of parameters of model (1) by the maximum likelihood method, using the file produced by the second module.

This procedure can also be applied to evaluate a single student, provided that the discrimination index and ability level are known for all the items.

From a technical viewpoint, the use of the system requires a room equipped with personal computers connected with a server. The execution program for the test is loaded onto each PC, and once it is run, it constructs the test by drawing the items from the archive of items stored in the server. As soon as all the students have completed the test, the files of answers are transferred to the server and the relevant data are analysed.

3. A case study

The application that will be described here refers to a group of 135 students (from the Faculty of Economics at the University of Perugia) who took the examination for the course on descriptive statistics in February 1999. The examinations took place in three separate periods; within each period (lasting a maximum of three days), the students, who were divided into 6 groups ranging in size from 18 to 28, were given the same test composed of 100 items (with one exception which will be mentioned later). The test was changed from one period to the next.

There were basically three types of items: those concerning definitions, those involving ability to carry out calculations or apply formulas, and those aimed at establishing the understanding of concepts. The items shown in Figure 1 fall into the third category.

In the first two periods, the test included 100 items. The students were given 1 hour to complete the test, and they were allowed to go back and change their responses to items they had already answered.

For the third period, 11 items were added to the 100 base items. These additional items were repeats of previous questions with slight formal modifications. In addition, students were not allowed to go back and change their answers to previous questions. The purpose of this was to take into account, in some way, possible student guessing strategies.

At the end of the three test periods, each student took also an oral examination. The assessment was conducted knowing neither the results of the test nor the score obtained at the preliminary written examination concerning application to real data.

The bivariate distribution of the grades expressed as a score out of a maximum of thirty points obtained at the T/F test and at the oral examination is as follows:

Grades obtained at the T/F test	**Grades obtained at the oral examination**			
	< 23	23 - 26	> 26	**Total**
< 23	24	12	2	38
23 - 26	16	35	7	58
> 26	6	11	22	39
Total	46	58	31	135

In this table, independence can be safely rejected against the unrestricted alternative with a likelihood ratio test of 45.1 on 4 d.f. By applying the methods described in Dardanoni and Forcina (1998), it turns out that independence can also be clearly rejected against a specific form of concordance (which implies that all adjacent odds ratios are not smaller than 1). Furthermore, this form of concordance is accepted against the unrestricted alternative. In other words, the data support the idea of a fair agreement between the results of the two assessment procedures.

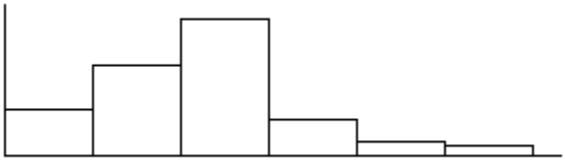
4. Discussion

We think that the system is a useful aid to the traditional examination: clearly, an interview where the examiner has available the results of the T/F test makes it possible to go more deeply into given topics and to probe any inadequacies which had emerged, permitting a more accurate assessment. A test such as that described cannot give complete “information about students’ statistical reasoning processes, their ability to construct or interpret statistical arguments, ...” (see Gal and Garfield, 1997, p. 7), which are very important goals of a sound assessment.

An additional point to be considered is the possibility that the system offers for examining the historical record of the items with the corresponding percentages for correct answers. This is useful to find out difficulties that students meet in the learning, and provides feedback to improve assessment as well.

The analysis of students’ responses showed that items involving the connection between the shape of the distribution and the value taken by descriptive statistics are particularly difficult. An example of this type of items is the following:

Consider the distribution represented by the histogram below:



The third quartile of the distribution may fall in the last but one class interval T F

This item requires connected understanding: students have to link conceptually histogram and distribution function, and then distribution function and quantiles. Only about 25% of the students responded correctly. Furthermore, cross-classification of students according to the responses and the scores obtained at the oral examination shows no significant association between the two variables; i.e. this item does not discriminate between higher and lower achievers at the oral examination. In

addition, cross-check by means of an equivalent item indicates that a large proportion of students responded by guessing.

Similar evidence can be drawn from the analysis of the responses to items requiring some reasoning on the structure of statistics and on their behaviour under data transformations. An example is given below:

Wages of the 10 employees of a small business are all increased at the rate of 4%.
Then, the standard deviation of the new distribution increases at the same rate T F

This item requires that student understands both the mechanism of the transformation (the new data come from a multiplication and not from an addition of a constant) and its effect on the statistic.

On the other hand, we found that items having a medium or low difficulty level (those responded correctly by a proportion of students ranging, say, from 50% to 75%) generally discriminate students quite well.

The above findings suggest that it is not convenient to propose items which require an elaborate thinking, where a yes or no reply is unlikely to represent the understanding of the underlying concepts.

A clear weakness of the system is the high probability of answering correctly by guessing. In this regard, the following device was tried out for measuring this phenomenon. As mentioned previously, for the third test period, control items were inserted in the test as repeats of certain items with slight changes. This allowed us to determine the relative frequency, f_i, of the discrepancies between the responses to equivalent items. This quantity was used to modify the probability of answering correctly according to the model $p_{ij} = [e^{\alpha_j(\theta_i-\beta_j)+\lambda f_i}]/[1+e^{\alpha_j(\theta_i-\beta_j)+\lambda f_i}]$, where λ is a positive parameter which measures the level of dependence of p_{ij} on f_i. Maximum likelihood estimates of θ_i under this model clearly reflect the level of f_i: students with higher f_i get lower grades.

The procedure above described can be improved in various directions. First of all, it is possible to present items in the form of multiple-choice questions, a form, perhaps, more suitable to gain insight on how students make mistakes and to discover misconceptions; secondly, for special topics, it can be interesting to provide “second level” items, i.e. questions to be proposed only to students who respond correctly to given items to investigate the rationale of the answers. A third direction is the implementation of an adaptive assessment approach, which could match the difficulty level of the test to the ability level of the single student.

REFERENCES

Dardanoni V. and Forcina A. (1998). A Unified Approach to Likelihood Inference on Stochastic Orderings in a Nonparametric Context, Journal of the American Statistical Association, **93**, pp. 1112-1123.

Gal, I. and Garfield, J.B. (1997). The assessment challenge in statistics education. Amsterdam: IOS Press.

Goldstein, H. and Lewis, T. (1996). Assessment: problems, developments and statistical issues. New York: Wiley.

Hambleton R. K. and Swaminathan H. (1985). Item Response Theory: Principles and Applications - Kluwer Nijhoff Publishing, Boston.

RÉSUMÉ

Dand cette étude on présente un procédé informatique pour l’évaluation du profit en statistique a l’aide de l’ordinateur. En autre, on illustre l’application du procédé dans le cadre d’un cours de statistique descritptive.