THE ROLE OF THE COMPUTER IN THE TEACHING OF STATISTICS

Tarmo Pukkila and Simo Puntanen
University of Tampere, Finland

1. Introduction

In this paper we consider the *nature* of a statistics course and discuss the role of the computer in it. In particular, we discuss the the course for the first year-students at the university level. In this context, for example the following questions are essential: How much effort should be put e.g. into handling the mathematical properties of central concepts like mean, variance and correlation? What is the real difference between a mathematics and a statistics course, or a course in computer science? In fact, is a subject called statistics needed any more? Can mathematicians handle formulae and computer scientists computers more efficiently than statisticians? However, there is surely no justification in calling a course a *real* statistics course, when all we know of it is that the formulae or means etc. are carefully handled. In our paper we handle the concept of a *real* statistics course and discuss the essential role of the computer in it.

The course which we are teaching is given in the Faculty of Economics and Administration at the University of Tampere. The computer has been used on a large scale in teaching basic statistics in the University of Tampere since the year 1975. Computer usage was begun applying the system *SURVO/71* in data-analytical problems. Pukkila (1978) and Pukkila and Puntanen (1980) are papers describing the commencement of computer use in the University. After promising experiences of its use in data-analytical practical statistical problems, use in teaching was extended in 1978 to cover also theoretical concepts (sampling distributions, testing of hypotheses, regression models, analysis of variance models, time series models, generation of data from these etc.). For the illustration of such theoretical statistical concepts the KONSTA system (Pukkila, Puntanen and Stenman 1984) was developed at the Department of Mathematical Sciences.

2. Aims of the Basic Course

When coming to the course most students have not yet had time to study especially deeply their main subjects, where later they have possibly to apply statistical methods. It might be argued in this kind of situation that the basic course in statistics comes too early in the students' program.

The basic course is obligatory for all students in the Faculty. This is in many cases the only reason why a student comes to the course, so that the attitudes of students towards the course are also not the best possible.

The goal in teaching is that after the basic course students should be able to cope with statistical problems eventually arising in their major subjects. Among other things this means that they should be able independently to carry out a brief empirical statistical research. In order to reach this goal,

the students should be able *to acquire data, to handle them, interpret results* and finally *to write a report*. Hopefully the course also arouses real interest for later statistical courses, especially among those studying an applied subject. On the other hand, the basic course has also to form a sound basis upon which to build later studies in statistics on the statistical curriculum.

3. Nature of Studies on the Basic Course

Basic course is intended to be as practical as possible. This does not mean that theoretical matters are not handled on the course. Nevertheless practical problems have a central role. For the students the handling of practical problems begins with the finding and formulating of statistical problems. It continues to obtaining data and analyzing them and ends with the writing of a report.

Let us consider some particular points in the teaching of statistics for first-year students. How much time should be given e.g. to the handling of the mathematical properties of basic concepts like mean, variance and correlation? Is such a course where these topics are treated in detail a real statistics course? What is the difference between a mathematics and a statistics course, or a course in computer science? In fact, is a subject called statistics needed any more? Can mathematicians handle formulae and computer scientists computers more efficiently than statisticians?

In trying to answer the above questions it must be remembered that the course we are talking about is a first-year course. This is important because the first statistics course has a somewhat different role from later courses. Attitudes to statistics are to a great extent created on the first course, and it can be said that in statistical research attitudes mean a great deal: If they are not "correct", then not even the correct formulae will help. Mathematical courses have longer traditions than those in statistics and all students have some idea of what mathematics is about, while statistics is a new subject to them.

There is surely no justification in calling a course a *real* statistics course, when all we know of it is that formulae of means etc. are handled carefully. Similarly a course which shows in detail how the computer can be used in statistical calculations is not necessarily a real statistics course. These conclusions can be drawn e.g. from one important aim of the first statistics course. After a course *a student should be able independently to carry out empirical research*. But knowing all formulae and computer commands does not necessarily imply that a student is able to write a report on his findings. These capabilities, however, are fundamental to the statistician's role. The writing of an accurate, understandable report is an essential part of the statistical research process.

It happens every now and then that a student comes and says: *"Here are the statistics I calculated. What next? Or do you sign my study book now?"* Here the main point, statistically seen, is *"What next?"* It is not enough for a statistician (not even for a first-year statistician) to

know all the correct formulae and their effects or to use the computer correctly.

A real statistics student should be able to *combine* different things in a sensible way; he should know the correct formulae, be able to calculate statistics and he should know *what to do* with the statistics he has calculated. And above all, he should be able to create statistical problems, i.e., he should have a statistical imagination. In simple terms: a mathematics student has to prove a theorem, a computer science student has to write a program which works in the desired way, a statistics student has to *solve a real statistical problem, a data matrix being connected with it*. These problems have rather different natures particularly as regards their levels of uniqueness. It is typical in an empirical statistical research that there are no unique problems with unique solutions. This is a matter that the first-year students should be faced with.

The computer is a necessary slave in instilling into students an actual feeling for empirical statistical research. This slave allows the statistical imagination to work freely without tedious calculations being a limiting factor. But not all decisions can he made by this fast slave, because its common sense is more limited than that of a well-practised statistician.

When speaking of the teaching of statistics it is useful to remember that statistics should help for example in making certain decisions. This means that statistical problems must be derived from real life. Real life research problems raise questions, and it is possible that statistics may help to answer these questions. Between questions and answers certain statistical methods are inevitably applied. This means that teaching statistics should not be dealing only with techniques for calculating means, variances, correlation coefficients etc. The technique should be combined with an empirical problem in the background. This is not easy especially on big courses, but a basic course in statistics should nevertheless contain a leitmotif. This backbone would also hopefully motivate students to study statistics.

Besides the empirical problems the basic course also involves many matters of a theoretical nature. Common to these is that for many students they are rather difficult to understand. These theoretical matters are connected in one way or another with the concept of sampling distribution. When teaching theoretical statistical concepts connected with sampling distributions we have at one stage repeatedly to use phrases like: *If we took from a certain population 100 simple random samples of a given size n and calculated the 95% confidence interval according to a certain formula for the population means μ from each sample, then about 95 of these intervals would include the right value of μ."* If the teacher has only chalk and blackboard available, then in the practical teaching situation only phrases like the above are possible. But nowadays other kinds of equipment are also available. Why not use them?

4. The Organization of Computer Usage in the Teaching of Statistics

In the University of Tampere the use of the computer was introduced on the basic courses in data-analytical problems in 1975. It was seen to be an

inconsistency while later statistical research obliged students almost invariably to use the computer, the basic course took no account of this. Our opinion was and is that in fact a course in statistics is the natural place to learn to use the computer in statistical problems.

An important impulse to the introduction of computer use on the basic course was the availability of the statistical data processing system *SURVO/71*. This was developed in the Computer Centre of the University. In 1981 *SURVO/71* was implemented on a DEC 2060, which replaced the old Honeywell system. The new system has about 100 lines for time-sharing terminals. The main feature of *SURVO/71* is that it is extremely simple to learn and use, and that the students can use it interactively.

SURVO/71 is used in studying empirical statistical problems on the basic course in statistics in the University of Tampere in such a way that every student solves some of the weekly exercises with the aid of the computer and using the system. These exercises are connected with a real data matrix from a practical research problem. The data, on the other hand, have been collected and stored in the memory of the computer during those years the computer has been used extensively in teaching. Students have also to collect their own data sets. It is quite essential that the data sets to be studied should be as interesting as possible. On the other hand, the data can well be messy, thus giving a taste of real data analysis.

In using the computer in certain weekly exercises the aim is that the real situation be imitated as closely as possible. For this reason these exercises are defined rather *loosely,* especially in later stages of the course. We know that in the real world the problems are not at all clear and for this reason it is a good policy to make students acquainted with problems which also on definition look like real problems. It seems that, especially in the beginning, many students feel confusion when they are given a loosely defined problem which demands a greater independence than for example the precisely defined and clear problems of mathematics they have been accustomed to handle in school.

On computers and *SURVO/71* the students follow two lectures each of two hours. After each of these two lectures every student has a supervised terminal exercise of one hour. For terminal and also other exercises the students are divided into about ten groups. After these lectures and terminal exercises, which take place at the very beginning of the course, they use the terminals in the University independently in their exercises.

Note that there must be two kinds of computer exercises: First those with only a few observations so that results calculated manually can he compared with the output of the computer, and secondly those concerning extensive real-life data.

At this point it must be said that the practical problems described above form only a part of the exercises done by students. They must also solve more traditional exercises weekly. Among these exercises there are also problems that teach students to make calculations by hand.

Then some words about KONSTA. As was already mentioned in the introductory section, KONSTA is an interactive computer system used mainly to illustrate basic concepts of statistics. The system is a collection of separate programs which are closely interconnected. The programs of KONSTA can be divided into four parts as follows:

- Generation of samples from desired populations and calculation of relevant statistics.
- Generation of observations from certain models such as standard linear model, time series, random walks, etc.
- KONSTA offers flexible possibilities to study the behaviour of various statistics when adding or removing observations. For example. one can generate pairs (x,y) such that x and y are random digits and calculate the regression and correlation coefficients after adding each observation.
- For handling data matrices there are particular programs planned to be especially useful in dynamic interactive data analysis when one wants to get quickly a rough idea of data and any slights changes in them.

5. Concluding Remarks

Now it may be asked what the computer has brought to teaching. Briefly, the answer is: It has essentially helped in making our course a *real* statistics course.

References

Pukkila, T. (1978). The utilization of the computer in the teaching of statistics at elementary level. In COMPSTAT 1978, Proceedings in Computational Statistics (pp. 422-428). Wien: Physica-Verlag.

Pukkila, T., & Puntanen, S. (1980). Computer usage on first statistics courses in the University of Tampere. In COMPSTAT 1980, Proceedings in Computational Statistics (pp. 145-151). Wien: Physica-Verlag.

Pukkila, T., Puntamen, S., & Stenman, O. (1984). The use of KONSTA in the teaching of statistics. Department of Mathematical Sciences, University of Tampere Finland, Report A 125, 118 pp.