COMPUTERS AND STATISTICAL ACTIVITY

Rolf Biehler
Institut für Didaktik der Mathematik (IDM)
Universität Bielefeld
Bielefeld, F.R.G.

1. INTRODUCTION

The scientific counterpart of school mathematics can no longer be just "mathematics", but must be the "mathematical sciences" including, for instance, informatics and statistics. This shorthand characterization will probably express a widely shared belief. In any case, it is the basic assumption of this paper and the research work that it will elaborate on. In order to develop perspectives for school mathematics, it is necessary and helpful to analyze trends in the development of the mathematical sciences which are related to the spread of the new technologies. A deeper philosophical analysis and conceptualisation of changes in general orientations, value systems, conceptual structures and in the tools for developing and applying mathematics is an important and demanding task for the didactics of mathematics.

Apart from its relevance in schools, science and society, statistics is a good case in point for studying the impact of the new technologies. It is a mathematical science which has quite a history in using computers, and the historical development of statistical soft-

In: H. G. Steiner & L. Bazzini (Eds.), Proceedings of the First Italian-German Bilateral Symposium on Didactics of Mathematics. Pavia, October 4 - 9 1988 (pp. 401-416). Quaderno/Consiglio Nazionale delle Ricerche/Progretto Strategico Tecnologico

ware reflects the general progress and problems of hardware and software development as well as the tension between system functionality and user orientation. Gnanadesikan and Kettenring (1988, 13) characterize statistics as a data science with close synergetic relations to mathematics and computer science. That computer science is mentioned on an equal standing with mathematics is already an important indication of change. Moreover, the relations between statistics, probability theory, mathematics, and the analysis of data have been complex, variable and controversial in the history of scientific statistics and its teaching. As these relations have recently experienced a new qualitative dynamics that is intimately related to the computing technology, a deeper study of these developments will be important and illuminating for the didactics of mathematics and statistics.

In the following, ideas are presented that rely on several sources, among others a study on the development of Exploratory Data Analysis and subsequent papers on didactical aspects of this part of statistics (Biehler 1982, 1988, 1988a) and the involvement in a current project (SOMA) that is concerned with analysing statistical software tools from the perspective of the didactics of mathematics. Nevertheless, the paper has a programmatic character in that observations and problem areas are discussed which deserve a further detailed study.

2. SOME RELEVANT CHANGES IN STATISTICS

There seem to be at least the following four domains where significant changes have occured in connection with the new technologies:

* analysis of multivariate data
* statistical graphics
* Monte Carlo simulation
* statistical modelling.

All four domains intersect with current reflections on the future structure and content of teaching statistics as well as with important aspects of informatics and general computer use.

2.1 Working environments and organizing statistical activity

Before I can come back to details of these domains, a very general dimension of change in statistics has to be made explicit which is essential to understand new developments. I should like to call it the shift from the theory ideal of mathematics to the engineering and technological aspects. A new emphasis is given to the construction and design of tools and what we shall call "integrated working environments" that materialize in statistical software. The concern for organizing statistical activity of real human beings in such working environments is complementary to this. Independent of concrete software, this trend has already become clear in the kind of Exploratory Data Analysis the American statistician John Tukey (1977) has been promoting (see also Biehler 1982). Statistical data analysis is conceptualized as an interactive, iterative, exploratory activity with data. Articles on styles of data analysis, on the ethics of a good data analyst, on adequate attitudes and rules of behaviour are being published. Tukey has even suggested that statistics should also be interpreted as a behavioural science. An early consideration of "human factors" was his suggestion of "quick and easy methods" and the definition of the practical power of a statisti-

cal test as the mathematical power multiplied with the probability of its use. Later, this concern expressed itself in research studies on human perception of statistical graphs (cf. Cleveland 1985, pp.229) and in fundamental research into designing human interfaces of statistical analysis systems as part of the fundamental research a discipline performs for the purpose of improving its tools.

2.2 Software trends

A recent trend in statistical software at PC-level consists of qualitative improvements of the human computer interface for the non-expert user, mainly through a replacement of formal languages by other means like menus and use of mouse for user input, for instance for selecting subsets of data directly in a graph or by simply "cutting, pasting and copying" in data tables, for instance in the software STATVIEW 512+, DATADESK or MACSPIN for the Macintosh computer. Nevertheless many basic requirements are not yet fulfilled if we look at a list Chambers put together 5 years ago:

" a. very flexible presentation of information to the user through a multiple-window display;

b. 'visual programming' interface to analytical operations to augment procedural languages;

c. dynamic graphical displays, such as the presentation of 3-dimensional data;

d. facilities to let analysts develop specialized analytical systems (e.g. with a menu-based user interface to specialized analytical techniques) from the general analytical system.

The new computing environment makes it possible to tailor multiple human interfaces to the same underlying analytical system, suitable for different levels of user sophistication or different appli-

cation areas." (Chambers 1983, pp.101)

Although it is quite a break-through that flexible, adaptable and extensible working environments for PC-level (like PC-ISP) are available that are similar in style to the prototype system S, such systems are fairly difficult for non-expert users because they are restricted to a command language user interface.

2.3 Theory and practice

The shift of emphasis from the theory ideal of mathematics to technological concerns has had an impact with regard to the relation between theory and practice in statistics. A main concern of theoretical statistics has been to prove that certain methods are optimal under certain model assumptions. In other words, we observe the conception that theoretical statistics provides an algorithm for a certain class of problems, and the practitioner has to decide whether the algorithm can be reasonably applied under his situational conditions, perhaps after some minor adaptations. For a while, statistical software has not been much more than a library of algorithms. With modern software tools, however, it is not just a collection of algorithms that is provided but an integrated, flexible and adaptable working environment. There is no longer (the fiction of) a more or less complete theoretical solution to a problem but a practical, only partly theory-based support for finding a reasonable solution in a feed-back process with the concrete situation and data. This is a kind of mathematics for which Fischer (1984) coined the term open mathematics.

The tendency of synthesizing working environments from elements of statistics, mathematics and computer science is visible in all four domains mentioned above: data base and spreadsheet envi-

ronments for manipulating multivariate data, graphical environments for constructing, varying, and editing statistical graphs, simulation and modelling environments for constructing situation-specific models and Monte Carlo simulations. In the following paragraphs I will briefly discuss new kinds and features of activity in such environments, which, in my opinion, deeply question what is currently considered as basic statistical activity and knowledge relevant to education. Moreover, the concept of a working environment with systems character is itself a challenge to the predominance of traditional linear-sequential types of teaching.

3. MULTIVARIATE DATA

Multivariate data have become the fundamental object of statistics. Most of the tables in newspapers, statistical yearbooks and other media have always been multivariate. In statistical practice, multivariate data are the standard case. New multivariate methods heavily rely on graphics (see 4.) and have enormously been stimulated by the possibilities of the new technologies. Statistical software reflects this situation: often the basic data structure is a table or matrix of data where rows represent cases and columns represent different variables.

It is important to realize that the application of elementary database and spreadsheet operations to a data table already provides many possibilities for answering and exploring elementary yet important questions posed to data sets. More sophisticated multivariate methods may be applied at a later stage. As an example, think of a table with traffic accidents where for every day of a year, the day in the week, the date, the number of accidents (with deaths,

with insured people), the deaths according to accidents inside cities, outside, highways etc. are represented.

The selection of subsets of variables or of cases for comparison can be an important activity. Sorting data tables and splitting columns according to a categorial variable are further examples. The additional exploratory power of a software like MACSPIN which permits to see 3-dimensional scatterplots rotating on the screen lies in the many operations for selecting and marking parts of the data according to the data appearance in the graph and by direct interaction with the graph.

Table arithmetic, where new variables are constructed with arithmetic operations from the old ones and added as further columns to the table, is important and well supported by a spreadsheet environment. The definition of new variables involves a specific type of mathematization or modelling in the context of statistics. The derived quantities are often those of interest for a further statistical analysis. Many tables in statistical yearbooks have their origin in table arithmetic applied to a table of initially measured quantities: statistical spreadsheet environments can be used to construct and re- construct situation-adequate tables.

Obviously, these features are closely related to important issues in informatics such as applications of spreadsheets and data bases. They also reinforce the table as a fundamental representation in the mathematical sciences. However, these features are quite different from usual curricular emphases on univariate statistics with its close relation to (univariate) probability.

4. STATISTICAL GRAPHICS

The computer graphics revolution in statistics has been described and analyzed in many papers. The re-integration of graphs into the system of statistical methods, their use for exploratory and communicative purposes, dynamic graphics and direct interaction with graphs are important tools for statistical work. The spread of hard- and software makes this potential available for many. However, the knowledge and competence concerning the statistical design and use of graphics lags far behind these technological possibilities.

It is profitable to emphasize the following features:

(1) Simple graphs like scatterplots attain a radically enhanced analytical power when they are embedded in flexible data base and spreadsheet environments. This is true even when the operations are very elementary from a mathematical and logical point of view; for instance

* selection of all kinds of subsets of data
* (non-linear) numerical transformations of scales, data and variables
* projections from higher dimensional data sets
* possibilities for enhancing the display with complex symbols instead of points
* operations on graphs like selecting and isolating subsets with interesting structure, identifying outlying points etc.

(2) Graphs have conquered a new, more fundamental role in the system of statistical methods; for instance, for revealing structure and anomalies in the initial exploration of data, before any numerical summaries like mean, variance, correlation coefficient etc.

shall be reasonably used. This applies to the scatterplot as the basic tool for two-dimensional data. Approaches like MACSPIN may conquer a similar fundamental role for higher dimensional data.

(3) The flexibility and scope of new software tools puts the user in the role of a co-designer of graphs, where new knowledge concerning statistical and perceptual aspects of a diversity of graphs are required as well as knowledge concerning the flexible use of a graphical software system.

These shifts profoundly question what we consider today as basic statistical knowledge and techniques at secondary level and beyond. At the same time, it has to be further explored and substantiated how far statistical graphs may open a path to statistical knowledge and thinking more easily than starting with mathematical techniques and theories in the traditional sense. For instance, does it pay to use MACSPIN as an introduction to multivariate statistics ?

5. MONTE CARLO SIMULATION

The developments sketched in the last two paragraphs increased the distance between methods of data analysis and the traditional probability based theory of statistics. On the other hand, however, Monte Carlo simulation has become increasingly used for maintaining, extending, and generalizing the probabilistic approach in statistical theory and practice. As Monte Carlo simulation provides results only with "statistical certainty", there is a price to pay from the standpoint of traditional mathematics. Nevertheless, the new developments provoke the radical question whether a didacti-

cal reconstruction of probability and statistics has to assign a much more fundamental role to simulation than has been considered up to now. Diaconis and Efron (1983) summarize essential impacts of the new computing facilities, where a million of arithmetic operations is often used to analyze 15 data points and the use of simulation plays an important role:

"The payoff for such intensive computation is freedom from two limiting factors that have dominated statistical theory since its beginnings: the assumption that the data conform to a bell- shaped curve and the need to focus on statistical measures whose theoretical properties can be analyzed mathematically."

Let us take the chisquared test for goodness-of-fit as an example. Instead of reyling on the theory of the chisquared statistics in the multinomial probability model, the following strategy is possible. If a certain numerical value of the chisquared statistics is observed, say Cobs, an ad hoc simulation can provide an estimation on how likely such an observation or a more extreme result would be. This strategy is equally well applicable to other plausible measures similar to the classical chisquared statistics. The strong argument for chisquared, i.e. that there exists a theory which reduces the practical comparison with real data to looking in a table of the family of chisquared distributions, is no longer valid.

A more radical application of this idea is the bootstrap method, which is even more computationally intensive. Traditionally the reliability or variability of a statistical estimation or numerical summary of data is analytically derived and calculated on the basis of an assumed theoretical probability model. In the bootstrap method, the set of real data observed is taken as an estimate of the underlying distribution without any other specific theoretical assumption. Afterwards, the simulation is based on this estimated model. In concrete terms, this means drawing samples with replace-

ment from the observed set of real data.

The paradigmatic application situation of bootstrap methods is a case where an initial and exploratory data analysis has led to a certain non-standard and situation-specific summary and description of the data, and the desire is to get an indication of the variability and reliability of the description. Let us ask what kind of tool is congenial to the generality of the bootstrap approach. Obviously, rather than a collection of standard algorithms, a system environment which includes the following features would be very helpful: a function that generates bootstrap samples for any set of real data and a facility that keeps an editable and re-executable record of the steps of an interactive exploratory analysis. The latter can be used to repeat the analysis on the bootstrap samples. Such record keeping facilities seem to be highly important for non- standard interactive statistical work. Among the existing software for PC-level it seems to be mainly PC-ISP that gives full support for this.

A major didactical perspective on Monte Carlo simulation has always been that its use, instead of analytical methods, may permit more interesting, realistic and complex applications of probability and statistics. Moreover, because of the generality of Monte Carlo simulation and its similarity to real experimentation, advantages for supporting students understanding of probability have been claimed (see Biehler 1988b). The new developments give new impact and support for such an approach.

6. STATISTICAL MODELLING

In many domains, computers have made it possible to generalize and extend "mathematical modelling" and its range of applica-

tions to "computational" or "symbolic modelling" using the new symbol systems of computer science to represent and study models. There are some further aspects that are particularly important in statistics:

(1) Monte Carlo simulation permits to take a wider class of probability models into consideration (non-Gaussian distributions);

(2) A much larger variety of functional dependencies can be handled (beyond linear models and elementary functions);

(3) intertwining the construction and validation of models more deeply with the analysis of data is a new possibility: new techniques for developing initial models through exploratory data analysis and techniques for diagnostic checking of provisionally assumed models have been developed;

(4) new emphasis is placed on-stochastic aspects of modelling: the functional dependencies and the question which variables are to be included in a model can get a more adequate treatment in multivariate situations ; descriptive models of single data sets attain more importance than stochastic models of chance mechanisms.

We will briefly look at one example. Although the general concept of non-linear regression (conditional expectation) was developed decades ago, computers have given new operative power to this concept. A good illustrative example is the "lowess" algorithm (the locally weighted regression scatterplot smoothing, see Chambers et al. 1983, pp.94). The lowess algorithm helps in situations where it is not reasonable to fit a predetermined mathematical function to a scatterplot of two variables x and y because the type of functional dependence can neither be derived from a theoretical hypothesis nor does the scatterplot show a clear visual structure. A fundamental operation would be to sclice the scatterplot into parts

and determine an "average point" for each part. The curve of the average points gives an indication how y changes with x on the average. Lowess is just a sophisticated version of this basic idea. For a large but finite number of points x_i in the range of x, lowess computes an "average point" y_i from considering the neighbourhood of the x_i's. The sequence of (x_i,y_i) represents a quasi-continuous functional relation which will be further studied in a graphical display together with the data. Replacing the fitting of analytically specified models by such smoothing methods seems to be a very general feature of recent change. A closer comparison to the classical method of least squares with regard to theoretical and historical background, computational effort, style of application and algorithmic representation would reveal many interesting aspects of the change of statistics, which are beyond the scope of this paper. Describing the change as a change in technique, however, would be inadequate, because algorithms like lowess unfold their practical power only in a working environment. In such a context it is possible to see the result of lowess graphically, to experiment with several parameters of lowess which yield curves of varying smoothness. Moreover, as lowess is often only an intermediate step, an environment will support follow-up activities in the process of modelling as a whole.

7. CONCLUSIONS

The objective of this paper was a brief analysis of some changes in statistics which are related to the spread of the computing technologies. Multivariate data, statistical graphics, Monte Carlo simulation and statistical modelling with and without computer support are changing domains that are relevant to statistical educa-

tion. Changes in emphasis, in what is considered fundamental and simple, in what is general and what is particular occured in all these domains. They question what is considered as basic knowledge and activity in statistical education. The new role of working environments for supporting statistical and mathematical activity, the mutual penetration of techniques, concepts, representations and tools from both computer science and mathematics for the benefit of a practical concern like statistics will certainly be relevant for reconsidering the perspectives of school mathematics as a whole.

REFERENCES

[1] BIEHLER, R., *Explorative Datenanalyse - Eine Untersuchung aus der Perspektive einer deskriptiv-empirischen Wissenschaftstheorie.* IDM-Materialien und Studien Bd. 24, Bielefeld: Universität Bielefeld, 1982

[2] BIEHLER, R., *Educational perspectives on Exploratory Data Analysis.* In: Morris, R. (ed.): Studies in Mathematics Education. Vol. 7: Teaching Statistics. Paris: UNESCO 1988

[3] BIEHLER, R., *Changing conceptions of statistics: a problem area for teacher education.* To be published in: Proceedings of the ISI Round Table Conference, Törökbalint, Ungarn July, 1988

[4] BIEHLER, R., *Computers in probability education.* IDM Occasional Paper 108. To be published in: Kapadia, R. (ed.). Chance Encounters - Probability in Education: A Review of Research and Pedagogical Perspectives. Dordrecht: Reidel 1989

[5] CHAMBERS, J.M., *The new future of data analysis*. In: Proc. 44th ISI session, 97-103, 1983

[6] CHAMBERS, J.M.; CLEVELAND, W.S; KLEINER, B.; TUKEY, P.A., *Graphical Methods for Data Analysis*. Belmont: Wadsworth and Boston: Duxbury Press, 1983

[7] CLEVELAND, W.S., *The Elements of Graphing Data*. Monterey: Wadsworth, 1985

[8] DIACONIS, P.; EFRON, B., *Computer-intensive methods in statistics*. In: Scientific American 248, 96-110, 1983

[9] FISCHER, R., *Offene Mathematik und Visualisierung*. In: mathematica didactica 7, 139-160, 1984

[10] GNANADESIKAN, R.; KETTENRING, J.R., *Statistics teachers need experience with data*. The College Mathematics Journal 19 (1), 12-14, 1988

[11] TUKEY, J.W., *Exploratory Data Analysis*. Reading: Addison- Wesley, 1977

Software

DATADESK; Velleman, P.F. & A.Y., Data Description Inc. Box 4555, Ithaca, NY 14852. Apple Macintosh. Student version: Kinko's Academic Courseware Exchange, 4141 Stock Str., Santa Barbara, CA 93110, USA.

MACSPIN Graphical Data Analysis Software; Donoho, A.W. & D.L., Gasko, M., D2- Software Inc. Austin Texas, USA, Apple Macintosh.

PC-ISP (PC Interactive Scientific Processor); Artemis Systems Inc., 1985, New York/London, Chapman and Hall, IBM-PC and compatibles S - cf. Becker, R.A./Chambers, J.M. (1984). S: An Interactive Environment for Data Analysis. (for UNIX systems)

STATVIEW 512+; Brainpower Inc., 2400 Ventura Boulevard, Calabasas, CA 91302, USA, Apple Macintosh