EXPLORATORY DATA ANALYSIS AND THE SECONDARY STOCHASTICS CURRICULUM

Rolf Biehler
Institut für Didaktik der Mathematik (IDM)
Universität Bielefeld – Federal Republic of Germany

Introduction

The emergence of Exploratory Data Analysis (EDA) (cf. Tukey 1977 as a main reference) presents a challenge to more traditional views, attitudes and value systems of statistics, which are often also the implicit basis of curricula and teaching approaches to statistics and probability. Simple examples, ideas and techniques of EDA are sometimes considered to be a new curriculum content, because it is hoped that they may replace the rather boring teaching of techniques of descriptive statistics by more interesting examples of real data analysis, where the students may become more actively involved in processes of discovering relevant features of the systems the data refer to (cf. e.g. Gnanadesikan et al. 1983, Landwehr/Watkins 1986).

More generally, the following paper will present some tentative ideas for the discussion on how the further evolution of curricula may react to changes in statistics which have become visible by the emergence of EDA. Problems, obstacles and new opportunities of development on the level of more or less "philosophical" conceptions of statistics, probability, and EDA will be discussed and considered as an important factor if there is an attempt at relating ideas from EDA to other areas of the curriculum, and to develop a more conscious attitude of teachers towards their subject. The views presented are based on detailed historical and epistemological studies of EDA and its relation to classical statistics (cf. e.g. Biehler 1982, 1985).

1. New appreciation of graphical representations

The new use of graphical representations as exploratory tools for statistics is probably the most important aspect which can be learnt from EDA, and which may radiate into other curriculum areas.

EDA is related to a more general movement in statistics towards using graphs as tools in research, as tools for analyzing data and not only as a means of communicating "the obvious to the ignorant". This latter attitude towards the role of graphs in statistics has been of long standing. Statistics saw its scientific character, among other things, in contrast to the unscientific age of mere graphical and descriptive statistics. Thus, even E. Pearson felt some need to defend his appreciation of graphics.

> *By suggesting that much can be learnt in this way (by graphs), I run the risk of being accused of encouraging slapdash methods of handling statistical data, a step against the tradition of 60 years' development of statistics as science.* (Pearson 1956, 143)

Such a somewhat negative view of graphics still dominates curricula and teacher attitudes. Especially attempts at teaching statistics which are concerned with a critical approach towards the use of statistics in mass media; conceive of graphics predominantly in terms of deception and "at best" in terms of communication, but not in terms of graphs as the most important device for discovering unexpected phenomena and structures in data, i.e. the attitude of EDA which J. Tukey already formulated as a program in 1962 (cf. Tukey 1962, 49). One illustrative example is the different evaluation of scale transformations, particularly, zooming in parts of graphs. Whereas Tukey (e.g. 1977, 146) uses this as a "powerful microscope" to see detailed structure, Huff (1954, 39) views the same possibility as a potential source of deception.

A most fundamental idea underlying EDA is that varying the representation and using multiple representations of data is a means of developing new knowledge or insights. This can be exemplified by switching from tables to graphs, from lists of numbers to stem-and-leafs, by reducing numbers to a discrete variety of symbols in statistical maps to make easy the exploration of overall structure, by constructing summary displays of batches like box plots which make possible an effective comparison of several batches.

Experiencing graphics in this spirit may also contribute to a critical understanding the Huff tradition is aiming at. The reason that graphical representations may be consciously experienced as tools which help placing the focus on particular aspects of the data, and do not merely present data "as they are".

From this perspective, a different and more conscious view may be taken of such classical displays as histograms. Historically, using "discrete" bar chats to present continuously varying data was quite an achievement and a discovery. Today, however, histograms are usually not introduced as tools to reveal structures in data which would be difficult to perceive with other tools, e.g., with plots, where the data have simply been plotted on a line. Instead, they are introduced as graphical representations for the mathematical notion of "frequency distribution" with a view towards probability.

EDA has developed several new formats of representation. Stem-and-leaf displays and boxplots are often rendered prominent with regard to the secondary curriculum.

The traditional curriculum on descriptive statistics may be transformed in the direction of EDA by using these displays in an investigative spirit (cf. the textbook by Landwehr/Watkins 1986). It would be essential, however, to give substantial support to the investigative attitude against the tendency of most didactical transpositions (Chevallard 1985) to reduce knowledge to techniques.

On the other hand, as already practiced in some books, these displays, or slight modifications of them, can also enrich those parts of curricula where probability and inference is the main goal. Visualizing chance variability by drawing several random samples and representing them by a collection of box plots may be such an application. Thus, the displays can be used to initiate a fruitful comparison, interaction and restructuring of experience

had with these diagrams in an exploratory setting, and in the context of ideal chance variability (cf. also Landwehr/Swift/Watkins 1984) who use box plots in an interesting way to teach confidence intervals.

The same displays and others may be useful in combination with statistical tests. Students can experience how test results are related to visual impressions of variability and difference. How this should be done is very much dependent on the conception of the statistics adopted. In R.A. Fisher's "Statistical Methods of Research Workers" (1925) we find the following remarkable passage:

> *The preliminary examination of most data is facilitated by the use of diagrams. Diagrams prove nothing, but bring outstanding features readily to the eye; they are therefore no substitute for such critical tests as may be applied to the data, but are valuable in suggesting such tests, and in explaining the conclusions founded upon them.* (quoted after Pearson 1956, 127).

From the rigid interpretation of classical statistics which emerged later, this use would not be acceptable: Hypotheses and test criteria should be chosen before looking at the data because of problems of multiplicity. From an EDA point of view, the practice suggested is maintained and extended, while at the same time admitting that the nominal significance levels cannot be interpreted in the usual sense. With real data, graphs are used to check and modify assumptions on which tests are based. Graphs are considered as containing potentially more information than is detectable by the test. From this perspective, it would be one-sided only to emphasize, with simulated random data, that graphs show all sorts of fine structure although there is "really" random noise.

Discussing these problems in detail is certainly beyond the secondary level, and it is an open problem to be reconsidered in statistical education to what degree two distinct "ethics" of classical statistics and EDA can, perhaps gradually, be developed in secondary education

2 Going beyond the one-dimensional case

Most secondary curricula on probability and statistics teach only one-dimensional data resp. random variables. If two-dimensional problems are taught at all, they receive a rather technical treatment and two-dimensional data mostly appear as if they belonged to a completely different ontological category than one-dimensional data. But, actually,

> *Most bodies of data involve observations associated with various facets of a particular background, environment, or experiment. Therefore, in a general sense, data are always multivariate in character* (Gnanadesikan 1977, 1).

The emergence and present application of EDA are deeply related to the emergence and dissemination of multivariate methods in today's statistical practice. Which role multivariate data may play in secondary education is an open question.

Multivariate data have already entered some schools through the backdoor of computer science courses where data bases are being explored. Besides, data security is topical and may gradually destroy the image of statistics and probability which is taught in the statistics classroom, that is that statistical data analysis is not concerned with individual cases but with "laws of average".

EDA presupposes going beyond this conception, making this possible at an elementary level at the same time. If box plots are used to compare several batches of data, this may be a first easy step in the multivariate direction. Also, stem-and-leaf displays do not erase the individual data values, but show a structured portrait of them. This facilitates recalling the object which belongs to the respective data values. This possibility can be extended by adding labels to some of the values in the plot, or by substituting the digits in the leaves by appropriate labels, or by symbols for the value of some second variable. These features constitute quite a different relation to data. To give an illustrative example: a medical practitioner once said that he like stem-and-leaf displays to show the values of an important medical variable for his patients, because they allowed him to more easily see the individual case and its relative position behind each number.

In more technical terms, stem-and-leaf displays take into account that all data are multivariate in character. They mediate between the one-dimensional and the multi-dimensional cases. This is also exemplified by the general heuristics to look for "explanations" (i.e. further potentially influential variables) if structures like outlies, gaps, popular values or several peaks have been discovered in a display.

To sum up, one-dimensional data are treated within an open context of potentially relevant further variables. This "ontology" of EDA is different from the common approach to probability. Usually, probability is introduced by using random devices like dice or coins where students are to learn that nothing can be predicted and that the variability of outcomes cannot be "explained" or related to other variables – in contrast to deterministic situations. There are certainly several good reasons in favor of beginning probability instruction with almost ideal random situations. But keeping totally random situations and totally deterministic situations strictly separated during the whole curriculum is certainly not desirable, because most real situations are a mixture of both. EDA is mainly concerned with such intermediate situations, where it is not clear at the beginning which aspects of the data should be interpreted or treated as random. EDA should be viewed as a new or further opportunity to bridge the gap between the two extremes of determinism and complete randomness.

In particular, it may be misleading to interpret difficulties of students, e.g. that they suspect relations between outcomes of a random experiment and other variables like specific spatio-temporal circumstances or operator skill, to be some type of "magical thinking" which ought to be extinguished in favour of appreciating probability and randomness. Perhaps the students' thinking can be interpreted as viewing the situation as an "open multivariate situation" similar to the thinking in EDA. Their inclination to seek connections may not be false in principle, but perhaps not appropriate in some situations. In fact, a recent philosophically oriented introduc-

tion to a collection of articles on EDA has the title: "Theories of Data Analysis: From Magical Thinking Through Classical Statistics" (Diaconis 1985), and the author draws close relations between EDA and magical thinking, defending at the same time this type of thinking, if it is controlled to a certain extent.

3. Reconsidering educational conceptions of statistics

It is often considered an important goal of statistics instruction to overcome the public image that "anything can be proved with statistics" ("lies, damned lies, statistics"). Going beyond merely describing data and beyond an arbitrary use of data to support arguments is often considered to be one of the main goals of statistics instruction. Teaching rudimentary ideas of confidence intervals and significance testing seems to be essential, because expressing "amounts of uncertainty" is often neglected in public debate. Also, students can, hopefully, learn from this in what sense and under what conditions something can be "proved" with statistics. If statistics instruction is predominantly viewed from this perspective, including ideas from EDA may seem to be rather counterproductive. Although it is true that EDA also goes beyond mere description, it does not do so, however, in the direction of inference and "proof", but rather in the direction of encouraging new hypotheses.

EDA does not produce the seemingly unequivocal, precise and "final" results of inferential statistics, but often a multiplicity of rather vague aspects with varying degrees of uncertainty, and the result obtained may be in part contradictory. Tukey calls such results cautiously "indications". The openness, the subjective and the hypothetical dimensions of scientific research and of knowledge, are not suppressed, but clearly shown. What a diagram may mean, or in what direction further research should proceed, requires communication and discourse. The subject matter expert is not "compelled towards insight" by a result of EDA, but accepted as a partner in communication. Communication is necessary between data analysis experts and subject matter experts, and within each group in order to bring together different experience to a common benefit. Placing a new emphasis on such features of applied mathematics within the curriculum is often considered educationally very important. Fischer (1984) coined the term "open mathematics" for this, and views EDA as a prototype (cf. also Fischer/ Malle/Bürger 1985).

For further clarification, let me refer also to a paper given at ICOTS I by John Bibby (1983, 241). He emphasized three "tensions" that had to be resolved while developing the Open University course "Statistics in Society". The first was the tension between statistics as an "exact science" (objective, rigorous, culture-free, technique-oriented) and statistics as a "social product" (produced as the outcome of human responses to a wide variety of conflict-laden situations). The third one was the tension between EDA/descriptive statistics and inferential statistics. Perhaps one should describe the two related tensions also somewhat differently. Statistics and data analysis are essentially also social activities. Communication and co-operation has a prominent role, in particular with regard to attaining truth and objectivity, a goal which cannot be attained by strictly obey-

ing logical rules for dealing with data alone. The question of whether teaching EDA seems to be as to what extent these features would and could have a place in the curriculum e.g. by means of classroom discussions, in contrast or in addition to teaching an idealized view of "statistics as an exact science".

Of course, such an attitude can to a certain degree also be practiced in connection with inferential statistics. But this would probably presuppose a different attitude, for instance, towards statistical tests and other methods, perhaps more in R.A. Fisher's spirit than in the spirit of the decision-theoretic interpretation of statistics.

References

Bibby, J. (1983). An Open University service course. In ICOTS I (pp. 238-249).

Biehler, R. (1982). Explorative Datenanlyse – Eine Untersuchung aus der Perspektive einer deskriptiv-emprischen Wissenschaftstheorie. IDM-Materialien und Studien Bd. 24. Bielefeld: Universität Bielefeld.

Biehler, R. (1985). Die Renaissance graphischer Methoden in der angewandten Statistik. In H. Kautschitsch & W. Metzler (Eds.), Anschauung und mathematishce Modelle (pp. 9-58). Schriftenreihe Didaktik der Mathematik UBW Klagenfurt Bd. 13. Wien: Hölder-Pichler-Temsky/Stuttgart: B.G. Teubner.

Chevallard, Y. (1985). La Transposition Didactique: Du Savoir Savant au Savoir Enseigné. Grenoble: La Pensée Sauvage.

Diaconis, P. (1985). Theories of data analysis: From magical thinking through classical statistics. In D.C. Hoaglin, F. Mosteller, & J.W. Tukey (Eds.), Exploring Data Tables, Trends and Shapes (pp. 1-36). New York: Wiley.

Fischer, R. (1984). Offene Mathematik und Visualisierung. mathematica didactica 7, 139-160.

Fischer, R., Malle, G., & Bürger, H. (1985). Mensch und Mathematik: Eine Einführung in dadaktisches Handeln und Denken. mannheim: BI Wissenschaftsverlag.

Gnanadesikan, R. (1977). Methods for Statistical Data Analysis of Multivariate Observations. New York: Wiley.

Gnanadesikan, R., Kettenring, J.R., Siegel, A.F., & Tukey, P.A. (1983). Symposium on Exploratory Data Analysis. In M. Zweng et al (Eds.), Proceedings of the Fourth International Congress on Mathematical Education (pp. 344-357). Boston: Birkhäuser.

Huff, D. (1954). How to lie with statistics. London: Gollancz.

Landwehr, J.M., Swift, J., & Watkins, A.E. (1984). Information from samples. Prelim. version Quantitative Literacy Project.

Landwehr, J.M., & Watkins, A.E. (1986). Exploring data. Palo Alto: Dale Seymour Publications.

Pearson, E.S. (1956). Some aspects of the geometry of statistics. Journal of the Royal Statistical Society, Series A 119, 125-146.

Tukey, J.W. (1962). The future of data analysis. Annals of Mathematical Statistics 33, 1-67.

Tukey, J.W. (1977). Exploratory data analysis. Reading (Mass.): Addison-Wesley.