

# **Probabilistic thinking, statistical reasoning, and the search for causes - Do we need a probabilistic revolution after we have taught data analysis?**

Revised and extended version of a paper presented at the Fourth International Conference on Teaching Statistics (ICOTS 4), Marrakech, Morocco, 25-30 July 1994

Rolf Biehler  
University of Bielefeld, Germany

## **1. Introduction**

The relation between probability and statistics in a curriculum is influenced by the cultural context (Holmes 1994). The conceptual difficulties with learning probability are sometimes used to recommend a curriculum emphasizing statistics with minimal probability (Garfield and Ahlgren 1988). On the other hand, statistical data analysis imports many more foreign elements into a mathematics curriculum than probability does, which may be a reason to recommend a concentration on probability. With regard to general education, most people agree that - if we teach probability - we should teach applied probability, i.e. emphasizing how probability can be used to model and understand some part of the real world. Therefore, questions concerning the meaning of probability are very important. How are probability models related to deterministic models and to thinking in terms of causal relations?

However, to which extent will the idea of a statistics curriculum with minimal probability be reasonable? Tendencies in statistics towards relativizing the role of probability seem to reinforce this inclination in favor of a minimum curriculum. However, there are indications that this may contribute to new learning difficulties, and we will raise the following basic questions to which the paper will provide some first answers and reflections:

(1) To what extent are learning difficulties with probability influenced/caused by a too simple (pure, objectivistic) educational conception of probability?

(2) To what extent are novices' difficulties with a successful use of data analysis methods influenced/caused by a too simple (probability-free) educational conception of data analysis?

## **2. Two cultures of thinking: Exploratory data analysis and probabilistic thinking**

My basic thesis is that two different cultures of thinking exist: outside school, and inside school as a basis for different courses or curricula. I will speak of probabilistic thinking on the one hand and, as a shorthand, of EDA thinking on the other hand. The latter will mean the way of thinking in Exploratory Data Analysis (EDA), which I consider to be a certain school in statistics (Tukey, 1977). The philosophy of EDA has made explicit several implicit orientations underlying the practice of data analysis. In this sense, EDA thinking is far more wide-spread than just within a small community.

An example of a curricular separation of probability and data analysis is the Quantitative Literacy Series, where the EDA part (Landwehr, 1986) can be used independently of the probability part (Newman, 1987) and vice versa. Subsequently, knowledge of both parts are

New address of the author: University of Kassel, [biehler@mathematik.uni-kassel.de](mailto:biehler@mathematik.uni-kassel.de). This paper was originally published in: Garfield, Joan (ed.) Research Papers from ICOTS 4, Marrakech, 1994. University of Minnesota.  
Available at: <http://www.mathematik.uni-kassel.de/didaktik/biehler/publications>

combined in the service of inferential statistics (Landwehr, 1987). There are other materials that are restricted to the EDA point of view (Ogborn, 1991) and most German textbooks on "stochastics" could well serve as embodiments of a probabilistic approach, whose fundamental ideas have been described by Heitele (1975) Several attempts have been developed to integrate the various aspects into a coherent whole. This paper favors a complementarist approach that assumes that data analysis and probability cannot or should not be reduced to one another, but rather that their relations have to be developed. We interpret David Moore's (1990) emphasis on *data* and *chance* as the two sources or objects of a curriculum in this sense, and we will come back to his approach later. We consider the complementarist approach superior to reductionist approaches or approaches that merely emphasis mathematical similarities between concepts in probability and data analysis (expected value - mean; probability - frequency etc.).

Since the 70s of this century, we witness a "data analysis revolution" that is partly anti-probabilistic. Tukey (1972, p. 51) expresses the attitude to probability in the American EDA tradition as "'Data analysis' instead of 'statistics' is a name that allows us to use probability where it is needed and avoid it where we should. Data analysis has to analyze real data." The French tradition in data analysis is somewhat more extreme: "*1<sup>st</sup> principle*. Statistics is not probability. Under the name of mathematical statistics, the authors have erected a pompous discipline which is rich in hypotheses that are never satisfied in practice" (Benzecri, 1980, p.3, my transl.)

As a first approximation, the basic difference can be expressed as follows: Probabilists seek to understand the world by constructing probability models, whereas EDA people try to understand the world by analyzing data. Although both partly need the methods and concepts of the other, a tension exists. This tension is not new. The history of probability and statistics has seen different kinds of relations and roles of probability in data analysis, and vice versa.

There was an elucidating debate in social statistics in Germany in the 20s of this century, which I would like to refer to. The discussion is interesting because it was not confined to technicalities, but included questions of causal analysis, and because it gave rise to an appreciation of dualistic goals in statistics.

Flaskämper (1927,1929) starts from a definition of statistics implying that the application of statistics is appropriate for those cases where the law of large numbers holds. These are situations with some stationarity and homogeneity. He diagnoses that this holds for many situations in the natural sciences, but not so often in the social sciences. (The idea of deliberate random sampling from populations was just under development and was mentioned as a special case.) The question was: Which social processes can be treated as random experiments where the law of large numbers holds? Where do we have time stability of averages? A paradigmatic case is the proportion of male and female births. Flaskämper criticizes the research program of social statistics that started with Quetelet to look for time stable averages (for example of suicide and death rates etc.). In this tradition, time stable indicators were used to compare and explain differences in these indicators within a population (for instance with regard to social stratification).

A basic metaphor taken from the 19th century statisticians was the idea of a system of constant and variable causes that influence an event. The law of large numbers holds if the variable causes cancel each other out and the effect of the "constant" causes reveals itself only

with large numbers. From this perspective, a theory of causal inference was constructed. Causes for individual events like a person's death or the reasons for specific accidents were distinguished from factors that influence a whole group of objects such as health conditions, which in turn influence death rates and the distribution of death causes (Zizek, 1928) The theory included the "dissection" of "collectives". A mean or proportion of the entire collective is dissected into a mean or proportion of each of the subsets of the dissection. The dispersion of this set of numbers can be accounted for by probability theory (a "homogeneous group"), or if not, is an indication of the influence of the variable that was used in the dissection. Statisticians, however, were already aware of the problem of confusing influential variables if the subsets were different not only in the discriminating variable, but also with regard to other variables.

Flaskämper argues for a liberation from probabilistic prejudice in favor of dualistic goals and objects of statistics. Situations where the law of averages holds and probabilistic ideas can be applied, and other situations where goals are more descriptive in nature. Interest can focus on the time dependence of social indicators (instead of an time stability). Furthermore, new summaries should be invented and applied to describe data distributions, because the mean as a summary may often not be appropriate. The average (the mean), having worn an ontological decoration of indicating the "real" and stable law behind random fluctuation, has become a summary statistics that has to hold its ground against competitors.

Dualistic goals in statistics have also been recognized in the recent data analysis revolution in the shape of exploration and confirmation. Also, the door was opened for new kinds of summaries and methods that need not be derived or justified from probabilistic assumptions.

Data analysts are criticized by probabilists for looking for patterns where none exist except patterns produced by chance. Probabilists, however, are accused of a somewhat complementary offense: "The stochastic spirit says, 'Are there auto accidents? Well, tough. Cover yourself from head to toe with insurance.' The deterministic spirit would say, 'Analyze the causes. Make changes. Pass laws. Get the drunks off the road.' Prudence says: do both. " (Davis, 1986, p. 20. Davis and Hersh criticize - in other words - that variation is described or modeled by probability models, and that questions of causal explanation are ignored.

Speaking of two cultures of thinking is related to considering "paradigms" in Kuhn's sense (1970) as constitutive for science. Among others, a paradigm contains techniques and methods, world views, attitudes and exemplars. Exemplars are prototypical examples showing to which cases and how a theory is applied. People who try to characterize EDA often use similar elements to describe essential features of their approach, emphasizing that technique alone is not sufficient for such a characterization. Probabilists do not form such a clear-cut group with shared convictions: On the surface, we find the basic split into personalists (subjectivists) and frequentists (objectivists); beneath that surface, a rich structure of different meanings can be reconstructed in history and in current practice.

The probabilistic developments in the sciences in the period between 1800-1930 have been summarized as a Probabilistic Revolution (Gigerenzer, 1989; Krüger, 1987 ; Krüger, 1987). This diagnosis of a Probabilistic Revolution is an interesting counter part to the data analysis revolution under way since the 1970s.

The Probabilistic Revolution has rendered probabilistic models indispensable in science. Quantum mechanics is often quoted as an example of a theory with an inherently indeterminate view of natural phenomena. But this ontological indeterminism, the conception of the irreducibility of chance is a much stronger attitude. The need for probability models can be attributed to human ignorance, combined with a hope that further hidden variables may be found that can explain more of the variability. However, the essence of the probabilistic revolution was the recognition that in several cases probability models are useful types of models that represent kinds of knowledge that would still be useful *even when further previously hidden variables were known and insights about causal mechanisms are possible*. In these cases, we can speak of an epistemological irreducibility of chance, leaving open the question of an ontological indeterminism.

Depending on the case, seemingly deterministic laws of the macro level can be interpreted as consequences of the law of large numbers of an inherently indeterminate micro level. In other cases, the traditional view of the determinists in the 19th century is plausible: Variation is modeled by causal factors; constant and variable causes act inseparably together and produce a result. In the long run, variable causes cancel each other out and the result of the constant causes becomes detectable. That an appreciation of probability models can be seen as independent from a conviction that chance is irreducible, or not, is reinforced by the recent debate on deterministic chaos (Berliner, 1992; Chatterjee, 1992) A major point is that the ontological debate of whether something "is" deterministic or not may not be useful, rather, a situation can be described with deterministic and with probabilistic models and one has to decide what will be more adequate for a certain purpose.

The possibility that two different, partly incompatible ways of thinking may exist is relevant for evaluating children's thinking. When children approach a situation regarded as a truly and purely random experiment by experts, they may nevertheless try out strategies and concepts whose origin are causal schemata. It would be too easy to see only misconceptions in this approach. This judgment of a misconception assumes that there is only one normative solution or one correct understanding of a situation, namely by means of a probability model.

We will relate and contrast the two approaches with regard to the following dimensions:

- *seeking connections* vs. guarding against chance artifacts, seeking connections in single cases vs. long run regularity;
- *explaining and describing variation* by causal and other factors, by probability models;
- *unique data sets vs. data generating mechanisms*;
- *look at the data or work with the model*.

Exaggerating the difference aims at working out problems and obstacles more clearly.

### **3. The culture of EDA thinking**

What are some of the highly valued goals and attitudes of EDA? *Seeking connections*. Looking for patterns in the world is a natural behavior of human beings. Diaconis (1985) relates this tendency to EDA: "...we can view the techniques of EDA as a ritual designed to reveal patterns in a data set. Thus, we may believe that naturally occurring data sets contain structure, that EDA is a useful vehicle for revealing the structure, and that revealed structure can sometimes be interpreted in the language of the subject matter that produced the

data"(p.2). If we do not check this structure, we may come close to magical thinking which we can define as "our inclination to seek and interpret connections between the events around us, together with our disinclination to revise belief after further observation"(p.1). A special argument by probabilists is that science should protect itself against so-called chance artifacts: It is claimed that people tend to see "structure" where there is really only chance, and that may mislead science.

Seeking connections is especially relevant in complex multivariate data sets. The EDA approach risks that many more connections are discovered than can be maintained by the proof of further experience. But progress in science comes first of all from exploration, and it is quite natural that not all hypotheses survive being tested by further data. Experts in EDA are generally aware of the limitations of exploratory results, and what kind of remedies are available for avoiding false and rash conclusions (Diaconis, 1985). There is the danger that non-experts misunderstand EDA as "anything goes in data snooping".

*Explaining and describing variation.* In the world of traditional descriptive statistics, the description and summarization of univariate distributions is a major topic. Histograms for continuous and bar charts for discrete variables are used, as well as several numerical summaries. At first glance, "EDA plots" such as stem-and-leaf displays and box plots seem to do nothing more than to enrich this repertoire. Numerical summaries and the distribution are not properties of any single case, but have to be interpreted as a property of the group of objects (as a whole) to which the numbers refer. This standpoint is a source of difficulties: "For a student to be able to think about the aggregate, the aggregate must be 'constructed'" (Hancock, 1992, p. 355). Individual-based reasoning is much more common, a phenomenon that we also discovered in our teaching experiment on EDA (Biehler, 1991).

Teaching descriptive statistics in the spirit of EDA, however, aggravates this problem. If we have a set of screws with diameter as a variable, individual cases are not very interesting. In contrast, in simple applications with univariate data, EDA influenced courses emphasize to notice outliers and look for explanations, generally some discriminating factor that makes a difference between the bulk of the data and the outlying observation. Other examples are an observed separation into groups that may also be explained by discriminating variables. The data to which elementary examples refer may be a class of cities or regions, or people, with some interest in the individuals themselves. When students explore how much pocket money they get or how long they travel to school, it is quite natural to explain differences between individuals by specific circumstances, and to be interested in an individual as related to the group (which rank does a certain person have in a collective?).

Introductions into EDA often use data sets that already contain several variables. An "analysis by group" is emphasized - taking the box plot as an important elementary tool for comparing groups. This can also be understood as looking for decomposing variation by an explanatory variable. Explaining variation can come down to the level of individual cases (objects, events, persons), which may explain a certain exception or difference. EDA plots are influenced by this approach: It is easy to identify the objects related to numbers in stem-and-leaf displays, outlying values are especially coded in box plots, interactive graphs support the identification and localization of groups and points. Whereas the object of usual statistics is conceived of as a collective, ensemble, or group, EDA is more open and groups have still to be formed and reshaped.

Looking for associations and relations between variables is one important goal of elementary EDA, and the use of scatter and line plots as well as two-way tables is a natural extension of questions students confront after having explored single variables.

As "interpretation" and "generating hypotheses" is considered as important, further variables or aspects of situations get involved. An approach with a sterile separation into one variable methods, two variable methods and many variable methods would run against this emphasis.

*Single data sets.* Predominantly, EDA focuses in the first place on the current data set. At the beginning of an analysis, it is not assumed that data are a sample from some clear-cut larger population, and generalization is tentative. The process by which the data have been generated may be very complex, heterogeneous and to a large extent unknown. As an extreme case, EDA is concerned with whole populations or unique data sets.

*Maxim: Look at the data.* This maxim comprises two meanings: *Look* at the data - implying the esteem of graphical displays and human pattern recognition capabilities - and *look* at the *data* - meaning a fundamental respect of real data. The latter is associated with a certain distrust in probability models or in modeling that is not related to real data. Oversimplified models may mislead the data analyst. In a sense, it is believed that data are "real", whereas models are tentative constructions. On the other hand, EDA people seem to appreciate subject matter knowledge and judgment as a background for interpreting data much more than traditional statisticians seem to do: Radical frequentists reject any knowledge except sample frequencies and the probability model.

#### **4. The culture of probabilistic thinking**

Interpretations of probability vary, and this causes teaching and learning problems. I have much sympathy with efforts to relate the various interpretations as parts of a general whole. Bayesians have claimed that their approach can be regarded as an extension of the classical frequentist approach, and an extreme position is de Finetti's (1974) argument that we can reinterpret a frequentist probability from the subjectivist point of view as a mere *façon de parler*. I would prefer Shafer's (1992) approach that turns de Finetti's upside down. The "special situation" of a repeatable experiment where we know long-run frequencies is the major reference situation. Degree of belief and support interpretations are also possible in this reference situation, and are here nothing but another aspect of the probability concept, which nevertheless has an objective meaning in the special reference situation. With regard to the situations we encounter in practice, the following can be stated: "Most applications of probability, including statistical inference, lie outside this special situation, of course. But we can think of these applications as of various ways of relating real problems to the special situation. Much statistical modeling amounts to using the special situation as a standard of comparison. Statistical arguments based on sampling or randomization depend on artificially generated random numbers which simulate the special situation. Bayesian analyses are arguments by analogy to the special situation"(Shafer, 1992, p.103).

We will characterize some features of probabilistic thinking, mainly related to the above special situations, in other words with regard to the objectivist aspects.

*Don't seek connections.* The precise maxim is: Don't seek connections in individual cases - but concentrate on regularities and stabilities in many repetitions instead. The law of large

numbers is regarded as a research strategic maxim. A famous characterization of situations to which probability applies is John Venn's "individual irregularity with aggregate regularity": Society was considered as an example similar to the ensembles of gas molecules in statistical physics.

Overcoming "magical thinking" has become a general pedagogical objective associated with teaching probability. Animistic ideas, magical luck agents, mysterious "hot hands" in base ball playing have to be dispelled from people's thinking. Probabilistic thinking is the pillar of enlightenment: Chance is not magic but computable - at least to a certain extent, namely with regard to long run behavior. Risks become calculable and decision under uncertainty can be rationalized.

*Explaining and describing variation* by probability models. A fundamentally new idea for many students is the shift from individual cases to systems of events: Whereas a theory or explanation of single events is not possible or interesting, long run distribution can be modeled, predicted and partly explained by probability models. David Moore (1990, p. 99) writes that one "goal of instruction about probability is to help students understand that chance variation rather than deterministic causation explains many aspects of the world." A prototypical example is provided by the Galton Board. The major phenomenon is that more balls can be found in the middle (favorable middle). This aspect can be explained by a chance mechanism respectively a chance model: the probability of the ball to go left or right is 0.5 at every nail. Together with the assumption that a ball's turn to right or left is independent from the previous turn, we can conclude that all possible paths are equally likely, and that the larger number of paths leading to the middle explains the phenomenon. The favorable middle cannot be explained by concentrating on deterministic causation of individual ball's paths.

*Not single data sets but chance generating mechanisms and their probability distribution.* The focus of interest in the Galton Board like in other random experiments is modeling the chance generating mechanism. In surveys, where random samples of a population are drawn, the real interest is in the population, and not in the sample. Modeling the chance mechanism is equivalent to getting knowledge about the distribution in the population - the random selection process is imposed on reality.

*Maxim: Work with the model.* In a sense, the model (the probability distribution) is the deeper, although not completely known reality, of which the data provide some imperfect image. This attitude resembles Platonist ideas where every real square is only an imperfect image of the real perfect ideal square. However, the maxim has a research strategic component as well: For a certain domain, theoretical modeling is favored over doing just "data analysis". Another aspect may be the respect for the power of mathematical methods, because a mere empirical study of frequency data is of limited power: "Chance variation can be investigated empirically, applying the tools of data analysis to display the regularity in random outcomes. Probability gives a body of mathematics that describes chance in more detail than observation can hope to discover" (Moore, 1990, p. 118).

If we look back at these controversies and obstacles, we may become sensitive to learning problems that may be caused by a too simple adoption of one of the extreme positions: a purely probabilistic attitude or a probability-free conception of data analysis. We have seen that "experts" themselves have had problems of understanding. However, present experts in general are better off than high school students. Two types of knowledge may

contribute to this superiority. They know the different types of thinking and the interface between the two, and in what circumstances to use the one or the other, and they use metaphors and experience from one domain for the benefit of the other.

## 5. Do we need a probabilistic revolution in teaching?

Falk and Konold (1992) refer to the Probabilistic Revolution and develop the thesis that in "learning probability, we believe the student must undergo a similar revolution in his or her own thinking" (p.151). This thesis is formulated independently of the question whether students have learnt data analysis before. It would imply that students have already learnt ways of thinking that may become an obstacle for accepting the probabilistic point of view.

It is elucidating to quote some of the difficulties people had historically in adopting a probabilistic point of view with regard to life insurance, which now has become a standard application of probability. Annuities and insurances were calculated on the basis of intuitions "that ran directly counter to those of the probabilists. Whereas the dealers of risk acted as if the world were a mosaic of individual cases, each to be appraised according to particular circumstances by an old hand in the business, the mathematicians proposed a world of simple, stable regularities that anyone equipped with the right data and formulae could exploit...They [the practitioners] had to replace individual cases with rules that held only *en masse*, and to replace seasoned judgment with reckoning ... The practitioners equated time with uncertainty, for time brought unforeseen changes in these crucial conditions; the probabilists equated time with certainty, the large number that revealed the regularities underlying the apparent flux" (Gigerenzer, 1989, pp. 25).

Research with children shows that they partly behave similarly to the non-probabilistic practitioners of old. Konold (1989) provides some examples in his research, where his subjects showed belief systems that he calls the "outcome approach". They have two features "(a) the tendency to interpret questions about the possibility of an outcome as requests to predict the outcome of a *single trial* and (b) the reliance on *causal* as opposed to stochastic explanations of outcome occurrence and variability" (p.65). With regard to the meaning of "probability = 70% for rain tomorrow" in a weather report, subjects interpreted the 70% as related to a measure of the strength of a causal factor such as the percentage of cloud cover or of humidity etc. A frequency interpretation as the rate of successful predictions was less frequent. Subjects should determine the probabilities of various sides of a bone that should be tossed similarly to a die. Students expressed reservations whether additional trials can contribute to estimating which side is most likely to land upright. In some experiments, some students did not even use provided frequency information for the requested judgment. Others were convinced that reliable information could be obtained from careful inspection of the bone rather than from conducting trials. Variations in the outcome of trials were attributed to the way the bone was tossed.

Konold resumes "It needs to be stressed that a formal probabilistic approach does not necessitate the denial of underlying causal mechanisms in the case of chance events ... In practice, however, a causal description is often seen as impractical if not impossible ... accepting a current state of limited knowledge, a probabilistic approach adopts a 'black-box' model according to which underlying causal mechanisms, if not denied, are ignored. The mechanistic model is not abandoned in the outcome approach" (p.70) "...The application of a

causal rather than a black-box model to uncertainty seems the most profound difference between those novices and the probability experts and, therefore, perhaps the most important notion to address in instruction" (p.92).

This situation may become even more precarious, if students have passed a course on EDA with all the emphasis on explanation, interpretation, causal relations on various levels and hypothesis generation. Consciously changing the point of view will become even more important.

Another recent example, which Moore (1990) and Konold and Falk (1992) briefly mention in their pleading for probabilistic thinking, are runs (streaks) in basketball player failures which can be still explained by a binomial distribution with probability  $p = 0.7$ . There is no need to look for specific causes like nervousness and other things. There has been a scientific discussion on this case. Tversky and Gilovich (1989) make an analysis of empirical data showing that the binomial model well explains the data: in particular the chance of a success in a shot is independent from the previous shot. As fans always detect patterns (hot hands), their conclusion is the following: " Our research does not tell us anything general about sports, but it does suggest a generalization about people, namely that they tend to 'detect' patterns even when none exist." (Tversky, 1989, p.21) We find a controversial interpretation of the status of their results. Whereas Gould (1989) interprets the results as if a hot hand did not exist, Hooke (1989) is right in claiming that an inference like "If the model 'explains' the data, then the model is correct and unique" is not reasonable. However, some teachers and researchers seem to share Gould's misinterpretation. Although Tversky and Gilovich do not make such strong ontological claims as Gould, they claim the non-existence of "pattern" other than chance pattern, whereas people tend to think that they have to invent explanations for a non-random pattern. However, the fact that the null-hypothesis is not contradicted does not exclude alternative models. A series of successes may be explained by some factor, even when a binomial model well describes the variation. As we have learned above, the fact that a probability model fits a system of events well is compatible with causal dependence of the individual event. From this point of view, the surprise is that the causes act in a way that a binomial model fits the data well. However, it could be the case that one can identify a variable on the basis of which a better prediction than the binomial model would be possible. The hot hand illusion has similarities to the gamblers' fallacy. This has been emphasized by Falk and Konold (1992, p. 158). However, emphasizing this difference is also important: The assumption of independence in the gambling context is initially plausible, whereas it is not plausible in a basketball context. Similarly, several researchers assume for certain that coin flipping is to be modeled by a binomial model, and children's response is judged against this background. However, coin flipping can also be done in a way that independence has to be rejected in favor of serial correlation, and physical theories can be developed to explain some aspects of coin flipping (DeGroot, 1986)

We have to deepen this aspect. When Richard von Mises (1972) tried to build foundations for a frequentist theory of probability, one of his axioms was the "principle of an impossible gambling system": No system of betting is successful in improving the gambler's chances (for a precise formulation and a proof from the assumption of formal independence, see Feller, 1968, pp. 198). However, this does not rule out other possibilities of improving chance by observing variables from which the roulette result is not independent. Recently,

people constructed pocket computers and machinery with which some physical data on a Roulette wheel (velocity of the ball and the wheel, their positions) were gathered and processed in order to achieve a better prediction than just the uniform distribution (Brass, 1985).

At first glance, the latter possibility seems to contradict von Mises' principle. In history, a related epistemological obstacle can be identified. When Francis Galton wanted to understand the process of heritage of body height, for instance, he was confronted with the following enigma. Height was normally distributed in the population, and this was usually explained by an analogy to error theory. "If the normal curve arose in each generation as the aggregate of a large number of factors operating independently, no one of them of overriding or even significant importance, what opportunity was there for a single factor, such as a parent, to have a measurable impact? At first glance the beautiful curve Galton had found in Quetelet's work stood as a denial of the possibility of inheritance" (Stigler, 1986, p. 272).

A solution was found in modeling the system by common bivariate distributions and by inventing the concept of correlation and regression. The stability of the normal distribution can be explained as having been produced by a chance process, where each parent is the source of normally distributed sons and daughters where the mean is dependent on the parents characteristics. Thus, an independent mixture of normal distributions is produced.

These ideas were further developed later on. Statisticians became interested in "sources of variation" and in "explaining variation" (by factors). Analysis of variance and regression analysis have become the techniques to do that. Interestingly, it is the "unexplained variation" that is modeled by probability. The modern practice is flexible with regard to which variables are explicitly included in a functional model and which are left over as residual variation. In this context, probability models become a residual category instead of "explaining variance" themselves.

The lesson I would like to learn from that is the following. The metaphors and examples (the ontology) used for introducing the probabilists' point of view in school seems to date back to the times before Galton and not to have made the step towards several variables which require a new standpoint. It is not surprising that problems with cognitive obstacles may worsen, if teaching EDA, on the other hand, opens the mind for all kinds of relations and dependencies between variables.

## **6. Relating causal and probabilistic aspects**

If we assume an antagonism of probabilistic and causal analysis as above, a probabilistic revolution seems to be even more necessary in teaching when we have taught EDA before. However, this antagonism of causal analysis and probability models is an inadequate and misleading picture, as Steinbring (1980) has worked out with regard to teaching probability. And I will argue that, vice versa, learning EDA may even contribute to overcoming this antagonism.

That students should switch to appreciating frequency information *instead* of making physical analyses of bones and other random devices may be misleading. This becomes very clear when people are asked to construct a chance device. A good deal of physics must be known to construct a good "purely random" Galton board with independence and equal probabilities. If we would like to make a die with equally likely sides, physical and geometric

homogeneity is important. Nobody would construct such a die by trial and error and frequency analysis alone, namely reshaping bones and testing whether the frequencies are o.k. In general, a mixture of physical and frequency analysis is reasonable. Using partly symmetrical square stones with six sides as dice gives students the possibility to generate hypotheses. For example, the hypothesis that opposite sides have equal but unknown probability is plausible. An estimation of the three unknown probabilities can be achieved by experimentation, but people are right in not rashly giving up the symmetry constraint. Riemer (1985) has suggested this approach for introductory probability education.

The metaphor of constant and variable causes is also applicable to the Galton Board: The whole physical device represents the "constant causes", and the "variable causes" are responsible for the variation of the individual paths of the balls. One can become aware of constant causes when changing them; for example by changing the angle (to the left/right; forwards/backwards) of the Galton board with regard to the ground. Usually, the effect of the change cannot be observed in individual trials but manifests itself in different long run distributions.

This pattern of thinking has been called "statistical determinism" and can be expressed by the following graph in Fig. 1.

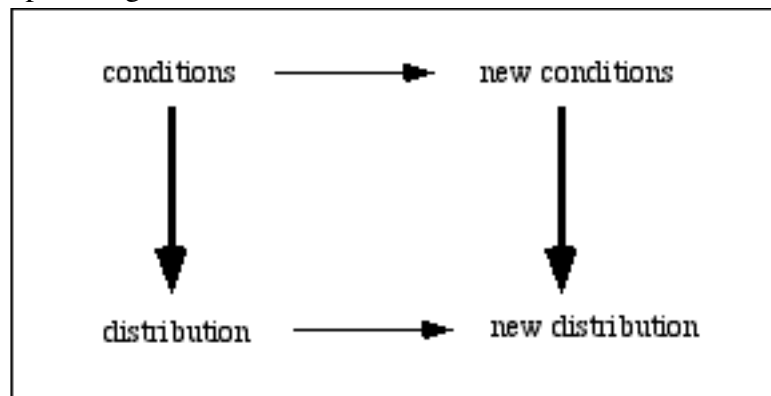


Fig. 1 Schematic view of statistical determinism: Conditions determine a certain distribution, a change of conditions results in a change of the distribution.

This scheme implies that a (probability) distribution is a property of something (a chance setup, a complex of conditions), and that this property can be influenced and changed by external variables.

The transition from an individual event to a system of events, which can be characterized by a distribution, has to be interpreted as a transition that can reveal new types of knowledge, new causes, explanations and types of factors that can not be detected at the individual level *also in many cases where a causal analysis of individual events would be informative*. These new factors may explain features of a group as a whole. Nevertheless, the perspective of the individual case should not be eliminated but maintained as a genuine complementary view of a situation. If the Galton board is analyzed in a classroom, it would be better to emphasize the possibility of explaining the favorable middle as a system feature by a chance model. Claiming that the path of a ball is really indeterministic by saying that the ball has "no memory" (a often-used metaphor for independence) may be misleading and run against the pupils' knowledge of physics.

We can view here several "levels of analysis", the level of *individual events*, of the *distribution in a system*, of the *background conditions*. These are levels that can be adopted by an analyst or not, depending on the purpose. These distinctions have been suggested as a means to improve students' understanding (Sill, 1992). It remains a problem that there is no general answer to the question, however, to which extent the system laws are relevant for the individual case. Playing roulette with its well-defined chance structure is much different from individual risk assessment, where no unique reference set (system) exists

In the context of chance devices, students tend to ask whether they can influence certain events: Is it possible that one can influence a spinner, a flipped coin, the path of a ball in a Galton board? Emphasizing individual irregularity in this case would not point ahead. Students should rather acquire a way of thinking about such situations as visualized in the following Fig. 1. The progressive way would be to point to the problem that we have to analyze whether the *distribution* has changed under the new conditions. Prototypical simple random experiments would be tilting a Galton board, changing to inhomogeneous dice etc.

The Galton board is used as a model device in education, and it is very important that students relate the above way of thinking to more realistic and relevant situations. How does a reduction of the maximum speed limit on a road affect the distribution of car velocity on that road? How is the distribution of diseases different near nuclear power plants? etc.

My thesis is that learning EDA can contribute to this way of thinking if data analysis examples are taught from this spirit. Empirical distributions can be compared under the old and under the new conditions. Let us take traffic accidents as another example: We can provide some causal explanations at the level of individual accidents in every instance. We can aggregate data for a longer period, however, and analyze how the number of accidents changed over the years, how different it is on weekdays and weekends, whether there is some seasonal variation. The aggregation makes changes in boundary conditions detectable, which may not be detectable at the individual level. Aggregating individual data or dissecting aggregated data are basic processes that gain their importance from the above perspective.

This practice of data analysis can contribute to similar ways of thinking with probability models. Vice versa, probability distributions model idealized data generating processes that may give some metaphorical explanation, for the practice of data analysis, even in those cases, where initial data exploration is not used for constructing (conjecturing) a probability model. Another point is the need for summarization to reveal structure. Regularities may show up only after summarization. This is emphasized in EDA in various ways. Why is this the case? A metaphorical understanding of the law of large numbers may help here.

*Symmetrical unimodal distributions* are something distinctive. They often indicate a "homogeneous" group and deviations from the unimodal symmetric appearance may have specific explanations such as assignable causes for "outliers", a mixture of 2 homogeneous groups in the case of bimodality. Separation into more groups may indicate a further explanatory variable. People can learn these rules by heart for practicing data analysis. However, a deeper understanding of the occurrence of symmetrical unimodal distributions is related to processes that can be visualized with a Galton board. This serves more as a metaphor, a reference standard in Shafer' s (1992), sense described above. One should not take this analogy too serious, and regarded more as a metaphor or model how data could have been generated.

In the Bayesian approach, the special reference standards are referred to as models for specifying degrees of belief or uncertainty ("as uncertain as drawing a white ball from a box with such and such a ratio of red and white balls"). Related to EDA, the variability of data in finite samples in reference situations is the relevant dimension of analogy ("what would be the conclusion/the pattern, if the data were a random sample of such and such a model"), We will discuss this in the next paragraph.

## **7. Inferential statistics and guarding against chance variation**

The aforementioned basic scheme of statistical determinism is often hidden in standard approaches to statistical inference. The distributions are reduced to the mean values and the question "are the mean values different" is posed - assuming that distributions are equal in all other respects (normally distributed with known variance). The problem of whether a difference is statistically significant steals into the foreground, masking the basic conceptual question of the difference of distributions.

This gives rise to discussing a basic limitation of the scheme in Fig. 1. The picture of a deterministic dependence of long run distributions from conditions in contrast to the problematic individual level masks the basic problem of everyday statistical analysis, namely that we have knowledge about an intermediate level: about samples (random or other). The conditions do not determine the sample completely; variation is an important and unavoidable feature in all finite sample sizes. If we want to conjecture the change of conditions from observing different distributions, we have already problems with deciding whether the distribution has changed.

The creation, respectively the distinction between the concepts of population and sample was a historical achievement that was not straightforward. A conceptual shift from long run stability to analyzing variation in finite samples is required. Inference statistics is concerned with special situations where the assumptions of random sampling can be made.

However, interpretation of data could be refined if students adopted an attitude from the perspective of "if the data were a random sample". This includes attitudes that the data could have turned out slightly different, that different summaries may be differently reliable, that summaries with small data are generally "unreliable". The use of bootstrap and resampling techniques to estimate the variability or significance of a discovered phenomenon can often only be interpreted in this metaphorical sense. It is useful, even when it is known that the data cannot be modeled as a random sample.

## **8. Some conclusions**

Moore (1990) has suggested a conception that is to help seeing the study of data and chance as a coherent whole. His approach is summarized in "the progression of ideas from data analysis to data production to probability to inference" (p.102). After having had experiences in data analysis, students have to do a major transition in concepts when they have to understand that, now, data sets are considered that are a sample and the interest shifts to the population. The production of random samples from populations and the randomization in experiments should be an intermediate step that consolidates the conceptual shift from data analysis to inference.

The above analysis aimed at pointing to further relating and interfacing the various cultures of thinking - being aware of obstacles and differences. One conclusion is that understanding probability should not only profit from the techniques, concepts and displays of data analysis that can be used for an empirical analysis of chance variation. A relation and distinction has to be established on the level of interpretations and intuitive models like influencing variables and various kinds of causes. A second conclusion I would like to draw is the following: The above progression of ideas could be interpreted as if a course ends with the methods of inference statistics. Several textbooks exemplify this inclination. Students are never confronted with a more unstructured situation where they have to do initial exploratory data analysis, and then decide about whether to construct a probability model or to critically question discoveries from the point of view of inference statistics. EDA, probability and inference statistics seem to be concerned with very different kinds of application with no overlap. This may lead to the problem that EDA experiences are cognitively stored in a separate compartment, and strategies and concepts of data analysis are never reflected and adjusted according to the concepts and experiences in probability and statistical inference. This compartmentalization is hardly a desirable outcome of education.

## References

- Benzecri, J. P., et al. (1980). *L'Analyse des Données. II. L'Analyse des Correspondances*. Paris: Dunod.
- Berliner, M. L. (1992). Statistics, probability and chaos (with comments). *Statistical Science*, 7(1), 69 - 122.
- Biehler, R. & Steinbring, H. (1991). Entdeckende Statistik, Stengel-und-Blätter, Boxplots: Konzepte, Begründungen und Erfahrungen eines Unterrichtsversuches. *Der Mathematikunterricht*, 37(6), 5-32.
- Brass, T. A. (1985). *The Newtonian Casino*. London: Longman.
- Chatterjee, S. & Yilmaz, M. R. (1992). Chaos, fractals and statistics. *Statistical Science*, 7(1), 49 - 68.
- Davis, P. J. & Hersh, R. (1986). *Descartes' Dream - The World According to Mathematics*. Brighton, Sussex: The Harvester Press.
- de Finetti, B. (1974, 1975). *Theory of Probability. Vol. 1 & 2*. New York: Wiley.
- DeGroot, M. H. (1986). A Conversation with Persi Diaconis. *Statistical Science*, 1(3), 319-334.
- Diaconis, P. (1985). Theories of data analysis. In D. C. Hoaglin, F. Mosteller, & J. W. Tukey (Ed.), *Understanding Robust and Exploratory Data Analysis* (pp. 1-36). New York: Wiley.
- Falk, R. & Konold, C. (1992). The psychology of learning probability. In F. Gordon & S. Gordon (Ed.), *Statistics for the Twenty-First Century. MAA Notes #26* (pp. 151-164). Washington, DC: Mathematical Association of America.
- Feller, W. (1968). *An Introduction to Probability Theory and its Applications*. New York: Wiley.
- Flaskämper, P. (1927). Die Statistik und das Gesetz der großen Zahlen. *Allgemeines Statistisches Archiv*, 16, 501-514.

- Flaskämper, P. (1929). Das Problem der "Gleichartigkeit" in der Statistik. *Allgemeines Statistisches Archiv*, 19, 205-234.
- Garfield, J. & Ahlgren, A. (1988). Difficulties in learning basic concepts in probability and statistics: implications for research. *Journal for Research in Mathematics Education*, 19(1), 44 - 61.
- Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J., & Krüger, L. (1989). *The Empire of Chance*. Cambridge: Cambridge University Press.
- Gould, S. J. (1989). The streak of streaks. *Chance*, 2(2), 10-16.
- Hancock, C., Kaput, J. J., & Goldsmith, L. T. (1992). Authentic inquiry with data: critical barriers to classroom implementation. *Educational Psychologist*, 27(3), 337-364.
- Heitele, D. (1975). An epistemological view on fundamental stochastic ideas. *Educational Studies in Mathematics*, 6, 187 - 205.
- Holmes, P. (1994). Teaching statistics at school level in some European countries. In L. Brunelli & G. Cicchitelli (Ed.), *Proceedings of the First Scientific Meeting of the International Association for Statistical Education* (pp. 3-12). Perugia: Università di Perugia.
- Hooke, R. (1989). Basketball, baseball, and the null hypothesis. *Chance*, 2(4), 35-37.
- Konold, C. (1989). Informal conceptions of probability. *Cognition and Instruction*, 6(1), 59-98.
- Krüger, L., Daston, L. J., & Heidelberger, M. (Ed.). (1987). *The Probabilistic Revolution. Volume 1: Ideas in History*. Cambridge, MA: MIT Press.
- Krüger, L., Gigerenzer, G., & Morgan, M. S. (Ed.). (1987). *The Probabilistic Revolution. Volume 2: Ideas in the Sciences*. Cambridge, MA: MIT Press.
- Kuhn, T. S. (1970). *The Structure of Scientific Revolutions. 2nd ed.* Chicago: University of Chicago Press.
- Landwehr, J. M., Swift, J., & Watkins, A. E. (1987). *Exploring Surveys and Information from Samples*. Palo Alto, CA: Dale Seymour.
- Landwehr, J. M. & Watkins, A. E. (1986). *Exploring Data*. Palo Alto, CA: Dale Seymour.
- Moore, D. S. (1990). Uncertainty. In L. A. Steen (Ed.), *On the Shoulders of Giants* (pp. 95-137). Washington, DC: National Academy Press.
- Newman, C. E., Obremski, T. E., & Scheaffer, R. L. (1987). *Exploring Probability*. Palo Alto, CA: Dale Seymour.
- Ogborn, J. & Boohan, D. (1991). *Making Sense of Data: Nuffield Exploratory Data Skills Project. (9 Mini-courses with teacher booklets)*. London: Longman.
- Riemer, W. (1985). *Neue Ideen zur Stochastik*. Mannheim: B.I. Wissenschaftsverlag.
- Shafer, G. (1992). What is probability? In D. C. Hoaglin & D. S. Moore (Ed.), *Perspectives on Contemporary Statistics. MAA Notes #21* (pp. 93-105). Washington, DC: Mathematical Association of America.
- Sill, H.-D. (1992). Zum Verhältnis von stochastischen und statistischen Betrachtungen. In *Beiträge zum Mathematikunterricht 1992* (pp. 443-446). Hildesheim: Franzbecker.
- Steinbring, H. (1980). *Zur Entwicklung des Wahrscheinlichkeitsbegriffs - Das Anwendungsproblem in der Wahrscheinlichkeitstheorie aus didaktischer Sicht. IDM Materialien und Studien Band 18*. Bielefeld: Universität Bielefeld, Institut für Didaktik der Mathematik.

- Stigler, S. M. (1986). *The History of Statistics - The Measurement of Uncertainty before 1900*. Cambridge, MA & London: The Belknap Press of Harvard University Press.
- Tukey, J. W. (1972). Data analysis, computation and mathematics. *Quarterly of Applied Mathematics*, 30(April), 51-65.
- Tukey, J. W. (1977). *Exploratory Data Analysis*. Reading: Addison-Wesley.
- Tversky, A. & Gilovich, T. (1989). The cold facts about the "hot hand" in basketball. *Chance*, 2(1), 16-21.
- von Mises, R. (1972). *Wahrscheinlichkeit, Statistik und Wahrheit*. 4. Auflage. Wien: Springer.
- Zizek, F. (1928). Ursachenbegriffe und Ursachenforschung in der Statistik. *Allgemeines Statistisches Archiv*, 17, 380-432