PROBABILISTIC AND STATISTICAL THINKING

Manfred Borovcnik, *University of Klagenfurt, Austria*

Mathematical concepts enable us to structure our thinking, corresponding models help us to structure reality. They supply us with tools to recognize and solve problems. Stochastic models are not mere images of reality that fit more or less. Right from the basics they have more the character of scenarios to explore reality. This circumstance and only indirect feedback about their success impede understanding of concepts and reasonable application of the like models. Accordingly, mis-conceptions are abundant and recipe application is ubiquitous. Stochastic thinking seems to be quite different from other types of thinking like causal thinking, or logical thinking. The educational discussion until the 90's coined the notion of 'probabilistic thinking', from the 80's the discussion shifted to the notion of 'numeracy' and 'statistical thinking'. By examples and figurative deliberations a multi-faceted image of probabilistic and statistical thinking will be given.

1 Thinking in scenarios – some examples

The scenario feature of probabilistic thinking will be illustrated by some examples, which will also shed light on the merits of the probabilistic approach.

Transparency of decisions

In the face of uncertainty, a single decision may be made more transparent if one allows for weighing the various possibilities. The competing decisions will get thereby (expected) values instead of actual costs or wins.

The problem dealt with here is, if one should take out a policy for a comprehensive insurance of one's car for the next year. The focus is not on mapping the situation precisely onto a model but on illustrating matters; the rough model should just highlight the situation and the purpose of modelling by probabilities. The following table will give the like costs of the decisions (insurance yes or no) under the prospective circumstances (no accident at all, total wreckage).

Cost [in Euro]		Decision	
		A_1 = Insurance yes	A_2 = no
Potential future	T_1 = No accident	1 000	0
	T_2 = Total wreckage	1 000	20 000

With hindsight, one can easily tell if it were better to take out a policy – if no accident happened, no insurance = A_2 is the better decision. If one minimizes the maximal cost, then A_1 = insurance is better. This corresponds to risk-avoiding behaviour. More

margin for innovative behaviour will be opened by introducing probabilities: if one is ready to 'weigh' the possible futures T_1 and T_2, e.g. by relative weights of 39 : 1 (this corresponds to a probability of 1/40 for the total wreckage), then the cost of decision A_1 still is 1 000, but the cost of A_2 has decreased to 500, hence it would be better not to take out a policy. Clearly, the actual decision will depend on the weights for an accident. Other weights will lead to other decisions. However, the decision is now transparent: if one can weigh his/her chances of such an accident by 39 to 1, then A_2 would be better. To free oneself of the burden to fix one's chances, it could be advisable to find the so-called break-even point, i.e. those relative weights at which the decision turns from A_1 to A_2. Here it is 19 to 1. If someone evaluates his/her own chance to be higher or lower than this break-even, he/she should decide accordingly. With the latter procedure there is no need to weigh someone's chances exactly.

Judgment of risks

Not only technical systems have a reliability of survival dependent from its components. One may derive probabilities for the whole system to operate from an elementary (or more sophisticated) assumption. The result has more or less the character of scenario figures and gains more relevance in the comparison of various changes to the system. This will give indications for which changes to promote.

The following problem is drawn from engineering applications. A system has 3 components, the reliability of each is 0,95 for a specific purpose – e.g. that they are working well for exploring the Titan at a special mission. The system works if B_1 and B_2 operate well, or if B_3 works, see Fig. 1

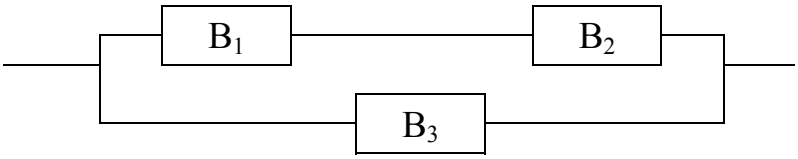


Fig. 1

What is the reliability, i.e. the probability of working well for the mission of the whole system? Is it better to take two full systems on the mission, or is it better to have each component doubled? How many complete systems, or, how many stand-by components for each one should be taken on the mission, if the reliability of the resulting system is required to be 0,99999 (whatever that should mean) ?

The standard solution treats the components as if they were independent of each other and have really the same reliability. Of course, they are *not* independent, and they have *not* the same reliability, and their reliability is *not* 0,95 (this is a *qualitative* statement of the involved engineers). Yet, the scenario is the only way to deal with the problem before the spacecraft is sent on its mission to the Titan. From this, one gets an idea of the reliability of the final system and if it pays the additional costs to build in a specific level of redundancy. A residual risk will always remain; the scenarios, however, will allow judging relative risks and costs of the various actions.

Fixing prices in the face of uncertainty

Expected values are the basis of fixing prices if the future is open to variation. The procedure necessitates weighing the various possibilities. One basis for weighing is taken here, namely to extrapolate risks of the past to the future, i.e. to use relative frequencies for the estimation of probabilities of the various risks at hand.

For a single car-owner e.g., the relative weights of a total wreckage have to be measured individually, the insurance company may rely on the statistics of accidents of the last years. For the sake of simplicity, we will again assume only two possibilities, the total wreckage and no accident, for all the policies. With 2% total wreckages from the past and 10 000 policies, the bookkeeping of the insurance company looks like:

200 total wreckages with a cost of 20 000 each amounts to payments of 4 000 000, i.e. 400 Euro per policy. This amount plus an equivalent for taking the risk plus expenses plus profit makes .., let us say 1 000 Euro per policy.

However, there is always a remaining risk for the company (not in our modelling here but in general) that the premium will not be sufficient. How high will be that remaining risk? How will it change if there are only 100 policies, or if there are 100 000? How will it change with various levels of the premium? Are all possibilities included in the scenario? No, the scenario is suitable only for a normal financial year, for the occurrence of catastrophes the insurance companies have the system of reinsurance, which reduces the remaining risk (of making high losses) by a higher number of policies.

Concluding from circumstantial evidence

At court, if no confession is available, the judges have to rely on circumstantial evidence. Doctors diagnosing various diseases have to rely on indications from blood tests, X rays, mammography results etc. in order to decide about medication or operation of a patient. There are ubiquitous situations where conditional probabilities could help to find the direction of further measurements to be taken best. Formal calculation is done according to Bayes' formula.

In what follows, we will refer to a blood test for diagnosing HIV with the following reliabilities: a person with the virus will be recognized by the diagnosing procedure (test positive) with a probability of 0,99 (in medical jargon this is called the sensitivity); a person not having the virus will be judged virus-free (test negative) with probability of 0,987 (the rate of false positives therefore is 0,013). For a person with positive result, how high is his/her risk to actually have the virus? If we apply the scenario of the person being representative of the whole population, we could use the prevalence of HIV (e.g. 0,02%) and come up with a probability of 0,0150. Judging the person to belong to a high-risk group with a prevalence of 10%, will result in a probability of 0,8943. For a discussion of the input probabilities and the interpretation of the resulting probabilities, see also Gigerenzer e. a. (1998).

How do we perform an adequate calculation of the actual probability of having the virus after a confirmation test when the first result was positive and the second then also was positive? Can one simply combine two applications of the Bayes' formula? (There is evidence that the reliabilities of the test conditional to the first positive result are not the same as before – in other words, if the test has wrongly gone positive it will more likely do the same a second time).

Which scenario is more applicable to the patient? Where do the reliabilities of the testing procedure come from? Do they also have only figurative character or have they come from a controlled experiment with blood samples of which the status of HIV (or not) was absolutely clear? Do we recommend the testing procedure for mass screening? What will be the consequences of mass screening? Is the testing suitable as diagnosing procedure? How can one improve the diagnosing procedure?

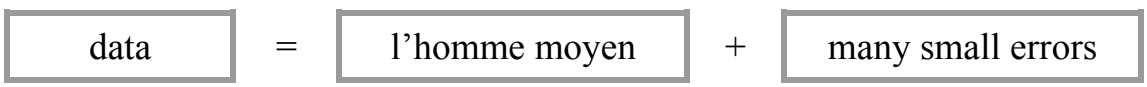
Again, the application has somehow the character of a scenario and gives more information about which action to do next. At the level of implementation of the diagnosing procedure e.g. it will allow a transparent deliberation of advantages (more precise information) and the relative 'cost' (from wrong positives and from wrong negatives). For a teaching approach including such questions, see Vancsó (2003, 2004). For an example related to mammography and resulting doubts if it should be introduced as screening procedure for 50+ women, see Hoffrage e.a. (2000).

2 Approaches to generalizing information

Singular data sets are prone to variation and therefore could convey everything. If someone seeks to adapt to future events, or, if someone wants stable descriptions of the "status", then what to rely on? There is a big need, there is also a desire to extract general features from singular data sets, i.e. to *generalize* the results found. Accordingly, there are a lot of strategies and models for this purpose. If one can derive other statements from the data at hand by logical argument, fine. If one can find out the exact conditions that will lead to a specific (desired) result by causal connection, fine. Even if the results are also due to some yet unexplained but *small* variation. However, how to find out that?

L'homme moyen

This figurative idea of Adolphe Quetelet (1835, see Stigler 1986) is intended to explain how a person gets his/her final outcome of a characteristic (e.g. height or head circumference etc.) by a value that represents the *l'homme moyen*. By errors of nature, however, this value is superimposed by many small errors. Quetelet transferred herewith the elementary error hypothesis of physics to a broader range of applications. This hypothesis finally leads to the normal distribution for the characteristic in the population (the central limit theorem of Laplace was known to Quetelet); a distribution that at times was rated to be a *law* of nature.



Remarkably, the addition of these errors amounts to the unexplained variation of modern views, which is *modelled* by randomness (a normal or some other distribution). The l'homme moyen represents the model for the data. In terms of generalization, it is the generalizable feature of the data, which is to be filtered out.

Other structural equations to split the data describe the generalizable part differently. All have their own interest:

data	=	signal pattern fit model explained common	+	noise deviation residual residual unexplained variation specific causes

Classical inference from data

Observed data are usually summarized to give

- predictions for future events
- a generally valid description of the population – e.g. a confidence interval

Both procedures include a risk of statements not being valid and necessitate estimating the magnitude of the variation. Furthermore, they require that the data generation process is a *random* sample of the population. In terms of the idea of l'homme moyen this means: If one knows the value of l'homme moyen and the magnitude of errors, i.e. the variation, then this amounts to generalizable information, i.e. one can predict future outcomes to be within

$$\text{l'homme moyen} \pm \text{variation}$$

If a person is within these bounds then it is due to no special cause, only *natural* variation is effective; if not then other specific (usually causal) explanation have to be searched for as he/she does not fit to the general case. In other words, there "must be" some other l'homme moyen type working for that person.

Wild and Pfannkuch (1999) find out that the tendency for searching for specific causes is very deep-seated and would lead people to seek for causes also in case that an individual's data are quite within the predicted bounds. This gives a more direct basis for earlier intervention e.g. in case of beginning decline of achievement say in sports (but also in a quality control setting).

If the process of data generation does not compare to the *naturalistic* process of elementary errors acting upon the l'homme moyen, then the data cannot be generalized and used in the way indicated before. In modern terms, we could say that in this case the data are not from a random sample of the population. The key to allow for generalizing findings from data is that they stem from a random sample.

Analogue deliberations may be made for the confidence interval method to generalize findings from data. Clearly, the context will be important to judge whether the random sample argument is valid and if the sample is actually taken from the target population to which the findings are to be generalized.

EDA approach towards inference from data

Exploratory data analysis is centred on the following structural equation:

data	=	pattern	+	deviations

So-called robust techniques should allow filtering out the pattern which means the procedure to find patterns should not be affected too much by some unusual, extreme data. Here, we are farthest from the idea of Quetelet. There is no process of natural superimposing small errors; they could even be very different to different elements of the population and quite big sometimes. In terms of the classical approach, there is no need for the sample to be random.

The justification for generalizing the split of data into a specific pattern and deviations comes from the knowledge of the context: The pattern should give an interesting insight into the context; the deviations may sometimes even shed more light onto the problems within the context. Furthermore, the aim of EDA is a multiple analysis of the data with the aim of several splits and resulting views on the underlying phenomena. The power for generalizing the findings is based on the *comparison* of the various patterns found. Deviations get more attention; they do not merely reflect small error entities "caused" only by nature. Accordingly, a lot of effort is also invested to interpret those deviations on the basis of context knowledge and explain or see why they are as they are. In terms of the 'cause split' of data, EDA looks for common causes in the pattern *and* for specific causes in the deviations.

The ANOVA approach to generalize findings from data

The analysis of variance approach is a standard but sophisticated technique to split the variation in the data into that from specific sources and the rest variation that is modelled by randomness (usually by the normal distribution). We will not go into technical details here (see e. g. Montgomery 1991) but will just discuss the structural equation for the data and its similarities to the l'homme moyen idea.

data	=	general mean	+	specific influences	+	unspecific influences
data	=	l'homme moyen	+	effects attributed to treatment	+	unexplained remainder
data	=	common cause	+	specific cause	+	random influence

For illustration, we will use the context of various teaching methods A, B, .. (=treatment) which might have an influence on some achievement score – the target variable of which data are available. The specific influences are attributed to the treatment A (or B, ..) a 'person' has actually got, the unspecific influences are modelled by randomness. The formal procedure of ANOVA allows deciding when the variation due to the specific influences is big enough as compared to the variation due to randomness in the remainder so that an influence of the treatments may be generalized from the data. With respect to Quetelet's figurative thinking, we separate the deviations of l'homme moyen into two parts, one is causally explained by the effect of the actual treatment, and a remainder that is not yet open to a causal explanation and that should also be small enough that it would not pay to search for further causal explanation.

3 Structuring of thinking

Concepts and models allow us to 'see' reality in a specific way – this acts as feedback also on our thinking, it structures our thinking insofar as it anchors analogies and figurative ideas and archetype models in our approach to reality. We have seen from the examples in section 1 that probability has a strong feature of building up scenarios for reality and not always to directly model reality. As a consequence, that has a deep impact on how learners can integrate probabilistic concepts into their repertoire as there is only indirect measures of how close the model depicts the real situation and also the *success* of a model is not easily to be judged. No wonder that mis-conceptions are abundant and deep-seated, i.e. are often not revised by pertinent education.

Probability is more than a material property

Normally, we learn by trial and error. By wrong decisions we lose and thereby we think about improvements and come up with better models and so forth. Here only one simple example is given to illustrate that matters with probability are much more complicated and feedback about success is by no means open to direct interpretation.

Take the two *Falk wheels* of fortune of Fig. 2 and give the choice which to spin. Obviously the left wheel has the bigger sector for winning (=1), therefore one should choose it. However, once you spin, you have not a high probability of winning, .. , and you might lose. What then? Was the choice wrong? The best action here is not awarded. Thus, you might speculate about the reasons why. At times, I checked this with my then seven years old daughter and right from the beginning she wondered about the special margin of the right wheel, …, and concluded that grown-ups should take the left but children had to take the right…(the author named the wheels after Ruma Falk).

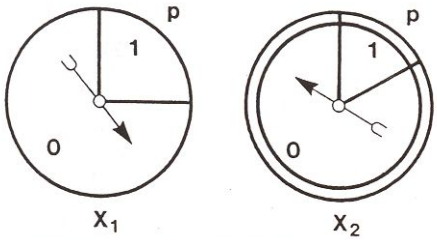


Fig. 2

With reference to the scenario use of probability in section 1 and with respect to the example above, relative frequencies as companion of probabilities are thus often more or less a metaphoric way to describe probability (for the genuine theoretical nature of probability see also Steinbring 1991). Yet they amount to a useful way to describe some features of the abstract concept, which is out of the reach of understanding. All the more it often may be used as a short cut to a mathematical derivation of probabilities by the method of simulation.

The idea of weighing the evidence

Relevant information about a situation under uncertainty may come from:

- Combinatorial multiplicities of possibilities, which are judged to be equally likely
- Frequencies of events in past or comparable series, which are judged to be 'similar'
- Personal judgement of involved risks

Core of the further information process to derive at a decision is the concept of expectation as illustrated in some of the examples of section 1. Quite often the three cases above involve – despite exact numerical values for the probability requested – *qualitative* aspects and the scenario character of the models used. Calculating (expected) values of tentative decisions in order to find a justification for the final decision made may be viewed as an exchange between 'money' (utility, cost) and uncertainty.

If someone is faced with a decision, the consequences of which lie in the future and cannot be foreseen (except a listing of all cases, which are considered to be 'possible'), then one has to find some optimality principle to signify one or some of the decisions as better than others. Minimizing maximum 'loss' works without introducing probabilities, ordering decisions by expected 'loss' requires the weighing of uncertainty. The potential of the method increases by an investigation of how sensitive the derived decision reacts to changes in the weighing process – again a plea for the scenario character of the probability approach.

Mis-Conceptions

With the indirect feedback to the probability approach it is not surprising that thinking in probabilities is not well integrated into the cognitive frame of individuals. Often, modelling a situation involves an attribution and separation of and between causal and random parts (see the discussion in section 2). The causal part of a

problem, if separated and explained, allows for more direct interventions and thus seems more promising.

For example, in Wild and Pfannkuch (1999), the actual score of a 2 out of 5 is compared to a probability of scoring of 70%. Despite the fact that it does not amount to a significant deviation from that probability (which means no intervention is necessary), people tend to seek for other, causative explanations of the low actual score. Once one has found such a causative explanation (e.g. temporary private problems of the player, a small physical injury, a quarrel in the team etc.), the track for success promising intervention is open. This in mind, people seem to extremely favour the causative approach (see Wild and Pfannkuch 1999).

In the tradition of Kahneman and Tversky (see e. g. Tversky and Kahneman 1972, or Kahneman e. a. 1982), a lot of misconceptions have been identified and described. Intuitive strategies like representativity, anchoring, and causal strategies constitute subsidiary strategies to surmount the difficulties in the process of weighing the uncertainty (for a discussion see Borovcnik and Peard, 1996). Among many others, here only the outcome orientation of Konold may additionally be referred to. According to it, information available to a person is more likely to be actually used in solving a problem, if it allows for a direct prediction of the outcome in quest, or, the problem is reformulated in order to allow for such a direct prediction (see Falk and Konold 1992). This fits quite well to Wild and Pfannkuch's observation above.

From the abundant mis-conceptions, one may conclude the difficulty of the venture to teach *probability* concepts and its necessity. This is also true for the education of the *statistics* part as not only Wild and Pfannkuch (1999) describe individual's problems to discriminate properly between causative and random parts in splitting and explaining variation in empirical data.

To use misconceptions effectively in teaching, it is not sufficient to confront learners with relevant situations which are prone to wrong approaches; instead of trying to revise wrong intuitions (which are very basic and deep-seated) one should build a bye-pass by re-presenting the situations by (very) simple material forms that allow for solutions – for promising examples in view of the Bayesian formula see Bea and Scholz (1995), Krauss e.a. (2002), or Hoffrage e.a. (2002).

Probabilistic and statistical thinking

From the examples in the first two sections, types of situations and types of thought that help with them are derived. In all cases to follow, a mingling between probabilistic and statistical thinking may be traced. With some twist of thought, the one or the other part predominates. Always the following strategies may be supportive in presenting (or in solving) the following standard situations: To give simple analogies, which are similar in characteristic features and which have illustrative potential for the solution. To give simulations which necessitate organizing clearly the model assumptions for the involved situation and effortlessly

yielding the solution. To re-formulate, or even to re-present the situation in a more basic manner.

One-off decision: The procedure of attributing (expected) values to decisions by an exchange between uncertainty and 'money' goes back to Christian Huygens 1657 (see Bentz 1983 or Freudenthal 1980). He developed his ideas in the context of lottery games and speaks of some uncertain lottery to 'be equally worth as' some amount of money given by a formula (for the expectation). Huygens himself already applied his concept of expected value also to insurances especially to life insurances. The less agreement on weighing the uncertainty, the more it gets important to supply additional justification to the weighing by investigating the consequences of different weights – the stronger becomes the scenario character of probability as was discussed in section 1.

To think about uncertain situations in terms of scenario like expected values and respect that as one tool amongst others to derive at transparent decisions, is a basic ingredient of stochastic reasoning. It cannot be stated clearly enough that the *values* for the decisions allow to signify some decisions as better than others *without* making it possible to predict (or even attempting to aim at predicting) the specific outcome in the 'future' – this seems counter-intuitive not only against the background of Konold's findings on the ubiquity of the outcome orientation in individuals.

Decision in the face of circumstantial evidence: Abundant situations require proceeding according to the actual values of conditional probabilities relative to new facts. Judgement before court by circumstantial evidence, or diagnosing procedures in medicine, are just two prominent examples of that kind of reasoning. Formally, the new conditional probabilities are calculated with the formula of Bayes. The never-ending disputes about the validity of prior probabilities (prior to the circumstantial evidence) indicate again the scenario-character of probability. Furthermore, the two involved 'directions' of conditional probabilities do have a complete different connotation from a causal standpoint: Whereas the conditional probability of e.g. having some virus to having a positive result on the diagnosing test *is causally* (and it is only by some errors possible that a negative result is achieved at), the backward direction of conditional probability from a positive diagnosis to actually having the virus is *merely indicative*. A fact, which is hard to accept for many due to a causative mis-interpretation of it. Findings from research on misconceptions tell that conditional probabilities are grossly overestimated in case of a possible causal interpretation, and are often even neglected (too small to be taken into consideration) when they are only indicative.

To think about pertinent situations in suitable terms of a Bayes model, amounts to basic ingredients of stochastic thinking. There are many endeavours to improve teaching on that issues. While it seems inadequate to transform us to Bayesian thinkers, it establishes a great progress in the teaching of probability to make us aware about the cognitive biases from causal re-interpretations. A great help with that comes from very basic material re-presentations, which demystify the causal

connotations that always come back to mind if the mathematics involved becomes too complicated. From all the endeavours, here only the approach of Krauss e. a. (2002) of natural frequencies is referred to.

'Natural' variation of randomness: Investigating the variation of data involves often a split of the data into explained und unexplained parts, or in causative and random parts, see section 1. As this separation is neither unique nor clear-cut, our intuition has to be backed up by simulation studies about what it would mean if the variation were only random. What are the implications of 'pure' randomness? The so-called 'square root of n' law may be demonstrated by such investigations and teaching experiments, see e.g. Riemer (1991), or Kissane (1981): If a target variable is the sum – or better the mean value – of other variables (that need not necessarily have the same distribution, e.g. 1 or 0 according to some dichotomous experiment, or the result of throwing a die), then

- a two sigma-rule becomes more accurate with increasing number of summands: approximately 95% of (simulated) values of the target variable are between the mean value plus or minus two times the standard deviation of the target variable,
- the standard deviation of the target variable decreases by a factor $1/\sqrt{n}$, n being the number of summing variables.

This is a manifestation of the elementary error hypothesis of physics of earlier times and a concrete example of the central limit theorem that states that the 'limiting' distribution for the (mean) target variable is a normal distribution.

To think about pure randomness' consequences in terms of a normal distribution and in terms of an ever-decreasing variation (by the square root of the number of 'trials') is the key to many a statistical procedure. To name only one, the estimation of a population mean by the mean of a random sample. The precision of that procedure is improved by larger random samples. The amount of improvement may be read off and adjusted to some accuracy required by investigating a sample, which is large enough. The model situation with the same summands (as in the simulation situation) being independently taken from the same population will also shed light on the importance of a *random* sample and not just an arbitrary sample to be taken in order that the law comes true.

4 Structuring reality

It would be too restricted to think of structuring reality by models as to depict the relevant features of a real-life situation, abandon less relevant ones, and to "filter out" a model that represents more or less an *image* of the original situation. Within that model one could then derive at mathematical solutions and re-interpret them into the context problem from the onset. This is *one* feature of probabilistic modelling only, one that is truly important, but there is more to say about probability models with respect to their more indirect and scenario character already described within the examples of sections 1 and 2.

Models and concepts allow to structure *thinking* and this in turn allows us to apply these models to reality structuring it and (partially) solving the problem there. It is worth to devote some extra thought on the objective side of models structuring *reality*. There are more ingredients that come to the fore with this focus, especially the interplay between causative and random parts of the variation of variables, which allows dealing with real problems.

The split of variation into causative and random parts

In empirical research, often causal influences for the variation of a target variable are searched for. For illustrative purpose, the reader should think of several alternative treatments, which could affect a target variable. The simplest model would be (with some known function f):

Target variable	=	f(treatment effect)

The variation of the target variable would then uniquely be determined by the treatment effect. However, there is a lot more sources of variation in the data for the target variable like plain measurement errors, variable external circumstances of the experiment, other attributes of the 'persons' which also could influence the final value, variation due to the specific persons that are sampled and investigated, etc. We have here a similar but more differentiated situation as in the l'homme moyen figure of Quetelet. We will give a simple structural equation of how the target variable emerges:

Target variable	=	a constant value	+	treatment effect	+	influential variable	+	random 'error'

Treatment and influential variables are *explanatory* variables, which explain the variation of the target variable by some causative (or associative) argument, the random part could represent further causal relations between values of some other variables of the person investigated and his/her value of the target variable. If one can establish tight relations of some of these with the target variable, then they could be integrated into the explanatory variables (see also Wild and Pfannkuch 1999). However, lacking more precise information about these variables necessitates dealing with them as if they were random.

The question if treatments are effective, i.e. different treatments have a 'substantially' different influence on the target variable, is now transformed to the question if the variation of the target variable is mostly 'explained' by the variation of the treatment effect, or if it is due to other influential variables, or even if it is due only to variation of those parts which are modelled to be random (and not yet open to causative explanation). However, the split into causative and random parts, the split of causative parts into treatment and other explanatory effects is not unique and may influence of course the final findings heavily.

The question when the causal part of the model is big enough to be judged as relevant is a technical one met by several specific significance tests, which should not worry us here. It is only important to state that these procedures rely on an investigation how pure randomness would influence the target variable.

An example for the structural equation 'in action'

The following example deals with the explanation of the target variable ‚body weight'. From many examples we have learned that there is quite a tight relation between weight and height of persons. Another explanatory variable is gender. The remainder is unexplained (by further influential variables) and modelled to be random. It could, for example, be explained further by the body type of a person (pyknic, leptosome etc.), or it could be explained by race, or by nutrition in early childhood etc.

Body weight	=	a constant value	+	gender effect	+	b × body height	+	random 'error'

Without the gender term, one would have a simple regression line for the model to describe the relation between weight and height. Separated between genders there is quite a different slope of the regression line. If the investigator does not include gender into the study, the variable would have the status of a *confounding* variable. A confounding variable changes or even completely reverses the relations found. Once, data are available on gender, the variable may be tested if it should be taken into the model; in case if it were a continuous variable like the height, it would be called *co-variate*. A covariate is simply a *candidate* to be included in the model for the data.

The remainder is not open to causative explanation as there are no specific data available. It will be modelled simply by pure randomness. The evaluation of explanatory variables if they should be integrated into the model for the data is done in comparison to the 'size' of the remainder: are the changes in variation due to candidates for explanatory variables big enough in comparison to the variation due to that remainder.

Neither is the interactive building of the model unique, nor are the components unique, which constitute the remainder. For an interpretation of the final model for the target variable, the separation of variation into causative (explained) and random (non explained) parts is essentially – the split is not ontological but only pragmatic. If this separation yields a relevant model, the causative interpretation may lead to promising interventions.

The formal procedure to separate the model entities is based on significance tests in the ANOVA or ANCOVA models; we will not go into the details. An intuitive understanding of this separation is at the core of anyone's reasoning who is involved or concerned with statements, which are backed up empirically by data from investigations.

Furthermore, the intuitive strive for looking for causative explanations for phenomena at hand, is not being tricked out by probabilistic reasoning. To the contrary, probabilistic models are a key factor to filter out causative elements of a problem to get more control over interventions on the target variable.

Two model situations to equalize other influential variables

Generalization of findings from samples to the population: To get reliable information about the mean of a population, a sample is taken from that population and the sample mean is taken as an estimate. If 'measuring' a single object of the population may be modelled merely by 'natural' variation of randomness, then all the properties of that randomness from the last section may be applied to an artificial summing or calculating the mean value of all the objects. Hence, the normal approximation and confidence intervals may be derived. This modelling is justified by a random sampling procedure to select objects for the sample to be 'measured'. It is the random sampling which guarantees – with the exception of some calculable risk – that the results are generalizable, i.e. that the confidence interval covers the 'true' mean of the population. In this sense, with random sampling the sample drawn is *representative* for the population (see also Borovcnik 1992).

To think that samples are representative of populations if the sampling is 'purely' random, is a key concept for the generalization of empirical findings. Randomness avoids all conceivable selective properties, which would lead to biased samples, or, randomness equalizes all selective properties, so that finally the sample is representative for the population. Any arbitrary procedure to select the sample is prone to systematic, unforeseeable and uncontrollable deviations.

Generalization of differences between two samples: Often, a comparison between two (or more) groups has to be made. The groups have got a different treatment – one e.g. a special medication for insomnia, the other only a placebo (a harmless substitute without a pharmaceutical substance). Is the medication more effective than placebo, is the decisive question. Of course, the two groups have to be as equal as possible with respect to all characteristics that could influence the effect of the treatment: People have not to know that they get the placebo, people in the treatment group should not represent the most persistent cases already proven insensitive to any treatment, and so forth. Strictly speaking, aught the groups to be drawn purely randomly from the population? With the exception of very few cases, this could not be fulfilled in practice.

However, if the investigated group as a whole does not differ substantially from the population, it is sufficient that a random *attribution* process establishes the subgroups for the different treatments, i.e. randomness decides to which group the next patient will belong. This random attribution should equalize all differences in the objects across the whole subgroups, which could have a causative influence on the target variable representing the 'success' of the treatment. Insofar, the random attribution should eliminate, or better equalize all causative elements that could make the

subgroups different (i.e. equalize the effect of all confounding variables) – the actually observed differences then may be attributed solely to the treatment and are thus generalizable.

5 Statistical Thinking

A brief introduction into the debate is given against the background that at all times the argument was used that there is more to probability and statistics than is contained in the mathematical version of the pertinent concepts – a special kind of thinking was advocated; several authors refer to an outstanding role of an interplay between intuitions and formal concepts (see Fischbein 1975, or Borovcnik 1992).

The educational debate on probabilistic and statistical thinking

The educational debate on probabilistic and statistical thinking is a long-ongoing one. Even in times as early as the 70's when the accent was heavily put on the mathematical and probabilistic part of the curriculum, a special type of thinking was argued to be behind the formal concepts – probabilistic *thinking*. It was not quite clear what could be understood by that but the argument was that there is some additive not yet included in the mathematical concepts. However, when put to a crucial test, either mystique arguments were used to describe and justify what probabilistic reasoning is, or it was reduced to key ideas in the mathematical development of the concepts.

Heitele (1975), e.g. gave a catalogue of fundamental ideas, which on the whole should constitute the various dimensions of probabilistic thinking. His list is reading like the titles of the chapters in a mathematical textbook on probability but could always also be attributed to some more general idea at second inspection:

- Calibrating the degree of confidence
- The probability space
- The rule of addition
- Independence
- Uniform distribution and symmetry
- Combinatorics to count equally likely cases
- Urn models and simulation,
- The idea of a sample to represent a population
- The idea of a random variable and its distribution
- Laws of large numbers.

There were a number of attempts to get a clearer image of the fundamental ideas behind to describe probabilistic thinking, e.g. Borovcnik (1997) endeavoured to arrange the ideas around the idea of information as a central hinge between individual's intuitions and the formal concepts of the mathematical theory:

- Probability as a special type of information about an uncertain 'issue'
- The idea of revising information when faced with new evidence
- To make transparent which information is used–also in the simulation of situations

- To condense information to a few numbers (thus eliminating randomness)
- To measure the precision of information
- To guarantee the representativity (=generalizability) of partial information
- To improve the precision of information

The roots of probabilistic reasoning in the psychological research go back to Kahneman and Tversky. They identify various tendencies in individual's behaviour to wrongly re-interpret problems and solve them in a way different to the accepted standard. Their approach of misconceptions had a great impact on further educational research. In an empirical investigation on children's behaviour, Green (1983) found abundant misconceptions that were to be expected according to Kahneman and Tversky. Scholz (1991) tried a constructive approach of a cognitive framework for individuals to allow for probabilistic reasoning in an adequate manner (i.e. to accept a standard interpretation of situations and end up with acceptable solutions). The work of Fischbein (1975) is devoted to develop instructive approaches towards developing a sound interplay between individual's intuitions and formal concepts of mathematics as a key to develop probabilistic reasoning.

From the 80's on, initiated by the EDA discussion started by Tukey (1977), the focus shifted towards statistical thinking and reasoning. The new motto then was numeracy, i.e. to learn to understand data and the information, which lies in them. The shift also involved more applications, real-life situations, all leading away from mathematics and probability – and also away from games of fortune. In the German debate at that time also procedures of formal statistical inference entered the stage and won much attention in the reform of curricula (the high goals of those times have now made place for more realistic ones, especially enabled by the simulation technique and the resampling idea for statistical inference). Numeracy and statistical reasoning internationally was promoted by big projects in the U S A, beginning with the Quantitative Literacy project, see e.g. Scheaffer 1991.

Numeracy and 'graphicacy' was targeted at simple but intelligent data analysis alongside the techniques of descriptive statistics and EDA. The role of the context where the data stem from, for the interpretation of results got increasing attention; it became undisputed that a sound analysis of data and results is not really possible without reference to the context. In the German debate, Biehler (1995) is seeking a balance between probabilistic and statistically loaded curricula, as courses biased towards data analysis would cause an all-too simple and probability free conception of stochastics.

However, asked what statistical thinking could be, no one gave a clear answer. This unsatisfactory circumstance was the starting point for Wild and Pfannkuch to integrate ideas from empirical research.

Statistical thinking as contrasted to probabilistic thinking is involved in all steps from the provisional problem out of a context, across all cycles of making the problem more precise and model it, to a final model substantiated by the data. Statistical

thinking is tightly related to the process of empirical research to filter out generalizable findings from empirical data. More or less, statistical thinking might be associated with strategies to increase knowledge.

The approach towards statistical thinking by Wild and Pfannkuch

Here, a brief discussion of statistical thinking along the lines of the approach by Wild and Pfannkuch (1999) will be given. According to their approach, four dimensions of that type of thinking should be regarded:

- The investigative cycle
- The interrogative cycle
- Types of thinking involved
- Dispositions

The investigative cycle is a systems analysis approach toward the initial (research) questions. It involves the components of

$$\text{Problem} \rightarrow \text{Plan} \rightarrow \text{Data} \rightarrow \text{Analysis} \rightarrow \text{Conclusions}$$

which might be run through several times for refinement.

In the *problem* phase it comprises to grasp the "system dynamics" of describing the target variable, i.e. to find 'all' relevant explanatory variables, and to reflect about possible confounding variables. Also, assumptions of relations, hypotheses on relations from relevant other investigations or theories, have to be used to clarify and "define problems". Omissions in that phase give rise to a bulk of flops in empirical research.

In the *planning* phase, the investigator has to deal with the development of a "measurement system" for all the included variables. A "design" has to be developed how the "sampling" and the "data management" actually has to be undertaken. A "pilot study" should give indications if the plan is adequate and practicable.

The *data* phase comprises "data collection, data management and data cleaning".

The *analysis* phase comprises "data exploration, planned analyses, unplanned analyses, and hypothesis generation".

The *conclusions* phase consists of "interpretation, conclusions, new ideas, and communication".

The interrogative cycle represents the strategic part of the investigation and includes the following phases (see Wild and Pfannkuch 1999 for details):

$$\text{Generate} \rightarrow \text{Seek} \rightarrow \text{Interpret} \rightarrow \text{Criticise} \rightarrow \text{Judge}$$

Types of thinking: Wild and Pfannkuch (1999) discern between general types of thinking and those fundamental to statistical thinking. For the general types, they list "strategic, seeking explanations, modelling, applying techniques". Specific to statistical thinking they list

- Recognition of need for data

- Transnumeration
  – changing representations to engender understanding […]
- Consideration of variation
  – noticing and acknowledging,
  – measuring and modelling for the purposes of
    prediction, explanation, or control
  – explaining and dealing with investigative strategies
- Reasoning with statistical models
- Integrating the statistical and the contextual
  – information, knowledge, conception

Dispositions amount to the psychological side of investigations and comprise the following attitudes: Scepticism, imagination, curiosity, openness (to ideas that challenge preconceptions), a propensity to seek deeper meaning, being logical, engagement, and perseverance.

Wild and Pfannkuch (1999) then proceed to the split of sources of variation in data into real (a characteristic of system) and induced (by data collection).

Finally they come up with the explanation of the regularities, the model found on basis of the data analysis process. They refer to statisticians who see the biggest contribution of their discipline in the isolation and modelling of 'signal' in the presence of 'noise'. Then they discuss about the relative and interchangeable character of random and causal influences. They refer to randomness as "just a set of ideas, an abstract model, a human invention which we use to model variation in which we can see no pattern". In that they come close to the scenario character of probability described here.

6 Conclusion

Questions central to probabilistic and statistical thinking have been raised. They should clarify that both types of thinking are mingling and are not easily be described. From the discussion, however, crucial components of these types of thinking should become clearer. There will be no simple answer also after further endeavours into that topic. Even if the ideas are not easily described, the foregoing exposition and the examples illustrate how pertinent thinking is organized and to which end it could serve. And how such thinking is blurred.

The role and eminent importance of data, the context where they stem from, and the attitude of empirical research should become quite clear from the discussion. The interpretation of probability statements as scenario figures assisting in a broader problem solving process to come up with a more transparent decision may become more accepted by the examples outlined in the paper. The splitting of variation in data into causative and random parts in the search of explaining, predicting, and controlling phenomena may be a guideline for further attempts to clarify the issues of statistical thinking.

References

Bea, W. and Scholz, R.: 'Graphische Modelle bedingter Wahrscheinlichkeiten im empirisch-didaktischen Vergleich', *Journal für Mathematik-Didaktik* 16, 299-327.

Bentz, H. J.: 1983, 'Zum Wahrscheinlichkeitsbegriff von Chr. Huygens', *Didaktik der Mathematik* 11, 76-83.

Biehler, R.: 'Probabilistic Thinking, Statistical Reasoning, and the Search for Causes – Do We Need a Probabilistic Revolution after We Have Taught Data Analysis?', in *Research Papers from the Fourth International Conference on Teaching Statistics*, Marrakech 1994, The International Study Group for Research on Learning Probability and Statistics, University of Minnesota.

Borovcnik, M.: 1992, *Stochastik im Wechselspiel von Intuitionen und Mathematik*, Bibliographisches Institut, Mannheim.

Borovcnik, M. 1997, 'Fundamentale Ideen als Organisationsprinzip in der Mathematikdidaktik', *Didaktik-Reihe der Österreichischen Mathematischen Gesellschaft* 27, 17-32.

Borovcnik, M. and Peard, R.: 1996, 'Probability', in A. Bishop e. a. (eds.), *International Handbook of Mathematics Education*, part I, Kluwer Academic Publishers, Dordrecht, 239-288.

Falk, R. and Konold, C.: 1992, 'The Psychology of Learning Probability', in F. Sheldon and G. Sheldon, *Statistics for the Twenty-First Century*, MAA Notes 26, The Mathematical Assocation of America, 151-164.

Fischbein, E.: 1975, *The Intuitive Sources of Probabilistic Thinking in Children*, D. Reidel, Dordrecht.

Freudenthal, H.: 1980, 'Huygens' Foundation of Probability', *Historia Mathematica* 7 (2), 113-117.

Gigerenzer, G., Hoffrage, U., and Ebert, A: 1998, 'AIDS Counselling for Low-Risk Clients, *Aids Care* 10 (2), 197-211.

Green, D. R.: 1983, 'A Survey of Probability Concepts in 3000 Pupils Aged 11-16 Years', in *Proceedings First International Conference on Teaching Statistics*, vol 2, Teaching Statistics Trust, 766-783.

Heitele, D.: 1975, 'An Epistemological View on Fundamental Stochastic Ideas', *Educational Studies in Mathematics* 6, 187-205.

Hoffrage, U., Lindsey, S., Hertwig, R., and Gigerenzer, G.: 2000, 'Communicating Statistical Information', *Science* 290, 2261-2262.

Hoffrage, U., Gigerenzer, G., Krauss, S., and Martignon, L. (2002). 'Representation facilitates reasoning: What natural frequencies are and what they are not.', *Cognition* 84, 343-352.

Huygens, C.: 1657, 'De ratiociniis in ludo aleae', in F. v. Schooten: *Exercitationes matematicae*, Leyden.

Kahneman, D. and Tversky, A.: 1972, 'Subjective Probability: A Judgement of Representativeness', *Cognitive Psychology*, 430-454.

Kahneman, D., Slovic, P. and Tversky, A.: 1982, *Judgement under Uncertainty: Heuristics and Biases*, Cambridge Univ. Press, Cambridge.

Kissane, B.: 1981, 'Activities in Inferential Statistics', in A. P. Shulte and J. R. Smart, *Teaching Statistics and Probability*, National Council of Teachers of Mathematics, Reston, Virginia, 182-193.

Krauss, S., Martignon, L., Hoffrage, U., and Gigerenzer, G.: 2002, 'Bayesian Reasoning and Natural Frequencies: A Generalization to Complex Situations' (submitted for publication).

Montgomery, D. C. : 1991, *Design and Analysis of Experiments*, J. Wiley & Sons, New York.

Quetelet, A.: 1835, *Sur l'homme et le dévelopment des ses facultés, ou Essai de physique sociale*, Paris.

Riemer, W.: 1991, 'Das '1 durch Wurzel aus n'-Gesetz – Einführung in statistisches Denken auf der Sekundarstufe I, *Stochastik in der Schule* 11, 24-36.

Scheaffer, R.: 1991, 'The ASA-NCTM Quantitative Literacy Project: An overview', in D. Vere-Jones (ed.), *Proceedings Third International Conference on Teaching Statistics*, International Statistical Institute, Voorburg, 45-49.

Scholz, R.: 1991, 'Psychological Research in Probabilistic Understanding', in R. Kapadia and M. Borovcnik, *Chance Encounters: Probability in Education*, Kluwer Academic Publishers, Dordrecht, 213-254.

Steinbring, H.: 1991, ‚The Theoretical Nature of Probability in the Classroom, in R. Kapadia and M. Borovcnik, *Chance Encounters: Probability in Education*, Kluwer Academic Publishers, Dordrecht, 135-166.

Stigler, S. M.: 1986, *The History of Statistics. The Measurement of Uncertainty before 1900*, Harvard Univ. Press, Cambridge, Mass.

Tukey, J. W.: 1977, *Exploratory Data Analysis*, Addison Wesley, Reading.

Vancsó, Ö. (ed.): 2003, *Matematika 10, 11*, Muszaki Kiado (Textbook in Hungarian).

Vancsó, Ö.: 2004, 'Inverse probabilities in everyday situation (Bayesian-type problems)', paper distributed in the *TSG 11: Research and Development in the Teaching and Learning of Probability and Statistics* (L. Jun, J. M. Wisenbaker org.) at ICME 10, Kopenhagen.

Wild, C. and Pfannkuch, M.: 1999, 'Statistical thinking in empirical enquiry', *International Statistical Review* 67, 223-265 (with discussion).