How wrong models become useful - and correct models become dangerous

Christian Hennig[1]

Seminar für Statistik,
ETH-Zentrum, CH-8092 Zürich, Switzerland

Abstract. The conceptual background of statistics and data analysis is considered from the viewpoint of constructivist philosophy. The relation of formal models to observable reality is discussed as well as the role of model assumptions and especially probability models in data analysis. I argue that approximate correctness of models is an ill-posed problem. The relation of model assumptions to observer-independent reality can never be assessed objectively. Instead, formal models are useful to learn about the derived statistical methods and to enable clear understanding between different observers. Some illustrations are given.

1 Introduction

All models are wrong - but some are useful. This famous statement of George Box is often cited to justify the work with models that clearly do not fit the circumstances exactly. But of course he did not say that wrong models are always useful and we are still left with the decision which wrong models are useful for a problem at hand, and in which sense they are useful. In classification and clustering, this problem arises, e.g., if we have to decide about the normal assumption or if we want to estimate the number of clusters and we wonder if a unique true number of clusters exists in reality.

In the present paper I will discuss the issue from a constructivist viewpoint. Constructivism is an epistemology which considers reality only dependent of its observers. Objective reality can never be observed, and perception is considered as a means of self-organization, not as a representation of objective reality. I give a brief introduction to constructivist philosophy in Section 2. This paper is a companion paper to Hennig (2002) and there is some intersection. But while Hennig (2002) discusses many aspects of the relation of constructivism to data analysis briefly, here I focus on the relation between formal (especially probability) models and reality.

The content of the paper may be somewhat provoking, and it should be said that it is not meant as the last word on the subject. It should simply serve to stimulate discussion about a too often neglected aspect of statistics and data analysis.

2 A short introduction into constructivism

A naive interpretation of perception can be characterized as follows: The reality outside is full of information, and to observe reality means to build a particular kind of representation of it in our brain. In the same manner science would be interpreted as building as objective representations as possible of reality in (more or less) formal language. From this point of view, the role of the observers is mainly passive. One problem of this interpretation is that the question of objectivity cannot be assessed in an objective manner. That is, to recognize the inadequacy of a particular observers' representation of reality, another observer is needed, whose objectivity again can be doubted.

The constructivist interpretation of perception is much different. Constructivists consider the role of the observer as active. Perception is interpreted as a means of self-organization of an individual, a social system, respectively. The reality outside acts as a perturbation of the observer, i.e., it forces the observer to perceive her reality in such a way that unfavorable situations such as touching a heated hot-plate are avoided. But it is by no means necessary that the observer-independent reality matches the personal reality, and the relation between these two realities cannot be observed.

There are two main principles of constructivism:

- There is no observation without observers. This means in particular that we have to take a look at the observer if we want to analyze observations.
- Observations are constructed in social dependence. Most of us probably feel that there is so much agreement between different observers that the first principle alone would not suffice to explain it. But the strong dependency of an individual on the society and in particular on the closest persons such as the parents in the first years of her life enforces an important role of communication and agreement for the construction of the personal perception. As a consequence, we also have to analyze communication if we want to analyze observations.

I refer to Watzlawick (1984) for a more comprehensive introduction into constructivism. There are various schools of constructivism, but they agree about the above mentioned principles. Some more literature can be found in Hennig (2002).

3 Reality, language and formal models

I use the term personal reality for the whole world of perception of a person. This includes sense impressions, thoughts and feelings, all of which are to be thought as interdependent. Personal reality changes in every moment, at least to some extent. It is useful to think about the relation between personal reality and language, because language is the most important carrier

of communication (at least in science) and the relation of formal models to reality is somewhat analogous to the case of language.

Imagine that you are asked to explain as precisely as possible your present perceptions (which cannot clearly be separated from thoughts and feelings). Probably, the first aspects which come to your mind are the aspects you can clearly describe in words: You are reading a book, the paper is white, here are letters etc. If you concentrate more deeply, then you will recognize that language is not able to describe precisely your detailed perceptions, e.g., how this particular letter "r" and the surface of the paper surrounding it are characterized, how they differ from other "r"s and other areas of this page, and what exactly your hands and your nose perceive while observing precisely this letter and this page. There are at least two reasons for this inability. Firstly, language is linear and digital as opposed to perception. Secondly, time is needed to give a static description, while perception is essentially dynamic.

In conclusion, I have two basic theses:

- Language and personal observations operate on distinct domains. Thus, language is never able to explain adequately the personal reality.
- The repercussion of language to personal reality is crucial. Once we have words, these words highlight and bundle some impressions and down-weight others, so that we are no longer able to observe reality in a way we would do without language.

The same theses apply to formal models with respect to social reality, which I will discuss now. Consider the scheme given in Figure 1. Social reality can be thought as to be made up by the whole amount of communication between human beings, of which language is an important part - and most accessible to scientific analysis. It is useful to distinguish different social realities of different social systems. For example, the reality of the social system of German classification researchers present at the annual GfKl meetings will differ from the reality of the system of teachers of a particular school. Obviously, individuals belong to more than one social system and the resulting social realities are not clearly separated.

While social realities can be analyzed by observing the system's discourse, such analysis has to take into account that the words of ordinary language are not well defined, and members of the same social system do not necessarily mean the same thing if they use, e.g., the term "democracy". At this point, scientific formal language, i.e., (often) mathematics enter. The main difference between mathematics and ordinary language is that the mathematical concepts are aimed to be uniquely defined. Science, at least where it is formalized, aims to be a social system where absolute agreement is possible because the concepts are cleaned from every connotation which is dependent on the individual and its social background. The problem is that the mathematical unification moves language one step further away from the primary personal observations. That is, observations have to be further reduced and

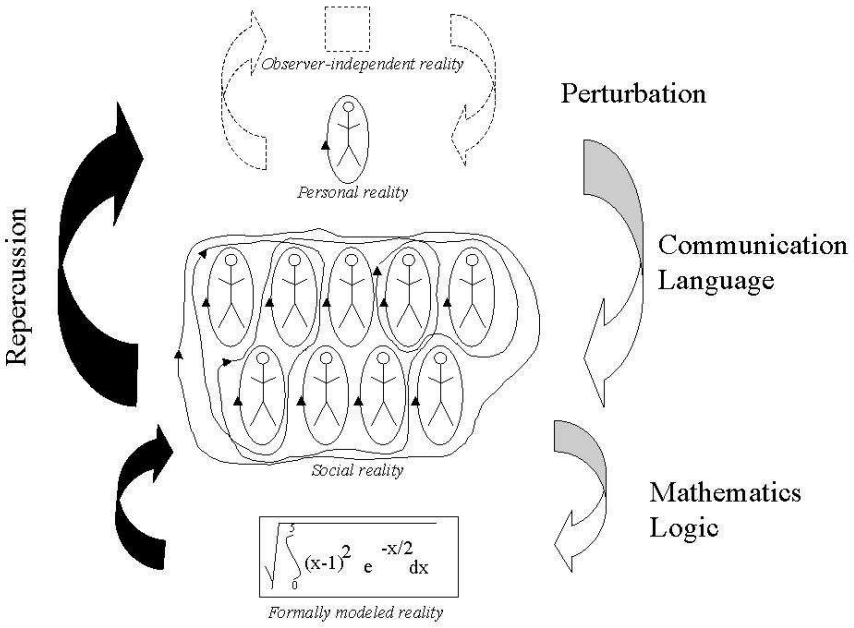

Fig. 1. "Semi-formal" model of the relation between reality and formal models.

communication has to deviate more from the original perception in order to fit into the mathematical framework. On the other hand, there is again repercussion. Formal models, once introduced in the scientific community, react on the discourse of the affected social systems, and they react also on the personal realities of the individuals who use them. For example, a "probability of rain" is nowadays accepted as understandable term in common language. As long as we believe that there exists an observer-independent reality which is affected by our actions, formal models as well as language will even react on this reality.

4 Are all models wrong?

4.1 General aspects

In the light of the previous section, all formal models are wrong in the sense that they operate on domains distinct from reality (no matter if observer-independent, personal or social). Interpreted in this way, they cannot even be approximately true (this is further discussed in the two following subsections). However, it is possible that individuals or social systems are influenced so much by the formalized discourse and the repercussion of the models, that they reduce their reality to the formalized aspects and, in the most extreme

case, that they are no longer able or willing to observe deviations. That is, models can match an observer's reality, but this does not say that they fit any observer-independent reality. Instead, it says something about the reduction of the perceptions of the observers. To formulate it provoking: While wrong models may be useful, "correct" models are dangerous.

In general, a researcher or a group of researchers suggesting a model will have a more complex perception of the phenomenon of interest. There will be modeled and non-modeled aspects. The act of modeling highlights the modeled aspects and down-weights the others. There is always the danger that the non-modeled aspects vanish from the scientific discourse. Thus, conscious ignorance is crucial for a reasonable work with models. By this I mean that the non-modeled aspects (including deviations between model and the researchers' perceived realities) either should be kept explicitly in the discussion, or that a clear and conscious decision is made that these aspects are not important with respect to the problem at hand.

The benefits and dangers of modeling are closely connected: Models enable understanding, clarification and communication of the researchers' perceptions and concepts, but this comes to the price that these perceptions must be adapted and reduced to the formal language.

4.2 The role of measurement

I stated that formal models cannot even be approximately true. It could be objected that it is possible, however, to calculate distances between empirical distributions of datasets and probability distributions or between dissimilarities on sets of objects and approximating ultra-metrics as generated by hierarchical clustering methods. Such distances could be interpreted in terms of a better or worse approximation of the data by the models.

It is important in this respect to consider the role of measurements. Measurement can be considered as a particular kind of language in the framework given above, which mediates between personal and unified scientific reality. Again it can be observed that measurements do not fit personal reality adequately, but that at the same time they react strongly on individual and social perception (an example for the implications of this repercussion is given in Hennig (2002)). The design and choice of measurements involves importance judgments, the construction of particular and often artificial environments and social negotiation. It is almost always more or less explicitly model based. That is, measurement can be interpreted as a social construction, and the question of the approximation of measures by models is to be distinguished sharply from the question of the approximation of personal or observer-independent reality.

Furthermore, the approximation of data by models can only be measured under some implicit model assumptions, which in itself can neither be verified nor approximated. For example, approximation of empirical distributions by probability models is only reasonable if the observations are assumed to be

i.i.d.. Complex enough dependency structures as well as non-identical distributions can by no means be excluded. The general idea of the approximation of data by probability models refers always to the impression that there is something like a (possibly very complicated) random mechanism generating the data, which can be expected to produce stable relative frequencies if repeated long enough. The following subsection discusses the question if there is evidence in favor of such a view.

4.3 Interpretation of probability models

The usual argumentation supporting the frequentist interpretation of probability says that apparently converging relative frequencies are an empirical fact (see e.g. von Mises(1936)). This is often illustrated by plots of stabilizing sequences of the relative frequency of events such as throwing a 6 with a dice with increasing number of throws. But this stabilization can be explained purely by the definition of the relative number of successes, where the increasing denominator enforces the decrease of the variation, while this does not allow any diagnosis of convergence. More formally, Fine (1973) proves that "apparent convergence" of relative frequencies follows from "apparent randomness". The validity of such a statement depends of course on the concrete formalization of these two concepts, which I do not discuss here. The meaning is that we would judge the outcomes of a sequence as too regular to be random if the relative frequencies of success would not seem to be convergent. Thus, apparent convergence is a consequence of how we assign the attribute "random" and has nothing to do with concrete data. By the way, this argument is an example of the mechanism of self-affirmation of well established scientific models, methods and results. The judgment of such a scientific concept as "experientially successful" is often based at least partly on the following technique: Whenever the application of the concept is apparently a failure, this is attributed to the violation of assumptions or other scientific working standards (in this case "too regular to be random"), while such problems are neglected in the case of success.

The laws of large numbers also do not work as a justification of the existence of converging relative frequencies of success. The most obvious reason is that the laws are given in terms of probability models. Thus, a valid interpretation of such models must be *assumed* to apply the laws to real data, and therefore they cannot provide a *foundation* for such an interpretation.

From this discussion I do not draw the conclusion that the frequentist interpretation of probability is useless and should be rejected. Not the interpretation itself is misled, but the belief that it should and could be connected to observer-independent reality in a scientific or logical manner. Such a connection can also not be established for any other interpretation of probability models, and every attempt to do so would suffer from similar defects.

De Finetti (1972) tries to give an operational justification for his subjectivist Bayesian interpretation of probability in terms of the betting be-

havior of individuals. But this does not mean that there is any objective relation of subjective probabilities to the states of mind of an individual, which are claimed to be modeled. If an individual does not obey de Finettis' betting rules, it is simply judged as "non-coherent", according to the above mentioned technique of self-affirmation. Hennig (2002), and, more detailed, Walley (1991) give arguments why the demand of Bayesian coherence is not necessarily rational.

Nevertheless I think that researchers, who work with probability models, should explicitly choose an interpretation of probability. Models are most useful to enable understanding and agreement among scientists, and this does not work if the interpretation is obscured.

A "constructivist frequentism" would mean that a researcher communicates clearly that she imagines the phenomenon of her interest as repeatable, and that she interprets the assigned probabilities as tendencies which would manifest themselves as limit values of converging relative frequencies in case that the imagined replications would take place. Some sources of variation are judged as non-essential (or non-observable). The researcher would call these sources "random". She would acknowledges that objective randomness is essentially non-observable and a correct representation of the (objective, personal, or social) real reasons for the variation is out of reach of formal modeling. Instead, a constructive frequentism would say something about how the researcher observes and how she judges the phenomenon, and this is accessible to scientific discussion. In the same manner, a constructivist Bayesianism and a constructivist interpretation of other concepts of probability can easily be imagined.

5 The benefits of incorrect model assumptions

It follows from the previous discussion that the (approximate) correctness of a model assumption for given data or a given setup is an ill-posed problem. Goodness-of-fit tests and graphical assessment can never check all important aspects of a model (this is also discussed in Hennig 2002). At most, data can give strong evidence against a model. This does not mean that "a method is not allowed if the model assumptions are not fulfilled", which would be a meaningless statement. Instead, the method should be rejected if the inspection of the data leads to the suspicion that the methods derived under the model assumption may lead to non-adequate conclusions. This holds, e.g., for mean and standard error under the occurrence of outliers (as long as there are no subject-matter reasons in favor of giving the outliers high influence), but not under seemingly rectangular distributed data.

This shows one of the two main uses of model assumptions: They allow us to learn about the methods. The other benefit of explicitly stated model assumptions is that they communicate clearly the researcher's perceptions and

judgments about the phenomenon of interest. This leads to understanding and often to agreement and new ideas.

To illustrate this, consider the use of normal mixture models for clustering. If a researcher assumes them, this means that

- she wants to explain the data as coming from a (usually) small number of homogeneous groups,
- she interprets a cluster as a bell-shaped point cloud with moderate tails,
- she wants to characterize the clusters only by their means and covariance matrices,
- she judges dependencies between points and the reasons for within-group variation as non-essential.

All these decisions, which cannot be justified from data or other observations alone, can now be discussed critically. This is one of the most important stages of scientific work, and at this stage, model diagnostics become useful. Often, the discovery of discrepancies between the data and the structures judged important by the researcher leads to interesting new findings. That is, model assumptions are often useful *because* they lead to the detection of clear deviations. If model based methods are a priori rejected because the assumptions are suspected not to hold, this key feature of model assumptions is ignored. (Of course, it is nevertheless often reasonable to replace unstable model-based methods by more stable ones.)

6 Concluding example

Based on constructivist epistemology, I developed an attitude towards models, for which the question of the correctness of the model is an ill-posed problem. Formal models and observer-independent, personal, and social reality operate on distinct domains. All attempts to verify objectively models or even probability interpretations such as the frequentist one are doomed to fail. Instead, formal models communicate the perceptions of researchers in a unified, clear manner, and they enable us to learn about data analytic methods.

Here is an example for the implications of this attitude: Often data are transformed in order to get a more normal-shaped distribution. According to the ideas given here, this is useless, as long as the only aim is to fulfill the normal assumption. A better rationale is as follows:

Data should be transformed in order to achieve correspondence between the treatment of the differences of the values by the data analytic method and the subject-matter meaning of these differences as judged by the researchers.

Suppose that a social scientist wants to compare two groups with respect to the answers to a question, for which five ordered alternatives are given. Often this is done by use of the Wilcoxon test, which is preferred over the t-test because the data is ordinal and non-normal. Suppose further that the

scientist argues that the linguistic difference between the first four categories is approximately equal while the fifth category is somewhat extreme. Under the assumption that the scientist's judgment is sufficiently precise, I suggest that a transformation of the categories to, say, 1, 2, 3, 4 and 8, followed by application of the t-test, will be clearly superior to the standard treatment, except under very extreme distributions of the answers.

While the Wilcoxon test seems to be "allowed" for ordinal data, it nevertheless treats the categories as numbers, whose effective differences are determined from the distribution. If there is additional information about these differences stemming from the personal or social reality of the scientists, it seems to be a bad choice to ignore this information only because there is no objective rationale how to translate it into numbers.

Further, we know from statistical theory that the t-test is approximately valid in i.i.d. settings with existing variances, that it is fairly good for light tailed distributions, and that it suffers mainly from extreme outliers and extreme skewness, which should not be present in this setup (but it would not be a good idea to score the highest category as 1000 and to apply the t-test, even if the subject-matter researcher would judge this to be the adequate value). Thus, we should rather trust the outcome of the t-test in this case than in some other experiments, where distributional shapes occur, which are very similar to the normal with moderately heavier tails. This is where the use of (well understood) model assumptions enters.

References

DE FINETTI, B. (1972): *Probability, Induction and Statistics.* Wiley, London.

FINE, T. L. (1973): *Theories of Probability.* Academic Press, New York.

HENNIG, C. (2002): Confronting Data Analysis with Constructivist Philosophy. In: K. Jajuga, A. Sokolowskij and H.-H. Bock (Eds.): *Classification, Clustering, and Data Analysis.* Springer, Berlin, 235–244.

VON MISES, R. (1936): *Wahrscheinlichkeit, Statistik und Wahrheit.* Springer, Vienna.

WALLEY, P. (1991): *Statistical Reasoning with Imprecise Probabilities.* Chapman and Hall, London.

WATZLAWICK, P. (1984) (Ed.): *The Invented Reality.* Norton, New York.