Confronting Data Analysis with Constructivist Philosophy

Christian Hennig[1]

Seminar für Statistik,
ETH-Zentrum (LEO),
CH-8092 Zürich, Switzerland

Abstract. This paper develops some ideas from the confrontation of data analysis with constructivist philosophy. This epistemology considers reality only dependent of its observers. Objective reality can never be observed. Perceptions are not considered as representations of objective reality, but as a means of the self-organization of humans. In data analysis, this leads to thoughts about the role of probability models (frequentist and epistemic), the necessity of subjective decisions (e.g. tuning constants), the nature of statistical predictions and the impact of the gathering of data to the reality.

1 Introduction

Some recent developments in epistemology, namely constructivist and postmodern theories, had a large impact in the social and educational sciences, but they are widely ignored up to now in the foundations and practice of many natural sciences including mathematics, statistics and data analysis. Data analysts are concerned with the generation of knowledge and with the question of how to learn about the reality from specific observations, which lies in the heart of constructivist epistemology. This is why I think it it fruitful to confront constructivist philosophy with data analysis, even though the rejection of the concept of "objective reality" by most of the constructivists seems off-putting to many researchers educated in the spirit of the natural sciences. This paper is meant as a short sketch of ideas to stimulate discussions.

2 Constructivist philosophy

2.1 A short introduction

In the literature there are various interpretations of "constructivism". One may distinguish "radical" from "social" and "methodological" constructivism (introductions and anthologies are e.g. Berger and Luckmann (1966), Watzlawick (1984), Gergen and Davis (1985), von Glasersfeld (1995)). There are three principles common to constructivist approaches to epistemology:

There is no observation without observer. There is no means to go beyond a persons observations, except by observations of others, observations of observations, respectively. Every judgment about the validity of an observation (which can be a belief or an inference as well) depends upon the observer, and upon the person who judges. In this sense, no objectivity is possible.

Observations are constructed in social dependence. As human beings, we are bound by language and culture. We do not only learn how to *communicate* perceptions, we learn even to *perceive* by means of language and culture. We are bound by material constraints as well, but this can only be observed through language and culture. Individuals get to socially recognized and accepted perceptions by interaction between their actions and the actions of their social systems. This process, starting from the first perceptions in the earliest childhood to sophisticated scientific experiments, is called the "construction of observations." It can be analyzed on the personal and social level.

Perception is a means of self-organization, not of representation. Constructivist epistemology rejects the hypothesis that observations and perceptions should be analyzed as somewhat biased representations of objective reality, because it is not possible to assess the difference between reality and representation independently of observers. Instead, perceptions are thought as a means for an individual (or a social system) to organize itself to fit (more or less) successfully to the constraints of its environment, which are recognized to exist, but not to be accessible objectively. Note that there also cannot be objectivity in the attribution of "success" to a process of self-organization. Values like this must be culturally negotiated.

Because the construction of observations involves individual actions, we can ascribe (more or less) responsibility for it to the individual. In particular, a social system is constituted by the way in which its members communicate their observations and beliefs. That is, all members influence the constructions which are valid for the social system, and the other way round.

2.2 Consequences

How can an epistemology influence the practice of a science? I think that the main contribution of the constructivist philosophy can be a shift of the focus of interest from some problems (e.g. "What is objectively true?") to others:

1. How do data analysts communicate and how does this construct their perception of reality?
2. What perception of reality gives rise to their models and methods? This is reversal to 1 and illustrates that the whole constructive process can be thought as circular.

3. In constructivism, alternative realities (personal as well as social) are possible. Given a construction of reality, how can alternatives be constructed, how and why is such a construction hindered, respectively?
4. What is the role of subjective decisions and how can the responsibility of the subjects be made visible?

According to its own standards, constructivism should not be seen as a "correct" or "wrong" philosophy. It is able to shift and broaden someone's view, but it has also been possible to develop many of the following ideas without an account to constructivism, as can be seen from some of the references given below.

3 A constructivist view of some aspects of data analysis

3.1 The role of probability models

For model based methods in the frequentist sense it is usually assumed that the data is generated by some random mechanism which can adequately be described by some probability model. It is known from robust statistics that methods based on a simple parametric model can be strongly misleading if the model holds only approximately. It does not need constructivist philosophy to recognize that no probability model, how general it may be without getting completely useless, can ever be verified by observations. Here is a short account of the difficulties of assessing model assumptions:

The practice of goodness-of-fit tests is a reversal of the usual logic of a statistical test: A probability model is accepted if the data do *not* show significant deviation from it, while usually only a *rejection* of the null hypothesis can be interpreted as a meaningful result of a test. Further, the principle of testing the goodness-of-fit of a model and accepting the model only in case of a non-significant result is a sure way to violate the model assumptions, because the chance for significance is, say, 5% under the model, but 0% for the resulting data. Graphical assessment of model assumptions shares, more informal, the same problems.

The assessment of the dependency structure of the data points is even more cumbersome. All statistical reasoning is based on repetitions, and this means that only periodical short-range dependency structures such as ARMA processes are accessible by observations. With 2000 data points, neither a dependency between only the fourth and fifth observation, nor regular dependencies of range larger than 1000 are observable. "I.i.d." may be distinguished from certain simple dependencies and heteroscedascities, but "dependent" can never be distinguished from "independent, but irregularly non-identical".

From the constructivist viewpoint, it does not make sense to attribute "reality" to something which is strictly not observable. Consequently, it neither makes sense to speak of any "objectively true" distribution, nor of the

approximation of it: How to observe the approximation of something which by itself is strictly not observable? Instead a model should be treated as a concept of the human mind which helps to structure the perceived reality.

A probability model may formalize a regular structure which a researcher perceives or presumes about the observed phenomena. To assume a model while not believing in its objective truth means to assess possible deviations from this structure as *not essential* with respect to the research problem of interest ("random" could mean that an observer judges the sources of variation as non-essential). This is a subjective decision which can be made transparent. Insofar, a model can be utilized to discuss different perceptions of a phenomena and to compare them with observations. But if the model assumptions are accepted without discussion, the sources of deviations vanish from the perception of the researchers, and this leads to a narrow view of reality.

Probability models cannot approximate an observer-independent reality as such, but they can approximate observations, i.e. data, by means of distances to empirical distributions. Such an approach is formalized by Davies (1995) and makes clear that there are always lots of different models adequate for a single dataset.

However, it makes sense to use models as "test beds" (Davies and Kovac (2001)) for methods. The true answer to an interesting real data analytic problem cannot be known (benchmark data are not "interesting" in this sense, because the truth about them must be assumed as known), and so it cannot be tested whether a data analytic method is able to find it. This means that formal models - not necessarily probabilistic - are useful to compare the quality of methods.

I often heard the objection against model-based methods that they should not be applied if it cannot be guaranteed that the model assumptions are fulfilled. This viewpoint is not compatible with constructivism, because it rests on a misunderstanding about such assumptions, which are not meaningful about objective reality, but about the perception of researchers. The advantage of model based methods is that the circumstances at which they work (or not) are made at least partially transparent. This can also be achieved by proceeding the other way round: Finding a good model for a given method, as suggested by Tukey (1962).

3.2 On the foundations of probability

It was stated in the previous section that probability models only reflect the perceptions and attitudes of the researchers. This could lead to the thought that probabilities should always be interpreted as epistemic, as done in the subjectivist Bayesian approach. But while an aleatory interpretation requires non-verifiable assumptions about the material reality, the epistemic interpretation requires non-verifiable assumptions about the states of mind of the individuals. For example, to observe behavior corresponding to epistemic

probabilities, it is necessary to postulate a linear scale of utility valid for different interacting individuals. Further, a kind of logical consistency must be demanded, which excludes that the individuals change their opinions during experiments in reaction to unpredicted events of any kind, which were not modeled in advance. Works on the foundations of probability such as Fine (1973), de Finetti (1974, 1975) and Walley (1991) include many arguments where such non-verifiable assumptions of the frequentist and subjective interpretations are discussed.

In conclusion, epistemic probabilities as models for beliefs of individuals are subject to analogous objections as those raised in the previous section against the frequentist models, as long as they are meant to approximate objectively true states of mind. And they share the same advantages if they are meant to illustrate the ground on which researchers act.

The decision between aleatory and epistemic probabilities should be a decision between *interests* of the researchers, namely the interest in modeling a unique reality shared by all involved individuals and the interest to model individually differing but internally consistent points of views. These both interests can today be constructed as essentially different. This difference evolved in the 19th century as a side-effect of an increasing recognition of individuals, while the former founders of probability theory apparently did not perceive a clear distinction between a ratio of successful to possible cases (as ancestor for a frequentist interpretation) and fair prices for gambles (as ancestors for epistemic interpretations). Constructivistically spoken the distinction between aleatory and epistemic probability did not exist (i.e., in the realities of the researchers) until the time of Laplace, and it does not make sense to argue about which of the two concepts Bernoulli or Bayes "really" meant. This is why the two concepts share the term "probability", resulting in lots of discussions about its "true meaning".

More elaborate concepts like imprecise probabilities (see e.g. Walley, (1991)) can incorporate more aspects of the perceptions of researchers, but they cannot result in a unique rational foundation of probability, because the acceptance of any unique formalization makes deviating perceptions invisible and is, in consequence, an obstacle for the progress of science (Feyerabend (1988) gives examples).

3.3 Subjective decisions in data analysis

In agreement with the constructivist point of view, Tukey (1997) justifies that different experts may draw different, equally reasonable conclusions from the same dataset. This is in contrast to the behavior of a majority of scientists concerned with statistics, even if the situation might be more tolerant in exploratory and graphical data analysis. Referees of methodological papers tend to demand a unique objectively justified choice of parameters such as tuning constants. Referees of applied papers like to have a single result based on a model which is claimed to be the only adequate one for the corresponding

type of analysis. The same holds for most of the clients of statistical consultation, who often hope that statistics answers their research questions without leaving any freedom for their own decisions. Many data analysts agree that e.g. robust statistics suffers from offering the user too many different estimation methods.

Of course, in many cases, the data analysis should help to make a decision, and only one action can be performed afterwards. But even in such cases it would be possible to make a responsible, subjective decision on the base of a variety of data analytic solutions, if the background and the arguments in favor of all these solutions are made transparent. Even a belief in the naive objectivist paradigm that there is a unique reality cannot hide the difficulty to recognize it and the possibility of quality criteria having more than one dimension.

I think that there are also some more questionable reasons for demanding uniqueness and objectivity. It seems that many individuals dislike to be responsible for the consequences of their actions. If I use the only objectively correct method to tackle a problem and something goes wrong, this must be due to unfortunate random outcomes or unpredictable events, but not to any responsible choice of my subjectively preferred model and method. Subject matter experts and data analysts often like to leave the responsibility for results at each other or at textbook authors. In not so few of my statistical consultations the subject matter experts left me with the ultimate problem to find a reference in the literature for what we had done, because they wanted to prevent us to take the whole responsibility.

A second point is that there may be fear that statistics loses its authority if it confronts the user with more than one correct answer. Thirdly, data analytic methods and in particular commercial statistical software packages are often meant for users without much data analytical knowledge and experience, and this leads to the thought that it is less dangerous to give them a single good parameter value instead of letting them choose what they want.

The constructivist approach does not mean that the choice of methods and models is arbitrary. It is crucial that the background and the arguments for subjective decisions are given as detailed as possible to gauge them. Lack of time or computing power, concentration on other aspects of a study, routine use of familiar methods or even lack of statistical knowledge are legitimate, honest reasons for such a choice. However, they may be criticized with every right by somebody who thinks to be able to do better. “Legitimate” does not mean “uniquely true”.

Presumably the most satisfying foundation of a choice of methods can be given in cooperation between subject matter experts with an alert interest in data analysis and data analysts with a lively interest in the subject matter. Experienced data analysts know that almost every data problem can give rise to a closely adapted, new and idiosyncratic treatment superior to the

application of standard methods, when time and resources suffice (Hampel (1998)).

Methodological projects would often benefit from resulting in a variety of well explained possibilities, e.g. defined by tuning constants. Usually optimal values of such constants (most famous the 5% significance level, used as tuning constant e.g. to define an outlier rejection rule) can only be found by the use of artificial optimality criteria, which are at least as difficult to justify as the constants itself. Instead, and more honestly, the constants could be chosen in accord to the beliefs and interests of the researcher as good as possible.

3.4 On predictions

My main point about statistical predictions of future events is that they always need to assume that the future equals the past in terms of the probability model. Keeping Section 3.1 in mind, this means that every possible difference between future and past has to be judged as non-essential by the researcher. This may be reasonable in some controlled technical experiments, but is usually a very restricted view in every setup where human decisions are involved. Often, e.g. in stock markets, the prediction itself influences the future. In Germany, in the seventies the need of nuclear power was advertised by the use of over-pessimistic predictions for the electricity consumption, neglecting totally the possibility to influence the reasons for the consumption instead of simply providing more electricity. From a constructivist viewpoint, there is the danger that an uncritically adapted model for prediction may obscure the perception of possibilities to change the behavior reflected in the model. It is more constructive to use the outcome of the model as an illustrative scenario which we may want to prevent (or, in other cases, to reach) instead of interpreting it as a realistic forecast.

3.5 Data change reality

Up to now I discussed situations where conclusions were to be drawn from given data. This has set aside a very important aspect, namely the construction of reality by means of the gathering of data and the decision to tackle problems by the use of data.

As an example consider the comparison of the quality of schools. If such a comparison should result in a ranking, it has to be carried out on the base of a one-dimensional ordinal criterion and usually on the base of numerical data. For example, unified tests resulting in a number of points can be performed.

This may have a strong impact on the considered reality and its perception. If the content of such a test is known at least approximately, schools and teachers will try to train their students to optimize the test results, no matter if the tested items correspond to the needs and talents of their particular groups of students. Further, not every capacity is equally easy to measure.

This may result in a down-weighting of abilities which are more difficult to assess by tests or to the invention of more or less questionable measures for them.

Such aspects should be taken into account in every study where the gathering of data is considered. Sometimes it will lead to the decision that it is not appropriate to consider the situation as a data analytic problem.

4 Conclusion

I gave a short introduction to constructivist philosophy and derived some, may be provoking, ideas connected with data analysis.

Human beings are dependent of structuring their thoughts, and they can organize themselves often well by inventing models, generating data, and analyzing them. They should, however, not forget that this changes their thoughts and perceptions and the thoughts and perceptions of others. Lots of interesting and important processes between individuals vanish if we only concentrate on looking at the data. Data analysis might benefit from taking the construction processes into account more consciously.

References

BERGER, P. L. and LUCKMANN, T. (1966): *The Social Construction of Reality.* Anchor Books, New York.

DAVIES, P. L. (1995): Data Features. *Statistica Neerlandica*, *49*, 185–245.

DAVIES, P. L. and KOVAC, A. (2001): Local extremes, runs, strings and multiresolution. *Annals of Statistics*, *29*, 1–47.

DE FINETTI, B. (1974): *Theory of Probability* (Vol. 1). Wiley, London.

DE FINETTI, B. (1975): *Theory of Probability* (Vol. 2). Wiley, London.

FEYERABEND, P. (1988): *Against Method* (revised version). Verso, London.

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

GERGEN, K. J. and DAVIS, K. E. (1985) (Eds.): *The Social Construction of the Person.* Springer, New York.

HAMPEL, F. (1998): Is statistics too difficult? *Canadian Journal of Statistics*, *26*, 497–513.

TUKEY, J. W. (1962): The future of data analysis. *Annals of Mathematical Statistics*, *33*, 1–67.

TUKEY, J. W. (1997): More honest foundations for data analysis. *Journal of Statistical Planning and Inference*, *57*, 21–28.

VON GLASERSFELD, E. (1995): *Radical Constructivism: A Way of Knowing and Learning.* The Falmer Press, London.

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

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