The game of red and blue revisited: propensities, subjective probabilities, and the goodness-of-fit paradox

Christian Hennig

Fachbereich Mathematik - SPST, Universität Hamburg

November 18, 2004

Abstract

"The game of red and blue" is an example given by Gillies (2000) to illustrate the superiority of his propensity interpretation of probability over the subjective interpretation of de Finetti. I will show that the falsifiability of probability models by means of statistical tests, which is crucial for Gillies' interpretation, is subject to serious difficulties, one of which is the goodness-of-fit paradox: The confirmation by a statistical (goodness-of-fit) test refutes the validity of a probability model. I conclude that the existence of objective probabilities is an irreducibly metaphysical assumption, i.e., it is strictly not testable by observations. On the other hand, I agree with Gillies that the example is a valid illustration of a shortcoming of the subjective interpretation, namely that the requirement that every change of a probability assignment has to follow Bayes' rule is too inflexible if data is observed that is inconsistent with the prior distribution. Instead of rejecting the two discussed interpretations of probability, I argue that probability modeling inevitably requires subjective decisions about accepting such untestable assumptions. The value of the models is their power to structure our thoughts and our perceptions of the uncertain, and not their correspondence with any objective reality.

1 Introduction

Probability models model random phenomena, i.e., situations with uncertain outcomes. In the present paper I argue in favor of the following theses:

- Statements about objective probabilities are not testable and falsifiable in the same manner as deterministic statements. The most important reason for this is that a validity statement for a probability model includes the statement that unexpected outcomes, which would lead us to reject the model, are possible with positive probability.

- All concepts of objective probability have an irreducible metaphysical core. Probabilistic modeling always involves subjective decisions about the acceptance of untestable assumptions.
- The subjectivist interpretation of probability hides rather hide than solves this dilemma: if a researcher follows it dogmatically, he is prevented from noticing aspects of the data that are too complex.
- Probability modeling is a good example of how formal models determine the perception of phenomena. Tests of such models are not possible independent of the model's own effect on perception.
- Propensity as well as subjectivist and other interpretations of probability can be adopted, but they all imply limitations of perception and this should be kept in mind. The adoption of an interpretation of probability is a decision to structure the perception. No objective or rational argument can generally favor one interpretation over the others.

In Section 2, "the game of red and blue" will be revisited to motivate the above position. "The game of red and blue" is an example given by Gillies (2000, p. 77-83) to illustrate the superiority of his propensity interpretation of probability to the subjective interpretation of de Finetti. The example goes back at least to Feller (1950, p. 67-95) and was used by Popper (1957a) to argue against the possibility of inductive logic.

Gillies' version of the propensity interpretation can be sketched as follows:

- A probability value p assigned to an outcome of a set of reproducible experimental conditions means that the experiment has the tendency ("propensity") to bring forth the outcome with approximately the relative frequency p under independent replication. Thus, this is a "long run"-interpretation as opposed to "single case"-propensities as advocated by Popper. Probability models for dependent sequences are interpreted in terms of independent repeatability of the whole sequence.
- Propensities are not operationally defined, but characterized by Kolmogorow's axioms, which are motivated by (but not deduced from) the properties of relative frequencies.
- The validity of a probability assignment can be falsified by statistical hypothesis tests.

There are various versions of propensity interpretations, the first of which is due to Popper (1957b). They all have in common that probabilities are interpreted as objective tendencies to generate the respective relative frequencies without using a limit of relative frequencies to define probabilities operationally, as frequentists do (e.g., von Mises, 1928).

Here is a sketch of de Finetti's (1970) subjectivist interpretation of probability, which goes back to the early thirties.

- A probability value p assigned to an outcome A of a future experiment[1] E by a subject S measures the degree of belief of S that A will be the result of E.
- More precisely, p is the supremum amount of money[2], which S thinks to be advantageous to pay in order to obtain an amount of 1 if E yields A and nothing in case of A^c.
- Subjective probabilities can be measured operationally with the following procedure: If S specifies her probabilities $p_1, p_2, \ldots$ for a set of outcomes $A_1, A_2, \ldots$, an opponent O is allowed to choose stakes $c_1, c_2, \ldots$ so that for each outcome A_j, S has to pay $p_j c_j$ to O, and O pays back c_j only if A_j happens. The stakes c_j may be positive or negative.
- The axioms of probability[3] follow from the demand of the coherence of S's probability assignments. This means that S has to choose $p_1, p_2, \ldots$ in order to prevent O from making a "Dutch book", i.e., choosing $c_1, c_2, \ldots$ in such a way that S loses money against O whatever the outcome of E is.
- "Learning from experience" is operationalized by allowing conditional bets, i.e., bets on some outcomes A, which are only evaluated if it is clear that another outcome B happened. Coherence then enforces Bayes' theorem on conditional probabilities, and "learning" means that the former conditional probabilities become the new unconditional ones after B is observed.
- Learning from experience is in most cases formalized in experiments that yield sequences of realizations of random variables $X_1, X_2, \ldots$ which are assumed to be exchangeable. This means that S assigns the same probability to all permutations of intersections of outcomes of the form $\{X_j \in A_k\}$, where j and k come from the same finite subset of the natural numbers. By de Finetti's representation theorem (de Finetti, 1937), exchangeability is equivalent to assuming that the $X_1, X_2, \ldots$ are independently and identically distributed according to a distribution P_θ conditionally on some (unobservable) parameter θ, which obeys some prior distribution $\pi(\theta)$. This means that the learning process can be fully described by the updating process of the distribution of θ according to new observations enforced by Bayes' theorem. De Finetti shows further that the posterior distributions of θ obtained

[1] An experiment in the subjectivist sense can be anything where the possible outcomes can be defined in advance. Especially, E may consist of a whole sequence of random experiments.

[2] I assume the linearity of the utility of money here as an idealization, which may be accepted at least if the total amount of money involved is relatively small. This issue is controversial and is discussed in almost all books on the foundations of probability.

[3] See Gillies (2000), p. 65 f., for a discussion of the uncountable additivity, which is not accepted by de Finetti.

> by the same sequence of outcomes (length converging to infinity) under different prior distributions converge toward a distribution which concentrates all its mass on eventually arbitrarily small intervals about a single limiting parameter θ_0. This provides a link to the propensity interpretation, which assumes the *objective* existence of such a parameter for independent sequences of repetitions of an experiment. For 0-1-sequences, this is simply the probability of "1" in a single go.

There are a lot of other interpretations of probability, for example the logical interpretation (e.g., Keynes, 1921, Jeffreys, 1931), and the frequentist interpretation (e.g., von Mises, 1928). Some interpretations require different systems of mathematical axioms, e.g., epistemological and aleatory versions of imprecise probabilities (e.g., Walley, 1993, Hampel, 2001). The reader is referred to Fine (1973), Gillies (2000) and to the cited original literature for a discussion of these interpretations. I agree with the arguments of Gillies (2000, Chapters 3-6) as to why he prefers the subjectivist interpretation over the logical one and the propensity interpretation over the frequentist one. Note that recent advocates of "objective Bayes" statistics in the tradition of Jeffreys (e.g., Bernardo and Smith, 1994) recognize the elicitation of subjective prior distributions from well informed experts as a philosophical ideal. They consider objective rules for initial probabilities as a reference standard, where elicitation is not possible or intersubjective agreement is important. Therefore they are philosophically closer to the subjectivist than to the logical interpretation.

Part of the discussion of the game of red and blue is the goodness-of-fit paradox, which implies that hypothesis tests cannot effectively check a probability model. In Section 3 I discuss whether proper interpretations of hypothesis tests are possible nevertheless. In Section 4 the subjective interpretation of probability is discussed in the light of the game of red and blue. I argue that the subjective as well as the propensity interpretation, if taken literally, leads to inevitable problems. However, in Section 5 I suggest a version of "probabilistic pluralism", which accepts different interpretations as models for our perception of random phenomena, provided the general limitations of formal modeling are kept in mind. It includes subjectivism and propensity and is motivated by constructivist philosophy. This means in particular that formal models are not to be interpreted as (even approximatively) matching objective reality, including "objective" requirements of a rational learning process of subjects. Instead, they should be taken as tools that structure (and thus change) our observations and therefore take part in the construction of what we call "reality."

2 The game of red and blue and the goodness-of-fit paradox

Here is how Gillies (2000, p. 78 f.) describes the "game of red and blue":

"At each go of the game there is a number s which is determined by previous results. A fair coin is tossed. If the result is heads, we change s to $s' = s + 1$, and if the result is tails, we change s to $s' = s - 1$. If $s' \geq 0$, the result of the go is said to be blue, whereas if $s' < 0$, the result of the go is said to be red. So, although the game is based on coin tossing, the results are a sequence of red and blue instead of a sequence of heads and tails. Moreover, although the sequence of heads and tails is independent, the sequence of red and blue is highly dependent. We would expect much longer runs which are all blue than runs in coin tossing which are all heads."

Note that a subjectivist could object that a "fair coin" is a construction that already assumes an objective interpretation of probability. However, let us assume that most subjectivists would agree with our probability assignments if they were allowed to test and inspect the sequence generating device intensively. If the initial value of s being -1 or 0 is decided by a coin toss, red and blue are exactly symmetrical. Gillies argument against de Finetti's subjectivist interpretation goes as follows:

"Two probabilists - an objectivist (Ms A) and a subjectivist (Mr B) - are asked to analyze a sequence of events, each member of which can have one of two values. Unknown to them, this sequence is in fact generated by the game of red and blue. ... Consider first the objectivist. Knowing that the sequence has a random character, she will begin by making the simplest and most familiar conjecture that the events are independent. However, being a good Popperian, she will test this conjecture rigorously with a series of statistical tests for independence. It will not be long before she has rejected her initial conjecture; and she will then start exploring other hypotheses involving various kinds of dependence among the events. If she is a talented scientist, she may soon hit on the red and blue mechanism and be able to confirm that it is correct by another series of statistical tests.

Let us now consider the subjectivist Mr B. Corresponding to Ms A's initial conjecture of independence, he will naturally begin with an assumption of exchangeability. Let us also assume that he gives..." equal probability $1/(n+1)$ to all possible numbers of blue results in a sequence of n goes (Gillies uses a different notation here). This assumption is made only for ease of computations, the argument given below also applies to any other exchangeable model. *"Suppose that we have a run of 700 blues followed by 2 reds. Mr B would calculate the probability of getting blue on the next go ... as* $701/704 \approx 0.996$. *... Knowing the mechanism of the game, we can calculate the true probability of blue in the next go, which is ... zero."* (It can easily be seen that s at the start of go 703 must be -2.) Gillies then further cites Feller (1950) showing that even in sessions as large as 31.5 millions of goes it happens with probability of about 0.7 that the most frequently occurring color appears 73 per cent of the times or more, in which case Mr B will estimate the subjectivist limit parameter θ_0 as at least 0.73 (here it is assumed that θ is prespecified as the parameter giving the probability of the color that turns out to occur most frequently). *"Yet, in the real underlying game, the two colors are exactly symmetrical. We see that Mr B's calculations using exchangeability will*

give results at complete variance with the true situation.

Moreover he would probably soon notice that there were too many long runs of one color or the other for his assumption of exchangeability to be plausible. He might therefore think it desirable to change his assumption of exchangeability into some other assumption. Unfortunately, however, he would not be allowed to do so according to de Finetti . . . ", because assuming exchangeability a priori means that with respect to all further Bayesian calculations only the number of observations being red or blue, but not their order, can be taken into account. Once assumed, exchangeability cannot be subsequently rejected. Otherwise, the coherence of probability assignments would be violated, as will be demonstrated in Section 4. As Gillies puts it, *"unless we know that the events are objectively independent, we have no guarantee that the use of exchangeability will lead to reasonable results."* This is the opposite of de Finetti's opinion that the assumption of objective independence can be made superfluous by subjective analysis using exchangeability.

An obvious objection to Gillies' criticism is the possibility to include deviations from exchangeability in a subjectivist Bayesian model. Mr B could put a prior probability of 0.9, say, on the exchangeability model and a probability of 0.1 on certain models specifying positive dependence between goes, which may fit the game of red and blue more or less adequately. Such an averaging over different models is in fact sometimes done in Bayesian statistics, though it is computationally complicated (e.g., see Hoeting, Madigan, Raftery and Volinski, 1999)[4]. If, after such prior modeling, the result of an observed sequence of experiments strongly indicates dependence, the posterior probability for exchangeability will be very small, and the subjectivist's conclusions will be much more realistic than the results under exchangeability given above. Gillies replies that *"what leads to so much complication is that on this approach it is necessary to consider all the possibilities in the very beginning of the investigation."* Since there is a very large number of possible kinds of dependencies, this is practically impossible. Further, Gillies reminds the reader that exchangeability, Bayesian conditioning and coherence lie at the heart of the Bayesian concept of "learning from experience". The modeling of complicated dependence structures may be justified from a pure subjectivist viewpoint, but it comes to the price of the loss of the attractive feature that different subjects with different prior models can be proved to arrive at consensus if the length of the random sequence converges to infinity. In Section 4 the discussion of the subjectivist approach to the game of red and blue will be continued.

Here is a criticism of the part of Gillies argument that concerns Ms A, the objectivist. Gillies seems to take for granted that the independence assumption or a certain model for dependence can be reliably rejected or confirmed by *"a series of*

[4]It is, however, difficult to find a simple model for deviations from exchangeability in the Bayesian literature so that a high probability, but smaller than 1, is assigned to an exchangeability model.

statistical tests". But doing subsequent statistical analyses after having applied a statistical test for assessing the model assumptions gives rise to serious problems, of which the "goodness-of-fit paradox" is the most striking one. The paradox can be stated as follows. Assume that the real distribution underlying a statistical experiment (which may be a sequence of repeated subexperiments) is Q. Assume further that a statistician carries out a statistical test ψ of level $\alpha = 0.01$, say, to confirm her null hypothesis that the real distribution is P, in order to base further statistical inference on the then confirmed assumption of P (P may or may not be equal to Q). Such tests to check an underlying model are called "goodness-of-fit tests" in the statistical literature, therefore the name "goodness-of-fit paradox". Let $\{\psi = 1\}$ be the event that ψ rejects the null hypothesis. Because ψ is an α-level test, $P\{\psi = 1\} = \alpha > 0$. Because P is used as a model assumption for the subsequent statistical analyses only if it is not rejected by ψ, the effective distribution R of the data for these statistical analyses is the true distribution Q conditional on the non-rejection of P, i.e., on $\psi = 0$. Thus, $R\{\psi = 1\} = 0 \neq \alpha$, and therefore $R \neq P$. This proves that, if statistical analysis based on the model assumption P are to be carried out only if P has been confirmed by a goodness-of-fit test, the effective distribution of the data for which P is assumed cannot be P, whether $P = Q$, i.e., P is the true underlying distribution of the *untested* data, or not. In other words: if we do not carry out a goodness-of-fit test, we do not know whether the model assumption holds. If we carry out such a test and *confirm* the model assumption, we can *disprove* its validity.

How does the goodness-of-fit paradox look like in the game of red and blue? An adequate independence test is the runs test (Lehmann, 1986, p. 176), which uses as its test statistic the number of runs of experiments yielding the same outcome. For example, the sequence with 700 blues and two reds has two runs, whereas a sequence of 700 blues followed by one red and one blue would have three runs. With a level $\alpha = 0.01$, the runs test rejects independence under 700 blues and two reds, but accepts it under 700 blues, one red and one blue. Under the distribution of sequences[5], for which independence is accepted by the runs test, a sequence of 700 blues and one red enforces blue in the 702nd go, while red would lead to a rejection of independence and is thus impossible. Thus, the resulting sequences are obviously dependent.

In the statistical literature it is well known that testing a model assumption, under which subsequent statistical inference is to be made, damages the validity of this assumption, see e.g. Harrell (2001, p. 56). Many standard textbooks on inferential statistics (e.g., Bickel and Doksum, 1977, p. 62) stress that null hypotheses and models should not be chosen dependent on the same data on which they are to be applied, because of the bias in subsequent analyses caused by preceding tests and also by decisions based on preceding informal data analysis (there is no fundamental difference between formal and informal decision making in this

[5]Note that under the propensity interpretation this distribution is over a thought experiment where series of 702 goes are repeated

respect). This implies that Ms A's strategy is not immune against Gillies' most important criticism of subjectivism: to avoid bias in the analyses, the objectivist has to specify her meta-model (all possible models to be tested) in the beginning of the investigation as well.

The effect of a concrete goodness-of-fit test can be assessed by simulations, as is done for normality tests in Easterling and Anderson (1978). However, I am not aware of any mention of the implications of the goodness-of-fit paradox in the discussion of the foundations of statistics with the single exception of Spencer-Brown (1957), which is seemingly ignored in recent statistics as well as philosophy.

3 Is hypothesis testing generally paradoxical?

In this section the meaning and the implications of the goodness-of-fit paradox are discussed, supplemented by some more general comments on hypothesis testing. It seems natural to me that the paradox gives rise to the following questions:

Question: Why is it relevant that the distribution *conditional on the acceptance of independence* is no longer independent? Should not the test apply to the *underlying* unconditional distribution, which may well be independent if the test confirms independence?

Answer: There are two different levels of answers to this question. From a practical point of view, consider the situation that a method of statistical inference, a confidence interval, say, is to be applied to the data conditional on the acceptance of independence, and that the known statistical properties of this method require the assumption of independence. The goodness-of-fit paradox results in a violation of the assumption, because effectively, the method is confronted with the conditional rather than the underlying distribution.

On a philosophical level, however, Ms A may still insist that she checks the independence model of the underlying process, regardless of the dependence caused in the conditional model. But there is a fundamental epistemological difference between checking deterministic predictions and probability models with a falsificationist attitude. The uncertainty about the precise outcome and the possibility of outcomes with a small probability are essential components of a probability model. It is dubious to take the realization of such a possibility as a reason to reject the model[6]. As long as a goodness-of-fit test is applied, the test (and not the underlying process) defines the sequences *that can be observed as generated by the model*, because we do not regard sequences for which independence is rejected as independent sequences. If

[6]One might ponder whether something similar to the goodness-of-fit paradox applies even to falsificationism of "deterministic" laws, because random measurement errors are often involved and handled by hypothesis tests.

full sequences of length 702, say, would be repeated, and Ms A would repeat her test (all the statistical theory of the tests is based on such an imagination), and would record all sequences for which independence was accepted, she could explore the resulting distribution afterward. She would then find dependence due to the goodness-of-fit paradox in these data, whether she would use them for subsequent analyses or not. My conclusion is that the conditional and not the underlying distribution is authoritative, because that is the one which is *observable* (the decision to which class we assign a phenomenon is part of the observational act). If we accept a model conditional on the result of a test, this test affects our perception of the tested phenomenon. Only if we accept a model without a test, we observe data directly from the underlying distribution - if we believe that there is one. This is hypothetical, because we have refused to test it.

Question: What about the practical relevance of the paradox? The violation of the assumption to be tested is based on only α false rejections of the null hypothesis. If α is small, can it be said that independence is fulfilled, if not exactly, then at least approximately?

Answer: The answer depends on what "approximately" means and what kind of analysis should be applied after the goodness-of-fit test. Usually, it is stressed that statistical analyses are based on model assumptions and that the analyses are only adequate if the assumptions are fulfilled. But model assumptions can never be verified: A "confirmation" by a goodness-of-fit test is not a verification. While α, the type I error of a statistical test, is controlled, the type II error - confirmation of a wrong model - can be arbitrarily large at least if flexible enough alternatives are considered. Thus, for all model-based methods, it is important *which* violations of the model tempt the method into misleading results, and whether such violations can be caused by the goodness-of-fit test paradox. In some situations this may indeed happen. For example, the omission of a variable with a non-significantly non-zero regression coefficient in linear regression can heavily bias the estimation of the other coefficients in unfortunate circumstances, and this becomes the worse, the more tests are used, see Harrell (2001). In other situations, the goodness-of-fit test may be harmless and may yield a distribution with which the method can cope very well, even if the real underlying distribution did not match the model assumption. In the game of red and blue, the runs test will reject the null hypothesis of independence for seven blues and two reds in a row at a p-value of 0.0099. Thus, if the real underlying distribution is independent, and Ms A uses the game of red and blue as an alternative, she will guess a probability of 0 for blue in the tenth go with a probability of about 1% (because in the game of red and blue, blue is impossible in this situation) instead of 7/9, the standard estimation under independence. This is the price for being able to reject the independence model when it does

not hold. The reader may judge whether this is harmless or not[7]. With 700 blues and 2 reds, as above, the type I error is smaller than 10^{-15}.

One could think that the influence of the paradox vanishes if α is chosen smaller and smaller. This is in principle true, but the smaller α is, the larger is the type II error, i.e., the probability to confirm a wrong model. Interestingly, Dawid (2004), who is concerned with a "de Finetti-Popper synthesis", namely the empirical falsification of a model belief of a subjectivist, proposes the so-called "Borel-criterion", which is mathematically equivalent to a hypothesis test with $\alpha = 0$, thus avoiding the goodness-of-fit paradox. But the avoidance is only theoretical, because his suggestions for attaining $\alpha = 0$ assume $n = \infty$, and a practical falsification for finite n will require $\alpha > 0$ as with a usual hypothesis test.

A formalization of the approximation of data by models can be obtained by defining distance measures between distributions (data can be seen as empirical distributions). This is discussed, e.g., in Davies (1995). Davies, however, avoids assuming the existence of objective probabilities. In his approach, the *observed data* are approximated and not an underlying true distribution; otherwise it could be observed that for arbitrary datasets, distributions with complex dependence structure can be constructed, which are not filtered out by any goodness-of-fit test, and which approximate the underlying objective distribution, if existent, arbitrarily badly in terms of any reasonable distance measure.

The consequence is that type I errors (which cause the goodness-of-fit paradox) can often be tolerated in practice, but not always. But this does only mean that the analyst is in a situation where his envisaged statistical method will work reasonably with large probability. It does not mean that there is any certainty that a real underlying distribution is met approximately - unless "approximation" is *defined* in terms of confirmation by a goodness-of-fit test.

Question: The paradox does not only apply to goodness-of-fit tests, but to statistical tests in general. Consider a one-sided t-test of the mean of a normal distributed random variable to be zero and assume that the null hypothesis holds for the underlying distribution. Conditional on acceptance of the zero mean by the test, data have neither a zero mean, nor are they i.i.d. normally distributed. Is hypothesis testing generally paradoxical and should it be abandoned?

Answer: I think that there is nothing wrong with hypothesis tests which are done at the end of a study, provided no analyses dependent on the outcome of these tests are done. If a new medicine is compared to an old one by means of a two-sample t-test of some measurement which is assumed to be i.i.d. normally

[7] A subjectivist, or any Bayesian, would say that this depends on the prior probability for the exchangeability model.

distributed among the patients, then the result of the test will be interpreted with respect to *new* patients. Therefore, the underlying distribution and not the distribution conditional on the test result is of interest. However, this is based on the untestable assumption that the underlying distribution is really normal (here, "approximate normality" in a well defined sense is sufficient) and that it is the same for the participants of the test and for new patients. In applied science, to the indignation of the philosophers, it is often useful to do things based on untestable assumptions.

Question: Should then goodness-of-fit testing be abandoned and should the scientists assume normality or independence without checking it?

Answer: Definitely not. There is a difference between the philosophical and the practical aspect. Philosophically, normality or independence cannot be checked without violating it. There is no observable objective distribution, not even in a falsificationist sense, because "critical" observation (with the option of rejection) does not leave the distribution unchanged. From a practical point of view, data analytic experience shows that analysis assuming an inadequate model is in most cases much worse than the slight violation of the model assumption by checking it. Informal graphical methods are often more informative than goodness-of-fit tests, but they do not prevent the goodness-of-fit paradox, since a researcher using these methods will also with positive probability reject a true model. Ms A is well advised not to trust the independence model.

Question: What about testing and making subsequent inference on different datasets? The goodness-of-fit tests and even a complicated model selection process may be applied to the first 1000 goes and the resulting model can then be used to analyze all later observations.

Answer: This avoids the goodness-of-fit paradox indeed and is often useful in practical model selection. However, it assumes that the first and the later dataset are independent of each other and follow the same model, which is untestable, so that the philosophical problem is only shifted (it would be somewhat ironic if an untested independence assumption would solve the paradox for an independence test).

There are some more difficulties with hypothesis testing, of which presumably multiple testing is the most notorious one. This is also relevant to Gillies' depiction of the game of red and blue, where Ms A is pushed into running even *"a series of statistical tests"*. On one hand, this is useful, because a single statistical test is only able to distinguish the null hypothesis from more or less specific alternatives (not every kind of dependence leads to "runs"), and they are useless if neither the null model, nor such distinguishable alternatives hold[8]. At the other hand, multiple

[8]It is easily forgotten that the alternatives against which a test is informative are not universal. Gillies (2000, p. 152) interprets the fact that the relative frequency of tails in some long coin

tests either raise the type I error of rejecting a true model or have to be corrected so that the power of every single test is reduced. Applicable corrections such as the Bonferroni correction (k tests of level α/k yield a probability of at most α of at least one rejection of the null hypothesis, cf. Holm, 1979) are almost always highly conservative, so that a researcher, who really wants his null hypothesis to be confirmed by a proper statistical method, has only to use so many tests that α/k gets smaller than any p-value[9]. Gillies (p. 155), confronted with the fact that one out of eleven independence tests applied to different "gambling systems" for coin tossing rejects his null hypotheses (which he does not want to be rejected), writes that *"our falsifying rule must be used with a certain "judiciousness""* and discusses that a single "falsification" out of a battery of tests should be tolerated under certain circumstances. But if he would give a formal rule for that, by this he would define a single combined test, and this is exactly what the Bonferroni correction does. Indeed, the Bonferroni combination with $\alpha = 0.05$ and $k = 11$ of Gillies' tests would not falsify independence, but would be rather conservative.

4 Is subjectivism the answer?

As mentioned in Section 2, subjectivists could model a possible deviance from exchangeability in principle, if they would specify alternative models and prior probabilities for this case before they start observing. Objectivists are in principle also not allowed to choose their hypothesis tests data-dependently, if they do not want the test distribution to be biased. Does this mean that Gillies' argument against the subjectivists is not valid, and that the subjectivists can even better cope with the game of red and blue, because there is no goodness-of-fit paradox in subjectivism? I do not think so.

Usually, not all strange phenomena that might occur in the data can be anticipated before the observations are made. Objectivists are not constrained so much by coherence considerations and change their models more easily, even paying the price that some analyses may be moderately biased afterward. An indication is that robust statistical methods, i.e., methods that can handle slight deviations from model assumptions, are much more widespread and successful in non-Bayesian than in Bayesian inference, because the possible deviation from a standard model does not have to be modeled precisely in an objectivist setting. The apparatus developed, e.g., in Huber (1981) and Hampel et al. (1986), is much easier to handle

tossing experiments does not deviate significantly from 0.5 as *"confirming the hypothesis of an unbiased coin and independent tosses."* But such a test is only powerful against probabilities different from 0.5 It assumes independence for the null hypothesis as well as for the alternative, and it does not provide any information concerning independence. Gillies knows better: on p. 155, he is concerned with real independence test.

[9] I admit that this does not always work; the minimum observed p-value becomes smaller with larger k.

than the Bayesian suggestions of Box (1980) and Berger (1994). Reasons for this can be found in Huber (1980).

The problem is not only computational but also philosophical. Robust Bayesian analysis requires that there is a "true" prior distribution of the subject covering all possible model deviations, though it may be unknown to the subject himself and unmeasurable by an operational elicitation procedure in finite time (cf. Berger, 1984). Subjectivists often criticize objectivists because they assume that there is a true frequentist probability or propensity, which is metaphysical. But the subjectivist assumption made about the subjects is metaphysical as well. The operational elicitation procedure proposed by de Finetti assumes an idealized situation because S (notation taken from the introduction) has to choose her probability assessments independently of any knowledge about the beliefs and the behavior of O. It is assumed that money (or whatever is at stake) has linear utility. The most problematic assumption is that the result of the measurement procedure is meaningful, i.e., that a belief of S exists which can be adequately expressed by values such as those elicited by the measurement procedure. The Bayesian model of probability beliefs is a mixture of a descriptive and a normative model. It is well known that people do not generally behave coherently, not even in the precisely defined Bayesian sense, see Walley (1991, p. 48) and the references given therein. But the operational measurements procedure forces them to do so, otherwise they are allowed to keep arbitrarily curious beliefs. I do not see any argument why there should be true underlying personal probabilities which are coherent (even if the concrete person would not act coherently unless forced), but not restricted in any other way. It is also not clear why the results of the operational procedure should be called "rational". Walley (1991, Chapter 5) discusses in depth that, under a lack of information, it is reasonable not to specify a precise betting quotient, namely the subjective probability, but to place bets on A up to a certain betting quotient p_- and to place bets on A^c only up to some $q_- < 1 - p_-$. This would mean that S could choose her betting quotients so that she has the feeling of making a good deal whenever O accepts some of her bets.

Is coherence[10] rational? Here is an elaboration of the incoherence which would arise if the subjectivist Mr B would reject his exchangeability model, which he had set up before the experiment, in the light of seemingly dependent observations in a situation like the game of red and blue. Let O now be the opponent of Mr B. Assume that Mr B had assigned probabilities to a sequence of 0-1 events, $X_1, \ldots, X_n$, judged to be exchangeable. Let P denote his whole initial model. Let $p = P(X_n = 1 | X_{n-1} = x_{n-1}, \ldots, X_1 = x_1)$ for some particular outcomes $x_1, \ldots x_{n-1}$. Choose these outcomes so that they seem to be incompatible with exchangeability in such a way that $X_n = 1$ seems to be favored compared with the exchangeability model, such as after, $(1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1)$. To operationalize the incoherence, O has to be allowed to choose conditional stakes before the experiment as well as later, once the values of $X_1, \ldots X_{n-1}$ have been observed. Let O choose

[10] Coherence is also required in Walley's (1991) alternative approach.

a stake $-S$ for the event $\{X_n = 1|X_{n-1} = x_{n-1}, \ldots, X_1 = x_1\}$ in the beginning, which means that the bet only takes place if $X_{n-1} = x_{n-1}, \ldots, X_1 = x_1$ will be observed. Then O has to pay pS to Mr B (I notate $S > 0$) and Mr B has to pay back S if $\{X_n = 1\}$. Suppose that O has held back at least another amount of S to bet again, unconditionally, after the $(n-1)$th observation.

After having observed $X_{n-1} = x_{n-1}, \ldots, X_1 = x_1$, Mr B realizes that he no longer believes in exchangeability and changes his betting quotient on $\{X_n = 1\}$ to $q > p$, because he now thinks that $\{X_n = 1\}$ is more likely than under his initial model. Now, O can generate a "Dutch book" by choosing stake S for $\{X_n = 1\}$, because this means that in case of $X_n = 0$ he gets $(q-p)S > 0$ from Mr B, and in case of $X_n = 1$, he gets $(1-p)S-(1-q)S = (q-p)S$ as well, thus Mr B is incoherent. But that is not the whole story. Assume that there is a real underlying (propensity) model Q with positive dependence among the observations, so that there is a good chance that Mr B observes something that lets him change his mind, and the real probability is $Q(X_n = 1|X_{n-1} = x_{n-1}, \ldots, X_1 = x_1) = q_0 \geq q$. If Mr B would now not change his probability from p to q, and O would place another stake of $-S$ on $\{X_n = 1\}$, then Mr B's expected loss would be $2q_0(1-p)S - 2(1-q_0)pS = 2(q_0-p)S$, which is at least twice the loss from the Dutch book. Even if O would not place any new stake on an unchanged p, Mr B would be better off in average if he would change his probability to q and O would choose the Dutch book. This means that in the situation of the game of red and blue, where Mr B's suspicion against exchangeability is justified, it is better for him to act incoherently than to follow his prior beliefs.

A subjective Bayesian might object that, if it is clear that $X_{n-1} = x_{n-1}, \ldots, X_1 = x_1$ would enlarge Mr S's probability for $\{X_n = 1\}$ compared to the exchangeability model, then exchangeability had already been the wrong model for Mr S's *prior* belief[11]. This is, in principle, a valid argument. However, in all but the simplest of setups, it is impossible to go through all possibilities, and this would also make the mathematics intractable. Simple models are preferred in most situations for reasons of interpretation, communication and understanding (in most cases, Bayesian modeling is *not* done for offering bets - and even the clients of bookmakers would prefer simple models). Furthermore, Mr B may change his mind not only because of the previous observations, but also because of new information from other sources than the modeled observations. It may therefore be more reasonable to start with an overly simple model and to modify it in case of a serious "prior-data conflict", even though this violates coherence[12]. "Robust Bayes" does not solve this problem, because it requires a specification of possible disturbances

[11] De Finetti (1937) was more optimistic (too optimistic, I think) about exchangeability, though. He wrote that *"it suffices that the conditions... do not present any apparent asymmetry"*. In objectivist probability modeling, analogously, independence models are chosen in most cases simply because the modelers have no idea where dependence might come from in the particular application. Independence is seldom tested if the experiment seems to be reasonably transparent.

[12] Even a very complicated model may be "overly simple", if conditional probabilities on some unexpected but possible events are not modeled separately.

of the model in advance. The incoherent but pragmatic approach is presumably taken by many practically oriented Bayesians. For example, Berger, 1984, suggests *"post-data modification"* of the prior without discussing the coherence issue.

This discussion should not be interpreted as proving that subjectivist modeling is wrong. The attractive feature of de Finetti's subjectivism is that it is a full, consistent theory for a subject S that *choses* to follow the Bayesian rules. Not only the assigned probabilities, but also this choice is subjective, and it has been shown that there are, at least in some situations, good arguments against this choice. I would not interpret the prior probabilities of S (or the consensus of a group of subjects) as a valid measurement of her real beliefs, but as an (often useful) construction. This construction is based on an interplay between S's (not exactly formalizable) beliefs and the Bayesian rules, i.e., the coherence requirement. Further important aspects are the format and the interpretability of the results of the analyses and the requirements of the (scientific) social system to which the subject belongs. A decision for exchangeability can be interpreted as the *decision to ignore the order of the observations*. It may be known that effects according to this order are not interesting with respect to a particular study, and this is an argument for exchangeability even if S would not exclude the occurrence of conditional dependence structures which would violate exchangeability. In general, the effect of prior modeling in Bayesian statistics is a prior weighting of possible different posterior outcomes, and this can be done not only because of belief, but also because of other reasons, e.g., to attain a desired balance between simplicity and accuracy in model selection. Further, if the subjectivist theory is not followed blindly and dogmatically, occasional remodeling in case of extreme prior-data conflicts or new information may seem acceptable (as well as goodness-of-fit testing and model checking by objectivists), even if it is incoherent with the prior choices.

5 Constructivist pluralism

I conclude this paper by sketching my own view of the discussions on the foundations of statistics. I call it "constructivist pluralism". It is pluralist in the sense that it accepts many interpretations of probability (including the ones discussed in the previous sections) as reasonable choices, but it does not accept any absolutist claim that one interpretation can be proved to be the best and no other is needed. It is constructivist in the sense that it is inspired by constructivist philosophy, by which I mean a philosophy based on the principles that

- there is no observation without an observer, no observation can be understood without reference to the observer, and observations can be described as *constructed* by the observer, by defining them through the act of perception,
- observations are constructed in social dependence, that is, social processes have to be analyzed to understand the act of perception (which, in particular,

includes classification by language),

- perception is a means of self-organization, not of representation.

In the literature there are various versions of constructivism. "Radical" and "social" constructivism are relevant for my exposition (introductions and anthologies are e.g. Berger and Luckmann, 1966, Watzlawick, 1984, von Glasersfeld, 1995, Gergen, 1999), whereas "methodological" and "mathematical" constructivism are not meant here. I do not discuss these interpretations in detail, but I think that the radical and social approach can be connected by acknowledging that there is a psychological as well as a social component of construction and none of these can be understood without the other.

The versions of constructivism indicated above agree that there is no objectivity in the sense of an observer-independent access to reality. This does not mean that it is claimed that no observer-independent reality exists. Constructivism does not dispute that it may be reasonable and useful imagine a reality outside of a personal brain and body and to use concepts which refer to such a reality, especially because it is a basic experience that we cannot change our perceptions as we like and we are clearly dependent on some outside world, be it as we perceive it or not. However, such concepts are necessarily metaphysical. This means that science, as far as it is concerned with the reality outside our brains and bodies, has to rely on metaphysical assumptions. I think that it is crucial to understand that, from a constructivist viewpoint, we are responsible for our constructed perceptions (because we construct them actively), but this does not mean that they are arbitrary. We perceive (i.e., we construct) strong physical and social limits of our freedom to construct arbitrary observations. That something is constructed does not mean that it does not really exist. The opposite statement makes sense: construction is the only accessible mode of existence.

Interpretations of probability are links between mathematical objects and the "real world", i.e., constructions of the "world outside" in personal and social reality[13]. A constructivist contribution to the discussion about the foundations of probability can be a theory about the relation of mathematical models to reality (see Hennig, 2002, for more implications of constructivism for statistics and data analysis). In Hennig (2003), I sketched such a theory by the scheme in Fig. 1. The main ideas are:

- Our perceptions and how we process them make up our personal reality. I have direct access only to my personal reality. I may assume that there is an observer-independent reality acting as an important condition of my perceptions, but I cannot subtract my acts of perception from it[14]

[13]This applies to subjective probabilities as well - as long as somebody else and not myself specifies them.

[14]The arrow from personal to observer-independent reality in Fig. 1 has a dual meaning: (a) It is a personal construction how we imagine such an observer-independent reality; (b) our actions change this objective reality.

- The social reality (which can be split up into the realities of different social systems) is made up of any communication (especially language) between persons. With respect to a personal reality, social reality is "outside" (while perceived communication is "inside"). Our communication is conditioned by our personal viewpoints, so that our personal reality affects the social reality.

  The repercussion of communication to personal reality is crucial. Once we have words, these words highlight and bundle some impressions and downweight others, so that we are no longer able to observe reality in a way we would do without language[15].

- The main difference between mathematics and ordinary language is that the mathematical concepts are aimed to be unambigously designed. 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 that is dependent on the individual and its social background. Thus, formal modeling is a mode of communication that aims at absolute agreement (or at least full understanding in case of disagreement). As between personal and social reality, there is a circular dependence between social reality and formalized reality. Concepts of social reality, e.g., ordinal language, are transferred into mathematical models, and the modeling process changes the communication about the concepts and therefore the social perception. Between personal reality and formalized reality, a similar cycle is at work.

- The different levels of reality operate on different domains. If we aim to transfer concepts from personal perception to communication and from usual language to formal models, they are fundamentally changed[16] (newly constructed, it could be said). A model is not a representation, but a construction which changes the perception of the modeled phenomenon, and by this it changes the phenomenon itself, since we cannot distinguish between the perceived phenomenon and the phenomenon itself.

The link between constructivism and the previous sections is that the goodness-of-fit paradox and the exchangeability assumption (and, more fundamental, any empirical check of a propensity model and the operational subjectivist measurement procedure) have been shown to be ways of *constructing* a perception of probability instead of *representing* something which is objectively there[17].

[15] I would interpret the debate between radical constructivists (von Glasersfeld, 1995) and social constructionists (Gergen, 1999) so that radical constructivists mainly concentrate on the construction of the personal reality and emphasize its effect on the social reality, while the social constructionist view is the reverse.

[16] Consider as an example the personal perception of a cow, the word "cow" and a cow as a well defined unit in some ecological model.

[17] Another illustration, which is related to the goodness-of-fit paradox, is given by Fine (1973, p. 91 f.). He shows that the apparent convergence of sequences of relative frequencies of some experimental outcome follows from apparent randomness under reasonable formal definitions

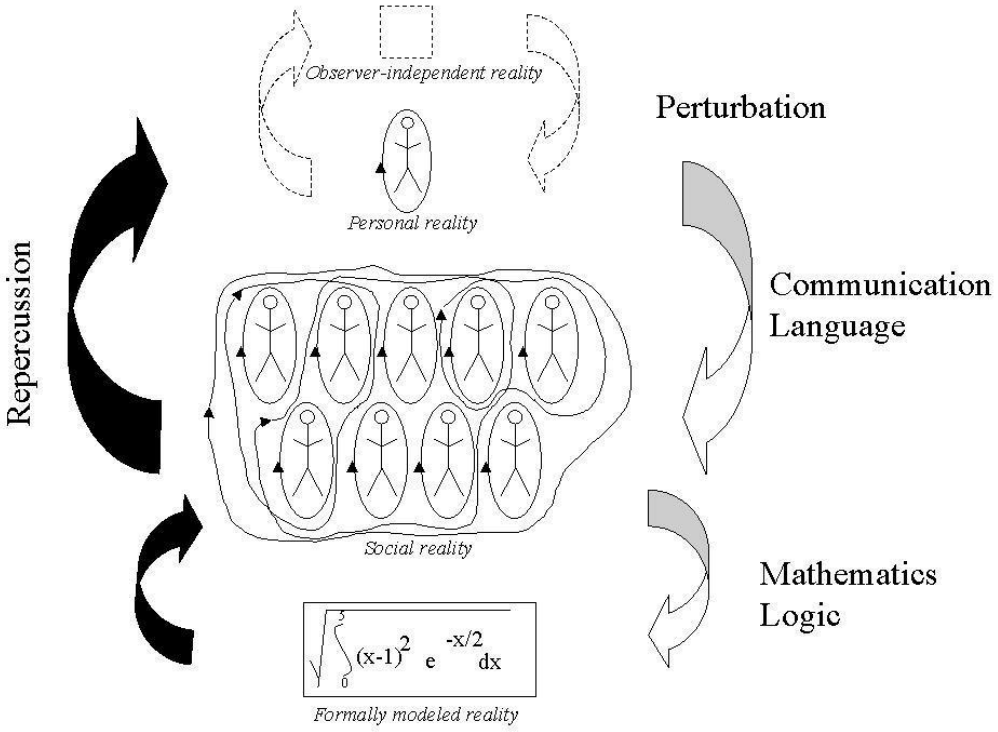

Figure 1: Informal model of the relation between reality and formal models.

I think that the arguments given in the previous chapters can be accepted even if a skeptical reader does not accept constructivism as a whole. For the purposes of the present paper it is not necessary to agree with the constructivist point of view, according to which simple material concepts such as a "tree" and basic mathematical facts are constructed. They are linked more directly to our daily actions so that their construction is more stable. If, for this reason, readers insist on an underlying reality or truth of such concepts, but not of probability, their point of view would still be compatible with the main points in this paper.

The term "objectivity" can be given various meanings. In the first meaning, as used above, objectivity refers to something which is strictly not observable and therefore metaphysical. While science has to start from such metaphysical assumptions, I think that it is generally accepted that they can be disputed and that science should be based as far as possible on observations and negotiated

of "apparent convergence" and "apparent randomness" for finite n. The frequentists use such apparent convergence to motivate their interpretation of probability. A possible interpretation is that, whenever we see a sequence of relative frequencies that does not seem to converge, we reject the randomness of the sequence of experiments. Thus, apparent convergence of relative frequencies is not an objective empirical fact (as von Mises suggests), but a mathematical consequence of our construction of randomness.

agreement.

A second meaning of "objectivity" often used in science is intersubjective agreement (as broad as possible). I view intersubjective agreement as a main and legitimate aim of science. Techniques used in science (including social science) such as logic, mathematical formalism and reproducible measurement procedures serve this aim. It is, however, important to recognize that personal and social reality have to be transformed to be fitted by such techniques. This second kind of objectivity, which science can more or less attain, does not guarantee by any means a correspondence between scientific statements and observer-independent reality[18]. Because science also aims at reliable predictions and successful actions, formalization in science is not arbitrary, but has to be connected to perception, i.e., different social and personal realities. This connection itself cannot be formalized or unified[19]. This is a fundamental limitation for mathematical modeling in science and, more generally, for scientific agreement[20]. The recent "objective Bayes" version of logical probability (Bernardo, 2003) aims at objectivity in this second sense. The problem with this is that the adaptation of a formal model to a complex real phenomenon enforces decisions to leave out, downweight or unify some aspects present in the personal or social perception, and such decisions can never be determined by objective guidelines. From the point of view of applied data analysis, I find Bayesian statistics much more convincing when the prior distribution is made up by a well-discussed formalization of useful available knowledge, which can never be justified indisputably, than when some ritualized "objective" reference prior is chosen, which should formalize the absence of information. Therefore, somewhat paradoxically, there may be stronger agreement (i.e., objectivity in the second sense) about the chosen methodology, when the modeling includes admittedly subjective decisions.

As a third meaning, Gillies (2000) seems to use "objective" in most places for *"a feature of the material world"* outside our brain and body. Because the construct of a "material world" is common in science, it is perfectly reasonable to construct a concept of probability which we place in the material world, and from which we can predict and explain some of the states of the observed world. This leads to a propensity interpretation with which nothing is wrong, unless it is claimed that the metaphysical assumptions can be "falsified" or even "confirmed" in an

[18] Agreement about scientific statements alone is no compelling argument in favor of correspondence to observer-independent reality. Scientific methods are constructed *in order* to attain agreement, and, thus, nothing else is needed to explain it. It could even be argued that formalized scientific reality is *further away* from any observer-independent reality (if its existence is taken for granted) than social and personal reality, because these two steps of transformation lie between formalized reality and the "reality outside", see Fig. 1.

[19] There could be a formalization, but then the connection of *this* formalization to perceived reality would be questionable etc.

[20] A means to total scientific agreement could be repercussion, i.e., change of social and personal realities in order to fit scientific explanations. However, this would also not lead to total agreement, because it is not be compatible with the encouragement of critical thought, which is still a principle of science for good reasons.

objective (first meaning) way.

Such a propensity, as defined by Kolmogorow's axioms (given that we agree about the arguments in favor of these axioms, which I do, see Gillies, 2000, p. 160 f.), is taken here as a mathematical model for some phenomena in the outside world that appears non-deterministic to our perception. No statement is possible about whether they are *really* non-deterministic. Once this limitation is accepted, the propensity interpretation can even be applied to situations which we do not construct as essentially random, but which are much easier to analyze by probability models than it would be to strive for an adequate formalization of every single influence factor. Certain economic time series provide an example, as long as a simple probability model does not seem to be inadequate from a data analytic viewpoint[21]. It could be said that such phenomena are *constructed as random* in the formalized reality, but not in the social reality made up by communication between scientists in ordinary language. The latter, however, may be intractable by precise scientific methods. A theory of adequacy checks, which are formally similar to hypothesis tests, is given by Davies (1995), who carefully prevents the notion of objectively (first meaning) existing probabilities, and describes probability models as tools of *"distanced rationality"*.

To summarize, "objective probabilities" should not be meant as "objectively existing (in the first sense)", but as "something that we perceive (thus construct) as a part of the material world", i.e. objective in the third sense.

De Finetti and the subjectivists claim that *"objective probability does not exist"*. They are right as far as they mean the first meaning of objectivity. It seems to me that philosophically interested scientists are often attracted by subjectivism and pushed away from objective probability, because the advocates of a propensity or frequentist interpretation suggest far too often that the first sense of objectivity is meant.

The subjectivists model personal degrees of belief instead of objective probabilities, which are metaphysical. But this does not solve the fundamental problem of mathematical modeling. While subjectivists do not make metaphysical assumptions about the material world, they formalize rational thinking and behavior. In order to do this in a unified manner, they are also forced to make dubious assumptions, see Section 4.

My thesis is that the gap between propensity models and non-deterministic phenomena in the material world and the gap between subjectivist probability models and rational behavior and thinking of real human beings are essentially the same. Such gaps necessarily open between mathematical models and the reality at which they are directed, and they can only be bridged by metaphysical assumptions.

[21] Gillies (2000) suggests his own version of probabilistic pluralism, according to which the propensity interpretation should be used in the natural sciences and the subjective interpretation should be used in the social sciences, including economics.

The crucial difference between objectivist and subjectivist interpretations of probability is that different things are modeled. Objective probabilities are about non-deterministic phenomena in the material world, subjective probabilities are about expected utilities of actions under uncertainty. Assigned correctly to these different aims, none of them is more or less valid than the other from a philosophical point of view[22]. There may also be room for logical and imprecise probabilities, but they have to face the fundamental limitations of mathematical models as well. Most problems in the foundations of probability are caused by inadequate claims of absolute superiority of some interpretations, by the lack of insight in the inevitability of metaphysical assumptions and subjective decisions[23], and by the lack of insight in the constructive character of formal modeling.

References

Berger, J. O. (1984) The robust Bayesian viewpoint. In: Kadane, J. B. (ed.): *Robustness of Bayesian Analyses*, Elsevier, Amsterdam, 64-125.

Berger, J. O. (1994) An overview of robust Bayesian analysis (with Discussion). *Test*, 3, 5-124.

Berger, P. L. and Luckmann, T. (1966) *The Social Construction of Reality.* Anchor Books, New York.

Bernardo, J. M. (2003) Discussion of Kardaun, O. J. W. F. et al. "Reflection on Fourteen Cryptic Issues Concerning the Nature of Statistical Inference". *International Statistical Review* 71, 307-314.

Bernardo, J. M. and Smith, A. F. M. (1994) *Bayesian Theory.* Wiley, Chichester.

Bickel, P. J. and Doksum, K. (1977) *Mathematical Statistics.* Holden-Day, San Francisco.

Box, G. E. P. (1980) Sampling and Bayes' inference in scientific modelling and robustness (with discussion). *Journal of the Royal Statistical Society, Series A*, 143, 383-430.

Davies, P. L. (1995) Data features. *Statistica Neerlandica* 49, 185-245.

Dawid, A. P. (2004) Probability, Causality and the Empirical World: A Bayes-de Finetti-Popper-Borel Synthesis. *Statistical Science* 19, 44-57.

[22]While I claim that subjectivist and objectivist probabilities model different levels of reality, they are obviously related and may compete in many situations. I agree with Hacking's (1975) thesis that the idea of probability was originally dual. Historical and social processes led to differentiation, but kept the similarities as well.

[23]Subjectivists are prone to this lack of insight as well, if they locate subjectivity only in the choice of the prior distribution, but not in the acceptance of particular formal rules for rational behavior.

de Finetti, B. (1937) Foresight: Its Logical Laws, its Subjective Sources. English translation in: Kyburg, H. E. and Smokler, H. E. (eds.): *Studies in Subjective Probability*, Wiley 1964, 93-158.

de Finetti, B. (1970) *Theory of Probability*, vol. 1 and 2. English translation, Wiley 1974.

Easterling, R. G. and Anderson, H. E. (1978) The effect of preliminary normality goodness of fit tests on subsequent inference. *Journal of Statistical Computing and Simulation* 8, 1-11.

Feller, W. (1950) *Introduction to Probability Theory and its Applications*. Wiley, New York.

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

Gergen, K. J. (1999) *An Invitation to Social Construction*. Sage Publications, Thousand Oaks.

Gillies, D. (2000) *Philosophical Theories of Probability*. Routledge, London.

Hacking, I. (1975) *The Emergence of Probability*. Cambridge University Press.

Hampel, F. R. (2001) An outline of a unifying statistical theory. In: de Cooman, G., Fine, T. and Seidenfeld, T. (eds.): *ISIPTA 2001. Proceedings of the Second International Symposium on Imprecise Probabilities and their Applications*, 205-212. Shaker.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. and Stahel, W. A. (1986) *Robust Statistics: The Approach Based on Influence Functions*. Wiley, New York.

Harrell, F. E. jr. (2001) *Regression Modeling Strategies*. Springer, New York.

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

Hennig, C. (2003) How wrong models become useful - and correct models become dangerous. In: Schader, M., Gaul, W., Vichi, M. (eds.): *Between Data Science and Applied Data Analysis*, 235-245. Springer, Berlin.

Hoeting, J. A., Madigan, D., Raftery, A. E., Volinsky, C. T. (1999) Bayesian model averaging: A tutorial (with discussion). *Statistical Science*, 14, 382-417.

Holm, S. (1979) A simple sequentially rejection multiple test procedure. *Scandinavian Journal of Statistics*, 6, 65-70.

Huber, P. J. (1980) Discussion of G. E. P. Box "Sampling and Bayes' inference in scientific modelling and robustness". *Journal of the Royal Statistical Society, Series A*, 143, 418-420.

Huber, P. J. (1981) *Robust Statistics*. Wiley, New York.

Jeffreys, H. (1931) *Scientific Inference*. Cambridge University Press.

Keynes, J. M. (1921) *A Treatise on Probability*. MacMillan, 1963.

Lehmann, E. L. (1986) *Testing Statistical Hypotheses*, 2nd ed. Wiley, New York.

Popper, K. R. (1957a) Probability Magic or Knowledge out of Ignorance. *Dialectica* 11, 354-374.

Popper, K. R. (1957b) The Propensity Interpretation of the Calculus of Probability, and the Quantum Theory. In S. Körner (ed.): *Observation and Interpretation. Proceedings of the Ninth Symposium of the Colston Research Society, University of Bristol*, 65-70 and 88-89.

Spencer-Brown, G. (1957) *Probability and Scientific Inference*. Longman, London.

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

von Mises, R. (1928) *Probability, Statistics and Truth*. 2nd revised English edition, Allen and Unwin 1961.

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

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