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

Christian Hennig,
Department of Statistical Science, University College London,
chrish@stats.ucl.ac.uk

November 29, 2006

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. It is shown in the present paper 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. This suggests that the existence of objective probabilities is an irreducibly metaphysical assumption, i.e., it is neither testable nor falsifiable by observations. On the other hand, the example is indeed a valid illustration of a shortcoming of the subjective interpretation. If subjectivists start with a priori-probability assignments assuming exchangeability of events, it is shown that incoherent behavior becomes rational under certain circumstances. Instead of rejecting the two discussed interpretations of probability, I argue that probability modeling inevitably requires the acceptance of untestable assumptions about reality. 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.

Note: This version appears exclusively as a research report of the Department of Statistical Science. A revised version of the part on propensities and the goodness-of-fit paradox (with an extended discussion of Gillies' propensity interpretation) is accepted for publication in Philosophia Mathematica (Hennig, 2007; a preliminary version is available on my webpage `http://www.homepages.ucl.ac.uk/~ucakche/`), but it doesn't contain the discussion of subjective Bayes and most of the general Chapter 6 of the present report.

Keywords: Hypothesis tests, constructivism, interpretations of probability, Bayes, coherence, rationality, constructivism

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. Two reasons are that the statement that a probability model holds implies the statement that unexpected outcomes, which would lead us to reject the model, are possible with positive probability, and that probability models are applied to whole observed sequences of events, e.g., to define independence between events, and there is only one observation of the whole sequence.

- The propensity concept of objective probability has an irreducible metaphysical core[1]. "Objective" probabilistic modeling involves subjective decisions about the acceptance of untestable assumptions.

- The subjectivist interpretation of probability does not solve this dilemma. The relation of its assumption that probability assignments should be coherent to actual personal beliefs is unclear and coherence is not necessarily rational.

- The propensity as well as the subjectivist concept (and presumably all other interpretations of probability, though this is not discussed in depth in this paper) do not only model but also affect and change perceived reality.

- The adoption of an interpretation of probability is a decision to adopt a certain structure for the perception of random phenomena. Any attempt to tie a probability model too closely to the idea of a true underlying reality has to fail because of the basic problem of formal modeling: the relation of a formal model to the modeled non-formal reality cannot be formally modeled and tested. This does not only apply to objective probabilities, but also to the relation of subjective probabilities to the true personal beliefs which are attempted to be modeled.

In Section 2, the basic elements of Gillies' propensity interpretation of probability and de Finetti's subjectivist interpretation are briefly reported. In Section 3, "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, 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, 67-95) and was used by Popper (1957a) to argue against the possibility of inductive logic.

[1] Among the objective probability concepts not discussed in detail in the present paper, this holds as well for von Mises (1928) frequentism as argued, among others, by de Finetti (1970), but not for classical probability definitions based on relative frequencies in purely finite settings.

Part of the discussion of the game of red and blue is the goodness-of-fit paradox, which implies that hypothesis tests in itself violate the probability model they are meant to confirm and therefore cannot check such a model. In Section 4 I discuss whether proper interpretations of hypothesis tests are possible nevertheless. In Section 5 the subjectivist interpretation of probability is discussed in the light of the game of red and blue. I argue that the subjectivist as well as the propensity interpretation leads to inevitable problems as long as they are considered as referring to something that exists objectively so that this existence can be checked by observations (I understand many subjectivists in a way that they claim that this holds for the personal degree of belief modeled by subjectivist probability). However, in Section 6 I sketch a version of "probabilistic pluralism", motivated by constructivist philosophy, which accepts different interpretations as models for our perception of random phenomena, provided the general limitations of formal modeling are kept in mind. This means 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 Propensities and subjective probabilities

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 propensities applicable to singular cases as advocated by Popper (1990; note that this is different from Popper's earlier works), for which empirical testability is out of question. 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[2] 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[3], 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 an amount of $p_j c_j$ 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[4] 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 a 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

[2] 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.

[3] 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.

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

> distributions of θ obtained 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 1991, Hampel 2001). The reader is referred to Fine (1973), Gillies (2000) and to the cited original literature for a discussion of these interpretations. Gillies (2000, Chapters 3-6) gives arguments why he prefers the subjectivist interpretation over the logical one and the propensity interpretation over the frequentist one, with which I agree. 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.

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

Here is how Gillies (2000, 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 (prior distribution). *"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 5. 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)[5]. 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 at 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 5 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 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 $\alpha > 0$ 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)[6]. 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 level-α 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 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

[5] I have, however, not succeeded 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.

[6] Note that in almost all practically relevant settings, a hypothesis test with $\alpha = 0$ is useless because the probability to reject the null hypothesis is then 0 even if the alternative holds.

for which P is assumed (which is the distribution over a long run of data sets for which P is accepted) 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 cannot check in a falsificationist manner whether P holds. If we carry out such a test and *confirm* P, we can *disprove* that P is the distribution of a long run of observed data sets.

A crucial point in this argument is that P is a probability model for a whole data set, not only for a single go. "Independence" as used by Gillies is defined in terms of the probability calculus and applies to the whole sequence of goes. But this means that there is only one observation to check P, namely the whole data set. Consequently, there is no long run (be it long but finite or infinite) to which a propensity interpretation of the independence model could refer.

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 full data sets 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, there is an obvious dependence between the goes in these sequences.

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, 56). Many standard textbooks on inferential statistics (e.g., Bickel and Doksum 1977, 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 (which refers to what I call violation of the model assumption by the goodness-of-fit paradox) in subsequent analyses caused by preceding tests. Often, model checks by goodness-of-fit tests are replaced by informal graphical methods such as normal probability plots or time series plots in applied data analysis, but this does not prevent the goodness-of-fit paradox, because such graphical methods lead to the rejection of a true probability model with positive probability as well, with the only difference that this probability cannot be explicitly computed because of the intuitive non-formal nature of these methods.

The effect of a goodness-of-fit test on subsequent analyses which assume a seemingly confirmed distribution P 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.

4 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 goes *conditional on the acceptance of independence* are no longer independent? Should not the test apply to the *underlying* unconditional distribution, which may well be independent if the test confirms independence?

Answer: 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 after application of the test, 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. 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 full sequences of length 702, say, were repeated, and Ms A repeated her test (all the statistical theory of the tests is based on such an imaginary repetition), and recorded all sequences for which independence had been 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 relevant, because that is the one which is *observable*. The decision to which class we assign a phenomenon is part of the observational act and is necessarily an integral part of the idea of a repetition and a long run, as long as propensities are not seen as something generally unobservable and therefore entirely metaphysical.

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 could observe data directly from the true underlying distribution - given that we believe that there is one.

Question: What about the practical relevance of the paradox? The violation of the assumption to be tested is based on only a proportion of 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 cause misleading results, and whether such violations can be caused by the application of goodness-of-fit tests. In some situations this may indeed happen. For example, the omission of a variable with a non-significant non-zero regression coefficient in linear regression can heavily bias the estimation of the other coefficients in unfortunate circumstances, and this becomes 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 under the real underlying distribution the goes are 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.

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

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 by the probability model and not an underlying true distribution.

The consequence of this discussion 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 presumably work reasonably[8]. It does not mean that there is any certainty that a real underlying distribution is met approximately by the accepted null hypothesis P - 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 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 (see the first question). 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 some 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?

[8] I am almost sure that the following result could be proven: Be Q a Bernoulli model for independent goes with true probability p for heads in a single go. Let R be the distribution Q conditional on confirmation of independence by the runs test. Let C be the usual confidence interval to the level β calculated as if the underlying distribution were Q. Then $R\{p \in C\} > \beta$. In words, the probability that the true parameter is inside the confidence interval is *larger* under R than under the erroneously assumed Q. Therefore, the goodness-of-fit paradox, while distorting Q, has no negative consequences on the succeeding statistical analysis in this case except of the confidence level becoming conservative.

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. 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.

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 the 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[9]. On the other hand, multiple 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 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[10]. 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

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

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

he gave 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.

5 Is subjectivism the answer?

As mentioned in Section 3, subjectivists could model a possible deviance from exchangeability in principle if they specified 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 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, 48) and the references given therein. The operational measurement procedure forces them to do so - and as long as they are coherent, there are no other restrictions and the people are allowed to keep arbitrarily unreasonable beliefs. I do not see any convincing argument why there should be true underlying personal probabilities which are coherent (even if the concrete person did 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[11] rational? Here is an elaboration of the incoherence which would arise if the subjectivist Mr B rejected 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 prediction from an exchangeability model, such as after, $(1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1)$. To operationalize the violation of de Finettis coherence demand, 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 a stake $-T$ 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 $X_{n-1} = x_{n-1}, \ldots, X_{n-1} = x_{n-1}$ will be observed. Then O has to pay pT to Mr B (I assume $T > 0$) and Mr B has to pay back T if $\{X_n = 1\}$. Suppose that O has held back at least another amount of T 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 T for $\{X_n = 1\}$, because this means that in case of $X_n = 0$ he gets $(q-p)T > 0$ from Mr B, and in case of $X_n = 1$, he gets $(1-p)T - (1-q)T = (q-p)T$ as well, thus Mr B is incoherent. But this 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 now did not change his probability from p to q, and O placed another stake of $-T$ on $\{X_n = 1\}$, then Mr B's expected loss would be $2q_0(1-p)T - 2(1-q_0)pT = 2(q_0 - p)T$, which is at least twice the loss from the Dutch book. Even if O did not place any new stake on an

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

unchanged p, Mr B would be better off on average if he changed his probability to q and O chose 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 $X_{n-1} = x_{n-1}, \dots, X_1 = x_1$ would increase Mr S's probability for $\{X_n = 1\}$ compared to the exchangeability model, then exchangeability was already the wrong model for Mr S's *prior* belief[12]. 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). An important argument of de Finetti to favor the subjective approach over the frequentist (and propensity) one is that his subjectivist definition is operational - probabilities are observable and not metaphysical. This is not compatible with claiming after the occurrence of an inconsistency that the "true" subjectivist probability had not properly been observed in advance. Such a claim may lead to a posteriori adjustment of the prior probability - and this means that observing betting behavior with the subjectivist framework in mind actually *changes* the observed probabilities.

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[13]. "Robust Bayes" does not solve this problem, because it requires a specification of possible disturbances 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 *chooses* 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) artificial construction. This construction is based on an interplay between S's (not exactly formalizable) beliefs and the Bayesian rules,

[12] 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.

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

i.e., the coherence requirement. Further important aspects are the format and the interpretability of the results of the data analyses for which the Bayesian computations and assignments are used 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 did not exclude the occurrence of conditional dependence structures which would violate exchangeability. Through de Finetti's representation theorem, exchangeability guarantees the representation of the a posteriori-distribution by a distribution over a single interpretable parameter (the probability for "1" in a sequence of 0-1 goes) and exchangeability is often chosen because such an interpretation is desired. 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.

6 Discussion: probability models and reality

To summarize the main points, it has been shown that hypothesis tests cannot be used to check objectively interpreted probability models in a falsificationist sense. More generally, if (as usually done and necessary to define "independence") a whole sequence of observations is modeled by a probability model, there is only a single observation (namely the whole sequence) at hand to judge the adequacy of the model. Every "long run" interpretation of probability can therefore at best be very weakly linked to the observations.

On the other hand it has been shown that in a subjectivist setup incoherent assignments of probability can be rational, given that the initial assignment of a priori probabilities involves exchangeability assumptions or other simplifications (as usually made in the applied literature) that can get in conflict later with the then observed data. If subjectivists insist on coherence and claim that in these cases the initial probability assignment has not been "true" but there has been a true coherent a priori-probability which has not been observed because of simplicity considerations, the subjective interpretation loses its operationality and has to be treated as dealing with an unobserved and unobservable hidden truth - which was the main target of de Finetti's criticism of objectivist interpretations. Note further that exchangeability plays a key role in the Bayesian concept of learning from experience, and therefore the adoption of complex models that can deal with seemingly correlated data (where the correlation does not come in a handy format including an exchangeable error term or something similar) would spoil some other arguments that have been used to favor the (subjectivist) Bayesian approach. For the subjectivist interpretation, the link between an a priori probability model and the observable reality (about personal beliefs) becomes very weak as well, even though it is at least conceivable that a subjectivist performs the Herculean task of assigning probabilities to all possible outcomes of a sequence of moderate length so that all possible deviations from the idea of exchangeability are taken into account a priori. Even such a subjectivist, however, is forced to ignore every new knowledge becoming available in the course of the random experiment from other sources than those modeled in advance in order to behave coherently, which does not seem to be very rational.

Several good arguments are around as well (e.g., in Gillies, 2000) to dispute strong links to observable reality for the frequentist approach, Popper's (1990) single case propensity approach and the objectivist Bayesian approach.

Instead of looking for a probability concept that is a "better" reflection of reality than those mentioned above, I suggest to acknowledge that the very idea of probability as formalization of uncertainty always has to involve non-observable components such as the possibility or probability of events that did not happen or the concept that some events had been unlikely though they actually happened.

If probability interpretations are seen as concepts to structure our perception of ideas that do not necessarily refer to something objectively existing (be it personal beliefs), it can further be acknowledged that such interpretations *changes* our perception and therefore the observed reality. Here is one example in the context of independence tests. Assume that a researcher wants to analyse several series of observations using an independence assumption, and we test this assumption for the different series using the runs test, say. In some cases independence will be rejected. She will then consider the series for which independence has been rejected as belonging to a different "class" from the others, she will start to look for irregularities in the experiments leading to these special series and quite often she will find some, while the "tested-as-independent" series are accepted as they are[14].

It has been demonstrated in the present paper how the checking procedure for long run-propensities changes what is actually observed as a "run". In every application of a long run-concept it is necessary to define what a repetition is. This always involves building up a class out of non-identical events, and this can never be fully objective (as de Finetti, 1970, pointed out). It can also be seen how the demand of coherence changes people's behavior in many cases or at least their observable probability assignments (the a priori assignments of the subjectivist who takes all possible outcomes sensibly into account in

[14] Another illustration, which is related to the goodness-of-fit paradox, is given by Fine (1973, 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 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.

the beginning also seem to me rather a construction enforced by the Bayesian methodology than a proper representation of what she believed before).

In this sense I see probability interpretations as methods to construct particular perceptions of reality[15]. With this in mind, different interpretations give rise to different perceptions, none of which is in itself right or wrong. Long run-concepts support (and change) our perception of phenomena in the "outside world" (and are "objective" only in this sense) which we construct as "regular" or "repetitions", but can by no means claim that they are the only legitimate way to formalize such phenomena. The subjective concept supports (and changes) our perception of rational processing of degrees of belief, but alternative concepts neither of rationality nor of modeling belief can be ruled out.

Therefore my attitude toward the foundation of probability is pluralist, but in a different sense than that of Gillies (2000). While it makes sense to think about the internal consistency of probability interpretations, a final decision about "better" or "worse" concepts cannot be made by philosophers. Every scientific problem using probability concepts needs a new decision in which way we want our perception to be structured and changed and in which aspects of uncertainty we are interested. This is necessarily subjective and not general. The attitude discussed here puts me in the vicinity of constructivist philosophy (see, e.g., Berger and Luckmann 1966, Watzlawick 1984, von Glasersfeld 1995, Gergen 1999) and is in fact inspired by it. It may be worthwhile to elaborate further what constructivism has to offer to the analysis of mathematical models of reality and to the foundations of probability in particular.

References

[1] Berger, J.O. (1984). The robust Bayesian viewpoint, in J.B. Kadane (ed), *Robustness of Bayesian Analyses*, Elsevier, Amsterdam, pp. 64-125.

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

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

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

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

[6] 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.

[15] I would apply this idea to more general mathematical models of reality. This is discussed in more detail in Hennig (2002, 2003).

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

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

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

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

[11] 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.

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

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

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

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

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

[17] 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.

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

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

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

[21] Hennig, C. (2007). Falsification of propensity models by statistical tests. To appear in *Philosophia Mathematica.*

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

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

[24] 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.

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

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

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

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

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

[30] 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*, pp. 65-70 and 88-89.

[31] Popper, K.R. (1990). *A World of Propensities*. Thoemmes.

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

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

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

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

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