

# Falsification of propensity models by statistical tests and the goodness-of-fit paradox

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

August 16, 2006

## Abstract

Gillies (1973) introduced a propensity interpretation of probability in which probabilities are linked to experience by means of a falsifying rule for probability statements. The present paper makes two main contributions. (1) It is argued that general statistical hypothesis tests should be allowed as falsification rules instead of the restricted rules proposed by Gillies. (2) The “goodness-of-fit paradox” is introduced, which can be stated as follows: the confirmation of a probability model by a statistical (goodness-of-fit) test refutes the validity of the model.

The paradox is illustrated with an analysis of “the game of red and blue”, which has been used by Gillies (1973, 2000) and Popper (1957a). Several possibilities to interpret the paradox and to deal with it are given. The validity of Gillies’ propensity interpretation is discussed in the light of the results of the paper. The conclusion is that the propensity approach is useful, but the connection of propensity models to the observed reality is weaker and needs more subjective decisions of the researcher than expected by Gillies.

**Keywords:** Hypothesis tests, interpretations of probability, Neyman-Pearson theory

## 1 Introduction

The present paper belongs to the realm of the foundations of probability, which seems to be a somewhat unusual topic in *Philosophia Mathematica* up to now. However, probability theory is a branch of mathematics, and therefore the foundations of probability treat the relation of a particular class of mathematical models to reality, which should be a legitimate part of the philosophy of mathematics.

Gillies (1973) introduced a propensity interpretation of probability in which probabilities are linked to experience by means of a falsifying rule for probability statements (FRPS), which will be discussed and modified in the present paper.

In general, propensity approaches interpret probabilities as objective strength of the tendency (“propensity”) of a situation or experimental condition to bring forth a certain outcome. In so-called “long run propensity theories”, a probability of an event corresponds to the relative frequency of successful outcomes under potentially infinitely many repetitions of the experiment (it is required that arbitrarily many repetitions are possible in principle, according to general natural laws). Gillies’ interpretation belongs to this class, as well as the early propensity theory of Popper (1957b). A particular feature of Gillies’ approach is that he explicitly requires the repetitions to be independent. Note that this does not rule out probability models for dependent sequences of random experiments such as Markov chains. They can be interpreted in terms of independent repeatability of the whole sequence.

As an alternative to “long run” theories, “single case” propensity interpretations have been proposed, e.g., by Fetzer (1983), Popper (1990) and Miller (1994). Propensity theories can be distinguished from frequentist objective probability interpretations (e.g., von Mises, 1928), which are operationalist, while propensities are theoretical quantities which are not directly observable. This avoids some important problems of the frequentist approach, namely the necessity to observe “approximately infinitely long” sequences (under the propensity interpretation, it suffices that they are “possible in principle”) and the precise specification of gambling systems under which probabilities are unchanged (see, e.g., chapter 5 of Gillies, 2000). However, the connection of the propensity concept to experience is less obvious than for the frequentist interpretation.

Gillies (1973) suggests the following link: if a high probability is assigned to certain outcomes of an experiment, such outcomes can be treated as predicted with practical certainty. The model can then be falsified if such a prediction is not fulfilled in reality. This idea is equivalent to a statistical hypothesis test, although Gillies only allows particular hypothesis tests as FRPS. In Section 2, I will argue that Gillies’ definition of an FRPS is too restrictive and more general tests should be allowed. This also strengthens the connection between Gillies’ interpretation and the Neyman-Pearson theory of statistical tests.

As Gillies himself already noticed, there is an essential difference between his FRPS and traditional “Popperian” falsification rules for deterministic models. Because in statistical tests there is always a certain probability of rejecting a true null hypothesis ( $H_0$ ), repeated application of the FRPS in an attempt to falsify a certain model by data generated from exactly that model will inevitably lead to some erroneous falsifications of the true model. Though it may be argued that erroneous falsifications of deterministic models may happen as well because of random measurement errors, the positive probability of rejecting a true probability model is more serious. It leads to the “goodness-of-fit paradox”, which I will introduce in Section 3, and which essentially refers to the fact that it can be proved in a certain sense that a probability model *is violated by successful application of the FRPS*, i.e., if it is *confirmed* by the FRPS. Therefore, it could be argued that a probability model, which is confirmed by the FRPS, doesn’t hold anymore even if it held before. The goodness-of-fit paradox will be illustrated by “the game of red and blue”, which 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.

In Section 4, in a summarizing discussion, I will discuss the validity of statistical hypothesis tests as falsification rules for propensity models in the light of the goodness-of-fit paradox, some possible methods to deal with the paradox and some connected issues such as multiple testing. I will argue that Gillies' propensity approach (with the modification given in Section 2) can be maintained if it is interpreted in a more modest way, namely as a valid description of how scientists can think about uncertainty in a rational way. It should be acknowledged that there are some inevitably subjective and metaphysical elements in modelling uncertainty in terms of propensities, but such elements cannot be prevented by any interpretation of probability.

Though especially the Sections 2 and 3 may seem to be quite critical on Gillies' approach, I'd like to emphasize in advance that I find the basic principle of giving a long run propensity interpretation using hypothesis tests as falsification rules quite fruitful. I agree with many arguments given by Gillies (2000) to support his interpretation and to highlight shortcomings of others, though I won't discuss them in the present paper.

The term "subjective" is used several times in the present paper and it is emphasized that subjective decisions (e.g., to decide about a test statistic and a rejection region) are necessary. I use "subjectivity" here in a quite broad sense, meaning any kind of decision which can't be made by the application of a formal rule of which the uniqueness can be justified by rational arguments. Note that "subjective decisions" in this sense can (and often do) take into account subject-matter knowledge, and can be agreed upon by groups of experts after thorough discussion, so that they could be called "intersubjective" in many situations and are certainly not "arbitrary". However, even in such situations different groups of experts may legitimately arrive at different decisions.

## 2 Tests as falsification rules, and the role of alternative hypotheses

### 2.1 Preliminaries

The elements of Gillies' interpretation are:

- a probability space obeying Kolmogorov's axioms,
- a specification of conditions of a real experiment, which can in principle be repeated independently an arbitrary number of times<sup>1</sup>, and

---

<sup>1</sup>Gillies admits that not all conceivable conditions of an experiment can be held fixed - at least the point in time and space can't be repeated identically -, and he doesn't require that all details of the experiment are specified, only a subset chosen by the scientist. Furthermore, a so-called "spacing condition" may be required which forces the repetitions to be separated

- the FRPS, which can be used to falsify or confirm a probability model using a finite amount of data generated by the specified experiment.

In Section 9 of Gillies (1973), the FRPS is defined as follows.

**Definition:** Let  $X$  be a random variable (RV) with values in some space  $\mathcal{R}$  (which in most cases will be  $\mathbb{R}$ ), distributed according to a distribution  $P$  with a density  $f$  (notation:  $\mathcal{L}(X) = P$ ). The distribution  $P$  is *falsifiable*, if it is possible to partition  $\mathcal{R}$  into disjoint sets  $C$  and  $C^c$ ,  $\mathcal{R} = C \cup C^c$ , where

- (i)  $P(C) = \alpha \leq \alpha_0$  where  $\alpha_0$  is some small “critical probability”, e.g.,  $\alpha_0 = 0.05$  (I’ve chosen a notation here that looks more similar to the usual notation of statistical tests than Gillies’ notation),
- (ii)  $\max_{x \in C} f(x) / \max f = l < l_0$ , where  $l_0 < 1$  is a small critical value,
- (iii)  $\max f$  is “in some sense” representative (or at least not hugely atypical) for the values of  $f$ . Gillies (1973) does not define more precisely what is meant by “in some sense”, but he gives examples. From the discussion on his p. 170 it can be seen that Gillies intends (iii) to entail  $f(x) > l \forall x \in C^c$ .

$P$  as a model for  $X$  is *falsified* if the observed realization of  $X$  is in  $C$ .

Part (i) defines a conventional statistical hypothesis test, while (ii) and (iii) are side conditions that restrict the statistical tests qualified for being chosen as FRPS (and the corresponding potentially falsifiable distributions) to such tests that Gillies finds “intuitively justified”. Note that, as long as we don’t allow randomized tests, a general statistical hypothesis test is defined by the specification of a test statistic and a rejection region  $C$ . In situations where the choice of the test statistic is not of major interest, I will sometimes refer to rejection regions alone without explicitly defining the corresponding test.

I will argue below that general hypothesis tests should be allowed and the restrictions (ii) and (iii) should be dropped.

Figure 1 illustrates Gillies’ definition. A  $\mathcal{N}(0, 1)$ -distribution (normal distribution with mean 0 and variance 1) fulfills the conditions of a falsifiable distribution with  $\alpha_0 = 0.05$ , where  $C = C_1 = (-\infty, u_{0.025}] \cup [u_{0.975}, \infty)$ ,  $u_\beta$  being the  $\beta$ -quantile of the  $\mathcal{N}(0, 1)$ -distribution.  $\max_{x \in C_1} f(x) / \max f$  is obviously small and  $\max f$  is, according to Gillies (1973, p. 174), “representative” enough for the values of  $f$ . Therefore (ii) and (iii) can be accepted as fulfilled.

The region  $C_2 = [u_{0.475}, u_{0.525}]$  shown in bright gray is a rejection region of a valid hypothesis test with  $\alpha = 0.05$  as well, but it doesn’t fulfill Gillies’ conditions, because  $C_2$  contains the largest density values. According to Gillies, it would be “counter-intuitive” to reject a  $\mathcal{N}(0, 1)$ -distribution in case of  $X \in C_2$  because he considers the values where the density is high as the most typical values of a distribution.

---

enough from each other in order to be interpretable as independent, see Gillies (1973, p. 95).

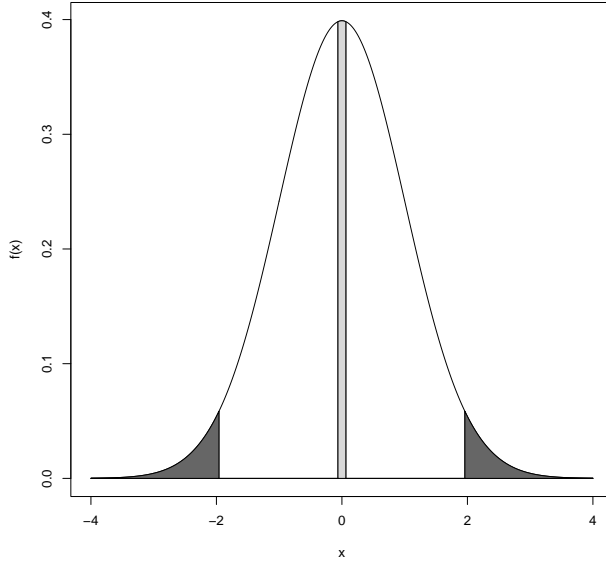


Figure 1:  $\mathcal{N}(0, 1)$ -distribution with rejection region  $C_1$  in the tails (dark gray), which can be used for Gillies' FRPS, while the rejection region  $C_2$  (bright gray) is not allowed by Gillies' definition because  $\max_{x \in C_2} f(x) / \max f = 1$ .

Note that, because  $f(x) > l \forall x \in C^c$  is required, all sets  $C$  corresponding to potential FRPSs have the form  $(-\infty, u_{\alpha/2}] \cup [u_{1-\alpha/2}, \infty)$ . In particular, one-sided tests are ruled out.

Therefore, given an RV and its distribution, the choice of the statistical test to construct an FRPS seems to be quite restricted. However, the choice of the RV itself is not restricted by Gillies. The reason can be illustrated by the following example (Gillies, 1973, p. 193). Consider standard coin-tossing where the  $n$  tosses are modelled as independent and the probability for heads and tails is 0.5 each. This means that every possible sequence  $\tilde{x}$  of tosses has the same probability  $f(\tilde{x}) = 0.5^n = \max f$ . Therefore, this distribution is not falsifiable according to Gillies' definition. To be able to falsify this model nevertheless, Gillies suggests to consider the RV which gives the numbers of heads. This RV has a falsifiable distribution, given that  $n$  is not too small. The smallest density values occur for results close to 0 and to  $n$ , and depending on  $\alpha_0$  and  $l_0$ , an FRPS based on a suitable rejection region  $C$  containing the tails of this distribution can be defined.

Gillies (1973) discusses in some length (Chapter 11) why he thinks that the choice of the test for his FRPS should be restricted and why more general tests, particularly as derived by the Neyman-Pearson theory, should be ruled out. His main arguments are:

- (a) Counter-intuitive tests (see above) should be avoided.
- (b) According to the Neyman-Pearson approach, the choice among the many possible tests of a particular  $H_0$  has to be based on an alternative hypothesis, but Gillies doesn't want his FRPS to depend on the choice of an alternative because
  - (i) the choice of an alternative hypothesis leads to a certain arbitrariness, and
  - (ii) in many practical situations, there is no alternative hypothesis of particular interest against which the  $H_0$  should be tested.

Here is an overview of my arguments why general hypothesis tests should be allowed for defining an FRPS and alternative hypotheses should be taken into account, as it is done in Neyman-Pearson theory. Note, however, that I don't suggest to stick to Neyman-Pearson optimality theory in all cases, because it may be advantageous, depending on the practical situation, to use tests which are good against many alternative hypotheses rather than optimal against a single one<sup>2</sup>. More details will be given in the following subsections.

1. (Subsection 2.2) Gillies' restricted tests can distinguish the  $H_0$  from some, but not from all alternative models. This means that, according to Gillies, a probability model cannot be falsified even if it is clearly wrong, given that the true distribution belongs to the class of alternatives that can't be distinguished from the  $H_0$  by the given test.
2. (Subsection 2.3) The concept of an alternative hypothesis and of the power of a test (namely the probability not to reject the  $H_0$  given that the alternative is true) is quite useful to understand what a given test (which may or may not be Gillies' proposal) essentially does, and it is actually necessary to explain why it doesn't make sense to choose  $\alpha_0 = 0$ . Such a choice would look quite attractive from Gillies' theory alone, but is certainly not sensible.
3. (Subsection 2.4) The fact that Gillies allows arbitrary RVs undermines his own restrictions. By using transformations of RVs, tests can be constructed that fulfill Gillies' conditions but are equivalent to the counter-intuitive tests that he wants to rule out.

## 2.2 Wrong models that cannot be falsified by the FRPS

The core of the arguments 1 and 2 above is that any hypothesis test, whether it is derived from a particular alternative hypothesis or not, can only distinguish the  $H_0$  from particular alternatives, but not from them all. The following definition is based on the idea that evidence against a distribution (or class of

---

<sup>2</sup>Technically, for every reasonable test, an alternative hypothesis can be constructed against which the test is optimal, but many reasonable tests are not mainly motivated by the particular alternative hypothesis ensuring optimality, which can be quite counter-intuitive in some cases.

distributions) in favour of an alternative distribution (or class) is given empirically by sets which have a small probability under the first distribution and a higher probability under the second one.

**Definition:** A class of distributions  $\mathcal{Q}$  is *distinguishable* from a class of distributions  $\mathcal{P}$  on the same probability space  $\mathcal{R}$ , if there is a set  $C$  and  $0 \leq \alpha < 1$  such that

$$P(C) \leq \alpha, Q(C) > \alpha \forall P \in \mathcal{P}, Q \in \mathcal{Q}. \quad (1)$$

Such a set  $C$  defines an *unbiased* test of the  $H_0$  that the true distribution belongs to  $\mathcal{P}$  against  $\mathcal{Q}$  and therefore *distinguishes*  $\mathcal{P}$  from  $\mathcal{Q}$ .  $\alpha$  is the *level* of this test,  $Q(C)$  the *power* at  $Q$  and  $\min_{Q \in \mathcal{Q}} Q(C)$  the *minimum power*.

For a given  $C$ , the classes  $\mathcal{P}$  and  $\mathcal{Q}$  of *all* distributions for which (1) holds, are called the *maximum classes distinguished by  $C$* .

Note that the Neyman-Pearson theory is about maximizing the power for a fixed level. We will not stress optimality here, but it is obviously desirable to have high power and low level at the same time.

Whatever the set  $C$  is, as long as  $P(C) < 1$  for a given distribution  $P$ , a distribution  $Q$  on  $\mathcal{R}$  can be found with  $Q(C) = 0$ . Therefore, obviously, the class of all possible distributions on  $\mathcal{R} \setminus \{P\}$  cannot be distinguished from  $\mathcal{P} = \{P\}$ . More general,  $C$  with  $P(C) = \alpha$  can't distinguish  $Q$  with  $Q(C) \leq \alpha$  from  $P$ .

This has an important practical implication. If the  $H_0$  is  $P$  and the true distribution is  $Q \neq P$ , but  $Q$  is not distinguished from  $P$  with a given  $C$  corresponding to an FRPS obeying Gillies' conditions, it is not possible to falsify  $P$ . This holds even in cases where it may be quite obvious from a data analytic point of view, with enough observations, that the true distribution is rather  $Q$  than  $P$ .

An example is given in Figure 2. Gillies suggests the set  $C_1$  for his FRPS, but if  $P$  is  $\mathcal{N}(0, 1)$  and  $Q$  is  $\mathcal{N}(0, 0.09)$ ,  $P(C_1) > Q(C_1)$  and  $C_1$  cannot distinguish  $Q$  from  $P$  (it distinguishes  $P$  from  $Q$ , i.e., it can reject  $Q$  *in favour of*  $P$ , though). This holds for all sets of the form  $(-\infty, u_{\alpha/2}] \cup [u_{1-\alpha/2}, \infty)$  and therefore for all possible variants of the FRPS. Therefore, if  $Q$  is the true distribution,  $P$  can never be falsified even though  $Q$  is obviously different enough from  $P$  that the difference could be easily found by data analytic means if data from enough independent repetitions were available.

However,  $Q$  can be distinguished from  $P$ , though not by sets obeying Gillies' conditions. Instead, the set  $C_2$ , which Gillies found counter-intuitive, distinguishes  $\mathcal{N}(0, 0.09)$  from  $\mathcal{N}(0, 1)$  and is actually optimal for doing this according to the Neyman-Pearson theory. Therefore, there is a practical reason in a particular situation to consider  $C_2$  as rejection region of  $\mathcal{N}(0, 1)$ , and it shouldn't be ruled out by definition. I don't agree with Gillies counter-intuitivity argument anyway, because if the variance is large, observations which are extremely close to 0 are as unusual as observations which are in the tails, given that the two sets are suitably defined (with  $P(C_1) = P(C_2)$  as in the given example), and I don't see why of two "unusual sets" with the same small probability one

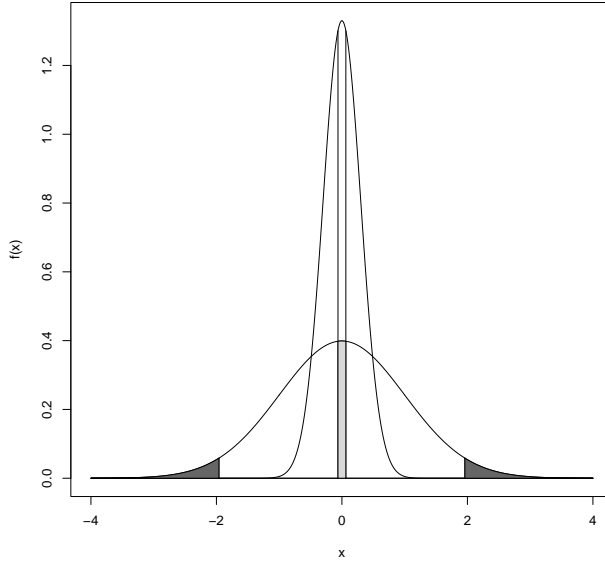


Figure 2:  $\mathcal{N}(0,1)$ -distribution with same sets  $C_1$  (dark gray) and  $C_2$  (bright gray) as in Figure 1, and  $\mathcal{N}(0,0.09)$ -distribution.

should be generally “better” (in the sense of confirming  $H_0$ ) than the other.

It can be seen as a downside of the argument elaborated here that my suggestion to drop restrictions allows the scientist a much stronger subjective decision, because the resulting test depends on the choice of the alternative hypothesis. However, if different tests can be of reasonable interest in different situations, it is not convincing to prevent subjectivity by imposing one particular test by definition.

In Section 2.4, the idea of favouring large density values will be made even more suspect.

### 2.3 “Alternatives” and “power” are needed

For a given rejection region  $C$ , the maximum classes distinguished by  $C$  have been defined above. This gives a description of what  $C$  essentially does. For example,  $C_1$  in Figure 1 distinguishes distributions with more probability in the tails (or in one of them) not only from  $\mathcal{N}(0,1)$ , but from all distributions with  $P(C_1^c) \geq 0.95$ ,  $C_1^c = (u_{0.025}, u_{0.975}]$ . This is fine if the scientist, in a given situation, considers the fact that most of the probability mass is concentrated on  $C_1^c$  as an essential characteristic of her model  $P = \mathcal{N}(0,1)$ . As with all mathematical models of reality, the scientist usually doesn’t believe precisely in her model  $P$ , but considers it as a somewhat reasonable approximation of reality.



$C_1^c$  doesn't give evidence in favour of  $\mathcal{N}(0, 1)$  alone, but in favour of the class of distributions with  $P(C_1^c)$  large enough. It depends on the problem and on the scientist, whether she is rather interested in this aspect of the true underlying distribution, or in something else, for example in testing whether the probability for values extremely close to 0 is small enough (which would be tested by  $C_2$ ). Therefore thinking about alternatives leads to a deeper understanding of the aspects of a model that are confirmed if the model is successfully tested.

Derived from the concept of the "alternative" is the concept of "power". The power of a test against a particular alternative is a measure of the quality of the test - it should not only have a small probability of rejecting a true  $H_0$ , but also a large probability of rejecting a  $H_0$  which is wrong. Therefore,  $\alpha$  held fixed, it is desirable to have the power as large as possible. But this depends inevitably on the alternative under which the power is computed.

Consider the following example. A scientist tries to confirm the model  $P$  of  $n$  RVs  $X_1, \dots, X_n$  independently identically distributed (i.i.d.) with  $\mathcal{L}(X_i) = \mathcal{N}(0, 1) \forall i$  by application of an FRPS. Even following Gillies' restrictions, he has to decide about a suitable test statistic, and there are several possibilities. The standard choice would be  $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ ,  $\mathcal{L}(\bar{X}) = \mathcal{N}(0, \frac{1}{n})$ , which is falsifiable. But because  $\mathcal{L}(X_1) = \mathcal{N}(0, 1)$ , which is falsifiable as well,  $X_1$  could as well be chosen as a test statistic, ignoring all further observations. How to choose between these statistics? While it could be argued that it is obviously silly to take just one observation into account, in fact the choice depends on the alternative that the scientist has in mind. In terms of power, for large enough  $n$ ,  $\bar{X}$  is much better to distinguish from  $P$  any i.i.d. model where  $\mathcal{L}(X_i)$  has an expected value  $\neq 0 \forall i$ . The use of  $X_1$  as test statistic is better to distinguish from  $P$  models for independent RVs where  $\mathcal{L}(X_1)$  has an expected value  $\neq 0$ , but  $E(X_i) = 0$  for  $i \geq 2$ . However, this is quite unlikely to be a realistic alternative in a situation in which the scientist initially had chosen a model with identical distributions as  $H_0$  and the first observation  $X_1$  is not known to be special in any sense.

There is a second reason why power considerations are needed for falsification rules. If the test level  $\alpha$  is larger than 0, then the application of the FRPS is inconsistent, as Gillies (1973, p. 188) notes, in the sense that its frequent application to true models necessarily leads to a portion of about  $\alpha$  wrong falsifications. From this point of view it would be attractive to have  $\alpha$  as small as possible. So why can't  $\alpha$  be chosen as  $10^{-7}$ , corresponding to a very small, unlikely rejection region, or even  $\alpha = 0$ ? The reason is that if  $\alpha$  becomes smaller and smaller,  $C$  becomes smaller and smaller as well. More precisely, if a sequence of tests of  $H_0$  is constructed with levels  $\alpha_1 > \alpha_2 > \alpha_3 > \dots$ , then  $C_1 \supset C_2 \supset C_3 \supset \dots$  from the corresponding rejection regions, given that all the tests are constructed following the same principle<sup>3</sup>.

, and therefore its probability under *any* distribution becomes very small, i.e., the power of the test against *any* alternative. It becomes almost impossible

---

<sup>3</sup>"The same principle" could either mean Gillies' FRPS, or Neyman-Pearson tests against the same alternative.

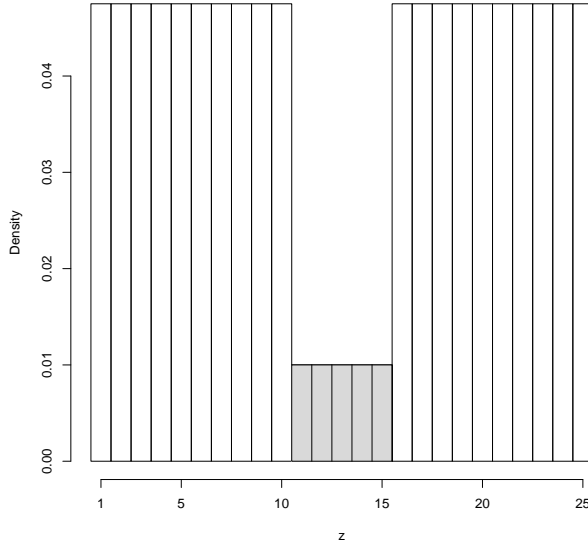


Figure 3: Density of  $Z = Y \circ X$  as defined in Section 2.4. The bright gray set corresponds to  $C_2$  in Figure 1.

to reject a wrong  $H_0$ , and therefore there is a trade-off between level and power. In the extreme case,  $\alpha = 0$  implies that the  $H_0$  can never be falsified under alternatives that have the same zero-probability sets as  $H_0$  (which holds, e.g., if  $H_0$  and the alternative are distributions with strictly non-zero densities on the real line, such as normal distributions). This particular argument doesn't need the specification of an alternative, but an alternative is needed to base the choice of  $\alpha$  in real life situations on power considerations.

Of course, all of this is well known in principle, but it should be recalled to understand that the proposal of a test to falsify a probability model without specification of an alternative hypothesis rather ignores an important feature of the resulting test than could be seen as an achievement.

## 2.4 The FRPS doesn't rule out counter-intuitive tests

After the preceding discussion, it may be surprising to learn that Gillies' approach actually doesn't rule out "counter-intuitive" tests like the one based on  $C_2$  in Figure 1. This means that the restrictions (ii) and (iii) don't achieve what they are supposed to achieve, and they are therefore superfluous, whether one agrees that "counter-intuitive tests" should be ruled out or not.

The reason is that the choice of the RV is still flexible. Let  $X$  be an RV with  $\mathcal{L}(X) = \mathcal{N}(0, 1)$ . Let  $Y$  be defined as follows:  $Y(x) = 1$  for  $x \in (-\infty, u_{0.0475}]$ ,

$Y(x) = 2$  for  $x \in (u_{0.0475}, u_{0.095}]$ ,  $Y(x) = i$  for  $x \in (u_{0.0475(i-1)}, u_{0.0475i}]$ ,  $i = 2, \dots, 10$ . Thus,  $(-\infty, u_{0.475}]$  is partitioned into 10 subsets with a probability of 0.0475 for each of them under  $\mathcal{N}(0, 1)$ , and  $Y$  assigns  $1, \dots, 10$  to these subsets. In the same manner, partition  $(u_{0.475}, u_{0.525}]$  into 5 subsets with a probability of 0.01 for each of them, and let  $Y$  assign  $11, \dots, 15$  to these subsets. Partition  $(u_{0.525}, \infty]$  into 10 subsets with a probability of 0.0475 for each of them, and let  $Y$  assign  $16, \dots, 25$  to these subsets. Now define the discrete RV  $Z = Y \circ X$ . The distribution of  $Z$  is shown in Figure 3. This is a falsifiable distribution in Gillies' sense, and the standard FRPS for  $\alpha_0 = 0.05$  falsifies the shown distribution for the set  $\{11, \dots, 15\} = Y(C_2)$ . This means that the distribution of  $Z$ , which is derived from  $\mathcal{L}(X) = \mathcal{N}(0, 1)$ , is falsified if  $X \in C_2$ , which is exactly what Gillies wanted to prevent. Analogous constructions are possible whenever the probability mass of  $X$  isn't concentrated on too few values.

A possible objection against this argument could be that  $Y$  is a counter-intuitive transformation and should be ruled out somehow, but I generally like to prevent arguments referring to "general intuition", because intuition is often not general, and even general intuition may be misled. Furthermore I am not aware of any suggestion of a definition of admissible transformations.

A further argument against restrictions based on density values would point out that in measure theory density values depend on the choice of the underlying measure. Furthermore they are not uniquely defined on zero mass sets and quite unstable, i.e., they can vary hugely under very small changes of the distribution.

## 2.5 Response to Gillies' arguments

Here is a concluding response to Gillies' arguments in favour of the restrictions (ii) and (iii) in Section 2.1.

- (a) Gillies' "intuition" is not shared by everyone and intuition is generally easily misled. The tests that Gillies wants to rule out are needed in some practical situations.
- (b-i) Subjective choices of the scientist neither can nor should be prevented - and actually they are necessary even in Gillies' own approach.
- (b-ii) Whenever a statistical test is defined, it can distinguish some but not all possible alternatives from the  $H_0$ . It adds to the understanding of the test to make this clear. Omnibus tests don't exist.

## 3 The goodness-of-fit paradox

### 3.1 Statement of the paradox

The goodness-of-fit paradox in a general situation can be stated as follows. Assume that the true distribution underlying a statistical experiment (which may be a sequence of repeated sub-experiments) is  $Q$ . Assume further that a statistician is interested in finding out whether the null hypothesis  $P$  can be

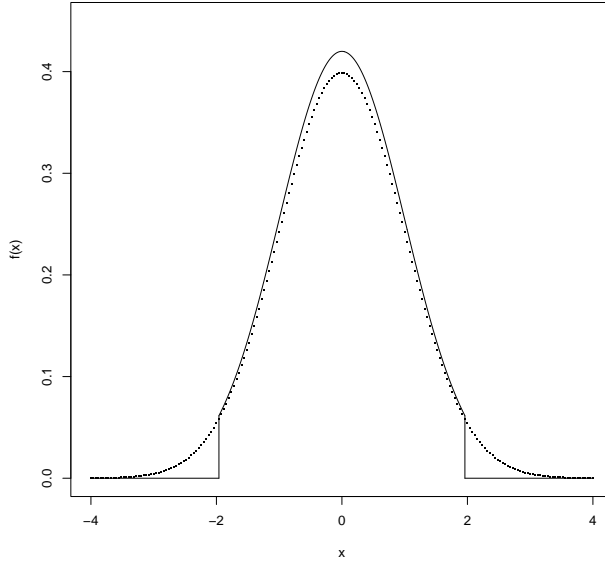


Figure 4:  $\mathcal{N}(0, 1)$ -distribution (dotted) and  $\mathcal{N}(0, 1)$ -distribution conditional on  $C_1^c$  in Figure 1 (solid), i.e., on non-falsification of  $\mathcal{N}(0, 1)$ .

confirmed as a model for the experiment ( $P$  may or may not be true, i.e., equal to  $Q$ ). Note that the following discussion holds as well if the  $H_0$  is a class  $\mathcal{P}$  of distributions, for example all normal distributions or, if the experiment consists of repeated sub-experiments, all distributions assuming independence of the sub-experiments. Therefore the statistician carries out a statistical test of level  $\alpha > 0$  defined by a test statistic and a rejection region  $C$  of the  $H_0$  that  $Q = P$ , and she accepts  $P$  as a model for the experiment only if she observes  $C^c$ , i.e., if the test doesn't falsify  $P$ . Such tests to check an underlying model are called “goodness-of-fit tests” in the statistical literature, therefore the name “goodness-of-fit paradox”. Because our test is an  $\alpha$ -level test,  $P(C) = \alpha > 0$ . Because  $P$  is accepted as a model for the underlying experiment only if it is not falsified, the distribution  $R$  of the data for which  $P$  is accepted is the true distribution  $Q$  conditional on the non-rejection of  $P$ , i.e., on  $C^c$ . Thus,  $R(C) = 0 \neq \alpha$ , and therefore  $R \neq P$ . This proves that, if  $P$  is accepted as a model only if it has been confirmed by a goodness-of-fit test, the effective distribution of the data for which  $P$  is assumed (which is, according to the propensity interpretation, the distribution over a potentially infinitely 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.

An illustration is given in Figure 4, which compares the  $\mathcal{N}(0, 1)$ -distribution with the truncated distribution that results from conditioning  $\mathcal{N}(0, 1)$  on non-falsification by the test defined by  $C_1$  in Figure 1.

The conceptual reason for the occurrence of the goodness-of-fit paradox is that, following Gillies' approach, a probability model is ruled out based on the occurrence of an event, of which the possibility to occur with nonzero probability is an essential part of the model. Unfortunately, as seen in Section 2.3, falsification rules based on sets with zero probability under  $H_0$  don't make sense for most models, because they don't have any power against alternatives of interest. Therefore, the paradox can't be prevented with any practically useful falsification rule. 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  $n = \infty$ , which is impossible in practice, and therefore a practical falsification for finite  $n$  will require  $\alpha > 0$  as with a usual hypothesis test.

This is different from falsification of deterministic models, where no such paradox occurs. However, it could be argued that in reality the situation with deterministic models is not different, because experiments to confirm or falsify them are always subject to random variations and therefore statistical rules derived from statistical measurement error models have to be used to decide how much deviation from the model's predictions can be tolerated, which, again, is subject to the goodness-of-fit paradox.

It may be argued that the term "paradox" is not justified, because the model refers to the underlying data generating process, which is modelled by the unconditional distribution, and not to the observed data conditional on non-falsification which cause the so-called "paradox".

Here are two different possible interpretations of the situation. I argue that the term "paradox" is justified if the first interpretation is adopted, but there is a problem with the second one as well. The interpretations are distinguished by whether the fact that a hypothesis test is carried out to confirm or falsify the model is considered to be a part of the experimental conditions, which, according to Gillies' theory, define the propensities (recall the beginning of Section 2.1).

- (a) *The test is considered to be a part of the experimental conditions.* This means that, whenever we carry out an experiment to obtain observations (usually a series of observations), we apply the goodness-of-fit test to confirm the model  $P$ , and we accept the model only if it is confirmed by the test. In this situation, if we repeat the experiment independently and we take into account all repetitions for which we confirm  $P$ , we will observe a sequence of outcomes that follows the underlying distribution (which may or may not be  $P$ ) *conditional on non-falsification*, i.e., the sequence is subject to the paradox.

- (b) *The test is not considered to be a part of the experimental conditions.* In this case, formally, the model  $P$  does not refer to the conditional distribution, but to the data generating process unaffected by the falsification rule. Therefore,  $R \neq P$  (using the notation above) is just a mathematical result, but by no means paradox, and the question of interest is only whether or not  $Q = P$ . Whether it is justified to interpret the test as not belonging to the experimental conditions, however, depends on how the scientist actually handles the situation in practice. If, in fact, every outcome of the experiment is tested, then interpretation (a) is a valid description of how the data modelled by  $P$  are obtained, while (b) is incomplete. (b) can be considered as valid if the test is just carried out once, but then  $P$  is used as a model for *further* data from the experiment *without testing*. For those further data, the paradox is prevented. The problem with this is that it is based on the assumption that the further data are independent of the data used to test  $P$ , and identically distributed. *This assumption is essentially untestable*, because such a test would involve all available data and accordingly would induce another goodness-of-fit paradox.

This leaves us with the following choice: either the paradox, i.e., the disturbance of the validity of a model by confirming it, is accepted, or we rely on the untestable metaphysical assumption that new data is identically distributed independently from the data we use to test the model assumption, because every attempt to test this leads us back into the paradox.

Note that in many practical situations there is only one data set at hand for which a model is constructed, without the possibility of repeating the whole experiment. Choice (a) - test all data sets - and (b) - test only the first data set - are then indistinguishable. The occurrence of the paradox in this situation depends on whether further statistical analyses that assume the model to hold are carried out conditionally on non-falsification.

There is a third choice to deal with the paradox.

- (c) Choose a model which makes testing with  $\alpha = 0$  possible, i.e., a truncated distribution corresponding to a model conditional under non-falsification such as the truncated normal from Figure 4. Because  $C_1$  has probability zero under this model, a falsification rule based on  $C_1$  prevents the paradox, and has still power against all alternatives with  $P(C_1) > 0$ . The problem with such a choice is that the model is then obviously not determined by considerations about the underlying phenomenon to be modelled alone, but also by the subjective choice of a rejection region by the scientist (which will usually be motivated by power considerations under another model, namely an unconditional one), while propensity interpretations of probability are intended to give an account of *objective* probabilities. However, it could be argued that choice (a) above involves the same problem without being explicit about it and choice (b) is based on a metaphysical assumption which to adopt is a subjective decision as well.

Unfortunately, if the test statistic is based on a series of observations and

not on a single one, truncation of the distribution of the test statistic usually induces dependence among the individual observations (this will be illustrated in the following section), and the resulting model is much more difficult to handle and to analyze than an untruncated normal model, say. Nevertheless, investigation of such truncated distributions (distributions conditional on non-falsification) is certainly a promising direction for statistical research.

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 (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 informal 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). A further example is discussed in the next section.

Note that the three possible practical choices (a), (b) and (c) above correspond to different approaches to statistical analyses. Choice (a) corresponds to the mainstream approach to check a model first (either by goodness-of-fit tests or by graphical diagnostics) before applying further statistical analyses. Choice (b) corresponds to so-called cross-validation or bootstrap techniques (see, e.g., Efron and Tibshirani, 1993), where only a part of the data is used to decide about the model which is then applied to the rest of the data. Generalization of a model to future observations which are not tested again corresponds to choice (b) as well. Investigations as described above (Easterling and Anderson, 1978) or “adjustments for model choice” as discussed for example in Chapter 4 of Harrell (2001) acknowledge that subsequent statistical analysis are done under the conditional model instead of the original one, and are therefore connected to choice (c).

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.

### 3.2 An example: The game of red and blue

“The game of red and blue” is an example used by Gillies (2000, p. 78 f.) to demonstrate the benefits of his propensity interpretation compared to the subjectivist interpretation of probability according to de Finetti (1970). In particular, he argues that objective independence, which can be confirmed or falsified by statistical tests, is necessary to apply Bayesian models of exchangeability (i.e., that the observations are not independent, but their order doesn’t matter, as advocated by de Finetti for various situations) successfully.

Here is how Gillies describes the experiment:

*“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.”*

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  $\theta_0$  as at least 0.73 (here it is assumed that  $\theta$  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 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. Gillies’ criticism can be disputed by subjectivists by pointing out that it is, in principle, possible to start with a subjectivist model that assigns some probability to violations of the exchangeability assumption. However, this seems to be quite difficult from a computational point of view and I have not succeeded in finding any simple model in the Bayesian literature that allows for believing in exchangeability with a high probability, but smaller than 1, and for assigning the rest of the probability mass to deviations from exchangeability.

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 this gives rise to the goodness-of-fit paradox.

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 successive blues and two reds has two runs (one run being the 700 blues, the other one the two reds), 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 for 700 blues and two reds, but accepts it for 700 blues, one red and one blue. Under the distribution of full data sets for which independence is accepted by the runs test, i.e., the distribution conditional on non-rejection, 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. Confirming independence by the runs test induces dependence.

Note that this example, as most real statistical analyses using frequentist

or propensity models, incorporates two different cases of repetition: a coin is repeatedly tossed,  $n = 702$  times, say, to define the experiment (independently, but the observed outcome is only independent under the  $H_0$  and dependent under the game of red and blue). But the interpretation of the distributions of the vector of outcomes of the whole experiment, of the runs test statistic, and of the outcomes conditional on non-falsification refers to independent repetitions of the whole experiment, i.e., all  $n$  tosses. In many cases, when further data is not available, we only have a single observation of this vector for our analyses. Any  $H_0$  to be tested with reasonable power in this situation needs to involve some independent repetition on some level (be it directly, as for the independence model, or indirectly, as for the game of red and blue) to generate a large enough “effective sample size” based on which it can be falsified. In most non-i.i.d. models used in statistics, e.g., time series or regression models, the random variation term is assumed to be i.i.d. for this reason. This means that the idea of independence is not only needed to justify exchangeability in subjectivist analyses, but also to ensure falsifiability (with effective sample sizes larger than 1) in objectivist analysis.

As outlined in Section 3.1, there are two main approaches to deal with the goodness-of-fit paradox. One approach (choice (b) above) is to test the  $H_0$  on a part of the data and then to assume it for further data, which are assumed to be equally and independently from the test data, without testing these new assumptions.

The alternative is to proceed with the data, conditional on non-falsification, as if it were generated by the  $H_0$ . This corresponds to choice (a) above, but the investigation of the resulting behaviour under the conditional distribution is rather connected to choice (c).

To explore the practical consequences of the goodness-of-fit paradox in the given situation, I carried out some simulations (1000 simulation runs with  $n = 100$  goes each) to compare the statistical properties of data generated by  $H_0$  (independent Bernoulli tosses with  $p = 0.5$  and  $p = 0.2$ ) with the distribution conditional on non-falsification by the runs test. The runs test has been used with  $\alpha = 0.05$  in the two-sided version, i.e.,  $H_0$  is rejected by too few or too many runs, corresponding to seemingly too large positive or negative correlation between goes, and in the one-sided version against the alternative of positive correlation between goes.

The following statistics have been computed for all simulated data sets: 0.9- or 0.95-confidence intervals for  $p$ , the standard estimators for  $p$  and for the estimation of the probability that “heads” is followed by another “heads” (i.e., conditional on heads in the preceding go). Theoretically, this should be  $p$  as well because of independence. The coverage probabilities of the confidence intervals and the means and mean squared errors of the estimators have been computed over all data sets, and only over the data sets for which the runs test did not reject independence. The latter corresponds to the conditional distribution of the statistics under non-falsification of independence.

The results show that it depends on what the model is used for whether the paradox affects the statistical analyses or not. No significant differences between

the coverage probabilities of the confidence intervals have been found between the three distributions (unconditional, conditional on non-falsification by the two-sided runs test and by the one-sided runs test) and between the means and mean squared errors of the estimators of  $p$ . Therefore, these analyses seem to be hardly affected by the goodness-of-fit paradox. The situation changes for the estimation of the probability that “heads” is followed by another “heads”. Conditional on non-falsification by the one-sided runs test, it was estimated (on average over 1000 simulation runs) as 0.491, which is significantly smaller than the corresponding simulated value for the unconditional distribution of 0.498. Similar results are observed with  $p = 0.2$  (0.184 for conditional/one-sided, 0.195 for unconditional - conditional/two-sided was between these values in both cases). This corresponds to the intuition, because the one-sided runs test excludes some sequences of goes with long runs, where “heads” has been followed by “heads” often. Note that in this particular situation it doesn’t make too much sense in practice to estimate the probability that “heads” is followed by another “heads”, because the goes are assumed to be independent by the model. The simulation serves as a toy example for situations where a more sophisticated model is used to estimate conditional probabilities by simulation, as it is done frequently, e.g., in meteorology.

Another way of looking at the practical implications is to take the “game of red and blue”-model as an alternative into account. Still assuming that the  $H_0$  of independence holds, the runs test will falsify it if under  $n = 9$ , seven blues and two reds are observed in a row (i.e., two runs). The p-value for this, i.e. the probability for only two runs conditionally on seven blues and two reds in any order, is 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 for “blue” in the next go erroneously based on the game of red and blue with a probability of about 1%, and this means, in the given situation, that she will judge “blue” as impossible (because in the game of red and blue, blue is impossible in this situation) instead of  $7/9$ , the standard estimation under independence<sup>4</sup>. 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<sup>5</sup>. With 700 blues and 2 reds, as above, the p-value is smaller than  $10^{-15}$ .

## 4 Further discussion

### 4.1 How to deal with the paradox?

In the present paper, some aspects of Gillies’ concept of proposing particular statistical hypothesis tests as falsification rules for probability models have been discussed. In Section 2 it has been argued that general hypothesis tests should

---

<sup>4</sup>This discussion is conditioned on the fact that the total number of “blues” is seven.

<sup>5</sup>A subjectivist, or any Bayesian, would say that this depends on the prior probability for the exchangeability model.

be allowed as falsification rules and the restriction given by Gillies are superfluous. The concepts of an alternative hypothesis and the power of a test against that alternative, which are ruled out in Gillies' approach, have turned out to be useful. In Section 3, the goodness-of-fit paradox has been discussed, which means that it can be shown that a statistical model confirmed by a hypothesis test doesn't hold anymore *because of the confirmation* (where "the model holds" refers to the distribution conditional on confirmation). The validity of the arguments in Section 3 doesn't depend on whether general hypothesis tests are allowed or only those obeying Gillies' restrictions.

Several approaches can be taken to deal with the goodness-of-fit paradox (the first three corresponding to the approaches (a)-(c) given in Section 3.1).

- (a) The application of a falsification rule could be interpreted as part of the experimental conditions. This defines the distribution conditional on non-falsification  $R$  as the "true" probability model.  $R$  is equal to  $Q$ , the true distribution before applying the falsification rule, conditional on  $C^c$ , where  $C$  is the rejection region. This model is unequal to the initially tested model  $P$ , as has been shown under the assumption  $\alpha > 0$ .

$P$  could then still be maintained as a reasonable approximation to the true distribution  $R$ . We have  $P(C) = \alpha$ , which is small, and  $R(C) = 0$ . Whether or not  $P$  is a good approximation to  $R$  depends on two aspects:

- (i) How strongly is the use that is made of the model affected by basing analyses on  $P$  instead of  $P$  conditional on  $C^c$ ? In many situations, this can be simulated or worked out theoretically. Some examples have been given in Section 3.2, in most of which using  $P$  seemed to be harmless. However, different situations are known in the statistical literature. 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, see Chapter 4 of Harrell (2001).
- (ii) How good is  $P$  as an approximation of  $Q$  on  $C^c$ ? The discussion under (i) implicitly assumed that  $Q = P$ . But the only thing that is actually known from the successful application of our falsification rule about  $Q$  is that  $Q$  gave rise to  $C^c$ , which allows the presumption that  $Q(C^c)$  is rather large, as is  $P(C^c)$ . Therefore, from confirmation of  $P$  by a statistical test, we know nothing more than that  $P(C) = \alpha$  small and  $R(C) = 0$ . Based on the data, no further assessment is possible whether  $P$  is a reasonable approximation, or even equal to  $Q$  on  $C^c$ .

An important objection can be made at this point. Couldn't further tests or informal tests such as graphical diagnostics be carried out to confirm  $P$  to be a reasonable model for the true  $Q$  on the non-falsification region  $C^c$  of the first test? In other words, couldn't it make sense to test  $P$  against different alternatives to find out more about it? Note that Gillies (2000) seems to have this in mind when he

writes, e.g., on p.79 about “*a series of statistical tests*” to be carried out to find out, eventually, that the data in question are generated by the game of red and blue.

The answer is: yes, in principle. However, if  $k$  tests are carried out and  $P$  is taken as falsified if any of these (or any prespecified number) rejects  $P$  significantly, the whole multiple test procedure can be formalized as a single combined falsification rule leading again to a rejection region  $C$ , which may look more complex than the rejection regions of the single tests, but mathematically, in terms of the discussion given in the present paper, it is by no means essentially different from a single test. Furthermore, multiple testing results in a loss of power for every single set, given that the overall test level is fixed. A standard illustration for this is the so-called Bonferroni correction for multiple testing, which is based on the fact that in case of a combination rule that rejects  $P$  if any of the  $k$  tests rejects it, an overall level of  $\alpha$  is kept if every single test is carried out on a level of  $\alpha/k$  (see Holm, 1979, for discussion and a more sophisticated version). The more tests used, the weaker every single one becomes. Eventually, under a combination of several tests, there is a combined rejection region  $C$ , and whether  $P$  is a good approximation for  $Q$  on  $C^c$  cannot further be assessed in terms of falsification. (Usually, the choice of  $P$  as a plausible model will be motivated by mostly informal subject-matter considerations, and these considerations are the only further means to justify  $Q \approx P$  on  $C^c$ .)

- (b) If the application of the falsification rule is not interpreted as part of the experimental conditions, the unconditional distribution  $Q$  is defined as the “true model”. Under this interpretation,  $H_0 : Q = P$  can be tested on some test data without causing the goodness-of-fit paradox for new data independent of the test data. The price is that it is not allowed to test  $H_0$  or the independence of the test data on the new data. Again, this can only be justified by subject-matter considerations (independence is usually justified by arguments like “we don’t see any obvious source of dependence”, which are clearly quite weak).

This is the usual approach to significance testing in applied research. In an investigation whether a new drug is more effective than a placebo, for example, the result is derived from data for some patients, but it is interpreted in a way generalized to further patients. Whether these further patients are independent of the test patients and follow the same distribution remains untested.

- (c) Truncated models could be used which can be tested with  $\alpha = 0$ . Apart from the problems already raised in Section 3.1, this is subject to the discussion under (a-ii): how good is  $P$  as an approximation of  $Q$  on  $C^c$ ?
- (d) It could be suggested that goodness-of-fit testing should be abandoned

altogether, which means that Gillies' propensity interpretation with falsification rule has to be rejected. It can be expected that this point of view is taken by many Bayesians (who dislike the whole principle of significance testing, see, e.g., Howson and Urbach, 1993), among others. I don't want to discuss alternative interpretations of probability in detail here. I just mention some Bayesian literature (Box, 1980, Berger, 1984, Dawid, 2004), in which the need to decide about the validity of a model based on real data (be it informal diagnostics or test-like falsification procedures) is acknowledged. From a data analytical point of view, causing the goodness-of-fit paradox by model diagnostics is in almost all cases less dangerous than to assume an untested model.

## 4.2 Propensities with falsification rule - a valid interpretation?

Concerning the validity of Gillies' propensity interpretation with falsification rule, I see two main implications of the present paper.

1. The connection between experience and probability models established by the falsification method is weaker than Gillies seems to believe. Assuming approach (a) above, the effect of the falsification rule is only that we know that the probability of a rejection region under the resulting distribution is in fact zero if we test a model which assigns a small probability  $\alpha$  to this region. This confirms the model to some extent as a reasonable approximation if the fact that we don't expect observations from the rejection region is an aspect of  $P$  which is of major importance to us.
2. Though propensity interpretations refer to objective probabilities in the "real world", it is impossible to exclude a strong subjective contribution to finding a model. Subjective considerations are needed to choose the significance test(s) used for confirmation or falsification, to justify  $Q \approx P$  on the region of non-falsification and, assuming approach (b) above, to make further untestable assumptions.

Having these implications in mind, I still believe that Gillies' interpretation is very useful, especially considering the strong arguments that can be raised against alternative interpretations of probability (this doesn't imply that I reject all other interpretations). While I am much less optimistic than Gillies about the possibility to observe strong information about whether there is some true probability in the real world and what the "true model" is, I acknowledge that Gillies gives a convincing account of the interpretation of the construct "probability" that researchers have in mind when thinking about objective uncertainty. It seems that "falsification" of this construct is only possible based on rejection regions with small probability - the core idea of the statistical hypothesis test. The goodness-of-fit paradox, the possibilities to deal with it and the subjective freedom to choose models and alternatives against which to test the models re-

fer to the essentially unsolvable problems of data analysis based on objectivist models.

I won't discuss alternative interpretations of probability here, but I think that these problems can be traced back to more basic difficulties with mathematical modelling of reality, which occur in different ways also with other interpretations. For a brief account of this point of view, see Hennig (2002, 2003). Overviews of criticisms of alternative interpretations are given by Fine (1973), Gillies (2000). The approach of Davies (1995) to interpret probability models as approximations to data instead of objective features of the real world, and to define so-called "adequacy regions" which are equivalent to particular non-falsification regions in the sense of the given paper seems to be quite similar to the more modest (compared to Gillies) interpretation of long run propensities that I have in mind.

## 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] Bickel, P.J. and Doksum, K. (1977). *Mathematical Statistics*. Holden-Day, San Francisco.
- [3] 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.
- [4] Davies, P.L. (1995). Data features. *Statistica Neerlandica* 49, 185-245.
- [5] Dawid, A.P. (2004). Probability, Causality and the Empirical World: A Bayes-de Finetti-Popper-Borel Synthesis. *Statistical Science* 19, 44-57.
- [6] de Finetti, B. (1970). *Theory of Probability*, vol. 1 and 2. English translation, Wiley 1974.
- [7] 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.
- [8] Efron, B. and Tibshirani, R. (1993) *An Introduction to the bootstrap*. Chapman and Hall.
- [9] Feller, W. (1950). *Introduction to Probability Theory and its Applications*. Wiley, New York.
- [10] Fetzer, J. H. (1983). Probabilistic Explanations, in P. Asquith and T. Nickles (eds.) *PSA 1982, Vol. 2*, Philosophy of Science Association, East Lansing, MI, pp. 194-207.
- [11] Fine, T.L. (1973). *Theories of Probability*, Academic Press, New York.

- [12] Gillies, D. (1973). *An objective Theory of Probability*. Methuen & Co., London.
- [13] Gillies, D. (2000). *Philosophical Theories of Probability*. Routledge, London.
- [14] 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.
- [15] Harrell, F.E. jr. (2001). *Regression Modeling Strategies*. Springer, New York.
- [16] 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.
- [17] 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.
- [18] Holm, S. (1979). A simple sequentially rejection multiple test procedure. *Scandinavian Journal of Statistics*, 6, 65-70.
- [19] Howson, C. and Urbach, P. (1993). *Scientific Reasoning: The Bayesian Approach*. 2nd edition. Open Court, Chicago.
- [20] Lehmann, E.L. (1986). *Testing Statistical Hypotheses*, 2nd ed. Wiley, New York.
- [21] Miller, D. W. (1994). *Critical Rationalism. A Restatement and Defence*, Open Court, Chicago.
- [22] Popper, K.R. (1957a). Probability Magic or Knowledge out of Ignorance. *Dialectica* 11, 354-374.
- [23] 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.
- [24] Popper, K.R. (1990). *A World of Propensities*. Thoemmes.
- [25] Spencer-Brown, G. (1957). *Probability and Scientific Inference*. Longman, London.
- [26] von Mises, R. (1928). *Probability, Statistics and Truth*. 2nd revised English edition, Allen and Unwin 1961.