Surprise and Evidence in Statistical Model Checking

Abstract

There is considerable confusion about the role of p-values in statistical model checking. To clarify that point, I introduce the distinction between measures of surprise and measures of evidence which come with different epistemological functions. I argue that p-values, often understood as measures of evidence against a null model, do not count as proper measures of evidence and are closer to measures of surprise. Finally, I sketch how the problem of old evidence may be tackled by acknowledging the epistemic role of surprise indices.

1 Introduction: surprise, evidence and p-values

In statistical practice, p-values are often used to assess the tenability of a "null" (or default) model H_0 in the light of observed data. The precise role of p-values, however, is not clear, despite their widespread popularity in the empirical sciences. They are often cited as a basis for the rejection of a null model or, vice versa, for claiming that the evidence against the null model is not sufficiently strong to warrant rejection of the null. More dramatically, they are often confounded with posterior probabilities of a null model, e.g. when a p-value of 0.04 obtains, practitioners without a sufficient mathematical education often tend to assert that "the null model has a probability of 0.04". Although this is a well-known fallacy – p-values do not give posterior probabilities – practitioners often commit it. This leaves the question open how to understand p-values most adequately. Philosophers of science, on the one hand, have tried to integrate p-values into a philosophy of statistical inference. Most notably, Deborah Mayo uses them to explicate *severe tests*, the basic notion of her philosophy of inference.[1] Given that Mayo identifies evidence for a model with the survival of severe tests, p-values take the function of measuring evidence in her framework. On the other hand, statisticians have researched on the link between p-values and Bayesian measures of evidence.[2] Others try to calibrate them in a suitable way as to understand them as measures of surprise.[3] In total, there is no consensus about

[1] In a special, simplified case, the p-value of the null model can be understood as the *degree of severity* with which a specific alternative model passes a severe test (cf. Mayo 2004, p. 111, Mayo and Spanos 2006).

[2] Cf. Casella and Berger 1987, Berger and Sellke 1987 for the relationship between posterior probabilities and p-values and Sellke, Bayarri and Berger 2001 for the relationship to Bayes factors.

[3] Cf. Bayarri and Berger 1997, 1999 and 2000, Robins et al. 2000, Sellke et al. 2001

the proper role of p-values in statistical inference. However, due to the ubiquitous occurrence of p-values in scientific research reports, this question is of keen and abiding interest. I believe that a clarification of the inferential role of p-values has to address the distinction between surprise and evidence: What do we expect from a fruitful and valuable concept of evidence? What is, on the other hand, the role of surprise in statistical inference? I would like to answer this question by rethinking the functions of *surprise* and *evidence* in section 3: Whereas measures of evidence form an objective basis for the rejection and acceptance of models, surprise indices have a heuristic value that is brought to bear in an exploratory model analysis. This investigation is supplemented by a closer look at the properties of p-values in sections 2 and 4. Consequently, p-values cannot count as genuine measures of evidence – they are much closer to measures of surprise. Finally, the above results are applied to gaining a novel understanding of the problem of old evidence in confirmation theory. For reasons of simplicity, all considerations are restricted to full parametric models.

2 P-values: A closer look

It is not clear what information p-values, as ubiquitous as notorious in statistical practice, really convey. Most frequently, they are interpreted as measures of evidence, as measures of the discrepancy between the data and the null model. To see his in an example, consider a standard statistical experiment. We have a family of statistical models, parametrized by a real parameter ϑ. For instance, we flip a coin several times. The parameter $\vartheta \in [0, 1]$ then denotes the propensity of the coin to fall "heads", so that $\vartheta = 0.5$ means that the coin is fair whereas $\vartheta = 1$ means that the coin always comes out "heads". Now, we might want to examine the fairness hypothesis $H_0 : \vartheta = 0.5$ and we perform repeated coin tosses. The p-value then indicates how far H_0 is tenable in the light of the result of the incoming data. For calculating the p-value, we have to choose a statistic X – i.e. a function of the data – that summarizes the results of the coin tosses.[4] For instance, such a statistic X could be the number of heads occurring in the trial. Furthermore, p-values are based on a "distance statistic" T that measures the discrepancy between the null model H_0 and the observed data. Assume that $X = x_0$, i.e. x_0 heads were actually observed. Then, the p-value sums up the (H_0-)likelihoods of those possible values of X that fit the null model to

[4]Since X is supposed to contain all relevant content, it is tacitly assumed to be *sufficient* with regard to the parameter of interest ϑ, i.e. X captures all information about ϑ that is contained in the full data Y.In other words, sufficiency demands that the full distribution of the data Y conditional on the sufficient statistic does no longer depend on ϑ: $P(Y = y|X = x, \vartheta) = P(Y = y|X = x)$.

a lower degree than the observed value x_0:

$$p_{\text{obs}}(x_0) := P_{H_0}(\{T(X) \geq T(x_0)\}) \tag{1}$$

In other words, the bigger the distance between the observed data and the model H_0, the smaller the p-value. P-values yield the probability that, given H_0 is true, more extreme results will be observed. Or put even another way, they yield the probability that the fit between data and model is worse than the actually obtained fit.[5] Figure 1 illustrates this for $H_0 = N(0,1)$ and $p_{\text{obs}} = 0.05$ when T is identified with the likelihood function. It is now suggestive to conclude that a low p-value suggests poor evidence for H_0 and that a high p-value suggests good evidence for H_0, independent of which alternative model is considered. The validity of this conclusion will be

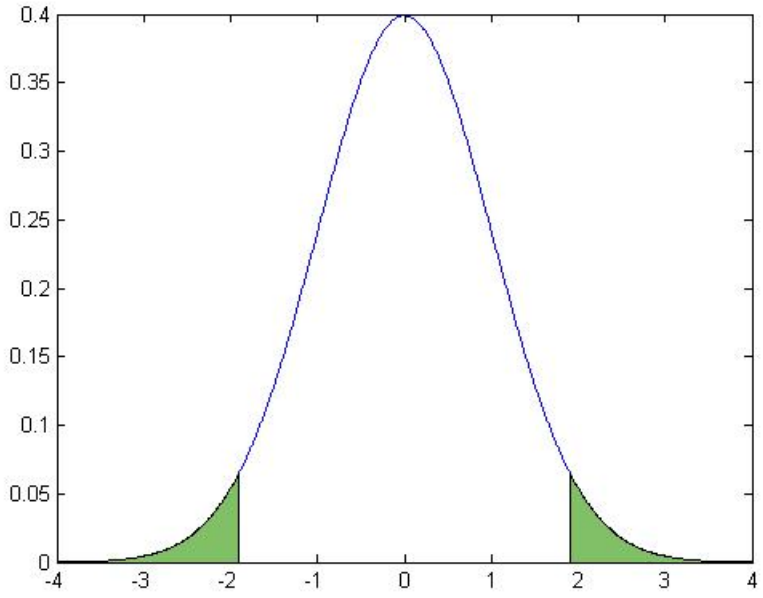

Figure 1: The 2,5%-tails of the standard normal distribution $N(0,1)$.

questioned in the subsequent sections, however. At this point, I would like to

[5]Equation 1 tells us that the statistic X should be minimally sufficient, i.e. representable as a function of any other sufficient statistic. To see this, consider again the coin toss experiment. Assume the null model $H_0 : \vartheta = 0.5$ and identify T with the likelihood function. Of course, both the entire data vector $X := (X_1, \ldots, X_n)$ as well as the sample mean $\overline{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i$ are sufficient statistics. But $T(X)$ will be constant (all sequences of zeros and ones are equally likely) whereas $T(\overline{X}_n)$ will mirror our intuitions that extreme sequences like '11111...' diverge to a higher degree from the null model than an average sequence. This is due to the fact that minimal sufficient statistics as $\overline{X}_n$ cumulate all "informationally equivalent" results in a single point of the sample space. Therefore we should base the p-values on minimally sufficient statistics. This point was brought to my attention by Teddy Seidenfeld, it is also contained in the third chapter in Seidenfeld 1979.

draw the reader's interest to the choice of the distance function. There arise two kinds of situations when evaluating the tenability of a null (or default) model H_0: First, H_0 may be compared to alternative models. Second, no specific models may compete with H_0. This distinction corresponds to the one-sided and two-sided hypothesis testing problems. In other words, in a one-sided testing problem the distance function T has a specific departure direction measures whereas in the two-sided problem, no such direction is given. Presupposing the existence of a distance function without a specific departure direction, however, is far from trivial. Building on the intuition that the less likely a result under the null model, the more it diverges from the null, the likelihood function constitutes a natural choice for such situations. Then, the p-value is the sum of probabilities of those elements of the sample space that are equally or less likely than the observed value x_0.[6]
After this brief introduction to p-values, the next section deals with the the demarcation between surprise and evidence and the identification of their epistemic functions.

3 Surprise and evidence

Evidence about a parameter of interest ϑ is required for making inferences about that parameter, for giving sensible estimates of ϑ and for deciding to work with this rather than that value of ϑ. Evidence measures transform the data as to provide the basis for inferences about ϑ. They are supposed to be suitable for public communication in the scientific community, i.e. they must be objective and free of subjective bias and distortion. For instance, we all have different opinions on the plausibility of certain models, but we should not disagree on the strength of evidence in favor of this and against that model. Unless we agree on the weight of evidence, we may be unable to listen to nature's verdict on competing hypotheses. Miscellaneous researchers should be able to draw on strength of evidence in their inferences and conclusions. Hence, I take those constraints – evidence as an objective, publicly communicable concept – to be the common ground of our inquiry. To a certain extent, both aspects of "evidence" are alluded to in the similarity to the word "evident". We require an objective, quantitative method to represent the information which the data convey about the unknown parameter.

This conditions affects above all candidate measures of evidence that depend on the *sample space* because partitions of the sample space are particularly open to subjectively grounded dissent. Different sample spaces are best illus-

[6]Neither is it obvious that there is only one sensible choice for the probability measure in the evaluation of $\{T(X) \geq T(x_0)\}$. For instance, Bayesians could choose either the prior or the posterior predictive distribution of H_0 or even suggest further calibration. Cf. Bayarri and Berger 2000, Robins, van der Vaart and Ventura 2000.

trated by simultaneously measuring an experimental result with two different instruments of which one has a wider measuring range than the other one. However, assume that for the actual result, it makes no difference which of the two instruments we use; only if other values had been observed, they would have diverged. We say that the instruments are associated with different sample spaces. In such a situation, we have the strong intuition that the evidential content of the data is not affected by the choice of the instrument: taking the other instrument would have actually yielded the same observation, hence our inferences should be the same.[7] The diverging range of the two instruments does not seem to be relevant for the evaluation of the actual trial. Indeed, that would violate the conditions for objectivity and public communicability of evidence which I outlined at the beginning of the section.

A natural move to circumvent such problems consists in the restriction of evidence to the *comparison* of different hypotheses about a parameter value. Indeed, it is the reply which I advocate. Our concerns about evidence address the question whether this rather than that value is a good estimate or basis for further analysis, and so on. Evidence then measures the degree to which the data favor a certain model (e.g. $\vartheta \leq 0$) over another model (e.g. $\vartheta > 0$). Then, it is natural to quantify evidence by means of likelihood ratios and Bayes factors which do fulfil the requirements for objectivity and independence of the sample space. The likelihood ratio of two models H_0 and H_1 is defined as

$$L(H_1, H_0, x) := \frac{P(x|H_1)}{P(x|H_0)} \tag{2}$$

When those likelihoods cannot be computed directly, e.g. because H_0 and H_1 are composite models with parameter ϑ, a more generalized version is given by the Bayes factor, the ratio between prior and posterior odds

$$B(H_1, H_0, x) := \frac{P(H_1|x)}{P(H_1)} \frac{P(H_0|x)}{P(H_0)} = \frac{\int P(\vartheta|H_1)\, P(x|\vartheta, H_1)\, d\vartheta}{\int P(\vartheta|H_0)\, P(x|\vartheta, H_0)\, d\vartheta} \tag{3}$$

Even posterior probabilities can quantify evidence, on the condition that there is an authoritative way of assigning prior probabilities in the specific problem. There is a variety of desirable properties which a reasonable measure of evidence is supposed to possess, e.g. invariance under transformation and reparametrization of the data. Indeed, Subhash Lele shows in his (2004) that under these and a lot of further reasonable constraints, the likelihood ratio of two models emerges as the optimal evidence function.[8] Indeed, such an understanding of evidence opens the way to a lot of fruitful applications

[7] Compare Howson and Urbach 1993, 192-193, and Royall 1997, 68-71.

[8] Cf. Lele 2004, in particular pp. 192-196. Lele also discusses cases where the likelihood ratio is not directly applicable, e.g. in the presence of nuisance parameters, and the problem of the sensitivity to outliers.

(see, for instance, Royall 1997 and 2000). Lele's result and Royall's applications have a positive character – they show how a comparative understanding of evidence can enhance and improve our statistical practice. However, they do not show, at least not in a rigorous sense, that it is meaningless to speak of "evidence for/against H_0" simpliciter. People who assert this usually use an argument which Royall has baptized "Fisher's disjunction":

> "Either an exceptionally rare chance has occurred, or the theory (i.e. the null model, the author) [...] is not true."[9]

In other words, results that are very unlikely under the null model count as strong evidence against the null model and justify dismissal. Note, by the way, the connection to p-values: The more the actual result diverges from the null model, the higher the p-value and the stronger, according to Fisher's disjunction, the evidence against the null. I am now going to show that it is impossible to make sense of Fisher's disjunction without introducing alternative models.

We have to decide whether the chance in the above quote has to be relatively rare (compared to the other possible outcomes) or absolutely rare, i.e. the probability of the observed outcome falls below a certain threshold. A result x_i with $p(x_i) = 0.01$ is very unlikely, but it can be very likely compared to the other results, and vice versa. To reject a hypothesis based on an absolute interpretation of Fisher's disjunction yields absurd results: think of a distribution with a large, finite number of points where any observation is very unlikely. According to the absolute interpretation, all possible results would provide strong evidence against the model. That is clear nonsense since this very strong conclusion would be warranted independent of what we observe. Hence, if we want to make sense of Fisher's disjunction, it seems mandatory to assume that a *relatively* unlikely outcome is a guide to evidence against a model or justifies dismissal of the null. Nevertheless I believe that reliance on relative unexpectedness cannot work either. Here is my argument.

There is a very general problem – measures of relative unexpectedness involve the likelihood of results that were not observed, thereby depending on the sample space and violating the objectivity conditions. We have already seen that such a dependence is fallacious – think of the two measuring instruments. But there is a more specific problem, too. Measures of relative unexpectedness (or surprise) are relative to a statistic that summarizes the data in an adequate way. The choice of the statistic, however, reveals implicit assumptions about the target of the investigation. Consider again the repeated toss of a coin. Our null model H_0 asserts that the coin is fair, i.e. all sequences of heads and tails are equally likely under H_0. Since all results are

[9] Fisher 1959, 39.

equally likely, a measure of relative unexpectedness returns the same value, regardless of the actually observed sequence. Therefore the full data cannot be the right statistic when we aim at substantial conclusions.[10] It appears natural to count only the number of heads and tails that occurred in the trial, because we believe that the order of outcomes does not matter at all. But then we have identified a parameter of interest because that particular statistic is minimally sufficient with regard to the propensity of the coin to fall heads. Thus, when we want to determine what an "exceptionally rare chance" could be, we have to identify a parameter of interest ϑ and to choose a statistic that is minimally sufficient with regard to ϑ.[11] In other words, there is no exceptionally rare chance as such – any such chance is relative to the choice of a statistic that determines *the way in which it is exceptional*. For instance, we might observe much more heads than tails, leading to a surprising result under the null model. But even when a relatively expected result obtains – heads and tails are roughly balanced – the data could have a pattern that casts heavy doubt on the independence assumption regarding the single trials. Such a pattern (like '10101010') would also be very unexpected under the null model, but it is not detected by counting the number of heads. Hence, we do not judge the tenability of H_0 "in general", without recourse to a specific parameter or comparison to alternatives – we always examine a certain way the data could be surprising. Thus, when applying Fisher's disjunction, we are asking specific questions about a parameter as "why that value of ϑ rather than another one?". The choice of the statistic reveals a comparative question. This contradicts, of course, the aim to speak of evidence for or against a model simpliciter, without recourse to alternatives.

To summarize: Fisher's disjunction and the inference from relatively unlikely results to evidence against the hypothesis neglect (a) that relative unexpectedness must be based on a (minimally sufficient) statistic and (b) that the choice of the statistic specifies a particular question of inference. We have seen that evidence has to be an objective and intersubjectively accessible concept in order to be communicable in scientific journals. That condition urges us towards a comparative understanding of evidence, as the failure of Fisher's disjunction makes clear. Measures which take into account counterfactual considerations as dependence on the sample space violate this constraint. For those measures different interpretations must be given. They might be classified as measures of *surprise* or *relative unexpectedness*. The following section goes back to p-values and connects them to the epistemo-

[10] Compare again Seidenfeld 1979, 80. Seidenfeld also discusses Fisher's disjunction, but under the (equivalent) label of "significance tests".

[11] In the present case, it appears at superficial sight that there can be only one parameter of interest. But some model families have two or more parameters, e.g. mean and variance in the case of the normal distribution. For instance, the sample mean is minimally sufficient for the population mean, but not for the population variance.

logical functions of measures of surprise.

4 P-values and measures of surprise

Given the foregoing distinction between surprise and evidence, we can now see that p-values are better interpreted as measures of surprise than as measures of evidence. First, they depend on a specific partition of the sample space, i.e. on the likelihood of outcomes that have a higher discrepancy to the null than the actually observed one. This is clear from equation (1). Therefore p-values can not be as objective as it is required for measures of evidence. Second, we have seen in the previous section that measures of evidence ought to be comparative. P-values can accommodate that by determining a specific direction of departure. For instance, we might have a normally distributed population with known variance 1, and we might want to examine whether the population mean exceeds 0. Hence, we choose the standard normal model $H_0 : N_{0,1}$ as our null model and check whether the p-value indicates a significant departure towards a higher population mean μ. Taking the sample mean $\overline{X}_n$ as a minimally sufficient statistic, assume that we obtain $\overline{X}_n = 0.6$ with a sample size of $n = 10$. Then, familiar transformations yield a p-value of 0.029:

$$\begin{aligned} p_{\text{obs}}(0.6) &= N_{0,1}\{\overline{X}_n \geq 0.6\} = N_{0,\frac{1}{n}}[0.6; \infty] \\ &= N_{0,1}[1.897; \infty] = 0.029 \end{aligned}$$

According to all statistical practitioners, such a low p-value constitutes a significant departure from the null. But what exactly does that mean? It cannot mean that *all* population means greater than 0 are strongly favored over the null model. We quickly recognize that by looking at the alternative model $\mu = 2$ which diverges much more from the data than the null model $N_{0,1}$. So, a low p-value probably means that *some* alternative models with population mean greater than zero are favored over the null, e.g. the model $\mu = 0.6$ (see also figure 2). In other words, it depends on the peculiar alternative models (or the direction of departure) whether p-values constitute evidence against the null or not; the p-value itself is independent of which alternatives we have in mind. In other words, p-values alone cannot carry information about weight of evidence, or in other words, they do not give not enough information to assess the tenability of a model. Such assessments are relative to the alternative models proposed. Moreover, in the above example the p-value is extremely low whereas a sensible measure of *evidence* as the likelihood ratio gives at best *moderate* evidence against H_0. Even for the maximally favored model $\mu = 0.6$, we get only $P(\overline{X}_n|\mu = 0.6)/P(\overline{X}_n|\mu = 0) \approx 6.049$ which can by no means count as strong evidence against H_0. The low p-value deceives us into disbelieving the null model although the weight

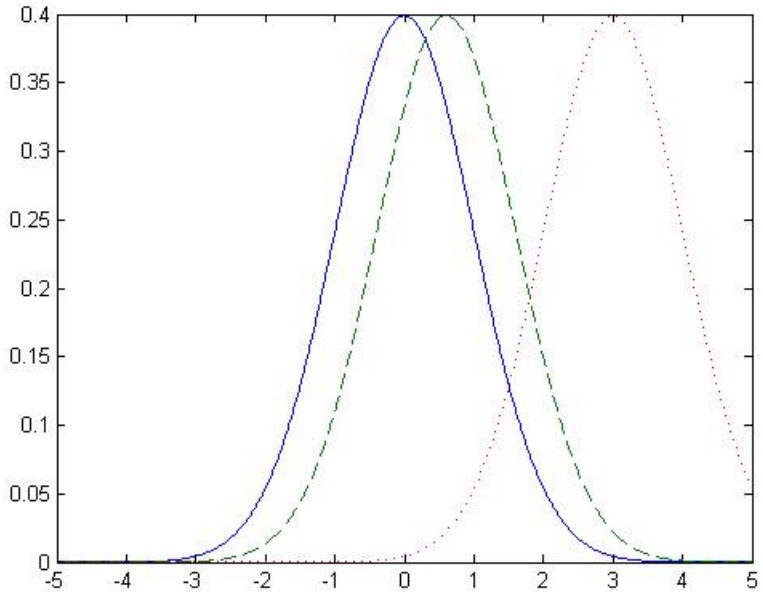

Figure 2: The null model $N_{0,1}$ (straight line) versus the alternative models $N_{0.6,1}$ (dashed line) and $N_{2,1}$ (dotted line)

of evidence, as measured by the likelihood ratio, is far from being conclusive against H_0, even if we choose the model that makes the data most likely. Hence, p-values give an overly pessimistic evaluation of the null model and cannot serve as an evidential basis of our decisions and inferences.

Still, it is an open question whether there are other applications for p-values. Since two decades, statisticians have been researching on the connection between p-values and Bayesian measures of evidence. Indeed, there is a compatibility result when a specific direction of departure from the null model is distinguished. Take again a normal distribution N_{μ,σ^2} with known variance σ^2 and unknown mean μ. The rivalling models are $H_0 : \mu \leq 0$ and $H_1 : \mu > 0$. Then, the statistic $\overline{X}_n := \frac{1}{n}\sum_{k=1}^{n} X_i$ is minimally sufficient with regard to μ. Casella and Berger (1987) show that the p-value of with regard to H_0, $\overline{X}_n$ and $T(x) := x$ provides a *lower bound for the posterior probability of* H_0, taken over a certain class of prior densities π that assign equal weight to both models. In mathematical terms,

$$\inf_{\pi} P(\mu \leq 0|x_0) = p_{\text{obs}}(x_0) := P(X \geq x_0|\mu = 0) \tag{4}$$

or, equivalently,

$$1 - p_{\text{obs}}(x_0) = \sup_{\pi} P(\mu > 0|x_0) \tag{5}$$

(Theorem 3.2. in Casella and Berger 1987, 108.[12]) In other words, under suitable (and "impartial") prior assignments, the null model is at least as

[12]Casella and Berger even derive this result for any distribution family that is indexed

likely as the p-value indicates. That result explains why Bayesian posterior probabilities and p-values are often conflated. In the Casella/Berger case, the p-value $p_{\text{obs}}(x_0) = P(X \geq x_0|\mu = 0)$ sums up the probability of those values where the evidence in favor of H_1 is greater than at the actually observed value x_0. In a similar vein, under a suitably narrow class of prior distributions π and alternatives H_1, p-values can be calibrated as to provide lower bounds on Bayes factors (see Sellke, Bayarri and Berger 2001):

$$\inf_{\pi} B(H_1, H_0, x) = -ep_{\text{obs}}(x) \log p_{\text{obs}}(x) \tag{6}$$

Hence, we see how p-values depend on proper measures of evidence, offering a lower bound for the posterior probability of the null. In practice, this can be very useful: instead of a cumbersome and computationally expensive Bayesian analysis, a quickly performed computation of the p-values gives a rough idea of whether the null model is severely shaken by the data. The p-value is easy to calculate and avoids careful deliberation about prior probabilities etc. For instance, if the p-value is greater than 0.1, we know that the null model has *at least* a probability of 0.1 so that it remains a serious candidate. In other words, knowing the p-values can make more detailed investigations that aim at the dismissal of the null model superfluous. Rightfully, Bayesians often stress that the use of p-values in a Bayesian framework has merely auxiliary character; as soon as a full Bayesian analysis is possible, they cannot play any role. So, although p-values can give a rough idea about a the evidential content of the data, the actual computation of the strength of evidence or a rejection of the null model can seldom, if ever, be based on the calculation of p-values. In the latter case, this is particularly salient because they merely provide lower bounds for the posterior probability of and the evidence against the null model where an upper bound would be required. Hence, although no scientific report should cite the observed p-value in favor of rejecting the null model (as it is often done, unfortunately), working with p-values remains practically useful, having a heuristic value.[13]

Another function of p-values might consist in measuring *surprise* or *relative expectedness* in the data. Surprise has an epistemological function that is particularly important in exploratory model analysis. When a result turns out to be surprising under all possible parameter values, we are more or less forced to develop and to specify alternatives to the original model. Whereas, when a result is not at all surprising in a certain respect, we might decide to go on with the old model or draw our attention to other ways in which

by μ that is (1) symmetric around zero and (2) has a monotonely increasing likelihood ratio.

[13]Recall however, that these results hold for p-values with a specified direction of departure. When no such direction is specified, p-values grossly overstate the evidence against the null, and their interpretation becomes much more difficult, as it was shown by Berger and Sellke 1987.

the result could be surprising. In many cases, it is overly optimistic to assume that all relevant models are available *a priori*, i.e. before having a look at the data and proceeding with a model selection analysis. If all relevant models were known, measures of surprise would be pointless and could be directly replaced by model selection procedures.[14] Thus, surprise measures can help us to decide whether we should base the subsequent analysis on a comprehensive or a parsimonious set of rivalling models. Compare that kind of reasoning to a model selection analysis. Model selection and validation require reasons for favoring a model over a number of competitors – no successful model is thought to best the absolutely best one, but only a more reasonable approximation of the truth than its competitors. That selection procedure often requires a careful assignment of prior probabilities etc. and deliberation about the precise selection criteria. By contrast, an exploratory model analysis is only concerned with the set of models which will be subjected to the model selection procedures. That is the point where surprise measures enter the stage. They have a heuristic value in guiding our analysis. Indeed, the usage of p-values outlined above (as giving lower bounds on posterior probabilities and Bayes factors) appears to be driven by the same heuristic considerations.

When we interpret p-values as measures of surprise, distance to the model is usually measured by the likelihood function.

$$p_{\text{obs}}(x_0) := P_{H_0}(\{P_{H_0}(X) \geq P_{H_0}(x_0)\}) \tag{7}$$

Thus, the p-value sums up the probability of those outcomes that are less likely than the actually observed one. Note that the dependence on the sample space which p-values exhibit is not harmful for a measure of surprise since surprise and relative expectedness are psychological concepts which are not subject to strong objectivity requirements as the concept of evidence. Nonetheless, there are two basic problems with interpreting p-values as measures of surprise: on the one hand, p-values tend to overstate the evidence against the null model when they are based on the likelihood function.[15] On the other hand, they are not continuous in the probability density. Practically nearly undetectable changes and measuring inaccuracies can yield huge differences in the p-values which is clearly unacceptable.

Both problems suggest that we better calibrate p-values in a way that avoids these problems. Furthermore, we might wish to make them more amenable to numerical computation, e.g. using Monte Carlo methods (cf. Bayarri and Berger 1999). There are lots of suggestions about how to perform such calibrations, most of them fine-tuned to the use of surprise indices in a Bayesian

[14] For papers that describe how model selection can be done in a fully Bayesian way, see e.g. Wasserman 2000.

[15] Cf. Berger and Sellke 1987. This distinguishes surprise-measuring p-values which basically arise from a two-sided testing problem from p-values in the one-sided testing problem discussed above.

framework.[16]. Instead of giving a comprehensive review that would go beyond the scope of this article I would like to present a very recent proposal by John V. Howard, made in his (2007). Howard's proposal solves the above problems is very intuitive, too, saving the basic intuitions about p-values.

Howard 2007 takes the density function p as the distance function and suggests a calibration of p-values by trunctating the density function at the actually observed value x_0. Then, the integral over the trunctated density function constitutes the surprise index h:

$$\begin{aligned} s(x_0) &:= \int \min\{p(x_0), p(x)\}dx \qquad (8) \\ &= p_{\text{obs}}(x_0) + p(x_0) \int 1_{\{p(x)>p(x_0)\}}dx \end{aligned}$$

The relationship to p-values is obvious from the second line of equation 8

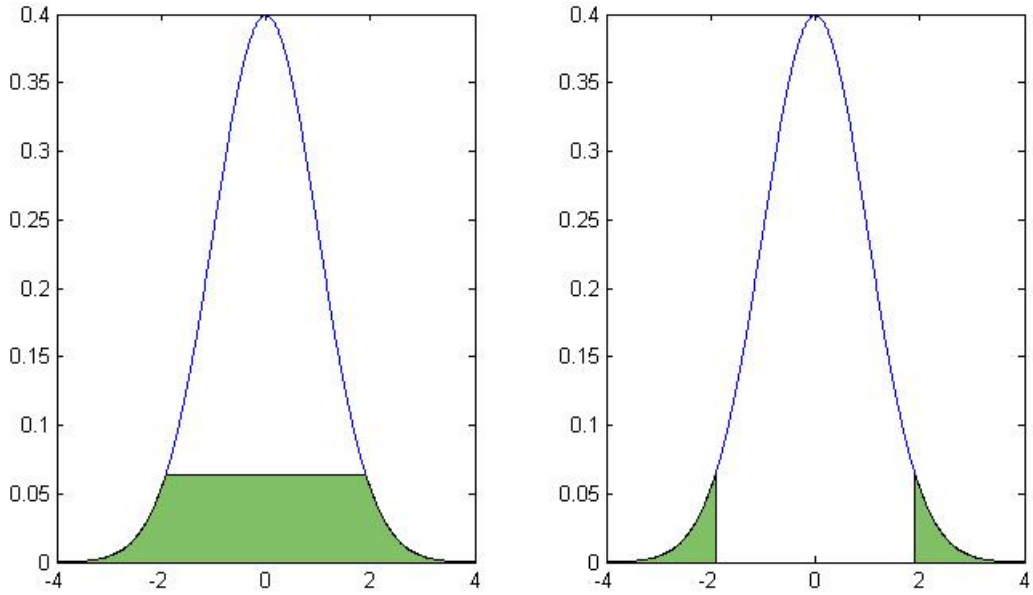

Figure 3: To the left, Howard's surprise index, graphically interpreted. To the right, the classical p-value.

and figure 3: s-values do not only focus on the probability of more "extreme" outcomes, they also consider other characteristics of the distribution. A low s-value assigns both as small absolute and a small relative probability to the actually observed result whereas p-values state that the observed result has a low absolute probability and that an unspecified alternative offers a better

[16]See, for instance, Bayarri and Berger 1998, 1999 and 2000, as well as Robins, van der Vaart and Ventura 2000

explanation. In this sense, s-values are closer to measuring surprise than p-values although both are closely related to each other. It is easy to see that Howard's surprise index s solves the problem of continuity and is also more conservative than the original p-values.

There are, of course, lots of different measures of surprise, and I do not want to argue for a specific measure. The vast number of different tasks in model exploration is unlikely to favor a single measure in all circumstances, and consequently, there is a large variety of surprise measures (see the Bayarri/Berger papers cited above, and also Good 1956 and Evans 1997). Indeed, I did not turn my attention to Howard's surprise measure for that reason. Rather, I wanted to show how p-values can be related to measures that are more closely attached to relative expectedness and surprise. Apart from that, in virtue of their connections to Bayesian measures of evidence, p-values still play a useful heuristic role in exploratory model analysis.

5 Bayesianism, surprise and the problem of old evidence

The penultimate section sheds a new light on a classical problem in confirmation theory: the problem of old evidence.[17] It arises when formerly known evidence is used to confirm a newly developed hypothesis. In other words, a new hypothesis H enters the space of hypotheses. Then, it seems to be fair that H can be confirmed by formerly known evidence E because H had no chance to be confirmed or disconfirmed by E prior to its development. But Bayesian conditionalization cannot account for the fact that the introduction of a new hypothesis changes our epistemic situation. The old evidence is already known and has probability 1 so that it cannot confirm anything. According to orthodox conditionalization, the new hypothesis must have been a part of the previous Bayesian supermodel even if we did not formulate it explicitly. Such a line of reasoning, however, seems to be far from reality and scientific practice. There is no Bayesian supermodel containing all hypotheses that could ever be formulated, and even if there were such a supermodel, we would not entertain it due to its complexity. Rather we restrict Bayesian model checking to a selection of sensible and fruitful models. Hence, Bayesian confirmation theory cannot account for the introduction of novel theories.

To see the problem in practice, recall that Einstein's General Theory of Relativity (GTR) was the first theory that successfully explained the perihelion

[17] For a detailed description of the problem and several attempted Bayesian solutions, see Earman 1992.

advance of Mercury in 1915. However, the Mercury data were already intensively studied in the nineteenth century by Leverrier and other astronomers and were considered as a major anomaly of Newtonian celestial mechanics. This went so far that a new, unobserved planet inside the Mercury orbit ("Vulcan") was postulated in order to account for the perihelion advance. According to Bayesian Conditionalization, the perihelion data E were old evidence so that $P(E) = 1$ at the time when Einstein developed the GTR. Hence, according to Bayesian conditionalization, E cannot raise the credence in GTR. This clashes with the viewpoint of most physicists that GTR was best confirmed by the old Mercury perihelion data and not by novel evidence, e.g. the deflection of starlight by the sun. More generally, scientists happily embrace solutions to penetrating old problems so that old evidence might have a special confirmatory power.

If we limit the scope of a Bayesian analysis to model *selection* and try to account for the epistemic role of surprise, we can possibly get a better understanding of the problem of old evidence. When data are surprising with regard to a model and when this unexpectedness does not disappear upon refinements of the model, we are inclined to focus our efforts on the development of new hypotheses. In the long run, resilient anomalies demand for the development of alternative models. Repeatedly detecting relatively unobserved results leads us into questioning our models and drives scientific model change. Bayesianism, on the other hand, has no place for surprise measures, as shown in the previous section. However, developing alternatives as a consequence of surprising, unexpected results better corresponds to the dynamics of science. Hence, a surprise index is not only valuable for exploratory data analysis, but also for guiding and driving the introduction of novel models that are not contained in a Bayesian supermodel.

The problem of old evidence shows that Bayesian conditionalization cannot be a comprehensive account of statistical model checking and development. Bayesian statisticians have conceded that for a long time, on the grounds that exploratory model analysis is placed outside a strictly Bayesian framework, just because it precedes and lays the groundwork for Bayesian model selection. Thus it is strange that confirmation theorists have been worrying for such a long time about the problem of old evidence, nourishing a hope that was already abandoned by statistical practitioners. With the surprise/evidence distinction in mind, we can devise a strategy to resolve the problem: Straightforward Bayesian analysis is adequate for model selection and validation, but it must be supplemented by an account of exploratory model analysis. Developing surprise indices and clarifying their epistemological role helps to accomplish the latter task.

6 Summary and conclusions

This paper has made several points concerning the role of surprise and evidence in statistical model analysis. First, I have outlined the confusion about the most adequate interpretation of p-values. Then I have elaborated a separation of surprise and evidence, the latter being an essentially comparative concept whereas the former is allowed to depend on the sample space. Evidence has its place in model selection and validation whereas quantifying surprise is useful in exploratory model analysis and choosing the right supermodel for a Bayesian analysis. Surprise indices, on the other hand, drive and guide the development of competitors to the first, tentative models, so that their function clearly differs from measures of evidence. For practitioners, it is important to keep that distinction in mind – surprise indices are not adequate for final decisions on the tenability of a model. That distinction was then brought to bear on the interpretation of p-values. It was shown that p-values cannot be measures of evidence whereas they (a) provide lower bounds on measures of evidence and (b) can be modified as to yield measures of surprise. An intuitive modification might be given by Howard's s-value. Finally, the surprise/evidence distinction gives a means of tackling the resilient problem of old evidence in confirmation theory. It is clear that this problem cannot be solved by an *evidential* appraisal of a certain model. But the epistemic role of surprise explains why purely Bayesian approaches come to their limits in early stages of model checking. Hence, the problem of old evidence does not pose a serious threat to Bayesian inference.

References

[1] M. J. Bayarri, James O. Berger (1997): "Measures of Surprise in Bayesian Analysis", *ISDS Discussion Paper 97-46*, Duke University.

[2] M. J. Bayarri, James O. Berger (1999): "Quantifying Surprise and Model Verification", in: J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith (ed.), *Bayesian Statistics 6*, 53-82. Oxford University Press, Oxford.

[3] M. J. Bayarri, James O. Berger (2000): "P Values for Composite Null Models", *Journal of the American Statistical Association* **95**, 1127-1142.

[4] James O. Berger, Thomas Sellke (1987): "Testing a Point Null Hypothesis: the Irreconcilability of P Values and Evidence", *Journal of the American Statistical Association* **82**, 106-111.

[5] George Casella, Roger L. Berger (1987): "Reconciling Bayesian and Frequentist Evidence in the One-Sided Testing Problem", *Journal of the American Statistical Association* **82**, 106-111.

[6] John Earman (1992): *Bayes or Bust?*. The MIT Press, Cambridge/MA.

[7] Michael Evans (1997): "Bayesian inference procedures derived via the concept of relative surprise", *Communications in Statistics* **26**, 1125-1143.

[8] R. A. Fisher (1959): *Statistical Methods and Scientific Inference*. Second Edition, Hafner, New York.

[9] I. J. Good (1956): "The Surprise Index for the Multivariate Normal Distribution", *Annals of Mathematical Statistics* **27**, 1130-35.

[10] John V. Howard (2007): *Significance testing with no alternative hypothesis: a measure of surprise*. Unpublished manuscript.

[11] Colin Howson, Peter Urbach (1993): *Scientific Reasoning: The Bayesian Approach*. Second Edition, La Salle: Open Court.

[12] Subhash Lele (2004): "Evidence Functions and the Optimality of the Law of Likelihood (with discussion)", in: Mark Taper and Subhash Lele (ed.), *The Nature of Scientific Evidence*, 191-216. The University of Chicago Press, Chicago & London.

[13] Deborah G. Mayo (2004): "Evidence as Passing Severe Tests: Highly Probable versus Highly Probed Hypotheses", in: Peter Achinstein (ed.), *Scientific Evidence*, 95-127. Johns Hopkins University Press, Baltimore/MD.

[14] Deborah G. Mayo, Aris Spanos (2006): "Severe Testing as a Basic Concept in a Neyman-Person Philosophy of Induction", *British Journal for the Philosophy of Science* **57**, 323-357.

[15] James M. Robins, Aad van der Vaart, Valérie Ventura (2000): "The asymptotic distribution of p-values in composite null models", *Journal of the American Statistical Association* **95**, 1143-1156.

[16] Richard Royall (1997): *Statistical Evidence: A Likelihood Paradigm*. Chapman & Hall, London.

[17] Richard Royall (2000): "On the Probability of Observing Misleading Statistical Evidence", *Journal of the American Statistical Association* **95**, 760-768.

[18] Richard Royall (2004): "The Likelihood Paradigm for Statistical Evidence (with discussion)", in: Mark Taper and Subhash Lele (ed.), *The Nature of Scientific Evidence*, 119-152. The University of Chicago Press, Chicago & London.

[19] Teddy Seidenfeld (1979): *Philosophical Problems of Statistical Inference: Learning from R. A. Fisher*. Reidel, Dordrecht.

[20] Thomas Sellke, M. J. Bayarri, James O. Berger (2001): "Calibration of P-Values for testing precise null hypotheses", *The American Statistician* **55**, 62-71.

[21] Larry Wasserman (2000): "Bayesian Model Selection and Model Averaging", *Journal of Mathematical Psychology* **44**, 92-107.