A NOTE ON P-VALUES INTERPRETED AS PLAUSIBILITIES
Author(s): Ryan Martin and Chuanhai Liu
Source: *Statistica Sinica*, Vol. 24, No. 4 (October 2014), pp. 1703-1716
Published by: Institute of Statistical Science, Academia Sinica
Stable URL: http://www.jstor.org/stable/24310965
Accessed: 14-04-2017 12:54 UTC

Statistica Sinica **24** (2014), 1703-1716
doi:http://dx.doi.org/10.5705/ss.2013.087

A NOTE ON P-VALUES INTERPRETED AS PLAUSIBILITIES

Ryan Martin and Chuanhai Liu

University of Illinois at Chicago and Purdue University

Abstract: P-values are a mainstay in statistics but are often misinterpreted. We propose a new interpretation of p-value as a meaningful plausibility, where this is to be interpreted formally within the inferential model framework. We show that, for most practical hypothesis testing problems, there exists an inferential model such that the corresponding plausibility function, evaluated at the null hypothesis, is exactly the p-value. The advantages of this representation are that the notion of plausibility is consistent with the way practitioners use and interpret p-values, and the plausibility calculation avoids the troublesome conditioning on the truthfulness of the null. This connection with plausibilities also reveals a shortcoming of standard p-values in problems with non-trivial parameter constraints.

Key words and phrases: Hypothesis test, inferential model, nesting, plausibility function, predictive random set.

1. Introduction

P-values are ubiquitous in applied statistics, but are widely misinterpreted as either a sort of Bayesian posterior probability that the null hypothesis is true, or as a frequentist error probability. Indeed, in 2012, media were reporting the discovery of the elusive Higgs boson particle (Overbye (2012)) and statistics blogs were pointing out how some journalists and physicists had misinterpreted the resulting p-values. Our objective here is to provide a new and simpler way to understand them, so that these misinterpretations might be avoided.

A prime reason for the frequent misinterpretation of p-values is that the standard textbook definition is inconsistent with people's common sense. The goal of this paper is to provide a more user-friendly interpretation. We show that the p-value can be interpreted as a plausibility that the null hypothesis is true. This "plausibility" is precisely defined within the inferential model (IM) framework of Martin and Liu (2013a), built upon two fundamental principles of Martin and Liu (2014a) for valid and efficient statistical inference. Consider the problem of testing a null hypothesis H_0 versus a global alternative H_1. We show that, under mild conditions, for any p-value (depending on H_0 and the choice of test statistic), there exists a valid IM such that the plausibility of H_0 is the p-value. In this sense,

the p-value can be understood as the plausibility, given the observed data, that H_0 is true. In the Higgs boson report, since the p-value is minuscule ($p \ll 10^{-6}$), one concludes that the hypothesis H_0 : "the Higgs boson does not exist" is highly implausible, hence, a discovery. This line of reasoning based on small p-values is consistent with Cournot's principle (Shafer and Vovk (2006)).

The word "plausibility" fits the way practitioners use and interpret p-values: a small p-value means H_0 is implausible, given the observed data. Evaluating plausibility involves a probability calculation that does not require one to assume that H_0 is true, so one avoids the questionable logic of proving H_0 false by using a calculation that assumes it is true. The use of IMs to provide probabilistic interpretations of classically non-probabilistic summaries is proving to be beneficial; see, for example, Martin (2014).

The remainder of the paper is organized as follows. Section 2 sets up our notation and gives the formal definition of p-value, with a brief discussion of its common correct and incorrect interpretations. The basics of IMs are introduced in Section 3, in particular, predictive random sets and plausibility functions. In Section 4 we prove that, given essentially any hypothesis testing problem, there is a valid IM such that the corresponding plausibility function, evaluated at the null hypothesis, is the p-value. There we highlight a similar connection between the IM plausibilities and objective Bayes posterior probabilities, and an apparently unrecognized shortcoming of p-values in problems with non-trivial parameter constraints. Two examples involving binomial and normal data are presented in Sections 4.3–4.4, and some concluding remarks are given in Section 5.

2. The P-value

2.1. Setup and formal definition

Let X denote observable data, taking values in $\mathbb{X}$. There is a sampling model $\mathsf{P}_{X|\theta}$, indexed by a parameter $\theta \in \Theta$, and the goal is to make inference on θ using the observed data $X = x$. Here both X and θ are allowed to be vector-valued, but do will not make this explicit in the notation. The hypothesis testing problem starts with a hypothesis, or assertion, about the unknown θ. Mathematically, this is characterized by a subset $\Theta_0 \subset \Theta$, and we write $H_0 : \theta \in \Theta_0$ for the null hypothesis and $H_1 : \theta \notin \Theta_0$ for the alternative hypothesis. The goal is to use observed data $X = x$ to determine, with some measure of certainty, whether H_0 or H_1 is true.

Consider the description of the p-value given by (Fisher, 1959, p.39), viewed as follows. If the observed $X = x$ gives small p-value, then one of two things occurred: relative to H_0, a rare chance event has occurred, *or* H_0 is false. The unlikeliness of the former drives us to conclude the latter. To put this in more standard terms, suppose there is a test statistic $T : \mathbb{X} \to \mathbb{R}$, possibly depending

on Θ_0, such that large values of $T(X)$ suggest that H_0 may not be true. The p-value is defined, for $X = x$, as

$$\mathsf{pval}(x) = \mathsf{pval}_{T,\Theta_0}(x) = \sup_{\theta \in \Theta_0} \mathsf{P}_{X|\theta}\{T(X) \geq T(x)\}. \tag{2.1}$$

When $\Theta_0 = \{\theta_0\}$, a point null, (2.1) simplifies to $\mathsf{pval}(x) = \mathsf{P}_{X|\theta_0}\{T(X) \geq T(x)\}$, the expression found in most introductory textbooks.

Intuitively, $\mathsf{pval}(x)$ compares the observed $T(x)$ to the sampling distribution of $T(X)$ when H_0 is true. If $\mathsf{pval}(x)$ is small, then $T(x)$ is an outlier under H_0 and we conclude that H_0 is implausible. Conversely, if $\mathsf{pval}(x)$ is relatively large, then the observed $T(x)$ is consistent with at least one $\mathsf{P}_{X|\theta}$, with $\theta \in \Theta_0$, so H_0 is plausible in the sense that it provides an acceptable explanation of reality.

2.2. Standard interpretations

Standard textbooks have adopted an equivalent though arguably more obscure interpretation. The standard textbook interpretation of p-value goes something like this:

> $\mathsf{pval}(x)$ is the probability that an observable X is "at least as extreme" as the x actually observed, assuming H_0 is true.

This leads to the common misinterpretation of p-value as a sort of Bayesian posterior probability of H_0. Lehmann and Romano (2005, Sec. 3.3) after laying out the details of the Neyman–Pearson testing program, have it that

> $\mathsf{pval}(x)$ is the greatest lower bound on the set of all α such that the size-α test rejects H_0 based on $T(x)$.

A danger here is that the conditioning on H_0 is hidden in the definition of size; users can potentially misinterpret $\mathsf{pval}(x)$ as the probability of incorrectly rejecting H_0 based on x.

Some statisticians have abandoned the use of p-values, advocating for other tools for measuring evidence supporting H_0 and/or testing H_0, such as confidence intervals; see, e.g., Berger and Delampady (1987, Sec. 4.3) and the discussion by G. Casella and R. Berger on that same paper. This preference for confidence intervals is fairly common in medical, social, and other applied sciences. Some journals, such as the *American Journal of Public Health*, have made concerted efforts to get authors to use confidence intervals rather than p-values (Fidler et al. (2004)). Still, confidence intervals are not free of their own difficulties, nor are Bayes factors (Kass and Raftery (1995)) or Bayesian p-values (Gelman, Meng, and Stern (1996); Rubin (1984)). We think a better or simpler way to understand the ubiquitous p-value is a valuable contribution.

3. Review of Inferential Models

3.1. Big picture

The inferential model (IM) framework produces exact prior-free probabilistic measures of evidence for/against any assertion about the unknown parameter; see Martin and Liu (2013a), Martin, Zhang, and Liu (2010), and Zhang and Liu (2011). This is accomplished by first making an explicit association between the observable data X, the unknown parameter θ, and an unobservable auxiliary variable U. Random sets are introduced to predict the unobservable U, and inference about θ is obtained via probability calculations with respect to the distribution of this random set. The IM framework has some connections with existing approaches, such as fiducial (Hannig (2009, 2013); Hannig and Lee (2009)), confidence distributions (Xie, Singh, and Strawderman (2011); Xie and Singh (2013)), Dempster–Shafer theory (Dempster (2008); Shafer (1976, 2011)), generalized p-values and confidence intervals (Tsui and Weerahandi (1989); Weerahandi (1993); Chiang (2001)), and Bayesian inference with default, reference, and/or data-dependent priors (Berger (2006)); Berger, Bernardo, and Sun (2009); Fraser et al. (2010); Fraser (2011); Ghosh (2011)).

IMs, fiducial, and Dempster–Shafer theory all introduce auxiliary variables into the inference problem. Both fiducial and Dempster–Shafer theory condition on the observed $X = x$, then develop a sort of distribution on the parameter space by inverting the data–parameter–auxiliary variable relationship and assuming that U retains its *a priori* distribution after $X = x$ is observed. The IM approach targets the (unattainable) best possible inference corresponding to the case where U is observed. Uncertainty about θ, after $X = x$ is observed, is propagated from the uncertainty about hitting the true U with a random set. In addition to accomplishing Fisher's goal of prior-free probabilistic inference, IMs produce inferential output which is valid for any assertion of interest (Section 3.3); fiducial probabilities are valid only for special kinds of assertions (Martin and Liu (2013a, Sec. 4.3.1)). Moreover, a general theory of optimal IMs, concerning efficiency of the resulting inference, may not be out of reach.

3.2. Construction

Following Martin and Liu (2013a), the IM construction proceeds in three steps.

A-step. This proceeds by specifying an association between X, θ, and U. Like fiducial, Dempster–Shafer, and Fraser's (1968) structural models, this can be described by a pair (P_U, a), where P_U describes the distribution (and also, implicitly, the support $\mathbb{U}$) of the auxiliary variable U, and a describes the data-generating mechanism driven by P_U. We write this as

$$X = a(\theta, U), \quad \text{with} \quad U \sim \mathsf{P}_U.$$

That is, if U is sampled from P_U and then plugged in to the function a for a given θ, then the resulting X has distribution $\mathsf{P}_{X|\theta}$. The association need not be described by a formal equation—it is enough to have a rule/recipe to construct X from a given θ and U; see e.g., Section 4.3. To complete the A-step, construct a sequence of subsets of Θ indexed by (x, u):

$$\Theta_x(u) = \{\theta : x = a(\theta, u)\}. \tag{3.1}$$

P-step. Based on the idea that knowing the unobserved value of U is "as good as" knowing θ itself, the goal of the prediction step is to predict this unobserved value with a predictive random set, denoted by $\mathcal{S}$. Certain assumptions are required on the support $\mathbb{S}$ and distribution $\mathsf{P}_{\mathcal{S}}$ of the predictive random set; see Section 3.3.

C-step. This step combines the observed $X = x$, which specifies a sub-collection of sets $\Theta_x(\cdot)$ in (3.1), with the predictive random set $\mathcal{S}$. The result is an x-dependent random subset of Θ:

$$\Theta_x(\mathcal{S}) = \textstyle\bigcup_{u \in \mathcal{S}} \Theta_x(u). \tag{3.2}$$

Evidence for/against an assertion $A \subseteq \Theta$ concerning the unknown parameter can now be obtained via the $\mathsf{P}_{\mathcal{S}}$-probability that $\Theta_x(\mathcal{S})$ is/is not a subset of A. More precisely, we evaluate

$$\mathsf{bel}_x(\cdot; \mathcal{S}) = \mathsf{P}_{\mathcal{S}}\{\Theta_x(\mathcal{S}) \subseteq \cdot\}, \tag{3.3}$$

the belief function, at both A and A^c, as a summary of evidence for and against A, respectively. Alternatively, we can report $\mathsf{bel}_x(A; \mathcal{S})$ together with

$$\mathsf{pl}_x(A; \mathcal{S}) = \mathsf{P}_{\mathcal{S}}\{\Theta_x(\mathcal{S}) \cap A \neq \varnothing\} = 1 - \mathsf{bel}_x(A^c; \mathcal{S}), \tag{3.4}$$

the plausibility function at A. It can be readily shown that $\mathsf{bel}_x(A; \mathcal{S}) \leq \mathsf{pl}_x(A; \mathcal{S})$ for any A and any $\mathcal{S}$. Then, as described briefly below, the pair $\{\mathsf{bel}_x(\cdot; \mathcal{S}), \mathsf{pl}_x(\cdot; \mathcal{S})\}$ is used for inference about θ; see Martin and Liu (2013a) for details.

Statistical inference based on the IM output focuses on the relative magnitudes of $\mathsf{bel}_x(A; \mathcal{S})$ and $\mathsf{pl}_x(A; \mathcal{S})$. An assertion A is deemed true (resp. untrue), given $X = x$, if both $\mathsf{bel}_x(A; \mathcal{S})$ and $\mathsf{pl}_x(A; \mathcal{S})$ are large (resp. small). Conversely, if $\mathsf{bel}_x(A; \mathcal{S})$ is small and $\mathsf{pl}_x(A; \mathcal{S})$ is large, then there is no clear decision to be made about the truthfulness of A, given $X = x$, so maybe one needs more data. The definition of "small" and "large" values of belief/plausibility functions are specified by the theoretical validity properties discussed in Section 3.3.

One can also construct frequentist procedures based on plausibility functions. For example, for $\alpha \in (0, 1)$, a $100(1 - \alpha)\%$ plausibility region for θ is defined as

$$\Pi_\alpha(x) = \{\theta : \mathsf{pl}_x(\theta; \mathcal{S}) > \alpha\}. \tag{3.5}$$

It is a consequence of Theorem 1 below that this region achieves the nominal $1-\alpha$ frequentist coverage probability.

Throughout it is assumed that $\Theta_x(u)$ in (3.1) satisfies $\Theta_x(u) \neq \varnothing$ for all (x,u) pairs. This boils down to there being no non-trivial constraints on the parameter space Θ. When this fails, one can usually take a dimension-reduction step, described in Martin and Liu (2014b), to transform to a problem where this assumption holds. Under this condition, it is sometimes convenient to evaluate the plausibility on the u-space as opposed to the θ-space as in (3.4). Given x and A, let

$$\mathbb{U}_x(A) = cl\{u : \Theta_x(u) \subseteq A\}, \tag{3.6}$$

where clB denotes the closure of the set B. If $\Theta_x(u) \neq \varnothing$ for all (x,u), as we have assumed, then belief and plausibility can be evaluated on the u-space as

$$\mathsf{bel}_x(A;\mathcal{S}) = \mathsf{P}_{\mathcal{S}}\{\mathcal{S} \subseteq \mathbb{U}_x(A)\} \quad \text{and} \quad \mathsf{pl}_x(A;\mathcal{S}) = 1 - \mathsf{P}_{\mathcal{S}}\{\mathcal{S} \subseteq \mathbb{U}_x(A^c)\}. \tag{3.7}$$

This formulation is used in the main result in Section 4.

3.3. IM validity

It is important that the IM's belief and plausibility functions are meaningful across similar studies. This type of meaningfulness is referred to as validity in Martin and Liu (2013a). Here, the IM is said to be valid if

$$\sup_{\theta \in A} \mathsf{P}_{X|\theta}\{\mathsf{pl}_X(A;\mathcal{S}) \leq \alpha\} \leq \alpha, \quad \forall\, A \subseteq \Theta, \quad \forall\, \alpha \in (0,1). \tag{3.8}$$

This means that, if A is true, then $\mathsf{pl}_x(A;\mathcal{S})$ is small for only a small proportion of possible x values, "outliers." Since it holds for all $A \subseteq \Theta$, a similar statement about $\mathsf{bel}_X(A;\mathcal{S})$ can also be made. Validity holds, without special modification, even for the scientifically important case of singleton A. In fact, for reasonably chosen predictive random sets (Martin and Liu (2013a, Corollary 1)), the latter "$\leq \alpha$" can be replaced by "$= \alpha$;" hence $\mathsf{pl}_X(A;\mathcal{S}) \sim \mathsf{Unif}(0,1)$ when $A = \{\theta_0\}$ is true. In Theorem 2 below we show that the p-value is a plausibility function at the null hypothesis. So (3.8) restates the familiar result that, if the null hypothesis is true, then the p-value is (stochastically dominated by) a $\mathsf{Unif}(0,1)$ random variable.

The validity property (3.8) holds if the predictive random set $\mathcal{S}$ satisfies certain conditions, no requirements on $\mathsf{P}_{X|\theta}$ or the association (P_U, a) are needed. Let $(\mathbb{U}, \mathscr{U})$ be the measurable space on which P_U is defined, and assume that $\mathscr{U}$ contains all closed subsets of $\mathbb{U}$. Martin and Liu (2013a) gives the following result.

Theorem 1. *The IM is valid for all assertions $A \subseteq \Theta$ if $\Theta_x(u) \neq \varnothing$ for all (x,u) and the predictive random set $\mathcal{S}$ satisfies the following:*

P1. *The support* $\mathbb{S} \subset 2^{\mathbb{U}}$ *of* $\mathcal{S}$ *contains* $\varnothing$ *and* $\mathbb{U}$, *and:*
  (a) *each* $S \in \mathbb{S}$ *is closed and, hence* P_U*-measurable, and*
  (b) *for any* $S, S' \in \mathbb{S}$, *either* $S \subseteq S'$ *or* $S' \subseteq S$.

P2. *The distribution* $\mathsf{P}_{\mathcal{S}}$ *of* $\mathcal{S}$ *satisfies* $\mathsf{P}_{\mathcal{S}}\{\mathcal{S} \subseteq K\} = \sup_{S \in \mathbb{S}: S \subseteq K} \mathsf{P}_U(S)$, $K \subseteq \mathbb{U}$.

Martin and Liu (2013a) show that a wide variety of predictive random sets are available for which P1–P2 hold, so that IM validity is rather easy to arrange. However, efficiency is a concern and, for this, they present a theory of optimal IMs.

4. P-value as an IM plausibility

4.1. Main result

On the association (a, P_U), the null hypothesis Θ_0, and the test statistic $T : \mathbb{X} \to \mathbb{T}$, we assume the following.

A1. For every (x, u) there exists θ such that $T(x) = T(a(\theta, u))$.

A2. $\sup_{\theta \in \Theta_0} T(a(\theta, \cdot))$ is a P_U-measurable function.

A3. $\mathsf{P}_U\{\sup_{\theta \in \Theta_0} T(a(\theta, U)) < t\} = \inf_{\theta \in \Theta_0} \mathsf{P}_U\{T(a(\theta, U)) < t\}$ for all $t \in \mathbb{T}$.

Here, A2 ensures the meaningfulness of the probability statement in A3, and holds generally under mild separability and continuity conditions, respectively, on Θ_0 and on T and a, while A3 makes precise the stochastic smoothness and stochastic ordering $T(X)$ should possess as a function of θ. Assumptions A2–A3 hold trivially for the important point-null case. It is also easy to check A3 in many common examples, e.g., if $X_1, \ldots, X_n$ are iid $\mathsf{N}(\theta, 1)$, and $T(X) = \bar{X}$, then $T(a(\theta, U)) = \theta + \overline{U}$, and A3 holds for any Θ_0 of the form $(-\infty, \theta_0]$.

Theorem 2. *If* A1–A3 *hold for the given association* (a, P_U), *hypothesis* Θ_0, *and test statistic* $T : \mathbb{X} \to \mathbb{T}$, *then there exists an admissible predictive random set* $\mathcal{S}$, *depending on* T *and* Θ_0, *such that the plausibility function* $\mathsf{pl}_x(\Theta_0; \mathcal{S})$ *equals* $\mathsf{pval}(x) = \mathsf{pval}_{T, \Theta_0}(x)$ *in* (2.1) *for all* $x \in \mathbb{X}$.

Proof. Without loss of generality, we reduce the baseline association $X = a(\theta, U)$, with $U \sim \mathsf{P}_U$, to $T(X) = T(a(\theta, U))$, again with $U \sim \mathsf{P}_U$. In this case, the A-step of the IM construction generates the collection of subsets:

$$\Theta_x(u) = \{\theta : T(x) = T(a(\theta, u))\}, \quad x \in \mathbb{X}, \quad u \in \mathbb{U}.$$

These sets are non-empty for all (x, u) by A1. For the P-step, we define a collection $\mathbb{S} = \{S_t : t \in \mathbb{T}\}$ of subsets of $\mathbb{U}$ with

$$S_t = cl\{u : \sup_{\theta \in \Theta_0} T(a(\theta, u)) < t\}, \quad t \in \mathbb{T}.$$

The sets are closed, are nested, and P_U-measurability follows from A2. Thus P1 in Theorem 1 holds. Define a predictive random set $\mathcal{S}$, supported on $\mathbb{S}$, with distribution $\mathsf{P}_{\mathcal{S}}$ satisfying

$$\mathsf{P}_{\mathcal{S}}\{\mathcal{S} \subseteq K\} = \mathsf{P}_U(S_{t_K^\star}) = \inf_{\theta \in \Theta_0} \mathsf{P}_{X|\theta}\{T(X) < t_K^\star\}, \quad K \subseteq \mathbb{U}, \tag{4.1}$$

where $t_K^\star = \sup\{t \in \mathbb{T} : S_t \subseteq K\}$; the last equality in (4.1) is a consequence of A3. For such as $\mathcal{S}$, the corresponding IM is valid. For notational consistency, set $A = \Theta_0$. The C-step proceeds as in the general case in Section 3.2, and, for any $x \in \mathbb{X}$, the plausibility function (3.7), evaluated at A, satisfies

$$\mathsf{pl}_x(A; \mathcal{S}) = 1 - \mathsf{P}_{\mathcal{S}}\{\mathcal{S} \subseteq \mathbb{U}_x(A^c)\} = 1 - \mathsf{P}_{\mathcal{S}}\{\mathcal{S} \subseteq S_{T(x)}\}. \tag{4.2}$$

The second equality in (4.2) needs justification. First, we have $S_{T(x)} \subseteq \mathbb{U}_x(A^c)$ since

$$\begin{aligned} u \in S_{T(x)} &\implies T(x) > \sup_{\theta \in A} T(a(\theta, u)) \\ &\implies T(x) = T(a(\theta, u)) \quad \exists\, \theta \notin A \\ &\implies \Theta_x(u) \subseteq A^c \\ &\implies u \in \mathbb{U}_x(A^c). \end{aligned}$$

It remains to show that $S_{T(x)}$ is the largest of the S_t's contained in $\mathbb{U}_x(A^c)$. For any small $\varepsilon > 0$, there exists $u \in S_{T(x)+\varepsilon}$ such that $T(x) \leq \sup_{\theta \in A} T(a(\theta, u))$; for this u, $\Theta_x(u) \not\subseteq A^c$, so $u \notin \mathbb{U}_x(A^c)$. We have verified (4.2), so

$$\begin{aligned} \mathsf{pl}_x(A; \mathcal{S}) &= 1 - \mathsf{P}_{\mathcal{S}}\{\mathcal{S} \subseteq S_{T(x)}\} \\ &= 1 - \inf_{\theta \in A} \mathsf{P}_{X|\theta}\{T(X) < T(x)\} \quad \text{[by (4.1)]} \\ &= \sup_{\theta \in A} \mathsf{P}_{X|\theta}\{T(X) \geq T(x)\}. \end{aligned}$$

The right-hand side is $\mathsf{pval}(x)$ in (2.1), completing the proof.

Corollary 1. *Under* A1, *if* $\Theta_0 = \{\theta_0\}$, *then the conclusion of Theorem* 2 *holds.*

Proof. Conditions A2–A3 hold automatically for singleton Θ_0 and suitable T.

4.2. Remarks

Dempster (2008, p.375) points out a similar connection between plausibility and p-value; specifically, he shows numerically how Fisher's p-value can be decomposed into two pieces—one piece corresponding to belief in H_0 and the other corresponding to "don't know"—the sum of which is our plausibility. His example is for the standard test for a Poisson mean based on a one-sided alternative hypothesis, and he claims no such a correspondence in general.

In the Bayesian setting, a search for "objective" priors often focuses on probability matching (e.g., Ghosh (2011)), that is, choose the prior such that the corresponding posterior tail probabilities and p-values are asymptotically equivalent. Given the connection between p-values and IM plausibilities, these objective Bayes posterior probabilities can also be interpreted as plausibilities. This is perhaps not surprising given that objective Bayes posterior distributions can be viewed as a simple and attractive way to approximate frequentist p-values (Fraser (2011)).

This connection between p-values and plausibilities also casts light on the argument in Schervish (1996) concerning the use of p-values as measures of evidence; see, also, Berger and Sellke (1987). He shows that, in general, p-values fail to satisfy that, for a given x, if $\Theta_0' \subseteq \Theta_0$, then the p-value for Θ_0' should be no more than the p-value for Θ_0. Theorem 2 explains this lack of coherence in that p-values for different hypotheses may be plausibilities with respect to different IMs with different scales. The same is true for Bayesian posterior probabilities for Θ_0' and Θ_0 if different priors are used for each testing problem, which would not necessarily be unusual.

In case $\Theta_x(u) = \varnothing$ for some pair (x, u), constructing an IM with plausibility function matching the p-value cannot be done as described in the proof of Theorem 2. Such a situation arises, for example, in a normal mean problem $\mathsf{N}(\theta, 1)$ with $\Theta = [-1, 1]$. If $X = -1$ is observed, then $\Theta_{-1}(u) = \{\theta \in [-1, 1] : -1 = \theta + \Phi^{-1}(u)\}$ is empty for $u > 1/2$. For such problems, Ermini Leaf and Liu (2012) present a modification of the IM approach that stretches the predictive random set just enough so that $\Theta_x(\mathcal{S})$ is not empty while maintaining validity. The result of this stretching is, in general, an increase in the plausibility function. The p-value depends only on the null hypothesis, so is not affected by parameter constraints. This is a shortcoming of the p-value, as evidence for a particular assertion should automatically become larger when the range of possible alternatives shrinks.

4.3. Binomial example

Consider a binomial model, $X \sim \mathsf{Bin}(n, \theta)$, where n is a known positive integer and $\theta \in (0, 1)$ is the unknown success probability. Inference on θ in the binomial model is a basic problem that is far from trivial (e.g., Brown, Cai, and DasGupta (2001)). In this case, the natural association is

$$F_\theta(X - 1) \leq U < F_\theta(X), \quad U \sim \mathsf{Unif}(0, 1),$$

where F_θ is the $\mathsf{Bin}(n, \theta)$ distribution function. There is no simple equation linking (X, θ, U) in this case, just a rule "$X = a(\theta, U)$" for producing X with given θ and U. We construct the p-value-based IM for a one-sided assertion/hypothesis.

Consider $A = (0, \theta_0]$ for some fixed $\theta_0 \in (0,1)$. If the null hypothesis is $H_0 : \theta \in A$, then the uniformly most powerful test rejects H_0 in favor of $H_1 : \theta \in A^c$ if and only if $T(X) = X$ is too large. With this choice of T, for the A-step, we write

$$\Theta_x(u) = \{\theta : T(x) = T(a(\theta, u))\} = \{\theta : F_\theta(x-1) \leq u < F_\theta(x)\}.$$

If $G_{a,b}$ denotes the $\mathsf{Beta}(a,b)$ distribution function, then we may rewrite $\Theta_x(u)$ as

$$\begin{aligned}\Theta_x(u) &= \{\theta : G_{n-x+1,x}(1-\theta) \leq u < G_{n-x,x+1}(1-\theta)\} \\ &= \{\theta : 1 - G^{-1}_{n-x+1,x}(u) \leq \theta < 1 - G^{-1}_{n-x,x+1}(u)\}.\end{aligned}$$

For the P-step, we construct the support $\mathbb{S} = \{S_t : t \in \mathbb{T}\}$, where, in this case, $\mathbb{T} = \mathbb{X} = \{0, 1, \ldots, n\}$. It is easy to see that

$$S_t = cl\{u : \sup_{\theta \in A} T(a(\theta, u)) < t\} = [F_{\theta_0}(t), 1].$$

When equipped with the measure $\mathsf{P}_{\mathcal{S}}$ in P2, determined by $\mathsf{P}_U = \mathsf{Unif}(0,1)$, the C-step produces a plausibility function for $A = (0, \theta_0]$, at the observed $X = x$, given by

$$\mathsf{pl}_x(A; \mathcal{S}) = 1 - F_{\theta_0}(x),$$

which is exactly the standard p-value for the one-sided test in a binomial problem.

4.4. Normal variance example

Consider a normal model, $\mathsf{N}(\mu, \sigma^2)$, and a sequence of independent samples $X_1, \ldots, X_n$. Here, $\theta = (\mu, \sigma^2)$ is unknown, but the goal is inference on σ^2, with μ a nuisance parameter. Following the general conditioning principles in Martin and Liu (2014b), we can focus on IMs determined by the minimal sufficient statistic,

$$\bar{X} = \mu + \sigma n^{-1/2} Z \quad \text{and} \quad (n-1)S^2 = \sigma^2 W,$$

where $Z \sim \mathsf{N}(0,1)$ and $W \sim \mathsf{ChiSq}(n-1)$, independent. This association involves two auxiliary variables but, since the goal is inference about the scalar σ^2, we can reduce the dimension. Write

$$\bar{X} = \mu + \frac{S}{n^{1/2}} \frac{Z}{\{W/(n-1)\}^{1/2}} \quad \text{and} \quad (n-1)S^2 = \sigma^2 W.$$

Since μ is a location parameter, it follows from Martin and Liu (2013b) that the first expression displayed above can be ignored, leaving the second as the marginal association for σ^2, which we now write as

$$T = \sigma^2 F^{-1}(U), \quad U \sim \mathsf{Unif}(0,1),$$

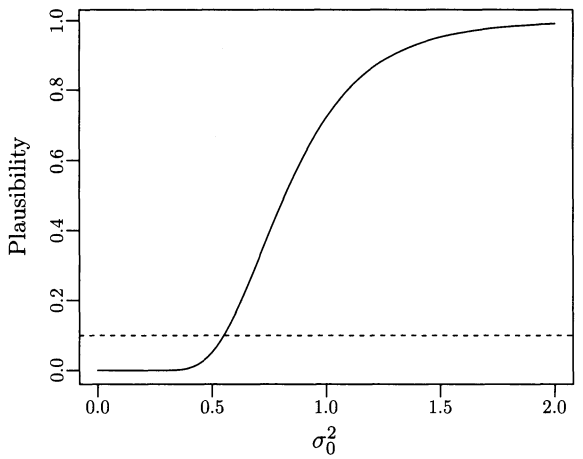

Figure 1. Plot of plausibility as a function of σ_0^2 in the normal variance example.

where $T = (n-1)S^2$ and F is the $\mathsf{ChiSq}(n-1)$ distribution function.

Consider testing $H_0 : \sigma^2 \leq \sigma_0^2$ versus $H_1 : \sigma^2 > \sigma_0^2$. For observed $T = t$, the standard test has p-value $\mathsf{pval}(t) = \mathsf{P}\{T \geq t\} = 1 - F(t/\sigma_0^2)$. It is straightforward to check that, with predictive random set $\mathcal{S} = [0, U)$, with $U \sim \mathsf{Unif}(0,1)$, the plausibility function is

$$\mathsf{pl}_t(\{\sigma^2 \leq \sigma_0^2\}; \mathcal{S}) = \mathsf{P}_{\mathcal{S}}\Big\{\mathcal{S} \ni F\Big(\frac{t}{\sigma^2}\Big)\Big\} = \mathsf{P}\Big\{U \geq F\Big(\frac{t}{\sigma_0^2}\Big)\Big\} = 1 - F\Big(\frac{t}{\sigma_0^2}\Big),$$

which is exactly the p-value. Moreover, the predictive random set above is "optimal" in the sense of Martin and Liu (2013a, Sec. 4.3.1), which provides some IM-based explanation for this test being the standard one in the statistics literature.

For an illustration, consider data presented in Problem 2-14 of Montgomery (2001) on the etch uniformity on silicon wafers taken during a qualification experiment. In this case, the sample size is $n = 20$ and the sample variance is $S^2 = 0.79$. If $\sigma_0^2 = 1$ so the goal is testing if $\sigma^2 \leq 1$, the p-value is 0.72, and the null hypothesis is quite plausible. More generally, we can plot the plausibility (or p-value) as a function of σ_0^2; see Figure 1. The horizontal line at $\alpha = 0.1$ characterizes a 90% plausibility lower bound for σ^2 defined by keeping all those σ_0^2 values with plausibility greater than 0.1; see (3.5).

5. Discussion

We have developed a new user-friendly interpretation of the familiar but often misinterpreted p-value. Specifically, we have shown that, for essentially

any hypothesis testing problem, under mild conditions, there exists a valid IM such that its plausibility function, evaluated at the null hypothesis, is exactly the usual p-value. This representation of p-values in terms of IM plausibilities casts light on a potential shortcoming of p-values that can arise in problems with non-trivial parameter constraints. In such cases, it is not clear how to modify the p-value, while modifications of the IM plausibility are readily obtained via the methods described in Ermini Leaf and Liu (2012).

There are a numerous alternatives to p-value in the hypothesis testing literature, popular, at least in part, because of the difficulties in interpreting p-values. For example, Jim Berger (and co-authors) have recommended converting p-values to Bayes factors, or posterior odds, for interpretation; for example, Sellke, Bayarri, and Berger (2001) make a strong case for their suggested "$-ep \log p$" adjustment. However, it is unlikely that p-values will ever disappear from textbooks and applied work, so compared to offering an alternative to the familiar p-value, it may be more valuable to offer a more user-friendly interpretation. To borrow an analogy Larry Wasserman used on his blog: many people are poor drivers, but eliminating cars is not the answer to this problem.

The connection between plausibility and p-value casts light on the nature of the IM output. IM belief and plausibility functions are understood in Martin and Liu (2013a) as measures of evidence given data. The fact that, in some cases, plausibility and p-value match up is useful, suggesting that one could reason with IM plausibilities as with p-values. The correspondence between plausibilities, p-values, and some objective Bayes posterior probabilities, suggests that the IM framework may in fact provide a unified perspective on robust, objective, probabilistic inference.

Acknowledgement

The authors are grateful for helpful suggestions from the Editor, an Associate Editor, and two referees. This work is partially supported by the U.S. National Science Foundation, grants DMS–1007678, DMS–1208833, and DMS–1208841.

References

Berger, J. (2006). The case for objective Bayesian analysis. *Bayesian Anal.* **1**, 385-402.

Berger, J. O., Bernardo, J. M. and Sun, D. (2009). The formal definition of reference priors. *Ann. Statist.* **37**, 905-938.

Berger, J. O. and Delampady, M. (1987). Testing precise hypotheses. *Statist. Sci.* **2**, 317-352. With comments and a rejoinder by the authors.

Berger, J. O. and Sellke, T. (1987). Testing a point null hypothesis: irreconcilability of P values and evidence. *J. Amer. Statist. Assoc.* **82**, 112-139.

Brown, L. D., Cai, T. T. and DasGupta, A. (2001). Interval estimation for a binomial proportion (with discussion). *Statist. Sci.* **16**, 101-133.

Chiang, A. K. L. (2001). A simple general method for constructing confidence intervals for functions of variance components. *Technometrics* **43**, 356-367.

Dempster, A. P. (2008). Dempster–Shafer calculus for statisticians. *Internat. J. Approx. Reason.* **48**, 265-277.

Ermini Leaf, D. and Liu, C. (2012). Inference about constrained parameters using the elastic belief method. *Internat. J. Approx. Reason.* **53**, 709-727.

Fidler, F., Thomason, N., Cummings, G., Fineh, S. and Leeman, J. (2004). Editors can lead researchers to confidence intervals, but can't make them think. *Psychol. Sci.* **15**, 119-126.

Fisher, R. A. (1959). *Statistical Methods and Scientific Inference.* Second edition, revised. Hafner Publishing Company, New York.

Fraser, D. A. S. (1968). *The Structure of Inference.* Wiley, New York.

Fraser, D. A. S. (2011). Is Bayes posterior just quick and dirty confidence? *Statist. Sci.* **26**, 299-316.

Fraser, D. A. S., Reid, N., Marras, E. and Yi, G. Y. (2010). Default priors for Bayesian and frequentist inference. *J. Roy. Statist. Soc. Ser. B* **72**, 631-654.

Gelman, A., Meng, X.-L. and Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statist. Sinica* **6**, 733-807. With comments and a rejoinder by the authors.

Ghosh, M. (2011). Objective priors: an introduction for frequentists. *Statist. Sci.* **26**, 187-202.

Hannig, J. (2009). On generalized fiducial inference. *Statist. Sinica* **19**, 491-544.

Hannig, J. (2013). Generalized fiducial inference via discretization. *Statist. Sinica* **23**, 489-514.

Hannig, J. and Lee, T. C. M. (2009). Generalized fiducial inference for wavelet regression. *Biometrika* **96**, 847-860.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *J. Amer. Statist. Assoc.* **90**, 773-795.

Lehmann, E. L. and Romano, J. P. (2005). *Testing Statistical Hypotheses.* Third edition. Springer, New York.

Martin, R. (2014). Random sets and exact confidence regions. *Sankhyā*, to appear; `arXiv:1302.2023`.

Martin, R. and Liu, C. (2013a). Inferential models: A framework for prior-free posterior probabilistic inference. *J. Amer. Statist. Assoc.* **108**, 301-313.

Martin, R. and Liu, C. (2013b). Marginal inferential models: prior-free probabilistic inference on interest parameters. Unpublished manuscript, `arXiv:1306.3092`.

Martin, R. and Liu, C. (2014a). Comment: Foundations of statistical inference, revisited. *Statist. Sci.*, to appear; `arXiv:1312.7183`.

Martin, R. and Liu, C. (2014b). Conditional inferential models: combining information for prior-free probabilistic inference. *J. Roy. Statist. Soc. Ser. B*, to appear; `arXiv:1211.1530`.

Martin, R., Zhang, J. and Liu, C. (2010). Dempster–Shafer theory and statistical inference with weak beliefs. *Statist. Sci.* **25**, 72-87.

Montgomery, D. C. (2001). *Design and Analysis of Experiments.* Fifth edition. Wiley, Hoboken, NJ.

Overbye, D. (2012). Physicists find elusive particle seen as key to universe. *The New York Times*, July 5:A1. `http://www.nytimes.com/2012/07/05/science/cern-physicists-may-have-discovered-higgs-boson-particle.html`.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Ann. Statist.* **12**, 1151-1172.

Schervish, M. J. (1996). P values: what they are and what they are not. *Amer. Statist.* **50**, 203-206.

Sellke, T., Bayarri, M. J., and Berger, J. O. (2001). Calibration of p values for testing precise null hypotheses. *Amer. Statist.* **55**, 62-71.

Shafer, G. (1976). *A Mathematical Theory of Evidence.* Princeton University Press, Princeton, N.J.

Shafer, G. (2011). A betting interpretation for probabilities and Dempster–Shafer degrees of belief. *Internat. J. Approx. Reason.* **52**, 127-136.

Shafer, G. and Vovk, V. (2006). The sources of Kolmogorov's *grundbegriffe. Statist. Sci.* **21**, 70-98.

Tsui, K.-W. and Weerahandi, S. (1989). Generalized p-values in significance testing of hypotheses in the presence of nuisance parameters. *J. Amer. Statist. Assoc.* **84**, 602-607.

Weerahandi, S. (1993). Generalized confidence intervals. *J. Amer. Statist. Assoc.* **88**, 899-905.

Xie, M. and Singh, K. (2013). Confidence distribution, the frequentist distribution of a parameter – a review. *Internat. Statist. Rev.* **81**, 3-39.

Xie, M., Singh, K. and Strawderman, W. E. (2011). Confidence distributions and a unifying framework for meta-analysis. *J. Amer. Statist. Assoc.* **106**, 320-333.

Zhang, J. and Liu, C. (2011). Dempster–Shafer inference with weak beliefs. *Statist. Sinica* **21**, 475-494.

Department of Mathematics, Statistics, and Computer Science, University of Illinois at Chicago, 851 S. Morgan St., Chicago, IL 60607, USA.

E-mail: rgmartin@uic.edu

Department of Statistics, Purdue University, 250 N. University St., West Lafayette, IN 47907, USA.

E-mail: chuanhai@purdue.edu

(Received April 2013; accepted December 2013)