Communications in Statistics—Theory and Methods, 41: 1478–1496, 2012

ISSN: 0361-0926 print/1532-415X online
DOI: 10.1080/03610926.2010.543302

Taylor & Francis
Taylor & Francis Group

Coherent Frequentism: A Decision Theory Based on Confidence Sets

DAVID R. BICKEL

Ottawa Institute of Systems Biology, Department of Biochemistry, Microbiology, and Immunology and Department of Mathematics and Statistics, University of Ottawa, Ottawa, Ontario, Canada

By representing fair betting odds according to one or more pairs of confidence set estimators, dual parameter distributions called confidence posteriors secure the coherence of actions without any prior distribution. This theory reduces to the maximization of expected utility when the pair of posteriors is induced by an exact or approximate confidence set estimator or when a reduction rule is applied to the pair. Unlike the p-value, the confidence posterior probability of an interval hypothesis is suitable as an estimator of the indicator of hypothesis truth since it converges to 1 if the hypothesis is true or to 0 otherwise.

Keywords Coherence; Coherent prevision; Confidence distribution; Decision theory; Fiducial inference; Foundations of statistics; Imprecise probability; Maximum utility; Minimum expected loss; Observed confidence level; Probability matching priors; Problem of regions; Significance testing; Upper and lower probability; Utility maximization.

Mathematics Subject Classification 62A01; 62F03; 62F15; 62F40.

1. Introduction

1.1. *Motivation*

A well-known mistake in the interpretation of an observed confidence interval confuses *confidence* as a level of certainty with "confidence" as the *coverage rate*, the almost-sure limiting rate at which a confidence interval would cover a parameter value over repeated sampling from the same population. This results in using the stated confidence level, say 95%, as if it were a probability that the parameter value lies in the particular confidence interval that corresponds to the observed sample. A practical solution that does not sacrifice the 95% coverage rate is to

Received July 9, 2010; Accepted November 22, 2010
Address correspondence to David R. Bickel, Ottawa Institute of Systems Biology, Department of Biochemistry, Microbiology, and Immunology, Department of Mathematics and Statistics, University of Ottawa, 451 Smyth Road, Ottawa, Ontario, K1H 8M5, Canada; E-mail: dbickel@uottawa.ca

report a confidence interval that matches a 95% *credibility interval* computable from Bayes's formula given some *matching prior* distribution (Rubin, 1984). In addition to precluding the error in interpretation, such matching enables the statistician to leverage the flexibility of the Bayesian approach in making self-consistent inferences, involving, for example, the probability that the parameter lies in any given region of the parameter space, on the basis of a posterior distribution firmly anchored to valid coverage rates. Priors yielding exact matching of predictive probabilities are available for many models, including location models and certain location-scale models (Datta et al., 2000; Severini et al., 2002). Although exact matching of fixed-parameter coverage rates is limited to location models (Welch and Peers, 1963; Fraser and Reid, 2002), priors yielding asymptotic matching have been identified for other models, e.g., a hierarchical normal model (Datta et al., 2000). For mixture models, all priors that achieve matching to second order necessarily depend on the data but asymptotically converge to fixed priors (Wasserman, 2000). Data-based priors can also yield second-order matching with insensitivity to the sampling distribution (Sweeting, 2001). Agreeably, Fraser et al. (2010) suggested a data-dependent prior for approximating the likelihood function integrated over the nuisance parameters to attain accurate matching between Bayesian probabilities and coverage rates. These advances approach the vision of building an objective Bayesianism, defined as a "universal recipe for applying Bayes theorem in the absence of prior information" (Efron, 1998).

Viewed from another angle, the fact that close matching can require resorting to priors that change with each new observation, cracking the foundations of Bayesian inference, raises the question of whether many of the goals motivating the search for an objective posterior can be achieved apart from Bayes's formula. It will, in fact, be seen that such a probability distribution lies dormant in nested confidence intervals, securing the above benefits of interpretation and coherence without matching priors, provided that the confidence intervals are constructed to yield reasonable inferences about the value of the parameter for each sample from the available information.

Except for confidence intervals that are conservative by construction, the condition of adequately incorporating any relevant information is usually satisfied in practice since confidence intervals are most appropriate when information about the parameter value is either largely absent or included in the interval estimation procedure, as it is in random-effects modeling and various other frequentist shrinkage methods. Likewise, confidence intervals known to lead to pathologies tend to be avoided. (Pathological confidence intervals often emphasized in support of credibility intervals include formally valid confidence intervals that lie outside the appropriate parameter space [Mandelkern, 2002] and those that can fail to ascribe 100% confidence to an interval deduced from the data to contain the true value [Bernardo and Smith, 1994].)

A game-theoretic framework makes the requirement more precise: for the 95% confidence interval to give a 95% degree of certainty in the single case and to support coherent inferences, it must be generated to ensure that, on the available information, 19:1 are approximately fair betting odds that the parameter lies in the observed interval. This condition rules out the use of highly conservative intervals, pathological intervals, and intervals that fail to reflect substantial pertinent information. In relying on a realized confidence interval to that extent, the decision maker ignores the presence of any recognizable subsets (Gleser, 2002), not only slightly conservative subsets, as in the tradition of controlling the rate of Type I

errors (Casella, 1987), but also slightly anti-conservative subsets. Given the ubiquity of recognizable subsets (Buehler and Feddersen, 1963; Bondar, 1977), this strategy uses pre-data confidence as an approximation to post-data confidence in the sense in which expected Fisher information approximates observed Fisher information (Efron and Hinkley, 1978), aiming not at exact inference but at a pragmatic use of the limited resources available for any particular data analysis. Certain situations may instead call for careful applications of conditional inference (Goutis and Casella, 1995; Sundberg, 2003; Fraser, 2004) or of minimum description length (Bickel, 2011b) for basing decisions more directly on the data actually observed.

1.2. *Direct Inference and Observed Confidence*

The above betting interpretation of a confidence posterior will be generalized in a framework of decision to formalize, control, and extend the common practice of equating the level of certainty that a parameter lies in an observed confidence interval with the interval estimator's rate of coverage over repeated sampling. (See [Shafer, 2010] for a review and extension of an alternative betting interpretation of probability.)

Many who fully understand that the 95% confidence interval is defined to achieve a 95% coverage rate over repeated sampling will for that reason often be substantially more certain that the true value of the parameter lies in an observed 99% confidence interval than that it lies in a 50% confidence interval computed from the same data (Franklin, 2001; Pawitan, 2001, pp. 11–12). This *direct inference*, reasoning from the frequency of individuals of a population that have a certain property to a level of certainty about whether a particular sample from the population, is a notable feature of inductive logic (e.g., Franklin, 2001; Jaeger, 2005) and often proves effective in everyday decisions. Knowing that the new cars of a certain model and year have speedometer readings within 1 mile per hour (mph) of the actual speed in 99.5% of cases, most drivers will, when betting on whether they comply with speed limits, have a high level of certainty that the speedometer readings of their particular new cars of that model and year accurately report their current speed in the absence of other relevant information. (Such information might include a reading of 10 mph when the car is stationary, which would indicate a defect in the instrument at hand.) If the betting interpretation holds for an interval given by some predetermined level of confidence, then coherence requires that it hold equally for a level of confidence given by some predetermined hypothesis.

Fisher's fiducial argument also employed direct inference (Fisher, 1945; Fisher, 1973, pp. 34–36, 57–58; Hacking, 1965, Chapter 9; Zabell, 1992). The present framework will depart from his in its applicability without exact confidence sets, in the closer proximity of its probabilities to repeated-sampling rates of covering vector parameters, in its toleration of reference classes with relevant subsets, and in its theory of decision. Since the second and third departures are shared with recent methods of computing the confidence probability of an arbitrary hypothesis, the main contribution of this article is the general framework of inference that both motivates such methods given an exact confidence set and extends them for use with approximate, valid, and non conservative set estimators and for coherent decision making, including prediction and point estimation.

This framework draws from the theory of coherent upper and lower probabilities for the cases in which no exact confidence set with the desired

properties is available. To allow indecision in light of inconclusive evidence, these non additive probabilities have been formulated for lotteries in which the agent may either place a bet or refrain from betting or, equivalently, in which the casino posts different odds to be used depending on whether a gambler bets for or against a hypothesis. Confidence decision theory will be formulated for this scenario by setting an agent's prices of buying and selling a gamble on the hypothesis that a parameter θ is in some set $\Theta' \subset \Theta$ according to the confidence levels of a valid set estimate and a nonconservative confidence set estimate that coincide with Θ'. As a result, the hypothesis that $\theta \in \Theta'$ has an interval of confidence levels rather than a single confidence level, that is, an interval of confidence posterior probabilities rather than a single confidence posterior probability. Equating the buying and selling prices reduces the upper and lower probability functions to a single confidence posterior, a probability measure on parameter space Θ, and thus reduces the interval to a point.

1.3. *Overview*

This subsection outlines the organization of the remainder of the article while offering a brief summary.

After preliminary concepts are defined (§2.1), Sec. 2.2 presents the new framework for confidence-based inference and decision. The family of probability measures (confidence posteriors) used in inference and decision can be stated in terms of coherent lower and upper probabilities and is thus completely self-consistent according to a widely accepted account of coherence derived from ideas of Bruno de Finetti (Sec. 2.3). This lays a foundation for decisions and for flexible inference about the truth of hypotheses without invoking the likelihood principle (Secs. 2.4, 2.5). The framework is compared to other versions of frequentist coherence based on upper and lower probabilities in Sec. 2.5.

While reporting an interval level of confidence in a hypothesis has the advantage of honestly communicating the inability of the data to determine a single confidence level, such intervals are less useful in situations requiring the automation of decisions. Under such circumstances, the family of confidence posteriors can be reduced to a single confidence posterior P^x by the use of exact or approximate confidence sets or by an automatic reduction rule (Sec. 3.1). The important special case of a scalar parameter of interest provides an arena for contrasting confidence posterior probabilities and p-values (Sec. 3.2). As will be seen, the observed confidence level $P^x(\vartheta \in \Theta')$ can differ markedly from the p-value for testing $\theta \in \Theta'$ as the null hypothesis not only in interpretation but also in numeric value.

Section 4 concludes the article by highlighting the main properties of the proposed framework.

2. Confidence Decision Theory

2.1. *Preliminaries*

2.1.1. *Basic Notation.* The values of $x \wedge y$ and $x \vee y$ are respectively the minimum and maximum of x and y. The symbols $\subseteq$ and $\subset$ respectively signify subset and proper subset. $1_{\Theta'} : \Theta \to \{0, 1\}$ is the usual indicator function: $1_{\Theta'}(\theta)$ is 1 if $\theta \in \Theta'$ and 0 if $\theta \notin \Theta'$. $|\Theta|$ is the number of members in Θ.

Angular brackets rather than parentheses signal numeric tuples. For example, if x and y are numbers, then $\langle x, y\rangle$ denotes an ordered pair, whereas (x, y) denotes the open interval $\{z : x < z < y\}$.

Given a probability space $\left(\Omega, \Sigma, P_\xi\right)$ indexed by the vector parameter $\xi \in \Xi \subseteq \mathbb{R}^d$, consider the random quantity X of distribution P_ξ and with a realization x in some sample set $\Omega \subseteq \mathbb{R}^n$. For notational convenience, partition the full parameter ξ into an interest parameter $\theta \in \Theta$ and, unless $\theta = \xi$, a nuisance parameter $\gamma \in \Gamma$, such that $\xi \in \Theta \times \Gamma$ and $P_{\theta,\gamma} = P_\xi$.

Except where otherwise noted, every probability distribution is a standard (Kolmogorov) probability measure. A *strictly incomplete* probability measure is a standard, additive measure with total mass less than 1.

Let $(\Theta, \mathscr{A})$ represent a measurable space and $\mathscr{B}([0, 1])$ the Borel σ-field of $[0, 1]$. The complement and power set of Θ' are $\overline{\Theta}'$ and $2^{\Theta'}$, respectively. The σ-field induced by $\mathscr{C}$ is $\sigma(\mathscr{C})$.

2.1.2. *Multimeasure and Multiprobability Spaces.* The following slight extension of probability theory, based on a multimeasure (Precupanu, 2008), facilitates a clear and precise presentation of the present framework. To prevent unnecessary confusion between single-valued probability and the specific type of multi-valued probability required, the former will be called *probability* in agreement with common usage, and the latter will be called *multiprobability*, a term defined below.

Definition 2.1. Given a measurable space $(\Theta, \mathscr{A})$ and a *multimeasure space*, the triple $\mathscr{M} = (\Theta, \mathscr{A}, \mathfrak{P})$ with a family $\mathfrak{P}$ of measures, the *multimeasure* $\mathscr{P}$ of $\mathscr{M}$ is a function $\mathscr{P}$ from $\mathscr{A}$ to the set of all closed intervals of $[0, \infty)$ such that $\mathscr{P}(A)$ is the convex hull of

$$\{P(A) : P \in \mathfrak{P}\}$$

for each $A \in \mathscr{A}$. The multimeasure $\mathscr{P}$ is said to be *degenerate* if $|\mathfrak{P}| = 1$ or *non degenerate* if $|\mathfrak{P}| > 1$.

Definition 2.2. The multimeasure $\mathscr{P}$ of a multimeasure space $\mathscr{M} = (\Theta, \mathscr{A}, \mathfrak{P})$ is a *probability multimeasure* if each member of $\mathfrak{P}$ is a probability measure. Then $\mathscr{M}$ is a *multiprobability space*, and $\mathscr{P}(A)$ is the multiprobability of event A for all $A \in \mathscr{A}$. The *expectation interval* $\mathscr{E}(L)$ of a measurable map $L : \mathscr{A} \to \mathbb{R}$ with respect to a probability multimeasure $\mathscr{P}$ on $\mathscr{M}$ is the convex hull of

$$\left\{\int L(\vartheta)\, dP(\vartheta) : P \in \mathfrak{P}\right\}.$$

Thus, the expectation interval of a scalar random variable with respect to a probability multimeasure is the smallest closed interval containing the expectation values of the random quantity with respect to the probability measures of the multiprobability space.

2.2. *Confidence Measures and Multimeasures*

Particular types of confidence sets form the basis of the multimeasure on which confidence decision theory rests.

Definition 2.3. A *set estimator* $\widehat{\Theta}$ for θ is a function defined on $\Omega \times \mathscr{R}$ for some nonempty interval $\mathscr{R} \subseteq [0, 1]$. A set estimator is called *valid* if its coverage rate over repeated sampling is at least as great as ρ, the nominal confidence coefficient:

$$P_\xi\left(\theta \in \widehat{\Theta}(X; \rho)\right) \geq \rho$$

for all $\xi \in \Xi$ and $\rho \in \mathscr{R}$. A set estimator is called *non conservative* if its coverage rate over repeated sampling is at no greater than the nominal confidence coefficient:

$$P_\xi\left(\theta \in \widehat{\Theta}(X; \rho)\right) \leq \rho$$

for all $\xi \in \Xi$ and $\rho \in \mathscr{R}$. A set estimator that is both valid and nonconservative is called *exact*. For some set $\mathscr{C}$ of connected subsets of Θ, a set estimator is called *nested* if it is a function $\widehat{\Theta} : \Omega \times \mathscr{R} \to \mathscr{C}$ such that, for all $x \in \Omega$, there is a $\mathscr{C}(x) \subseteq \mathscr{C}$ such that $\widehat{\Theta}(x; \bullet) : \mathscr{R} \to \mathscr{C}(x)$ is bijective, $\widehat{\Theta}(x; 0) = \emptyset$, $\widehat{\Theta}(x; 1) = \Theta$, and

$$\widehat{\Theta}(x; \rho_1) \subseteq \widehat{\Theta}(x; \rho_2) \tag{1}$$

for all $\rho_1, \rho_2 \in \mathscr{R}$ such that $\rho_1 \leq \rho_2$. Two nested set estimators $\widehat{\Theta}_1 : \Omega \times \mathscr{R} \to \mathscr{C}$ and $\widehat{\Theta}_2 : \Omega \times \mathscr{R} \to \mathscr{C}$ are *dual* if the ranges $\mathscr{C}_1(x)$ and $\mathscr{C}_2(x)$ of $\widehat{\Theta}_1(x; \bullet)$ and $\widehat{\Theta}_2(x; \bullet)$ induce the same σ-field, i.e., $\sigma(\mathscr{C}_1(x)) = \sigma(\mathscr{C}_2(x))$, for each $x \in \Omega$.

The desired multimeasure will be constructed from two confidence measures in turn constructed from dual nested set estimators.

Definition 2.4. Let $\widehat{\Theta} : \Omega \times \mathscr{R} \to \mathscr{C}$ denote a nested set estimator and $\mathscr{A}^x$ the σ-field induced by $\mathscr{C}(x)$, the range of $\widehat{\Theta}(x; \bullet)$ for each $x \in \Omega$. Then, for all $x \in \Omega$, $\widehat{\Theta}$ induces the probability space $(\Theta, \mathscr{A}^x, P^x)$ and the *confidence measure* or *confidence posterior* P^x, the probability measure on $\mathscr{A}^x$ such that

$$\Theta' \in \mathscr{C}(x) \implies \Theta' = \widehat{\Theta}(x; P^x(\Theta')). \tag{2}$$

The probability $P^x(\Theta')$ is the *confidence level* of the hypothesis that $\theta \in \Theta'$. If $\widehat{\Theta}$ is valid, nonconservative, or exact, then P^x and $P^x(\Theta')$ are likewise called valid, non conservative, or exact, respectively.

Definition 2.5. Consider the dual nested set estimators $\Theta_{\geq} : \Omega \times \mathscr{R} \to \mathscr{C}$, which is valid, and$\Theta_{\leq} : \Omega \times \mathscr{R} \to \mathscr{C}$, which is nonconservative. For every $x \in \Omega$, let $\mathscr{A}^x$ denote the common σ-field induced by each of the ranges of $\widehat{\Theta}_{\geq}(x; \bullet)$ and $\widehat{\Theta}_{\leq}(x; \bullet)$. If $P^x_{\geq}$ is the *valid confidence measure*, the confidence measure induced by $\Theta_{\geq}$, then $P^x_{\geq}(\Theta')$ is called a *valid confidence level* of the hypothesis that $\theta \in \Theta'$. For each $x \in \Omega$, the dual *non conservative confidence measure* $P^x_{\leq}$ and *non conservative confidence level* $P^x_{\leq}(\Theta')$ are defined analogously. On the multiprobability space

$$\mathscr{M}^x_{\geq,\leq} = \left(\Theta, \mathscr{A}^x, \left\{P^x_{\geq}, P^x_{\leq}\right\}\right), \tag{3}$$

called a *confidence multimeasure space*, the probability multimeasure $\mathscr{P}^x$ is called the *confidence multimeasure induced by* $\widehat{\Theta}_{\geq}$ *and* $\widehat{\Theta}_{\leq}$ *given some* x in Ω. Accordingly, the

confidence multilevel of the hypothesis that $\theta \in \Theta'$ is $\mathscr{P}^x(\Theta')$ for all $\Theta' \in \mathscr{A}^x$. By the definition of multiprobability, any hypothesis $\Theta' \in \mathscr{A}^x$ has a confidence multilevel of

$$\mathscr{P}^x(\Theta') = \left[P^x_{\geq}(\Theta') \wedge P^x_{\leq}(\Theta'), P^x_{\geq}(\Theta') \vee P^x_{\leq}(\Theta')\right]. \tag{4}$$

Various standard criteria used to judge confidence intervals (Bickel, 2010) provide guidance on the choice of a dual set estimator for inducing the confidence multimeasure.

Example 2.1 (Normal Distribution)**.** For n independent random variables each distributed according to $P_{\theta,\gamma}$, the normal distribution with mean θ and variance γ, the interval estimator Θ^α given by

$$\Theta^\alpha(x; \rho) = \left[p_x^{-1}(\alpha), p_x^{-1}(\rho + \alpha)\right]$$

for all $\rho \in [0, 1 - \alpha]$ is nested and is an exact $\rho(100\%)$ confidence interval for θ, where $\alpha \in [0, 1]$, $p_x(\theta')$ is the upper-tailed p-value of the hypothesis that $\theta = \theta'$, and p_x^{-1} is the inverse of p_x. Since Θ^α is both valid and non conservative, it is dual to itself, yielding the equality of the valid and non conservative confidence measures $P^x_{\alpha,\geq}$ and $P^x_{\alpha,\leq}$, each the distribution of

$$\vartheta = \bar{x} + T_{n-1}\hat{\sigma}/\sqrt{n},$$

where T_{n-1} is the random variable of the Student t distribution with $n - 1$ degrees of freedom. Hence, the confidence multimeasure $\mathscr{P}^x_\alpha$ induced by Θ^α is degenerate:

$$\left(\Theta, \mathscr{A}^x, \left\{P^x_{\alpha,\geq}, P^x_{\alpha,\leq}\right\}\right) = \left(\Theta, \mathscr{A}^x, \{P^x_\alpha\}\right)$$

If Θ' is an interval, then

$$P^x_\alpha(\Theta') = p_x(\sup \Theta') - p_x(\inf \Theta')$$

for all $x \in \Omega$ and $\Theta' \in \mathscr{A}^x$, from which it follows that the confidence measure P^x_α does not depend on the nested set estimator chosen and can thus be represented by P^x.

Special properties of degenerate confidence multimeasures are given in Section 3. The next example involves a nondegenerate confidence multimeasure.

Example 2.2 (Binomial Distribution)**.** Let P_θ denote the binomial measure with n trials, success probability $\theta \in \Theta$, and upper-tailed cumulative probabilities $p_{C,x}(\theta) = P_\theta(X > x) + CP_\theta(X = x)$ with $C \in [0, 1]$ as the correction factor usually set at $C = 1/2$. Consider the family $\mathscr{F}_C = \{\Theta^\alpha_C : \alpha \in (0, 1]\}$ of nested set estimators such that

$$\Theta^\alpha_C(x; \rho) = \begin{cases} \left[p^{-1}_{1-C,x}(\alpha), p^{-1}_{C,x}(\alpha + \rho)\right] & \rho \in (0, 1 - \alpha] \\ \emptyset & \rho = 0 \\ [0, 1] & \rho = 1 \end{cases}$$

for all $\alpha \in (0, 1]$, $\rho \in \Re = [0, 1 - \alpha] \cup \{1\}$, $x \in \{0, 1, \dots\} = \Omega$, where

$$p_{C',x}^{-1}(\alpha') = \theta' \iff p_{C',x}(\theta') = \alpha'. \tag{5}$$

Since the rates at which valid ($C = 0$) and nonconservative ($C = 1$) interval estimators cover θ are bound according to

$$P_\theta(\theta \in \Theta_0^\alpha(X; \rho)) \geq \rho,$$
$$P_\theta(\theta \in \Theta_1^\alpha(X; \rho)) \leq \rho,$$

the sets $\mathscr{F}_0$ and $\mathscr{F}_1$ are valid and non conservative families of nested set estimators, respectively, and for any $\alpha \in [0, 1]$, the valid set estimator Θ_0^α is dual to the non conservative set estimator Θ_1^α, thus inducing the valid confidence measure $P_{\alpha,0}^x$, the nonconservative confidence measure $P_{\alpha,1}^x$, and the confidence multimeasure $\mathscr{P}_\alpha^x$ on the σ-field $\mathscr{B}([0, 1])$ for each $x \in \Omega$. In order to weigh evidence in $X = x$ for the hypothesis that $0 \leq \theta' \leq \theta \leq \theta'' \leq 1$, Eq. (2) furnishes

$$P_{\alpha,C}^x\left(\left[p_{1-C,x}^{-1}(\alpha), p_{C,x}^{-1}\left(\alpha + \rho_{C,x}\right)\right]\right) = \rho_{C,x},$$

which in turn yields

$$\begin{aligned} P_{\alpha,C}^x([\theta', \theta'']) &= P_{\alpha,C}^x\left(\left[p_{1-C,x}^{-1}(\alpha), p_{C,x}^{-1}\left(\alpha + \rho_{C,x}''\right)\right]\right) - P_{\alpha,C}^x\left(\left[p_{1-C,x}^{-1}(\alpha), p_{C,x}^{-1}\left(\alpha + \rho_{C,x}'\right)\right]\right) \\ &= \rho_{C,x}'' - \rho_{C,x}', \end{aligned} \tag{6}$$

where

$$\rho_{C,x}' = p_{C,x}(\theta') - \alpha$$
$$\rho_{C,x}'' = p_{C,x}(\theta'') - \alpha.$$

Since α drops out of the difference, let $P_C^x = P_{\alpha,C}^x$. For any $\Theta' \in \mathscr{B}([0, 1])$, Eqs. (6) and (4) specify the confidence multilevel of the hypothesis that $\theta \in \Theta'$, e.g., if $P_{\alpha,0}^x([0, \theta'']) \leq P_{\alpha,1}^x([0, \theta''])$ and $1 < x < n$, then

$$\begin{aligned} \mathscr{P}^x([0, \theta'']) &= \left[p_{0,x}(\theta'') - p_{0,x}(\theta'), p_{1,x}(\theta'') - p_{1,x}(\theta')\right] \\ &= \left[P_{\theta''}(X > x), P_{\theta''}(X \geq x)\right], \end{aligned}$$

from which it follows that the generalized fiducial distributions of Hannig (2009, Example 2.1) are stochastically bounded by P_0^x and P_1^x. (The Poisson-distribution example of Dempster (2008) indicates that the extrema of the confidence multilevel equal the "belief" and "plausibility" values of Dempster-Shafer theory in the case of a scalar-parameter family of discrete probability distributions.) To illustrate the reduction of confidence indeterminacy with additional observations, the boundary values of $\mathscr{P}^x([1/4, 3/4])$ are plotted against n in Fig. 1 for the $\theta = 2/3$ case.

Remark 2.1. The restriction to σ-fields with events common to valid and non conservative confidence measures strongly constrains the choice of the estimators to ensure the ability to assign a confidence multilevel to any hypothesis of interest

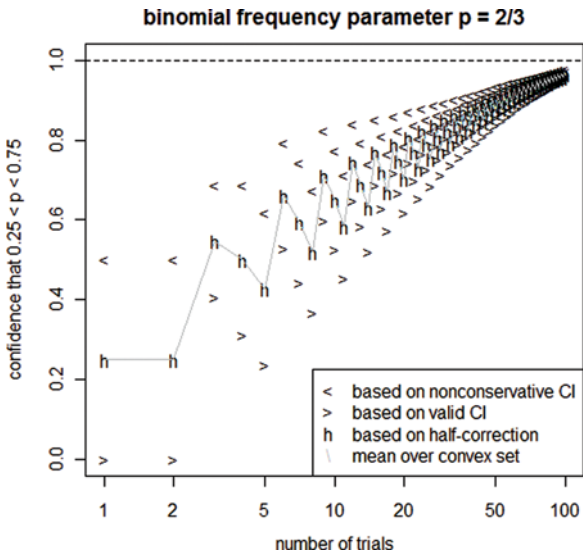

Figure 1. Confidence levels of the hypothesis that θ, the limiting relative frequency of successes, is between $1/4$ and $3/4$ as a function of n, the number of independent trials, with $\theta = 2/3$ as the unknown true value. In the notation of Example 2.2, the non conservative confidence level is $P_1^x([1/4, 3/4])$, the valid confidence level is $P_0^x([1/4, 3/4])$, and the half-corrected level is $P_{1/2}^x([1/4, 3/4])$. The confidence level averaged over the convex set is defined in Sec. 3.1. Sampling variation was suppressed by setting each number x of successes to the lowest integer greater than or equal to $n\theta$ instead of randomly drawing values of x from the $\langle n, \theta \rangle$ binomial distribution. (color figure available online.)

without a need for strictly incomplete probability measures. For further hypothesis flexibility, the σ-field $\mathscr{A}^x$ may be expanded via the replacement of the valid confidence measure in Eq. (3) by a probability distribution extending multiple confidence measures each derived from a different valid set estimator, in which case the nonconservative confidence measure would be replaced by an extension of multiple nonconservative confidence measures dual to the valid confidence measures. The valid confidence measures must be compatible with each other in the sense that the intersections of their events can have defined probabilities, and likewise for the nonconservative confidence measures. Extensions of this type are often necessary for the generation of Borel σ-fields of multi-dimensional parameter spaces. For instance, Polansky (2007, p. 24) derived conference measures for vector parameters of interest by means of a shape parameter indexing different asymptotically exact set estimators; α in Example 2.2 is a simple shape parameter.

2.3. *Coherence of Confidence Multilevels*

The confidence multimeasure $\mathscr{P}^x$ on confidence space $\mathscr{M}^x_{\geq,\leq}$ models the reasoning process of an ideal agent betting on inclusion of the true parameter value in elements of $\mathscr{A}^x$, the σ-field of $\mathscr{M}^x_{\geq,\leq}$, with upper and lower betting odds determined by the coverage rates of the corresponding valid and non conservative confidence sets. The coherence of the agent's decisions may be evaluated by expressing its betting odds in terms of upper and lower probabilities that lack the additivity property of

Kolmogorov's probability measures. Given the dual functions $u : \mathscr{A}^x \to [0, 1]$ and $v : \mathscr{A}^x \to [0, 1]$ such that

$$\begin{aligned} u(\Theta') + v(\Theta \backslash \Theta') &= 1, \\ u(\Theta' \cup \Theta'') &\geq u(\Theta') + u(\Theta''), \\ v(\Theta' \cup \Theta'') &\leq v(\Theta') + v(\Theta'') \end{aligned} \tag{7}$$

for all disjoint Θ' and Θ'' in $\mathscr{A}^x$, the values $u(\Theta')$ and $v(\Theta')$ are the *lower* and *upper probabilities* (Molchanov, 2005, Sec. 9.3) of the hypothesis that $\theta \in \Theta'$. The decision-theoretic interpretation is that $u(\Theta')$ is the largest price an agent would pay for a gain of $1_\theta(\Theta')$, whereas $v(\Theta')$ is the smallest price for which the same agent would sell that gain, assuming an additive utility function (Walley, 1991). The duality between u and v expressed as Eq. (7) means each function is completely determined by the other.

The function u is called the *lower envelope* of a family $\mathfrak{P}$ of measures on $\mathscr{A}$ if

$$u(\Theta') = \inf_{P \in \mathfrak{P}} P(\Theta')$$

for all $\Theta' \in \mathscr{A}$ (Coletti and Scozzafava, 2002, Sec. 15.2; Molchanov, 2005, Sec. 9.3). Since the lower envelope of a family of probability measures is a *coherent lower probability* (Walley, 1991, Sec. 3.3.3; Molchanov, 2005, Sec. 9.3) and since $\{P^x_{\geq}, P^x_{\leq}\}$ as specified in Definition 2.5 constitutes such a family, the agent weighing evidence for any hypothesis $\theta \in \Theta'$ by $\mathscr{P}^x(\Theta')$, with $\Theta' \in \mathscr{A}^x$, satisfies the minimal set of rationality axioms of Walley (1991). It follows that the agent avoids sure loss by making decisions according to the lower and upper probabilities

$$\begin{aligned} u(\Theta') &= P^x_{\geq}(\Theta') \wedge P^x_{\leq}(\Theta'), \\ v(\Theta') &= 1 - u(\Theta \backslash \Theta'). \end{aligned}$$

Conversely, the framework of Sec. 2.2 can be presented starting with de Finetti's prevision and the related concept of coherent extension (Walley, 1991; Coletti and Scozzafava, 2002) as follows. An intelligent agent first sets its prices for buying and selling gambles on the hypotheses corresponding to the elements of $\mathscr{C}$ according to the confidence coefficients of valid and nonconservative nested set estimators. Then it extends its prices or *previsions* to the family of the two probability measures on the σ-field induced by $\mathscr{C}$ in order to evaluate the probability of a hypothesis $\theta \in \Theta'$ for some Θ' in the σ-field but not in $\mathscr{C}$. This family in turn yields coherent lower and upper probabilities that equal the initial buying and selling prices whenever the latter apply, i.e., when the hypothesis is that $\theta \in \Theta'$ for some $\Theta' \in \mathscr{C}$. In this situation, a Dutch book cannot be made against the agent, i.e., a betting opponent cannot formulate a betting strategy such that the agent will suffer net loss regardless of the truth values of the hypotheses on which the agent and opponent place bets against each other (Bickel, 2010).

2.4. *Decisions Under Arbitrary Loss*

This section generalizes betting under 0–1 loss to making confidence-based decisions under any unbounded loss function. Confidence multilevels do not describe the

actual betting behavior of any human agent, but instead prescribe decisions, including amounts bet on any hypothesis involving θ, given that the agent will incur a loss of $L_a(\theta)$ for taking action a.

According to a natural generalization of the Bayes decision rule of minimizing loss averaged over a posterior distribution, action a' *dominates* (is rationally preferred to) action a'' if and only if

$$\forall \epsilon' \in \mathscr{E}(L_{a'}), \quad \epsilon'' \in \mathscr{E}(L_{a''}) : \epsilon' \leq \epsilon''$$
$$\exists \epsilon' \in \mathscr{E}(L_{a'}), \quad \epsilon'' \in \mathscr{E}(L_{a''}) : \epsilon' < \epsilon'',$$

where both expectation intervals (Definition 2.2) are with respect to the same confidence multimeasure $\mathscr{P}^x$. The confidence multimeasures impose no restrictions on agent decisions other than restricting them to non dominated actions.

This use of the confidence multimeasure in making decisions follows a previous generalization of maximizing expected utility to multi-valued probability. (Here, the utilities are expressed in terms of equivalent losses, as is conventional in the statistics literature.) Kyburg (1990, pp. 180, 231–234; 2003, 2006) and Kaplan (1996, Sec. 1.4) used the principle of dominance to make decisions on the basis of intervals of expected utilities determined by the expected utility of each probability measure: an action yielding expected utilities in interval A is preferred to that yielding expected utilities in interval B if at least one member of A is greater than all members of B and if no member of A is less than any member of B.

While multi-valued probabilities do not dictate how to choose one of the non dominated actions in situations that demand a choice equivalent to deciding between accepting a hypothesis or accepting its alternative, they may prove more practical when indecision can be broken by additional considerations, as Walley (1991, pp. 161–162, 235–241) explained. In the case of a human agent, Kyburg (2003) argued for selecting among non dominated actions on the basis of considerations that cannot be represented mathematically rather than selecting on the basis of an arbitrary prior distribution.

If a single-valued estimate of $1_{\Theta'}(\theta)$ is needed for some $\Theta' \in \mathscr{A}$, the *indeterminacy* $\sup \mathscr{P}^x(\Theta') - \inf \mathscr{P}^x(\Theta')$ can quantify a set estimator's degree of undesirable conservatism; some ways to eliminate such indeterminacy by replacing a confidence multimeasure with a confidence measure are mentioned in Sec. 3. If indeterminacy is removed, the above dominance principle reduces to the principle of minimizing expected loss; see Bickel (2009, 2010, 2011a).

2.5. *Likelihood Principle*

While in some cases the likelihood function can guide the construction of set estimators with desirable properties (Bickel, 2010), it plays no general role in confidence decision theory. Consequently, inference does not always obey the likelihood principle: some set estimators lead to values of evidential support and partial proof that depend on information in the sampling model not encoded in the likelihood function; cf. Wilkinson (1977).

An advantage of coherent statistical methods, in general, is the flexibility they give the researcher to simultaneously consider as many hypotheses and interval estimates for θ as desired. Although such versatility is usually presented as a consequence of the likelihood principle and Bayesian statistics, they are not needed to secure it once coherence has been established (Sec. 2.3). In the proposed

framework, the presence of multiple comparisons affects data analyses largely through the use of non additive loss functions (Bickel, 2011a).

That the proposed framework is not constrained by the likelihood principle distinguishes it from Peter Walley's W_1 and W_2, two inferential theories of indeterminate (multi-valued) probability intended to satisfy the best aspects of both coherence and frequentism (Walley, 2002). The coverage error rate of W_1 tends to be much higher than the nominal rate in order to ensure simultaneous compliance with the likelihood principle. Although the principle often precludes approximately correct frequentist coverage, more power can be achieved by less stringently controlling the error rate (Walley, 2002). Walley (2002) did not report the degree of conservatism of W_2, a normalized likelihood method. With a uniform measure for integration over parameter space, the normalized likelihood is equal to the Bayesian posterior that results from a uniform prior.

3. Confidence Posterior Distribution

An important realm for practical applications of the above framework is the situation in which inference may reasonably depend only on a single confidence measure P^x rather than directly on a confidence multimeasure $\mathscr{P}^x$. That is possible not only in the special case of degeneracy due to the availability of a suitable exact nested set estimator (Example 2.1), but can also be achieved either by transforming a nondegenerate confidence multimeasure to a confidence measure (Sec. 3.1) or by approximating a confidence measure (see Remark 3.1 below). In the ubiquitous special case of a scalar parameter of interest, a single confidence level of a hypothesis is a consistent estimator of whether the hypothesis is true under more general conditions than is the *p*-value as such an estimator (§3.2).

3.1. *Reducing a Confidence Multimeasure*

Interpreting upper and lower probabilities as bounds defining a family of permissible probability measures, Williamson (2007) argued for minimizing expected loss with respect to a single distribution within the family instead of using outside considerations to choose among actions that are non-dominated in the sense of Sec. 2.4. Consider the confidence multimeasure space $\mathscr{M}^x_{\geq,\leq} = \left(\Theta, \mathscr{A}^x, \left\{P^x_{\geq}, P^x_{\leq}\right\}\right)$ of confidence multimeasure $\mathscr{P}^x$ for some $x \in \Omega$. A much larger family $\mathfrak{P}$ of measures on $\mathscr{A}^x$ such that $\left\{P^x_{\geq}, P^x_{\leq}\right\}$ and $\mathfrak{P}$ have the same lower envelope u is the convex set

$$\mathfrak{P} = \left\{P^x_D : D \in [0, 1]\right\},$$

where $P^x_D = (1 - D)\, P^x_{\geq} + DP^x_{\leq}$, thereby forming the multiprobability space $\widetilde{\mathscr{M}}^x_{\geq,\leq} = (\Theta, \mathscr{A}^x, \mathfrak{P})$ and probability multimeasure $\widetilde{\mathscr{P}}^x$; cf. Smith (1961, Sec. 11), Wasserman (1990), and Paris (1994, pp. 40–42). The measure $P^x \in \mathfrak{P}$ selected according to some rule is called a *reduction* of $\widetilde{\mathscr{P}}^x$, and, by extension a reduction of $\mathscr{P}^x$.

Effective reduction of $\mathscr{P}^x$ to a single measure P^x can be accomplished by averaging over $\mathfrak{P}$ with respect to the Lebesgue measure. That average of the convex set is simply the mean of the valid and non conservative confidence measures:

$$P^x(\Theta') = \int_0^1 P^x_D(\Theta')\, dD = \left(P^x_{\geq}(\Theta') + P^x_{\leq}(\Theta')\right)/2 = P^x_{1/2}(\Theta') \tag{8}$$

for all $\Theta' \in \mathscr{A}^x$; recall that $P^x_{1/2} \in \mathfrak{P}$.

Other automatic methods of reducing a multimeasure to a single measure are also available. For example, the recommendation of Williamson (2007) to select the measure within the family that maximizes the entropy is minimax under convexity and Kullback–Leibler loss (Grünwald and Philip Dawid, 2004).

Example 3.1 (Binomial Distribution, Continued from Example 2.2)**.** As the gray line in Fig. 1 indicates, the mean measure P^x of the convex set (8) yields a confidence level between those of the valid and non conservative confidence measures, discarding the notable reduction in confidence non degeneracy from $n = 1$ to $n = 10$ as irrelevant for action in situations that do not permit indecision. This P^x is equivalent to the generalized fiducial distribution of Hannig (2009, Example 2.1, Choice 5). The approximate (half-corrected) confidence level also disregards non degeneracy information, yielding in this special case the same levels of confidence as does P^x. In contrast, the confidence multimeasure records the nondegeneracy as the difference between the agent's selling and buying prices of a gamble with a payoff contingent on whether or not $\theta \in [1/4, 3/4]$, a difference that becomes less important as n increases.

A confidence measure, whether reduced or derived from (approximately) exact confidence sets, minimizes expected loss according to Sec. 2.4. Thus, it generates optimal point estimators and optimal predictors in the same way as does a Bayesian posterior distribution (Bickel, 2010).

3.2. *Scalar Subparameter Case*

The equality between tail probabilities of confidence measures and p-values will be used to prove the consistency of estimating composite hypothesis truth, a property that holds under more general conditions for a confidence level than for a p-value.

3.2.1. *Confidence CDF as the p-Value Function.* If decisions are based on a single confidence measure of a scalar parameter of interest, then the cumulative distribution function of that measure is an upper-tailed p-value function.

Definition 3.1. Consider a function $p^+ : \Omega \times \Theta \to [0, 1]$ such that $p^+(x, \bullet) = p_x^+(\bullet)$ is a CDF for all $x \in \Omega$ and such that

$$P_\xi\left(p_X^+(\theta) < \alpha\right) = \alpha \tag{9}$$

for all $\theta \in \Theta$, $\xi \in \Xi$, and $\alpha \in [0, 1]$. Then, for any $x \in \Omega$, the map $p_x^+ : \Theta \to [0, 1]$ is called an *upper-tailed* p-value function for θ. Likewise, $p_x^- : \Theta \to [0, 1]$ is called a *lower-tailed* p-value function if

$$p_x^-(\theta) = 1 - p_x^+(\theta) \tag{10}$$

for all $\theta \in \Theta$ and for all $x \in \Omega$.

Uniformly distributed under the simple null hypothesis that $\theta = \theta'$, $p_x^-(\theta')$ and $p_x^+(\theta')$ are exact p-values of one-sided tests. Since Eq. (10) is an isomorphism between the two p-value functions, either element of the pair $\langle p_x^-, p_x^+ \rangle$ may be

conveniently designated by $p_X^{\pm}$. The *two-sided* p-value of the null hypothesis that θ is in a central region Θ' of Θ is

$$p_x(\Theta') = 2 \sup_{\theta' \in \Theta'} p_x^-(\theta') \wedge p_x^+(\theta')$$

for all $x \in \Omega$, reducing to the usual $p_x(\Theta') = 2p_x^-(\theta') \wedge p_x^+(\theta')$ for the point hypothesis that $\theta = \theta'$.

While the name *p-value function* used by Fraser (1991) has become standard in the scientific literature, *significance function* is also used in higher-order asymptotics (e.g., Brazzale et al., 2007). Singh et al. (2007) prefered the term *confidence distribution*, which for Efron (1993) and Schweder and Hjort (2002) instead refered to the confidence measure as a Kolmogorov probability distribution, though they do not call confidence levels probabilities of hypothesis truth. (Whereas any p-value function is isomorphic to a unique confidence measure as defined in Sec. 2.2, the p-value function can also be isomorphic to a strictly incomplete probability measure. That Wilkinson (1977) constructed a theory of incoherence based on such a measure underscores the need to sharply distinguish confidence measures from p-value functions.)

By the usual concept of statistical power, the *Type II error rate* of $p^{\pm}$ associated with testing the false null hypothesis that $\theta = \theta'$ at significance level α is $\beta^{\pm}(\alpha, \theta, \theta') = P_{\xi}\left(p_X^{\pm}(\theta') > \alpha\right)$ for any $\theta \gtrless \theta'$. For all $\alpha_1, \alpha_2 \in [0, 1]$ such that $\alpha_1 + \alpha_2 < 1$,

$$P_{\xi}\left(\alpha_1 < p_X^+(\theta) < 1 - \alpha_2\right) = 1 - \alpha_1 - \alpha_2,$$

implying that $\theta_X^+ : [0, 1] \to \Theta$, the inverse function of p_X^+, yields $\left(\theta_X^+(\alpha_1), \theta_X^+(1 - \alpha_2)\right)$ as an exact $100(1 - \alpha_1 - \alpha_2)\%$ confidence interval (Fraser, 1991; Efron, 1993; Schweder and Hjort, 2002; Singh et al., 2007).

Remark 3.1. In many applications, approximate p-value functions replace those that exactly satisfy the definition. For instance, Schweder and Hjort (2002) used a half-corrected p-value function like $p_{C,x}$ of Example 2.2 for discrete data. Other approximations involve parameter distributions with asymptotically correct frequentist coverage, including the asymptotic p-value functions of Singh et al. (2005), the distributions of asymptotic generalized pivotal quantities of Xiong and Mu (2009), some of the generalized fiducial distributions of Hannig (2009), and the Bayesian posteriors of Sec. 1.1. As with frequentist inference in general, asymptotics provide approximations that in many applications prove sufficiently accurate for inference in the absence of exact results (Reid, 2003).

3.2.2. *Confidence Levels Versus p-Values.* Although both confidence levels and p-values can be computed from the same p-value function, the following examples illustrate how they can lead to different inferences and decisions. Section 3.2.3 then demonstrates that the former but not the latter are consistent as estimators of composite hypothesis truth.

Example 3.2 (Point Null Hypothesis)**.** If $P^x(\vartheta < \bullet)$ is continuous on Θ, then $P^x(\theta = \theta') = 0$ for any interior point θ' of Θ. This means that given any alternative hypothesis $\theta \in \Theta'$ such that $P^x(\theta \in \Theta') > 0$, betting on $\theta = \theta'$ vs $\theta \in \Theta'$ at any

finite betting odds will result in expected loss, reflecting the absence of information singling out the point $\theta = \theta'$ as a viable possibility before the data were observed. (By contrast, the usual two-sided p-value is numerically equal to $p_x(\theta')$, which does not necessarily equal the probability of any hypothesis of interest.) If, on the other hand, the parameter value can equal the null hypothesis value for all practical purposes, that fact may be represented by modeling the parameter of interest as a random effect with non zero probability at the null hypothesis value. The latter option would define the confidence measure such that its cdf is a predictive p-value function such as that used by Lawless and Fredette (2005).

Example 3.3 (Beyond Statistical Significance)**.** Consider the null hypothesis $\theta' - \Delta \leq \theta \leq \theta' + \Delta$, where the non negative scalar Δ is a minimal degree of practical or scientific significance in a particular application. For instance, researchers developing methods of analyzing microarray data are increasingly calling for specification of a minimal level of biological significance when testing null hypotheses of equivalent gene expression against alternative hypotheses of differential gene expression (Bickel, 2011c; Bochkina and Richardson, 2007; Lewin et al., 2006; Van De Wiel and Kim, 2007). Bickel (2004) and McCarthy and Smyth (2009) in effect approached the problem with p-values of composite null hypotheses, in conflict with the confidence-posterior approach (Sec. 3.2.3). To circumvent arbitrary selection of Δ, Bickel (2011a) recently reported observed confidence levels of gene underexpression ($\theta < 0$) and gene overexpression ($\theta > 0$) after a correction of the null ($\theta = 0$) distribution for multiple comparisons.

3.2.3. *Consistency of Hypothesis Confidence.* More terminology will be introduced to establish a sense in which the confidence value but not the p-value consistently estimates the hypothesis indicator.

Definition 3.2. An indicator estimator $\hat{1}$ is consistent if, for all $\Theta' \in \mathcal{A}$,

$$\hat{1}_{\Theta'}(X) \xrightarrow{P_{\theta,\gamma}} 1_{\Theta'}(\theta)$$

for every $\gamma \in \Gamma$ and for every θ that is an element of Θ but not of the boundary of Θ'.

By the usual concept of statistical power, the *Type II error rate* of $p^{\pm}$ associated with testing the false null hypothesis that $\theta = \theta'$ at significance level α is $\beta^{\pm}(\alpha, \theta, \theta') = P_{\theta,\gamma}\left(p_X^{\pm}(\theta') > \alpha\right)$ for any $\theta \gtrless \theta'$. Commonly used in two-sided testing, the two-sided p-value of the null hypothesis that $\theta \in \Theta'$ is $p_x(\Theta') = 2\sup_{\theta' \in \Theta'} p_x^-(\theta') \wedge p_x^+(\theta')$ for all $\Theta' \subseteq \Theta$ and $x \in \Omega$.

The next two Propositions contrast the consistency of the confidence value with the inconsistency of the two-sided p-value.

Proposition 3.1. *Assume all one-sided tests represented by the p-value functions* $p^{\pm}$ *are asymptotically powerful in the sense that* $\lim_{n\to\infty}\beta^{\pm}(\alpha, \theta, \theta') = 0$ *for all* $\alpha \in (0, 1)$ *and for all* $\theta, \theta' \in \Theta$ *such that* $\theta \gtrless \theta'$. *The function* $\hat{1} : \mathcal{A} \times \Omega \to [0, 1]$ *is a consistent indicator estimator if* $P^x = \hat{1}_\bullet(x)$ *is a confidence measure corresponding to* $p^{\pm}$ *given* $X = x$ *for all* $x \in \Omega$.

Proof. By the definition of the boundary of a set Θ' as the difference between its closure $\bar{\Theta}'$ and its interior $\text{int}\,\Theta'$, the theorem asserts that, for all $\Theta' \in \mathscr{A}$, θ is either in $\text{int}\,\Theta'$, in which case the theorem asserts $P^X(\Theta') \xrightarrow{P_{\theta,\gamma}} 1$, or θ is in $\Theta \backslash \Theta'$, in which case the theorem asserts $P^X(\Theta') \xrightarrow{P_{\theta,\gamma}} 0$. Let $\mathscr{A}'$ represent the set of all disjoint open interval subsets of Θ'. Then,

$$\begin{aligned} P^X(\Theta') &= P^X\left(\text{int}\,\Theta' \cup \left(\overline{\Theta}' \backslash \text{int}\,\Theta'\right)\right) \\ &= P^X\left(\cup_{\Theta''' \in \mathscr{A}'} \Theta'''\right) + P^X\left(\overline{\Theta}' \backslash \text{int}\,\Theta'\right) \\ &= \sum_{\Theta''' \in \mathscr{A}'} P^X(\Theta''') + 0. \end{aligned}$$

Each term of the sum expands as

$$\begin{aligned} P^X(\Theta''') &= P^X((\inf \Theta''', \sup \Theta''')) \\ &= p_X^+(\sup \Theta''') - p_X^+(\inf \Theta''') \\ &= p_X^-(\inf \Theta''') - p_X^-(\sup \Theta''') \\ &= 1 - p_X^-(\sup \Theta''') - p_X^+(\inf \Theta''')\,. \end{aligned}$$

As the p-value functions are asymptotically powerful, $p_X^{\pm}(\theta') \xrightarrow{P_{\theta,\gamma}} 0$ for all $\alpha \in (0, 1)$ and for all $\theta, \theta' \in \Theta$ such that $\theta \gtrless \theta'$, with the result that each term may be written as a function of p-values that converge in $P_{\theta,\gamma}$ to 0:

$$\begin{aligned} P^X(\Theta''') &= \begin{cases} p_X^-(\inf \Theta''') - p_X^-(\sup \Theta''') & \theta < \inf \Theta''' \\ 1 - p_X^-(\sup \Theta''') - p_X^+(\inf \Theta''') & \theta \in \Theta''' \\ p_X^+(\sup \Theta''') - p_X^+(\inf \Theta''') & \theta > \sup \Theta''' \end{cases} \\ &\xrightarrow{P_{\theta,\gamma}} \begin{cases} 0 - 0 & \theta < \inf \Theta''' \\ 1 - 0 - 0 & \theta \in \Theta''' \\ 0 - 0 & \theta > \sup \Theta''' \end{cases} \end{aligned}$$

for all $\Theta''' \in \mathscr{A}'$. Summing the terms over $\mathscr{A}'$ yields

$$P^X(\Theta') \xrightarrow{P_{\theta,\gamma}} \sum_{\Theta''' \in \mathscr{A}'} 1_{\Theta'''}(\theta) = 1_{\Theta'}(\theta)$$

since $\theta \in \text{int}\,\Theta'$ implies that θ is in one element of $\mathscr{A}'$.

Remark 3.2. Polansky (2007, pp. 37–38) proved a similar proposition of consistency given a smooth distribution $P_{\theta,\gamma}$. A suitably transformed likelihood ratio test statistic is also a consistent indicator estimator under the standard regularity conditions (Bickel, 2011c).

Proposition 3.2. *Under the conditions of Theorem* 3.1, *the two-sided p-value* $p_X(\Theta')$ *is not a consistent indicator estimator.*

Proof. For any $\theta \in \Theta' \in \mathcal{A}$, the distribution of the two-sided p-value $p_X(\Theta')$ converges to the uniform distribution on $[0, 1]$ (Singh et al., 2007), violating consistency (Definition 3.2).

4. Concluding Summary

The confidence multimeasure $\mathcal{P}^x$ and the confidence measure or confidence posterior P^x bring both coherence and consistency to frequentist inference and decision making.

The coherence property established in Sec. 2.3 enables confidence-posterior decisions even in the absence of exact confidence sets. In addition, the reduction to a single confidence measure for inference and decision making shares the coherence of theories of utility maximization usually associated with Bayesianism (Sec. 3). In conclusion, the multilevel or level of confidence in a given hypothesis has the internal coherence of the Bayesian posterior or class of such posteriors without requiring a prior distribution or even an exact confidence set estimator.

More can be said if the parameter of interest is one dimensional, in which case the confidence level of a composite hypothesis is consistent as an estimate of whether that hypothesis is true, whereas neither the Bayesian posterior probability nor the p-value is generally consistent in that sense (Sec. 3.2.3). Specifically, the equality of the confidence level of $\theta \in \Theta'$ to the coverage rate of the corresponding confidence set guarantees convergence in probability to 1 if θ is in the interior of Θ' or to 0 if $\theta \notin \Theta'$ (Proposition 3.1).

Acknowledgments

Peer review furnished many useful comments that led to greater generality and clarity. The author especially thanks an anonymous referee for pointing out a flaw in the set estimator definition and for relating Examples 2.2 and 3.1 to Hannig (2009, Example 2.1). This work was partially supported by the Faculty of Medicine of the University of Ottawa and by Agriculture and Agri-Food Canada.

References

Bernardo, J. M., Smith, A. F. M. (1994). *Bayesian Theory*. New York: John Wiley and Sons.

Bickel, D. R. (2004). Degrees of differential gene expression: Detecting biologically significant expression differences and estimating their magnitudes. *Bioinformatics* 20:682–688.

Bickel, D. R. (2009). A frequentist framework of inductive reasoning. *Technical Report*, Ottawa Institute of Systems Biology, Ottawa, Ontario, Canada. (Available at arXiv:math.st/060237)

Bickel, D. R. (2010). Confidence posterior distributions. *Technical Report*, Ottawa Institute of Systems Biology, Ottawa, Ontario, Canada. (in preparation)

Bickel, D. R. (2011a). Estimating the null distribution to adjust observed confidence levels for genome-scale screening. *Biometrics* 67:363–370.

Bickel, D. R. (2011b). A predictive approach to measuring the strength of statistical evidence for single and multiple comparisons. *Can. J. Stat.* 39:610–631.

Bickel, D. R. (2011c). The strength of statistical evidence for composite hypothesis: Inference to the best explanation. *Statistica Sinica* DOI: 10.5705/ss.2009.125 (online ahead of print)

Bochkina, N., Richardson, S. (2007). Tail posterior probability for inference in pairwise and multiclass gene expression data. *Biometrics* 63(4):1117–1125.
Bondar, J. V. (1977). A conditional confidence principle. *Ann. Statist.* 5(5):881–891.
Brazzale, A. R., Davison, A. C., Reid, N. (2007). *Applied Asymptotics: Case Studies in Small-sample Statistics*. Cambridge: Cambridge University Press.
Buehler, R. J., Feddersen, A. P. (1963). Note on a conditional property of Student's t1. *Ann. Mathemat. Statist.* 34(3):1098–1100.
Casella, G. (1987). Conditionally acceptable recentered set estimators. *Ann. Statist.* 15(4):1363–1371.
Coletti, C., Scozzafava, R. (2002). *Probabilistic Logic in a Coherent Setting*. Amsterdam: Kluwer.
Datta, G. S., Ghosh, M., Mukerjee, R. (2000). Some new results on probability matching priors. *Calcutta Statist. Assoc. Bull.* 50:179–192.
Dempster, A. P. (2008). The dempster-shafer calculus for statisticians. *Int. J. Approx. Reason.* 48(2):365–377.
Efron, B. (1993). Bayes and likelihood calculations from confidence intervals. *Biometrika* 80:3–26.
Efron, B. (1998). R. A. Fisher in the 21st century, invited paper presented at the 1996 R. A. Fisher lecture. *Statist. Sci.* 13(2):95–114.
Efron, B., Hinkley, D. V. (1978). Assessing the accuracy of the maximum likelihood estimator: observed versus expected Fisher information. *Biometrika* 65(3):457–487.
Fisher, R. A. (1945). The logical inversion of the notion of the random variable. *Sankhya Ind. J. Statist. (1933–1960)* 7(2):129–132.
Fisher, R. A. (1973). *Statistical Methods and Scientific Inference*. New York: Hafner Press.
Franklin, J. (2001). Resurrecting logical probability. *Erkenntnis* 55(2):277–305.
Fraser, D. A. S. (1991). Statistical inference: likelihood to significance. *J. Amer. Statist. Assoc.* 86:258–265.
Fraser, D. A. S. (2004). Ancillaries and conditional inference. *Statist. Sci.* 19(2):333–351.
Fraser, D. A. S., Reid, N. (2002). Strong matching of frequentist and Bayesian parametric inference. *J. Statist. Plann. Infer.* 103:263–285.
Fraser, D. A. S., Reid, N., Marras, E., Yi, G. Y. (2010). Default priors for Bayesian and frequentist inference. *J. Roy. Stat. Soc. B* 72:631–654. (comment).
Gleser, L. J. (2002). Setting confidence intervals for bounded parameters. *Comment. Statist. Sci.* 17(2):161–163.
Goutis, C., Casella, G. (1995). Frequentist post-data inference. *Int. Statist. Rev.* 63(3):325–344.
Grünwald, P., Philip Dawid, A. (2004). Game theory, maximum entropy, minimum discrepancy and robust Bayesian decision theory. *Ann. Statist.* 32(4):1367–1433.
Hacking, I. (1965). *Logic of Statistical Inference*. Cambridge: Cambridge University Press.
Hannig, J. (2009). On generalized fiducial inference. *Statistica Sinica* 19:491–544.
Jaeger, M. (2005). A logic for inductive probabilistic reasoning. *Synthese* 144(2):181–248.
Kaplan, M. (1996). *Decision Theory as Philosophy*. Cambridge: Cambridge University Press.
Kyburg, H. E. (2003). Are there degrees of belief? *J. Appl. Logic* 1(3–4):139–149.
Kyburg, H. E. (2006). Belief, evidence, and conditioning. *Philosoph. Sci.* 73(1):42–65.
Kyburg, H. E. J. (1990). *Science and Reason*. New York: Oxford University Press.
Lawless, J. F., Fredette, M. (2005). Frequentist prediction intervals and predictive distributions. *Biometrika* 92(3):529–542.
Lewin, A., Richardson, S., Marshall, C., Glazier, A., Aitman, T. (2006). Bayesian modeling of differential gene expression. *Biometrics* 62(1):1–9.
Mandelkern, M. (2002). Setting confidence intervals for bounded parameters. *Statist. Sci.* 17(2):149–172.
McCarthy, D. J., Smyth, G. K. (2009). Testing significance relative to a fold-change threshold is a TREAT. *Bioinformatics* 25(6):765–771.
Molchanov, I. (2005). *Theory of Random Sets*. New York: Springer.

Paris, J. B. (1994). *The Uncertain Reasoner's Companion: A Mathematical Perspective*. New York: Cambridge University Press.

Pawitan, Y. (2001). *In All Likelihood: Statistical Modeling and Inference Using Likelihood*. Oxford: Clarendon Press.

Polansky, A. M. (2007). *Observed Confidence Levels: Theory and Application*. New York: Chapman and Hall.

Precupanu, A.-M. S. B. (2008). The Aumann-Gould integral. *Mediterr. J. Math.* 5(4):429–441.

Reid, N. (2003). Asymptotics and the theory of inference. *Ann. Statist.* 31(6):1695–1731.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Ann. Statist.* 12(4):1151–1172.

Schweder, T., Hjort, N. L. (2002). Confidence and likelihood. *Scand. J. Statist.* 29(2):309–332.

Severini, T. A., Mukerjee, R., Ghosh, M., (2002). On an exact probability matching property of right-invariant priors. *Biometrika* 89(4):952–957.

Shafer, G. (2010). A betting interpretation for probabilities and dempster-shafer degrees of belief. *Technical Report*, Rutgers University.

Singh, K., Xie, M., Strawderman, W. E. (2005). Combining information from independent sources through confidence distributions. *Ann. Statist.* 33(1):159–183.

Singh, K., Xie, M., Strawderman, W. E. (2007). Confidence distribution (cd) – distribution estimator of a parameter. *IMS Lecture Notes Monograph Series* 54:132–150.

Smith, C. A. B. (1961). Consistency in statistical inference and decision. *J. Roy. Statist. Soc. Ser. B Methodol.* 23(1):1–37.

Sundberg, R. (2003). Conditional statistical inference and quantification of relevance. *J. Roy. Statist. Soc. Ser. B. Statist. Methodol.* 65(1):299–315.

Sweeting, T. J. (2001). Coverage probability bias, objective Bayes and the likelihood principle. *Biometrika* 88(3):657–675.

Van De Wiel, M. A., Kim, K. I. (2007). Estimating the false discovery rate using nonparametric deconvolution. *Biometrics* 63(3):806–815.

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

Walley, P. (2002). Reconciling frequentist properties with the likelihood principle. *J. Statist. Plann. Infer.* 105(1):35–65.

Wasserman, L. (2000). Asymptotic inference for mixture models using data-dependent priors. *J. Roy. Statist. Soc. Ser. B Statist. Methodol.* 62(1):159–180.

Wasserman, L. A. (1990). Prior envelopes based on belief functions. *Ann. Statist.* 18(1):454–464.

Welch, B. L., Peers, H. W. (1963). On formulae for confidence points based on integrals of weighted likelihoods. *J. Roy. Statist. Soc. Ser. B Statist. Methodol.* 25:318–329.

Wilkinson, G. N. (1977). On resolving the controversy in statistical inference (with discussion). *J. Roy. Statist. Soc. Ser. B Methodol.* 39(2):119–171.

Williamson, J. (2007). *Probability and Inference*. Texts in Philosophy 2. London: College Publications, pp. 151–179.

Xiong, S., Mu, W. (2009). On construction of asymptotically correct confidence intervals. *J. Statist. Plann. Infer.* 139(4):1394–1404.

Zabell, S. L. (1992). R. A. Fisher and the fiducial argument. *Statist. Sci.* 7(3):369–387.