On Evidential Measures of Support for Reasoning with Integrated Uncertainty: A Lesson from the Ban of P-values in Statistical Inference

Hung T. Nguyen[1,2(✉)]

[1] Department of Mathematical Sciences, New Mexico State University, Las Cruces, USA
hunguyen@nmsu.edu
[2] Faculty of Economics, Chiang Mai University, Chiang Mai, Thailand

Abstract. In view of the recent ban of the use of P-values in statistical inference, since they are not qualified as information measures of support from empirical evidence, we will not only take a closer look at them, but also embark on a panorama of more promising ingredients which could replace P-values for statistical science as well as for any fields involving reasoning with integrated uncertainty. These ingredients include the recently developed theory of Inferential Models, the emergent Information Theoretic Statistics, and of course Bayesian statistics. The lesson learned from the ban of P-values is emphasized for other types of uncertainty measures, where information measures, their logical aspects (conditional events, probability logic) are examined.

Keywords: Bayesian statistics · Conditional events · Entropy inference procedures · Information measures · Information theoretic statistics · Integrated uncertainty probability logic P-values · Testing hypotheses

1 Introduction

The recent ban on the use of the notion of P-values in hypothesis testing (Trafimow and Marks [32]) triggered a serious reexamination of the way we used to conduct inference in the face of uncertainty. Since statistical uncertainty is an important part of an integrated uncertainty system, a closer look at what went wrong with statistical inference is necessary to "repair" the whole inference machinery in complex systems.

Thus, this paper is organized as follows. We start out, in Sect. 2, by elaborating on the notion of P-values as a testing procedure in null hypothesis signficance testing (NHST). In Sect. 3, within the context of reasoning with uncertainty where logical aspects and information measures are emphasized, we elaborate on why p-values should not be used as an inference procedure anymore. Section 4 addresses the question "What are the items in statistical theory

V.-N. Huynh et al. (Eds.): IUKM 2016, LNAI 9978, pp. 3–15, 2016.
DOI: 10.1007/978-3-319-49046-5_1

which are affected by the removal of P-values?". Section 5 points out alternative inference procedures in a world without P-values. We reserve the last Sect. 6 for a possible "in defense of P-values".

2 The Notion of P-values in Statistical Inference

It seems useful to trace back a bit of Fisher's great achievements in statistical science. The story goes like this. A lady claimed that she can tell whether a cup of tea with milk was mixed with tea or milk first, R. Fisher designed an experiment in which eight cups of mixed tea/milk (four of each kind) was presented to her (letting her know that four cups are mixed with milk first, and the other four are mixed with tea first) in a random sequence, and asked her to taste and tell the order of mixture of all cups. She got all eight correct identifications. How did Fisher arrive at the conclusion that the lady is indeed skillful? See Fisher [10], also Salbursg [29]. This kind of testing problem is termed Null Hypothesis Significance Testing (NHST), due to Fisher [9].

The important question is "Could we use P-values to carry out NHST?". You may ask "what is the rationale for using P-value to make inference?". Well, don't you know the answer? It could be the *Cournot's principle* (see, e.g., Shafer and Vork,[31], pp. 44+), according to which, it is practically certain that predicted events of small probabilities will not occur. But it is just a "principle", not a theorem! It does have some flavor of logic (for reasoning), but which logic? See also Gurevich and Vovk [14] , where two "interesting" things to be noted: First, to carry out a test, one just "adopts" a "convention", namely "for a given test, smaller values provide stronger impugning evidence"! And secondly, it is a fact that "every test statistic is equivalent to a unique exact P-value function".

3 Why P-values Are Banned?

Starting with NHST, the unique way to infer conclusions from data is the traditional notion of P-values. However, there is something fishy about the use of P-values as a "valid" inference procedure, since quite sometimes serious problems with them arised, exemplified by Cohen [6], Schervish [30], Goodman [13], Hurlbert and Lombardi [15], Lavine [16], and Nuzzo [28].

Having relied upon P-value as the inference procedure to carry out NHST (their bread and butter research tool) for so long, the Psychology community finally has enough of its "wrong doings", and without any reactions from the international statistical community (which is responsable for inventing and developing statistical tools for all other sciences to use), decided, on their own, to ban NHST-Procedure (meaning P-values), Trafimow and Marks [32]. While this is a ban only for their Basic and Applied Social Psychology Journal, the impact is worldwide. It is not about the "ban", it is about "what wrong with P-values?" that we should all be concerned. For a flavor of doing wrong statistics, see e.g., Wheelan [34].

Even, the ban gets everybody's attention now, what happend since last year? Nothing! Why? Even after the American Statistical Association issued a "statement" about P-values (ASA News [2]), and Wassertein and Lazar [33], not banning P-values (why not?), but "stating" six "principles".

What do you read and expect from the above "statement"? Some literature search reveals stuff like this. "Together we agreed that the *current culture* of statistical significance testing, interpretation, and reporting *has to go*, and that adherence to a minimum of six principles can help to pave the way forward for science and society". And in the *Sciences News*, for laymen, "P-value ban: small step for a journal, giant leap for sicence". See also, Lavine [16].

There are three theoretical facts which make P-values undesirable for statistical inference:

(i) P-values are not model probabilities.

First, observe that a hypothesis is a statistical model. The P-value $P(T_n \geq t|H_o)$ is the probability of observing of an extreme value t if the null hypothesis is true. It is not $P(H_o|T_n \geq t)$ even when this "model probability given the data" makes sense (e.g., as in the Bayesian framework where H_o is viewed as a random event). Note that when $P(H_o|T_n \geq t)$ makes sense and is available, it is legitimate to use it for model selection (a valid inference procedure from at least a common sense standpoint). In a frequentist framework, there is no way to convert $P(T_n \geq t|H_o)$ to $P(H_o|T_n \geq t)$. As such, the P-value $P(T_n \geq t|H_o)$, alone, is useless for inference, precisely as "stated" in the sixth principle of the ASA.

(ii) The reasoning with P-values is based on an invalid logic.

As mentioned by Cohen [6] and in the previous section, the use of P-values to reject H_o seems to be based on a form of Modus Tollens in logic, since after all, reasoning under uncertainty is inference ! and, each mode of reasoning is based upon a logic. Now, thanks to Artificial Intelligence (AI), we are exposed to a variety of logics, such as probability logic, conditional probability logic, fuzzy logics...which are logics for reasoning under various types of uncertainty. See a text like Goodman, Nguyen and Walker [12]. In particular, we could face rules that have exceptions (see e.g., Bamber, Goodman and Nguyen, [3]). The famous "penguin triangle" in AI can be used to illustrate well the invalidity of Modus Tollens in uncertain logics.

While we focus in this address on reasoning with P-values in probabilistic systems, perhaps few words about reasoning with more complex systems in which several different types of uncertainty are involved (integrated uncertain systems) should be mentioned. To create machines capable of ever more sophisticated tasks, and of exhibiting ever more human-like behavior, we need knowledge representation and associated reasoning (logic). In probabilistic systems, no additional mathematical tools are needed, since we are simply dealing with probability distributions, and the logic used is classical two-valued logic. For general integrated uncertain systems, new mathematical tools such as conditional events, possibility theory, fuzzy logics are needed. See, e,g., Nguyen and Walker [25], Nguyen and Walker [26], Nguyen [27].

(iii) As set-functions, P-values are not information measures of model support.

Schervish [30], while discussing the "usual" use of P-values to test hypotheses (in both NHST and Neyman-Pearson tests), "discovered" that "a common informal use of P-values as measures of support or evidence for hypotheses has serious logical flaws". We will elaborate on his "discovery" in the context of information theory.

Essentially, the reason to use P-values, in the first place, although not stated explicitly as such, to "infer" conclusions from data, is that they seems to be "information measures of location" derived from data (evidence) in support of hypotheses. Is that true? Specifically, Given a null hypothesis H_o and a statistic T_n and the observed value $T_n = t$, the P-value $p(H_o) = P(T_n \geq t|H_o)$, as a function of H_o, for fixed T_n and the observed value $T_n = t$, is "viewed" as a measure of support that the observed value t lends to H_o (or amount of evidence in favor of H_o) since large values of $p(H_o) = P(T_n \geq t|H_o)$ make it harder to reject H_o (whereas, small values reflect non-support for H_o, i.e., rejection). But this "practice" is always informal, and "no theory is ever put forward for what properties a measure of support or evidence should have".

What is an information measure? Information decreases uncertainty. Qualitative information is high if surprise is high. When an event A is realized, it provides an information. Clearly, in the context of "statistical information theory", information is a decreasing function of probability: the smaller the probability for A to occur, the higher the information obtained when A is realized. If A stands for "snowing", then when A occured, say, in Bangkok, it provides a "huge" amount of information $I(A)$. Put it mathematically (as in Information Theory, see e.g., Cover and Thomas, [7], $I(A) = -\log P(A)$. For a general theory of information without probability, but keeping the intuitive behavior that information should be a decreasing function of events, see, e.g., Nguyen [24]. This intuitive behavior is about a specific aspect of the notion of information that we are considering in uncertainty analysis, namely, *information of localization.*

In the context of testing about a parametric model, say, $f(x|\theta), \theta \in \Theta$, each hypothesis H_o can be identified with a subset of Θ, still denoted as $H_o \subseteq \Theta$. An information measure of location on Θ is a set-function $I : 2^\Theta \to \mathbb{R}^+$ such that $A \subseteq B \Longrightarrow I(A) \geq I(B)$. The typical probabilistic information measure is $I(A) = -\log P(A)$. This is the appropriate concept of information measure in support of a subset of Θ (a hypothesis). Now, consider the set function $I(H_o) = P(T_n \geq t|H_o)$ on 2^Θ. Let $H_o' \supseteq H_o$. If we use P-values to reject null hypotheses or not, e.g., rejecting H_o' (i.e., the true $\theta_o \notin H_o'$) when, say, $I(H_o') \leq \alpha = 0.05$, then since $H_o' \supseteq H_o$, we also reject H_o, so that $I(H_o) \leq \alpha$, implying that $H_o' \supseteq H_o \Longrightarrow I(H_o') \geq I(H_o)$ which indicates that $I(.)$ is not an information measure (derived from empirical evidence/ data) in support of hypotheses, since it is an increasing rather than a decreasing set function. *P-values are not measures of strength of evidence.*

4 Are Neyman-Pearson Testing Theory Affected?

So far we have just talked about NHST. How Neyman-Pearson (NP) testing framework differs from NHST? Of course, they are "different", but now, in view of the ban of P-values in NHST, you "love" to know if that ban "affects" your routine testing problems where in teaching and research, in fact, you are using NP tests instead? Clearly the findings are extremely important: either you can continue to proceed with all your familiar (asymptotic) tests such as Z-test, t- test, $\mathcal{X}^2$-test, KS- test, DF- test, ...or... you are facing "the final collapse of the Neyman-Pearson decision theoretic framework "(as announced by Hurlbert and Lombardi [15]! And in the latter (!), are you panic?

In accusing Fisher's work on NHST as "worse than useless", Neyman and Pearson embarked on shaping Fisher's testing setting into a decision framework as we all know and use so far, although science is about discovery of knowledge, and not about decision-making. The "improved" framework is this. Besides a hypothesis, denoted as H_o (although it is not for nullifying, but in fact for acceptance), there is a specified alternative hypothesis H_a (to choose if H_o happens to be rejected). It is a model selection problem, where each hypothesis corresponds to a statistical model. The NP testing is a decision-making problem: using data to reject or accept H_o. By doing so, two types of errors might be committed: false positive: $\alpha = P(\text{reject } H_o | H_o \text{ is true})$, false negative $\beta = P(\text{accept } H_o | H_o$ is false). It "improves" upon Fisher's arbitrary choice of a statistic to compute the P-value to reach a decision, namely a most powerful "test" at a fixed α-level. Note that while the value of α could be the same in both approaches, say 0.05 (a "small" number in $[0, 1]$ for Fisher), its meaning is different, as $\alpha = 5\,\%$ in NP approach (the probability of making the wrong decision of the first kind).

Let's see how NP *carry out* their tests? As a test is usually based on some appropriate statistic $T_n(X_1, X_2, ..., X_n)$ (though technically not required) where, say, $X_1, X_2, ..., X_n$ is a random sample, of size n, drawn from the population, so that we select a set B in the sample space of T_n as a rejection (critical) region.

The most important question is: What is the **rationale** for selecting a set B as a rejection region? Since data (values of the statistics T_n) in B lead to rejection of H_o (i.e., on the basis of elements of B we reject H_o), this *inference procedure* has to have a "plausible" explanation for people to trust! Note that an inference procedure is not a mathematical theorem! In other words, why a data in B provides evidence to reject H_o? Clearly, this has something to do with the statistic $T_n(X_1, X_2, ..., X_n)$!

Given α and a (test) statistic T_n, the rejection region R_α is determined by

$$P(T_n \in R_\alpha | H_o) \leq \alpha$$

Are P-values left out in this determination of rejection regions? (i.e., the rationale of inference in NP tests does not depend on P-values of T_n?). Put it differently: How to "pick" a region R_α to be a rejection region?

Note that the P-value statistic $p(T_n) = 1 - F_{T_n|H_o}(T_n)$ corresponds to a N-P test. Indeed, let α be given as the type-I error. Then, the test, say, S_n, which

rejects H_o if and only if $p(T_n) \leq \alpha$ has $P(\text{rejecting } H_o|H_o) = P(p(T_n) \leq \alpha)$ which will be $\leq \alpha$, if the random variable $p(T_n)$ dominates (first) stochastically the uniform distribution on $[0, 1]$ (so-called a "valid" P-value statistic).

Thus, in summary, the statistics used for significance tests (Fisher) and hypothesis tests (N-P) are the same, and hypothesis tests can be carried out via p-values (as the rationale for rejection), where significance levels (but not p-values) are taken as error probabilities. Thus, in improving Fisher's NHST setting, NP did not "improve" Fisher's intended inference procedure (i.e., P-values), so that both NHST and NP tests share the same (wrong) inference procedure. See Lehmann [17], Lehmann and Romano [18].

In this respect, I cannot resist to put down the following from Freedman, Pisani, and Purves [11], (pp. 562–563):

"Nowadays, tests of significance are extremely popular. One reason is that the tests are part of an impressive and well-developed mathematical theory... This sounds so impressive, and there is so much mathematical machniery clanking around in the background, that tests seem truly scientific-even when they are complete nonsense. St. Exupery understood this kind of problem very well:

"When a mistery is too overpowering, one dare not disobey" (*The Little Prince).*

The basic questions for inference in testing are these. What is the rationale for basing our conclusions (decisions for reject or accept) on a statistic T_n? Are all rejection regions should be determined this way? Based upon which logic (or rationale) we construct rejection regions, from which we reach decisions (i.e., from data to conclusions)?

With all the fancy mathematics to come up with, say, an asymptotic distribution of a statistic T_n (e.g., the Dickey-Fuller test for stationarity in AR(1) model), the goal is to say this. If the observed value of $T_n = t$ is "large", then reject H_o, where "large" is "defined" as the critical value c determined by $P(T_n \geq c|H_o) \leq \alpha$. Don't you see that is clearly equivalent to the P-value of T_n being less than α? Specifically:

$$t \geq c \Leftrightarrow P(T_n \geq t|H_o) \leq \alpha$$

Thus, this kind of inference (or logic!) is exactly the same as what Fisher had suggested for NHST. In other words, all fancy mathematical works aim at providing the necessary stuff for computing p-values! from which to jump to rejection conclusion, just like in NHST.

Freedman, Pisani, and Purves [11] seemed to feel something wrong about the "logic of p-values" /(In fact the "*logic of the z-test*"), pp. 480–481, but did not dare to go all the way to say that it's silly to use it to make inference. They said things like

"It is an argument by contradiction, designed to show that the null hypothesis will lead to an absurd conclusion and must be rejected". Wrong, we are not in binary logic! And

"P is not the chance of the null hypothesis being right", "Even worse, according to the frequentist theory, there is no way to define the probability of the null hypothesis being right".

Remember: if you must choose (or select) two things under uncertainty, you would choose the one with higher "probability" to be right, given your evidence (data). But you do not have it if it does not make sense to consider such a probability (since a hypothesis is not a random event), let alone with p-values. Perhaps, because of this "difficulty" that statisticians "play around" with a seemingly OK logic of p-values? and no one has catched it. Should we continue to use this wrong logic or try to find a better (correct) one? There is no possible choice anymore: the P-value logic is now officially banned!

It seems we cannot "repair" the p-value logic. We must abandon it completely. Clearly, the Bayesian approach to statistical inference does not have this problem. Note that, there are no *names* of tests in Bayesian statistics, since their is only one way to conduct tests (similar to estimation method), which is in fact a selection problem, based on Bayes factors (no test statistics, no sampling distributions!). The crucial thing is that, hypotheses are random events and hence it makes sense to consider "probabilities of hypotheses/ given data" which form a common sense reasoning for reaching conclusions.

However, the following observation seems interesting? Most of NP rejection regions are nested, i.e. of the form $(T_n > c)$, and as such their associated testing procedure is equivalent to using P-values whose threshold is taken as the size of the test. Thus, from a logical point of view, there is no difference between NHST and NP-testing as far as "inference" is concerned. However, there is something interesting here which could "explain" the meaning of NP-rejection regions. While the null hypothesis is rejected by using the "logic of P-values", this inference is controlled by the type-I error α (Noting that there is no such guarantee in NHST framework). In other words, while the inference based on P-value might not be "logical", it could be "plausible": decisions using P-values in NP-testing, say, by NP Theory are enforced by error probabilities. Specifically the use of P-values in NP-testing is in fact carried out *together* with error probabilities, and not *alone*. The implication is that decisions for rejecting hypotheses are controlled with specified error probabilities in advance.

5 Some Alternatives to P-values

Clearly, Bayesian selection (testing) is valid in this sense. The basic "ingredients" are priors and Bayes' theorem which are used precisely to obtain model probabilities for selection purposes. See a Text like Koch (2007).

With respect to "Statistics of the 21st century", David Draper (2009) already claimed that it is Bayesian statistical reasoning, in

Bayesian Statistical Reasoning:
An Inferential, Predictive and Decision-Making Paradigm for the 21st Century
(www.ams.ucsc.edu/~draper)

Well, not so fast! Even *now* with the possible "final collapse of the Neyman-Pearson decision theoretic framework and the rise of the neoFisherian" (Hurlbert and Lombardi [15]) the Bayesian statistics is not the unique "super power",

thanks to the existence (since 1959) of the *non-Bayesian Information Theoretic (IT) Statistics!*

Another approach to statistics which posseses also two similar ingredients, allowing us to reach model probabilities for selection, is the *Information Theoretic (IT)Statistics*. These are "a chosen class of possible models" (playing the role of prior distributions in Bayesian approach, but not subjectively in nature), and the notion of "cross entropy" (playing the role of Bayes' formula). See Kullback [21], Anderson [1], Burnham and Anderson [5], Konishi and Kitagawa [20], Cumming [8], Cover Thomas [7].

Testing of hypotheses is a special case of model selection. The IT approach to model selection is a sort of posterior analysis without subjective priors. What could be considered as priors (a priori) is our choice of a class of competing models for selecting the "best" model for prediction purpose. The Kullback-Liebler divergence (together with its estimate/ the AIC statistic) plays the role of Bayes' formula to arrive at model probabilities which serve as a valid inference for decisions on selections (as opposed to P-values).

Given data $X_1, X_2, ..., X_n$ from a population X, with, of course, unkown probability density $f(.)$, we consider a class of models $\mathcal{M} = \{M_j : j = 1, 2, ..., k\}$ where each $M_j = \{g_j(.|\theta_j) : \theta_j \in \Theta_j\}$, to be possible candidates for approximating $f(.)$.

Somewhat similar to Fisher's idea of finding a statistic to measure the incompatibility between the data and a hypothesis (here, a model), from which we can figure out a way to "reject" it, the IT approach proceeds as follows.

First, in general terms: Suppose we use a density $g(.)$ to approximate an unknown $f(.)$. How to measure the "lost of information?". Well, remember how you answer such a question in your simple linear regression? You use the coefficient of determination as a measure of how much the linear model captures the real variation of the true model, or equivalently, one minus that coefficient as how much information is lost when approximating the true model by a linear model. The IT approach is more general as it addresses directly to the models themselves, by considering a sort of distance between distributions. Note another similarity with P-values! A distance between the true model and an approximate one, measuring the loss of "information", could be used to judge whether the approximating model is reasonable.

As we will see shortly, this idea is much better than P-values, and serves as the fundamentals for valid statistical inference (in the sense that it can provide model probabilities for ranking alternatives, whereas P-values cannot).

Now, given densities f, g, there are many possible ways to define (real) distances between them (just as in functional analysis). We seek a kind of distance which measures a loss of information.

But what is *information*? it is right here that we need a theory of information!

Intuitively, information is a decrease of uncertainty. Uncertainty involves probabilities. Thus, any measure of information should be a function of probability (?). Clearly, an event A gives us less information when its probability $P(A)$ is high: how much information you learn for A="it will be real hot in April in

Chiang Mai next year"? Since $P(A) \simeq 1$, your information is zero (no surprise). How about $B =$"it could be snowing"? Well, big surprise, lot of information: $P(B) \simeq 0$, information is infinity! Thus, the information provided by the realization of an event A is of the form $I(A) = -\log P(A)$, i.e., a non increasing function of probability.

Based upon information theory, via Shannon's entropy, Kullback and Liebler (see Kullback, 1968) considered a "pseudo" distance (relative entropy, or divergence) between two probability density f and g as

$$I(f|g) = \int f(x) \log \frac{f(x)}{g(x)} dx = E[\log \frac{f(X)}{g(X)}]$$

measuring the loss of information when using g to approximate f. For a good explanation of this "loss of information", see Benish [4].

The expectation is of course with respect to the distribution f of the random variable X, so that sometimes we write $I(f|g) = E_f[\log \frac{f(X)}{g(X)}]$ to be explicit. It is a pseudo distance in the sense that $I(f|g) \geq 0$ and $I(f|g) = 0$ if and only if $g(.) = f(.)$.

The purpose to consider such a distance is to compare different $g(.)$ as possible candidates to be used as approximations of $f(.)$. Given, say, an i.i.d. sample $X_1, X_2, ..., X_n$ drawn from X (with true, unknown distribution f), we cannot compute, or even estimate $I(f|g)$, even if $g(.)$ is (completely) known (specified). Indeed,

$$I(f|g) = \int f(x) \log f(x) dx - \int f(x) \log g(x) dx$$

If $g(.)$ is known, then since

$$-\int f(x) \log g(x) dx = -E[\log g(X)]$$

this term can be estimated consistently, for large n, via the strong law of large numbers, by

$$-\frac{1}{n} \sum_{i=1}^{n} \log g(X_i)$$

But, the first term $\int f(x) \log f(x) dx$ is unknown. Fortunately, while it is unknown, it does not involve the candidate g, so that it is the same for any g. Thus, let $\int f(x) \log f(x) dx = C$, the comparison of $I(f|g)$ among various different g, only involves the term $-E[\log g(X)]$, namely

$$I(f|g) - C = -E[\log g(X)]$$

With this meaning of $I(f|g)$, clearly we seek

$$\arg \min_{g \in \mathcal{G}} [-E[\log g(X)]]$$

A problem arises when each $g(.)$ is a statistical model, i.e., $g(.)$ is only specified up to some unknown parameter (possibly vector) $\theta \in \Theta$: $g(.|\theta)$, so that we are facing

$$\min_{g \in \mathcal{G}}[-E[\log g(X|\theta)]]$$

which cannot be estimated anymore since θ is unklnown. We handle this problem by replacing $\theta \in \Theta$ by its MLE

$$\theta_n(X_1, X_2, ..., X_n) = \arg\max_{\theta \in \Theta} L_g(X_1, X_2, ..., X_n|\theta)$$

For i.i.d. sample $L_g(X_1, X_2, ..., X_n|\theta) = \prod_{i=1}^n g(X_i|\theta)$.

Thus, replacing $-E[\log g(X|\theta)]$ for $\theta \in \Theta$ by $-E[\log g(X|\theta_n(X_1, X_2, ..., X_n))]$ which is a random variable (since it depends on the values of the sample $(X_1, X_2, ..., X_n) = Y$). An estimate of it could be simply its mean. So, finally, we are led to estimating

$$E_Y E_X[\log g(X|\theta_n)]$$

from which, a reasonable selection criterion is

$$\max_{g \in \mathcal{G}} E_Y E_X[\log g(X|\theta_n)]$$

It turns out that, for large sample size n (see an appendix), $E_Y E_X[\log g(X|\theta_n)]$ can be estimated by $\log L_g(\theta_n|Y) - k$ where k is the number of estimated parameters (dimension of Θ), resulting in the so-called Akaike Information Criterion (AIC), for model g:

$$AIC(g) = -2 \log L_g(\theta_n|Y) + 2k$$

Note that $L_g(\theta_n|Y)$ is the value of the likelihood function of g evaluated at the MLE estimator θ_n.

Thus, for n large, the selection problem (containg hypothesis testing as a special case) is carried out simply by computing $AIC(g)$ for all $g \in \mathcal{G}$, and pick the one with smallest AIC value.

Let $AIC_{\min}$ be the smallest AIC value, corresponding to some model. Then, let

$$\Delta_g = AIC(g) - AIC_{\min}$$

and make a transformation, e.g., $\Delta_g \to e^{-\frac{\Delta_g}{2}}$ (as suggested by Akaike) to obtain "model likelihood), we arrive at Akaike's weights (of evidence supporting models)

$$w_g = \frac{e^{-\frac{\Delta_g}{2}}}{\sum_{g \in \mathcal{G}} e^{-\frac{\Delta_g}{2}}}$$

which are interpreted as evidence in favor of model g being the best approximating model in the chosen set of models $\mathcal{G}$, viewing as "model probability" given

the data, needed for explaining the rationale in the selection process (where P-values are lacking), noting that

$$w_h = \arg_{g\in\mathcal{G}}\, w_g \Leftrightarrow AIC(h) = \arg\min_{g\in\mathcal{G}} AIC(g)$$

How about "small sample size?". Remember how you consider small sample sizes in your introductory statistical courses? Here, we are talking about small size for AIC approximation to K-L statistic. As a "rule of thumb", the sample size n is considered as small, relative to the number of parameters k in the model g, when $n < 40\,k$. In that case, the AIC is "corrected" to be

$$AIC_c(g) = AIC(g) + \frac{2k(k+1)}{n-k-1}$$

6 Plausibilities in Inferential Models

In discussing this "crisis" in (frequentist) statistical inference with many colleagues, I received "mixed signals". Almost all agreed on the second "principle" of ASA's statement, namely " P-values do not measure the probability that the studied hypothesis is true", but stopped short of saying anything more! Some mentioned that ASA did not ban P-values. What does that means? Does that mean that you still can publish research papers using P-values in statistical journals (but not in psychology journals!)? I don't think so, since the statistical journals' editors have to take into account of public reactions (everybody was aware of the wrong doings/ not just the misuses of P-values from Sciences News), unless they can clarify their actions, scientifically. Almost all did not "feel" that NP testing is affected (and hence will "survive" this crisis) without explaning why, although Gurevich and Vovk [14] proved that "Every test statistic is equivalent to a unique (exact) p-function". Remember also: An inference procedure is valid only if it is based on a firm logical basis. The "future" will settle the matter soon, I guess.

Meanwhile, you could ask "Rather than throwing P-values away, can we find a way to save them so that they can contribute to statistical inference?" I happened to read Martin and Liu [22] in which they proposed a way to save P-values, in the framework of their theory of Inferential Models, Martin and Liu [23]. Below are the essentials.

Let the null hypothesis H_o be identified with a subset Θ_o of the parameter space Θ (in a statistical model $X \sim F_\theta(.), \theta \in \Theta$). For a test statistic T_n, the P-value of $T_n = t$ under Θ_o is extended to $Pv(\Theta_o|t) = \sup_{\theta\in\Theta_o} P(T_n \geq t|\theta)$. While $Pv(\Theta_o|t)$ is not the probability that Θ_o is true when we observe t, it could be equated to another concept of uncertainty, namely "plausibility", to be used as a new inference engine. But what is a *plausibility measure* (rather than a probability measure)? It is the capacity functional of a random set (see Nguyen, [27]). Specifically, if S is a random set with values in 2^Θ, then its capacity functional (or plausibility measure) $Pl_S(.) : 2^\Theta \to [0,1]$ is defined as $Pl_S(A) = P(S \cap A \neq \varnothing)$.

The result of Martin and Liu [22] is this. Any $Pv(.)$ on 2^{Θ} can be written as $Pl_S(.)$ for some random set S on 2^{Θ}. The construction of a suitable random set S is carried out within the Inferential Model framework, Martin and Liu [23]. Note that the computation of plausibilities does not require one to assume that the null hypothesis is true, as opposed to P-values, so that it does make sense to take $Pl_S(\Theta_o|t)$ as the plausibility that Θ_o is true. The intention is to transfer P-values to plausibilities (not to probabilities) and use plausibilities to make inferences. It remains to take a closer look at this proposal, especially with respect to the objections similar to those of P-values.

References

1. Anderson, D.R.: Model Based Inference in the Life Sciences. Springer, Heidelberg (2008)
2. ASA News, American Statistical Association releases statement on statistical significance and p-value, ASA News, March 2016
3. Bamber, D., Goodman, I.R., Nguyen, H.T.: Robust reasoning with rules that have exceptions: from second-order probability to argumentation via upper envelopes of probability and possibility plus directed graphs. Ann. Math. Artif. Intell. **45**, 83–171 (2005)
4. Benish, W.A.: Relative entropy as a measure of diagnostic information. Med. Decis. Making **19**, 202–206 (1999)
5. Burnham, K.P., Anderson, D.R., Selection, M., Inference, M.: A Practical Information Theoretic Approach. Springer, New York (2002)
6. Cohen, J.: The earth is round. Am. Psychlog. **49**(12), 997–1003 (1994)
7. Cover, T.M., Thomas, J.A.: Elements of Information Theory. Wiley, New York (2006)
8. Cumming, G.: Understanding the New Statistics. Routledge, New York (2012)
9. Fisher, R.A.: Statistical Methods for Research Workers. Oliver and Boyd, Edinburgh (1925)
10. Fisher, R.A.: Mathematics of the lady tasting tea, the world of mathematics. In: Newman, J.R. (ed.) (Part VIII): Statistics and the Design of Experiments, vo. III, pp. 1514–1521. Simon and Schuster (1956)
11. Freedman, D., Pisani, R., Purves, R.: Statistics. W.W. Norton, New York (2007)
12. Goodman, I.R., Nguyen, H.T., Walker, E.A., Inference, C.: Logic for Intelligent Systems: A Theory of Measure-Free Conditioning. Hardcover, North-Holland (1991)
13. Goodman, S.: A dirty dozen: twelve p-value misconceptions. Semin. Hematol. **45**, 135–140 (2008)
14. Gurevich, Y., Vovk, V., Fundamentals of P-values: introduction. Bull. Euro. Assoc. Theor. Comput. Sci. (2016, to appear)
15. Hurlbert, S.H., Lombardi, C.M.: The final collapse of the Neyman-Pearson decision theoretic framework and the rise of the neoFisherian. Ann. Zool. Fenn. **46**, 311–349 (2009)
16. Lavine, M.: Comment on Murtaugh. Ecology **93**(5), 642–645 (2014)
17. Lehmann, E.L.: The fisher, Neyman-pearson theories of testing hypotheses: one theory or two? J. Am. Stat. Assoc. **88**(424), 1242–1249 (1993)
18. Lehmann, E.L., Romano, J.P.: Testing Statistical Hypotheses. Springer, New York (2005)

19. Kock, K.R.: Introduction to Bayesian Statistics. Springer, Heidelberg (2007)
20. Konishi, S., Kitagawa, G.: Information Criteria and Statistical Modeling. Springer, New York (2008)
21. Kullback, S.: Information Theory and Statistics. Dover, New York (1968)
22. Martin, R., Liu, C.: A note on P-values interpreted as plausibilities. Stat. Sinica **24**, 1703–1716 (2014)
23. Martin, R., Liu, C.: Inferential Models. Chapman and Hall/CRC Press, Boca Raton (2016)
24. Nguyen, H.T.: Sur les measures d'information de type Inf. In: Nguyen, H.T. (ed.) Theories de l'Information, vol. 398, pp. 62–75. Springer, Heidelberg (1974)
25. Nguyen, H.T., Walker, E.A.: A history and introduction to the algebra of conditional events and probability logic. IEEE Trans. Man Syst. Cybern. **24**(2), 1671–1675 (1996)
26. Nguyen, H.T., Walker, E.A.: A First Course in Fuzzy Logic. Chapman and Hall/CRC Press, Boca Raton (2005)
27. Nguyen, H.T.: An Introduction to Random Sets. Chapman and Hall/CRC Press, Boca Raton (2006)
28. Nuzzo, R.: Statistical errors. Nature **506**, 150–152 (2014)
29. Salbursg, D.: The Lady Teasting Tea. A.W.H Freeman, New York (2001)
30. Schervish, M.J.: P values: what they are and what they are not. Am. Stat. **50**(3), 203–206 (1996)
31. Shafer, G., Vovk, V.: Probability and Finance: It's only a Game. Wiley, New York (2001)
32. Trafimow, D., Marks, E.: Basic and applieds. Soc. Psychol. **37**, 1–2 (2015)
33. Wassertein, R.L., Lazar, N.A.: The ASA's statement on P-value: context, process and purpose. Am. Stat. **70**, 129–133 (2016)
34. Wheelan, C.: Naked Statistics. W.W Norton, New York (2013)