confuse P values or ΔAIC with binary declarations. An argument against one is not necessarily an argument against the other. (3) Be careful interpreting a P value or ΔAIC as strength of evidence. That interpretation cannot be made formal and the connection between P, ΔAIC, and evidence must be recalibrated for each new problem. (4) Plot. Check models. Plot. Check assumptions. Plot.

Literature Cited

Berkson, J. 1942. Tests of significance considered as evidence. Journal of the American Statistical Association 37:325–335.

Lavine, M., and M. J. Schervish. 1999. Bayes factors: What they are and what they are not. American Statistician 53:119–122.

Murtaugh, P. A. 2014. In defence of P values. Ecology 95:611–617.

Schervish, M. J. 1996. P values: What they are and what they are not. American Statistician 50:203–206.

Ecology, 95(3), 2014, pp. 645–651

Recurring controversies about P values and confidence intervals revisited

Aris Spanos[1]

Department of Economics, Virginia Tech, Blacksburg, Virginia 24061 USA

Introduction

The use, abuse, interpretations and reinterpretations of the notion of a P value has been a hot topic of controversy since the 1950s in statistics and several applied fields, including psychology, sociology, ecology, medicine, and economics.

The initial controversy between Fisher's significance testing and the Neyman and Pearson (N-P; 1933) hypothesis testing concerned the extent to which the pre-data Type I error probability α can address the arbitrariness and potential abuse of Fisher's *post-data threshold* for the P value. Fisher adopted a falsificationist stance and viewed the P value as an indicator of disagreement (inconsistency, contradiction) between data $\mathbf{x}_0$ and the null hypothesis (H_0). Indeed, Fisher (1925:80) went as far as to claim that "The actual value of p ... indicates the strength of evidence against the hypothesis." Neyman's behavioristic interpretation of the pre-data Type I and II error probabilities precluded any evidential interpretation for the accept/reject the null (H_0) rules, insisting that accept (reject) H_0 does not connote the truth (falsity) of H_0. The last exchange between these protagonists (Fisher 1955, Pearson 1955, Neyman 1956) did nothing to shed light on these issues. By the early 1960s, it was clear that neither account of frequentist testing provided an adequate answer to the question (Mayo 1996): When do data $\mathbf{x}_0$ provide evidence for or against a hypothesis H?

Manuscript received 4 July 2013; revised 15 August 2013; accepted 16 August 2013. Corresponding Editor: A. M. Ellison. For reprints of this Forum, see footnote 1, p. 609.

[1] E-mail: aris@vt.edu

The primary aim of this paper is to revisit several charges, interpretations, and comparisons of the P value with other procedures as they relate to their primary aims and objectives, the nature of the questions posed to the data, and the nature of their underlying reasoning and the ensuing inferences. The idea is to shed light on some of these issues using the *error-statistical* perspective; see Mayo and Spanos (2011).

Frequentist Testing and Error Probabilities

In an attempt to minimize technicalities but be precise about the concepts needed, the discussion will focus on the hypotheses

$$H_0\colon \mu = \mu_0 \text{ vs. } H_1\colon \mu > \mu_0 \tag{1}$$

in the context of *the simple Normal model* $X_t \sim \text{NIID}(\mu, \sigma^2)$, $t = 1, 2, \ldots, n, \ldots$, where NIID stands for normal, independent, and identically distributed.

Fisher vs. Neyman-Pearson (N-P) approaches

In the case of the above null hypothesis, Fisher's significance and the Neyman-Pearson (N-P) hypothesis testing revolve around the test statistic

$$\tau(\mathbf{X}) = \frac{\sqrt{n}(\bar{X}_n - \mu_0)}{s} \overset{H_0}{\sim} \text{St}(n-1) \tag{2}$$

where $\text{St}(n-1)$ denotes a Student's t distribution with $(n-1)$ degrees of freedom, and

$$\bar{X}_n = \frac{1}{n}\sum_{t=1}^{n} X_t, \qquad s^2 = \frac{1}{n-1}\sum_{t=1}^{n}(X_t - \bar{X}_n)^2.$$

Fisher's significance testing ignores the alternative hypothesis in Eq. 1 and uses Eq. 2 to evaluate the P

value: $\mathbb{P}(\tau(\mathbf{X}) > \tau(\mathbf{x}_0); H_0) = p(\mathbf{x}_0)$, which is traditionally defined as the probability of obtaining a value of a test statistic $\tau(\mathbf{x})$ at least as extreme as the one observed $\tau(\mathbf{x}_0)$, assuming that H_0 is true. A P value lower then a designated threshold, say 0.05, is viewed as evidence against H_0. For historical accuracy, this needs to be viewed in conjunction with Fisher's *falsificationist stance* concerning testing in the sense that significance tests can *falsify* but never *verify* hypotheses (Fisher 1955). The subsequent literature, however, did extend the interpretation of P values to allow for large enough values to be viewed as *moderate* to *no evidence against* H_0; see Murtaugh (2013).

The same sampling distribution (Eq. 2) is used to define the Neyman-Pearson (N-P) Type I error probability: $\mathbb{P}(\tau(\mathbf{X}) > c_\alpha; H_0) = \alpha$, where c_α is the critical value for significance level α. This defines the t test

$$T_\alpha^{>} = \left\{ \tau(\mathbf{X}) = \frac{\sqrt{n}(\bar{X}_n - \mu_0)}{s},\ C_1(\alpha) = \{\mathbf{x} : \tau(\mathbf{x}) > c_\alpha\} \right\} \tag{3}$$

where $C_1(\alpha)$ denotes the rejection region and the superscripted $>$ denotes a one-sided test in the positive direction. The N-P approach differs from that of Fisher by justifying the choice of both $\tau(\mathbf{X})$ and $C_1(\alpha)$ on optimality grounds, i.e., the choices in Eq. 3 maximize the power: $\mathbb{P}(\tau(\mathbf{X}) > c_\alpha; \mu = \mu_1) = \pi(\mu_1)$, for $\mu_1 > \mu_0$ Note that the Type II error probability is $\beta(\mu_1) = 1 - \pi(\mu_1)$, for all $\mu_1 > \mu_0$ To evaluate the power, one needs the sampling distribution of $\tau(\mathbf{X})$ under H_1

$$\tau(\mathbf{X}) = \frac{\sqrt{n}(\bar{X}_n - \mu_0)}{s} \overset{\mu=\mu_1}{\sim} \mathrm{St}(\delta_1, n-1), \qquad \text{for } \mu_1 > \mu_0$$

where

$$\delta_1 = \frac{\sqrt{n}(\mu_1 - \mu_0)}{\sigma}$$

denotes the non-centrality parameter. It can be shown that test $T_\alpha^{>}$, as defined in Eq. 3, is optimal, uniformly most powerful (UMP); see Lehmann (1986). The power of a N-P test provides a measure of its generic [for any $\mathbf{x} \in \mathbb{R}^n$] capacity to detect different discrepancies from the null, given α.

A crucial difference between the P value and the Type I and II error probabilities is that the former is defined *post-data,* since it requires $\tau(\mathbf{x}_0)$, but the latter are defined *pre-data* since they only require n and the choice of α. Despite that, the P value is often viewed by practitioners as the observed significance level and recast the accept/reject rules into (Lehmann 1986): reject H_0 if $p(\mathbf{x}_0) \leq \alpha$, accept H_0 if $p(\mathbf{x}_0) > \alpha$, because the data specificity of $p(\mathbf{x}_0)$ seems more informative than the dichotomous accept/reject decisions.

P *value and the large* n *problem*

A crucial weakness of both the P value and the N-P error probabilities is the so-called large n problem: there is always a large enough sample size n for which any simple null hypothesis. H_0: $\mu = \mu_0$ will be rejected by a frequentist α-significance level test; see Lindley (1957). As argued in Spanos (2013), there is nothing paradoxical about a small P value, or a rejection of H_0, when n is large enough.

It is an inherent feature of a good (consistent) frequentist test, as $n \to \infty$ the power of the test $\pi(\mu_1)$, for any discrepancy $\gamma \neq 0$ from the null goes to one, i.e., $\pi(\mu_1) \underset{n\to\infty}{\to} 1$. What is fallacious is to interpret a rejection of H_0 as providing the same weight of evidence *for* a particular alternative H_1, irrespective of whether n is large or small. This is an example of a more general fallacious interpretation that stems from the fact that all rejections of H_0 are viewed as providing the same weight of evidence for a particular alternative H_1, regardless of the generic capacity (the power) of the test in question. The large n problem arises because, in light of the fact that

$$\delta_1 = \frac{\sqrt{n}(\mu_1 - \mu_0)}{\sigma}$$

the power depends crucially on n; it increases with $\sqrt{n}$. This renders a rejection of H_0 with a small n (low power) very different—in evidential terms—than one with a large n (high power). Hence, the claim that "the smaller the P value the more the evidence we have against the null hypothesis" (Murtaugh 2013) needs to be qualified. Indeed, the real problem does not lie with the P value or the accept/reject rules as such, but with how such results are transformed into *evidence for* or *against* a hypothesis H_0 or H_1.

The large n constitutes an example of a broader problem known as the *fallacy of rejection*: (mis)interpreting reject H_0 (evidence *against* H_0) as evidence *for* a particular H_1; this can arise when a test has very high power, e.g., large n. A number of attempts have been made to alleviate the large n problem, including rules of thumb for decreasing α as n increases; see Lehmann (1986). Due to the trade-off between the Type I and II error probabilities, however, any attempt to ameliorate the problem renders the inference susceptible to the reverse fallacy known as the *fallacy of acceptance*: (mis)interpreting accept H_0 (*no* evidence *against* H_0) as evidence *for* H_0; this can easily arise when a test has very low power; e.g., α is tiny or n is too small.

These fallacies are routinely committed by practitioners in many applied fields. After numerous unsuccessful attempts, Mayo (1996) provided a reasoned answers to these fallacies in the form of a post-data severity assessment.

Severity and the Fallacies of Acceptance/Rejection

Whether data $\mathbf{x}_0$ provide evidence for or against a particular hypothesis H depends crucially on the generic capacity (power) of the test to detect discrepancies from the null. This stems from the intuition that a small P value or a rejection of H_0 based on a test with low power (e.g., a small n) for detecting a particular discrepancy γ provides stronger evidence for γ than using a test with

much higher power (e.g., a large n). This intuition is harnessed by a post-data severity evaluation of accept/reject based on custom-tailoring the generic capacity of the test to establish the discrepancy γ warranted by data $\mathbf{x}_0$; see Mayo (1996).

Post-data severity evaluation

The severity evaluation is a *post-data* appraisal of the accept/reject and P value results with a view to provide an *evidential interpretation*; see Mayo and Spanos (2011). A hypothesis H (H_0 or H_1) "passes" a severe test T_α with data $\mathbf{x}_0$ if (i) $\mathbf{x}_0$ accords with H and (ii) with very high probability, test T_α would have produced a result that accords less well with H than $\mathbf{x}_0$ does, if H were false (Mayo and Spanos 2006).

The notion of severity can be used to bridge the gap between accept/reject rules and P values and an evidential interpretation in so far as the result that H passes test T_α provides good evidence for inferring H (is correct) to the extent that T_α severely passes H with data $\mathbf{x}_0$. The severity assessment allows one to determine whether there is evidence for (or against) inferential claims of the form $\mu_1 = \mu_0 + \gamma$, for $\gamma \geq 0$, in terms of a discrepancy γ from μ_0, which includes H_0 as well as any hypothesis belonging to the alternative parameter space $\mu_1 > \mu_0$.

For the case of *reject* H_0, the relevant claim is $\mu > \mu_1 = \mu_0 + \gamma$, $\gamma \geq 0$, with a view to establish the *largest discrepancy* γ from H_0 warranted by data $\mathbf{x}_0$. In this case, $\mathbf{x}_0$ in condition (i) accords with H_1, and condition (ii) concerns "results $\mathbf{x} \in \mathbb{R}^n$ that accord less well with H_1 than $\mathbf{x}_0$ does." Hence, the severity evaluation is

$$\begin{aligned}\text{SEV}(T_\alpha; \mathbf{x}_0; \mu > \mu_1) &= \mathbb{P}(\tau(\mathbf{X}) \leq \tau(\mathbf{x}_0); \mu > \mu_1 \text{ false}) \\ &= \mathbb{P}(\tau(\mathbf{X}) \leq \tau(\mathbf{x}_0); \mu \leq \mu_1) \qquad (4)\end{aligned}$$

where $\mathbb{P}(\tau(\mathbf{X}) \leq \tau(\mathbf{x}_0); \mu \leq \mu_1)$ is evaluated at $\mu = \mu_1$ since the $\text{SEV}(\mu < \mu_1)$ increases for $\mu < \mu_1$. Analogously, for accept H_0 (Mayo and Spanos 2006)

$$\text{SEV}(T_\alpha; \mathbf{x}_0; \mu \leq \mu_1) = \mathbb{P}(\tau(\mathbf{X}) > \tau(\mathbf{x}_0); \mu = \mu_1). \qquad (5)$$

It should be emphasized that what is important for interpretation purposes is not the numerics of the tail areas, but the coherence of the underlying reasoning.

Revisiting the P *value: a severity perspective*

To bring out a key weakness of the P value as a measure of evidence, let us relate it to the severity evaluation for reject H_0 by restricting the latter at $\gamma = 0$:

$$\begin{aligned}\text{SEV}(T_\alpha; \mathbf{x}_0; \mu > \mu_0) &= \mathbb{P}(\tau(\mathbf{X}) \leq \tau(\mathbf{x}_0); \mu \leq \mu_0) \\ &= 1 - \mathbb{P}(\tau(\mathbf{X}) > \tau(\mathbf{x}_0); \mu \leq \mu_0) \\ &\geq 1 - P(\mathbf{x}_0).\end{aligned}$$

This suggests that, for a small P value ($P = 0.01$), $1 - P(\mathbf{x}_0) = 0.99$, provides a lower bound for the severity assessment of $\mu > \mu_0$. Viewed from this vantage point, a small P value establishes the existence of *some* discrepancy $\gamma \geq 0$, but provides no information concerning *its magnitude*.

The severity evaluation remedies this weakness of the P value by taking into account the generic capacity of the test to output the magnitude of the discrepancy γ warranted by data $\mathbf{x}_0$ and test T_α. This, however, necessitates considering alternative values of μ within the same statistical model. This is because N-P testing is inherently testing within the boundaries of a statistical model, as opposed to mis-specification (M-S) testing which probes outside those boundaries, with the prespecified model representing the null; see Mayo and Spanos (2004).

Statistical vs. substantive significance

The post-data severity evaluation in the case of *reject* H_0 outputs which inferential claims of the form $\mu > \mu_1$ are warranted (high severity) or unwarranted (low severity) on the basis of test T_α and data $\mathbf{x}_0$. This provides the basis for addressing the *statistical* vs. *substantive* significance problem that has bedeviled practitioners in several fields since the 1950s. Once the warranted discrepancy γ^* is established, one needs to confer with substantive subject matter information to decide whether this discrepancy is *substantively significant* or not. Hence, not only statistical significance does not imply substantive significance, but the reverse is also true. A *statistically insignificant* result can implicate a substantively significant discrepancy; see Spanos (2010*a*) for an empirical example.

The severity perspective calls into question the use of *effect size measures*, based on "distance functions" using point estimators, as flawed attempts to evaluate the warranted discrepancy by attempting to eliminate the influence of the sample size n in an ad hoc way. Indeed, classifying effect sizes as "small," "medium," and "large" (Cumming 2011), without invoking subject matter information, seems highly questionable. In contrast, the post-severity evaluation accounts for the effect of the sample size n by taking into consideration the generic capacity of the test to output the warranted discrepancy γ in a principled manner, and then lets the subject matter information make the call about substantive significance.

More generally, in addition to circumventing the fallacies of acceptance and rejection, severity can be used to address other charges like the "arbitrariness" of the significance level, the one-sided vs. two-sided framing of hypotheses, the reversing of the null and alternative hypotheses, the effect size problem, etc.; see Mayo and Spanos (2011). In particular, the post-data severity evaluation addresses the initial arbitrariness of any threshold relating to the significance level or the P value by relying on the *sign* of $\tau(\mathbf{x}_0)$, and not on c_α, to indicate the *direction* of the inferential claim that "passed." Indeed, this addresses the concerns for the dichotomy created by any threshold; see Spanos (2011).

P Values and CIs

For the simple Normal model, the $(1-\alpha)$ CI for μ

$$\mathbb{P}\left(\bar{X}_n - c_{\frac{\alpha}{2}}\left(\frac{s}{\sqrt{n}}\right) \leq \mu \leq \bar{X}_n + c_{\frac{\alpha}{2}}\left(\frac{s}{\sqrt{n}}\right)\right) = 1-\alpha \quad (6)$$

is optimal in the sense that it has the shortest expected length. Its optimality can be demonstrated using the mathematical duality between Eq. 6 and the UMP unbiased test (Lehmann 1986)

$$T_\alpha = \left\{\tau(\mathbf{X}) = \frac{\sqrt{n}(\bar{X}_n - \mu_0)}{s},\ \ C_1(\alpha) = \{\mathbf{x} : |\tau(\mathbf{x})| > c_\alpha\}\right\}$$

associated with the hypotheses $H_0 : \mu = \mu_0$ vs. $H_1 : \mu \neq \mu_0$. The mathematical duality between hypothesis testing and CIs, however, has beclouded the crucial differences between the two types of inference procedures and led to several misleading claims, like (a) CIs are more informative than tests and (b) CIs avoid most of the weaknesses of tests. As argued by Murtaugh (2013): "*P* values and confidence intervals are just different ways of summarizing the same information." The truth is that these two procedures pose very different questions to the data and they elicit distinct answers.

CIs vs. hypothesis testing: the underlying reasoning

The key difference between hypothesis testing and CIs is that the sampling distribution underlying Eq. 6 does not coincide with Eq. 2, but instead takes the form

$$\tau(\mathbf{X};\mu) = \frac{\sqrt{n}(\bar{X}_n - \mu)}{s} \overset{\mu=\mu^*}{\sim} \mathrm{St}(n-1) \quad (7)$$

where $\tau(\mathbf{X};\ \mu)$ is a pivot (not a test statistic) and the evaluation does not invoke *hypothetical reasoning* ($\mu = \mu_0$), but *factual* $\mu = \mu^*$ (μ^* being the true value of μ, whatever that happens to be). Hence, a more pertinent way to write Eq. 6 is

$$\mathbb{P}\left(\bar{X}_n - c_{\frac{\alpha}{2}}(\frac{s}{\sqrt{n}}) \leq \mu \leq \bar{X}_n + c_{\frac{\alpha}{2}}(\frac{s}{\sqrt{n}});\ \mu = \mu^*\right) = 1-\alpha \quad (8)$$

which makes explicit the underlying reasoning. This crucial difference is often obscured by blurring the distinction between the null value μ_0 and the true value μ^* when deriving a CI by solving the *acceptance region*

$$C_0(\alpha) = \left\{\mathbf{x} : \left|\frac{\sqrt{n}(\bar{X}_n - \mu_0)}{s}\right| \leq c_\alpha\right\}$$

for μ_0, and then pretending that μ_0 stands, not for all its *unknown* values μ within that interval. What makes the blurring between μ_0 and the true value μ^* particularly elusive is that the mathematical duality ensures that under both modes of reasoning, hypothetical and factual, one is evaluating the same tail areas of $\mathrm{St}(n-1)$ for hypothesis testing and CIs. What is important for interpretation purposes, however, is not the numerics of the tail areas, but the coherence of the underlying reasoning and the nature of the ensuing inferences.

An important upshot of factual reasoning is that, *post-data*, one cannot attach a probability to the observed CI

$$\mathrm{OCI} = (\bar{x}_n - c_{\frac{\alpha}{2}}(s/\sqrt{n}) \leq \mu \leq \bar{x}_n + c_{\frac{\alpha}{2}}(s/\sqrt{n})) \quad (9)$$

because the post-data coverage probability is either zero or one; the factual scenario $\mu = \mu^*$ has played out and OCI either includes or excludes μ^*. Hence, one has no way to distinguish between more likely and less likely values of μ within an OCI using factual reasoning. Note that in hypothesis testing, post-data error probabilities, like the *P* value, are definable since the reasoning is hypothetical, and thus it applies equally post-data as well as pre-data. However, the mathematical duality enables one to use OCI as a surrogate test for two-sided hypotheses, by (illicitly) switching between the two different modes of reasoning.

Ironically, practitioners in several applied fields are happy to use this mathematical duality, but ignore the fact that some of the charges leveled at the *P* value apply equally to CIs. For instance, the CI in Eq. 8 is equally vulnerable to the large *n* problem because its expected length

$$E\left(\left[\bar{X}_n + \frac{s}{\sqrt{n}}c_{\frac{\alpha}{2}}\right] - \left[\bar{X}_n - \frac{s}{\sqrt{n}}c_{\frac{\alpha}{2}}\right]\right) = 2c_{\frac{\alpha}{2}}\left(\frac{\sigma}{\sqrt{n}}\right)$$

shrinks down to zero as $n \to \infty$; see also Murtaugh (2013). This calls into question various claims that OCIs provide more reliable information than *P* values when it comes to the relevant "effect size" (whatever that might mean).

Observed CIs and severity

The *post-data severity evaluation* can be used to bring out this confusion and shed light on the issues of distinguishing between different values of μ within an OCI. Hence, one cannot attach probabilities to inferential claims of the form

$$\mu > \bar{x}_n - c_{\frac{\alpha}{2}}\left(\frac{s}{\sqrt{n}}\right), \quad \text{and} \quad \mu \leq \bar{x}_n + c_{\frac{\alpha}{2}}\left(\frac{s}{\sqrt{n}}\right) \quad (12)$$

because the coverage probability is rendered degenerate post-data. On the other hand, severity can be used to evaluate inferential claims of the form

$$\mu > \mu_1 = \mu_0 + \gamma, \mu \leq \mu_1 = \mu_0 + \gamma, \quad \text{for some } \gamma \geq 0. \quad (13)$$

Thus, in principle one can relate the observed bounds

$$\bar{x}_n \pm c_{\frac{\alpha}{2}}\left(\frac{s}{\sqrt{n}}\right)$$

to these inferential claims

$$\mu_1 = \bar{x}_n - c_{\frac{\alpha}{2}}\left(\frac{s}{\sqrt{n}}\right)$$

and evaluating (Mayo and Spanos 2006)

$$\text{SEV}\left(\mu > \bar{x}_n - c_{\frac{\alpha}{2}}\left(\frac{s}{\sqrt{n}}\right)\right) \quad \text{or SEV}\left(\mu \leq \bar{x}_n + c_{\frac{\alpha}{2}}\left(\frac{s}{\sqrt{n}}\right)\right). \tag{14}$$

A moment's reflection, however, suggests that the connection between severity and the OCI is more apparent than real. This is because the reasoning underlying the severity evaluations in Eqs. 4 and 5 is *hypothetical,* evaluated under different values $\mu = \mu_1$, and *not* factual $\mu = \mu^*$. Indeed, the inferential claims and the relevant probabilities associated with SEV(.) in Eq. 4.7 have nothing to do with the coverage probability for μ^*; they pertain to the relevant inferential claims as they relate to particular discrepancies

$$\gamma = \left(\tau(\mathbf{x}_0) \pm c_{\frac{\alpha}{2}}\right)\left(\frac{s}{\sqrt{n}}\right)$$

in light of data $\mathbf{x}_0$.

CIs vs. hypothesis testing: questions posed

Inference procedures associated with hypothesis testing and CIs share a common objective: learn from data about the "true" ($\mu = \mu^*$) statistical model $M^*(\mathbf{x}) = \{f(\mathbf{x};\theta^*)\}$, $\mathbf{x} \in \mathbb{R}^n$ yielding data $\mathbf{x}_0$. What about the questions posed?

The question posed by a CI is: How often will a random interval $[L(\mathbf{X}), U(\mathbf{X})]$ cover the true value μ^* of μ, whatever that *unknown* value μ^* happens to be? The answer comes in the form of a $(1 - \alpha)$ CI using *factual* reasoning.

The question posed by a test is: how close is the prespecified value μ_0 to μ^*?

The answer comes in the form of an optimal test whose capacity is calibrated using the pre-data error probabilities. A closer look at the test statistic

$$\tau(\mathbf{X}) = \frac{\sqrt{n}(\bar{X}_n - \mu_0)}{s}$$

reveals that it is effectively a standardized distance between μ^* and μ_0, since $\bar{X}_n$ is an excellent estimator of μ^* and $\bar{x}_n$ is assumed to have been generated by $M^*(\mathbf{x})$.

Revisiting Akaike-type Model Selection Procedures

Akaike (1973) introduced *model selection* within a prespecified family

$$M(m) := \{M_{\theta_i}(\mathbf{z}) = \{f(\mathbf{z};\theta_i), \theta_i \in \Theta\}, \quad \mathbf{z} \in \mathbb{R}_Z^n, \ i = 1, 2, \cdots m\} \tag{15}$$

vwhich relies on minimizing a distance function based on the estimated log-likelihood (viewed as a goodness-of-fit measure) and a penalty function relating to the number of unknown parameters $\boldsymbol{\theta}_i$ associated with each model $M_{\theta_i}(\mathbf{z})$.

The objective function is

$$\text{AIC}(i) = -2\ln\mathcal{L}(\mathbf{z};\hat{\theta}_i) + 2K_i, \quad i = 1, 2, \cdots, m \tag{16}$$

where $\mathcal{L}(\mathbf{z};\theta_i) \propto f(\mathbf{z};\theta_i)$, $\theta_i \in \Theta$ is the likelihood function and K_i is the number of unknown parameters in $\boldsymbol{\theta}_i$. It can be viewed as trading goodness-of-fit/prediction against parsimony (simplicity). The primary aim is to *rank* all the models in $M(m)$ in terms of the estimated distance function, which is often interpreted as a metric of support; see Burnham and Anderson (2002).

In the particular case of nested regression models

$$M_{\theta_i}(\mathbf{z}) : y_t = \beta_0 + \sum_{j=1}^{i}\beta_j x_t^j + u_t, u_t \sim \text{NIID}(0, \sigma^2), \quad i = 1, 2, \cdots, m \tag{17}$$

the AIC takes the specific form $\text{AIC}(i) = n\ln(\hat{\sigma}_i^2) + 2K_i$, $i = 1, 2, \ldots, m$, where

$$\hat{\sigma}_i^2 = \frac{1}{n}\sum_{t=1}^{n}\left(y_k - \hat{\beta}_0 - \sum_{j=1}^{i}\hat{\beta}_j x_t^j\right)^2.$$

Evaluating the AIC(i) for all $i = 1, 2, \ldots, m$, yields a *ranking* of the models in $M(m)$, and the smallest is chosen.

Using goodness-of-fit/prediction as the primary criterion for "ranking the different models," however, can potentially undermine the reliability of any inference in two ways. First, goodness-of-fit/prediction is neither necessary nor sufficient for *statistical adequacy*: the model assumptions like NIID are valid for data $\mathbf{Z}_0$. The latter ensures that the actual error probabilities approximate closely the nominal error probabilities. Applying a 0.05 significance level test when the actual Type I error is closer to 0.60 can easily lead an inference astray! Indeed, the appropriateness of particular goodness-of-fit/prediction measures, such as $\ln\mathcal{L}(\mathbf{z};\hat{\boldsymbol{\theta}}_i)$, is questionable when $M_{\theta_i}(\mathbf{z})$ is statistically misspecified; see Spanos (2007).

One might object to this argument on the grounds that all inference procedures are vulnerable to statistical misspecification. Why single out Akaike-type model selection? The reason is that model validation based on thorough M-S testing to secure statistical adequacy (Mayo and Spanos 2004) is in direct conflict with such model selection procedures. This is because model validation will give rise to a choice of a particular model within Eq. 17 on statistical adequacy grounds, assuming Eq. 15 includes such an adequate model. This choice would render model selection procedures redundant and often misleading because the highest ranked model will rarely coincide with the statistically adequate one, largely due to the *second* way model selection procedures could undermine the reliability of inference. As shown below, the *ranking* of the different models is inferentially equivalent to N-P testing comparisons with a serious weakness: model selection procedures ignore the relevant error probabilities. If the implicit error probabilities are too low/high, that could give rise to unreliable inferences. In addition, if no statistically adequate model exists within Eq. 17, M-S testing would confirm that and no choice will be made, but model

selection procedures would nevertheless indicate a highest ranked model; see Spanos (2010*b*) for empirical examples.

AIC vs. N-P testing

At first sight, the Akaike model selection procedure's reliance on minimizing a distance function, combining the log-likelihood and the number of unknown parameters, seems to circumvent hypothesis testing and the controversies surrounding P values and accept/reject rules. Indeed, its simplicity and apparent objectivity made it a popular procedure among practitioners.

Murtaugh (2013) brings out the connections between P values, CIs, and the AIC, and argues that: "Since P values, confidence intervals, and ΔAIC [difference of AIC] are based on the same statistical information, all have their places in modern statistical practice. The choice of which to use should be stylistic, dictated by details of the application rather than by dogmatic, a priori considerations."

This argument is misleading because on closer examination, minimizing the AIC does not circumvent these problems and controversies. Although proponents of AIC generally discourage comparisons of only two models, the *ranking* of the different models by the AIC is inferentially equivalent to *pairwise comparisons* among the different models in $\{M_{\theta_i}(\mathbf{z}), i=1, 2, \ldots, m\}$, using N-P testing with a serious flaw: it ignores the relevant error probabilities; see Spanos (2010*b*).

To illustrate the connection between the AIC ranking and N-P testing consider a particular pairwise comparison between the following two models within Eq. 15:

$$M_0 : y_t = \beta_0 + \beta_1 x_t + u_t;$$
$$M_1 : y_t = \beta_0 + \beta_1 x_t + \beta_2 x_t^2 + \beta_3 x_t^3 + u_t. \qquad (18)$$

Let us assume that the AIC procedure selects model M_1, i.e.,

$$[n \ln(\hat{\sigma}_0^2) + 2K_0] > [n \ln(\hat{\sigma}_1^2) + 2K_1] \Rightarrow$$
$$(\hat{\sigma}_0^2/\hat{\sigma}_1^2) > \exp([2(K_1 - K_0)]/n). \qquad (19)$$

One can relate this AIC decision in favor of M_1 to the rejection of H_0

$$H_0\text{: } \beta_2 = \beta_3 = 0, \text{ vs. } H_1\text{: } \beta_2 \neq 0, \quad \text{or } \beta_3 \neq 0 \qquad (20)$$

by the F test

$$F(\mathbf{Z}) = \left([\hat{\sigma}_0^2 - \hat{\sigma}_1^2]/\hat{\sigma}_1^2\right)\left(\frac{n - K_1}{K_1 - K_0}\right),$$
$$C_1 = \{\mathbf{z} : F(\mathbf{z}) > c_\alpha\} \qquad (21)$$

where c_α denotes the critical value for significance level α. This suggests that the AIC procedure amounts to rejecting H_0 when $F(\mathbf{z}) > c_{\text{AIC}}$, for

$$c_{\text{AIC}} = \left(\frac{n - K_1}{K_1 - K_0}\right)\left[\exp\left(\frac{2(K_1 - K_0)}{n}\right) - 1\right]$$

e.g., when $n = 100$, $c_{\text{AIC}} = 1.94$, implying that the actual Type I error is $\alpha_{\text{AIC}} = 0.15$; using α_{AIC}, one can derive the implicit power function for the above F test. This indicates that the ranking of M_1 higher than M_0 by AIC involves a much higher significance level than the traditional ones. In this sense, the AIC implicitly allows for a much higher probability of rejecting the null when true. More generally, the implicit error probabilities associated with the AIC procedure are at best unknown, calling into question the reliability of any inferences. These results can be easily related to those in Murtaugh (2013) between ΔAIC and the relevant P value: $\mathbb{P}(F(\mathbf{Z}) > F(\mathbf{z}_0); H_0)$.

Summary and Conclusions

The paper focused primarily on certain charges, claims, and interpretations of the P value as they relate to CIs and the AIC. It as argued that some of these comparisons and claims are misleading because they ignore key differences in the procedures being compared, such as (1) their primary aims and objectives, (2) the nature of the questions posed to the data, as well as (3) the nature of their underlying reasoning and the ensuing inferences.

In the case of the P value, the crucial issue is whether Fisher's evidential interpretation of the P value as "indicating the strength of evidence against H_0" is appropriate. It is argued that, despite Fisher's maligning of the Type II error, a principled way to provide an adequate evidential account, in the form of post-data severity evaluation, calls for taking into account the power of the test.

The error-statistical perspective brings out a key weakness of the P value and addresses several foundational issues raised in frequentist testing, including the fallacies of acceptance and rejection as well as misinterpretations of observed CIs; see Mayo and Spanos (2011). The paper also uncovers the connection between model selection procedures and hypothesis testing, revealing the inherent unreliability of the former. Hence, the choice between different procedures should not be "stylistic" (Murtaugh 2013), but should depend on the questions of interest, the answers sought, and the reliability of the procedures.

Acknowledgements

I would like to thank D. G. Mayo for numerous discussions on issues discussed in this paper.

Literature Cited

Akaike, H. 1973. Information theory and an extension of the maximum likelihood principle. Pages 267–281 *in* B. N. Petrov and F. Csaki, editors. Second International Symposium on Information Theory. Akademia Kiado, Budapest, Hungary.

Burnham, K. P., and D. R. Anderson. 2002. Model selection and multimodel inference. Second edition., Springer, New York, New York, USA.

Cumming, G. 2011. Understanding the new statistics: effect sizes, confidence intervals, and meta-analysis. Routledge, London, UK.

Fisher, R. A. 1925. Statistical methods for research workers. Oliver and Boyd, Edinburgh, UK.
Fisher, R. A. 1955. Statistical methods and scientific induction. Journal of the Royal Statistical Society B 17:69–78.
Lehmann, E. L. 1986. Testing statistical hypotheses. Second edition. Wiley, New York, New York, USA.
Lindley, D. V. 1957. A statistical paradox. Biometrika 44:187–192.
Mayo, D. G. 1996. Error and the growth of experimental knowledge. University of Chicago Press, Chicago, Illinois, USA.
Mayo, D. G., and A. Spanos. 2004. Methodology in practice: statistical misspecification testing. Philosophy of Science 71:1007–1025.
Mayo, D. G., and A. Spanos. 2006. Severe testing as a basic concept in a Neyman-Pearson philosophy of induction. British Journal for the Philosophy of Science 57:323–357.
Mayo, D. G., and A. Spanos. 2011. Error statistics. Pages 151–196 *in* D. Gabbay, P. Thagard, and J. Woods, editors. The handbook of philosophy of science, volume 7: philosophy of statistics. Elsevier, Amsterdam, The Netherlands.
Murtaugh, P. A. 2014. In defence of *P* values. Ecology 95:611–617.
Neyman, J. 1956. Note on an article by Sir Ronald Fisher. Journal of the Royal Statistical Society B 18:288–294.
Neyman, J., and E. S. Pearson. 1933. On the problem of the most efficient tests of statistical hypotheses. Philosophical Transactions of the Royal Society A 231:289–337.
Pearson, E. S. 1955. Statistical concepts in the relation to reality. Journal of the Royal Statistical Society B 17:204–207.
Spanos, A. 2007. Curve-fitting, the reliability of inductive inference and the error-statistical approach. Philosophy of Science 74:1046–1066.
Spanos, A. 2010*a*. Is frequentist testing vulnerable to the base-rate fallacy? Philosophy of Science 77:565–583.
Spanos, A. 2010*b*. Akaike-type criteria and the reliability of inference: model selection vs. statistical model specification. Journal of Econometrics 158:204–220.
Spanos, A. 2011. Misplaced criticisms of Neyman-Pearson (N-P) testing in the case of two simple hypotheses. Advances and Applications in Statistical Science 6:229–242.
Spanos, A. 2013. Who should be afraid of the Jeffreys-Lindley paradox? Philosophy of Science 80:73–93.

Ecology, 95(3), 2014, pp. 651–653

Rejoinder

Paul A. Murtaugh[1]

Department of Statistics, Oregon State University, Corvallis, Oregon 97331 USA

I thank the editors of *Ecology* for their interest in my paper, and the discussants for their extensive comments. I found myself agreeing with most of what was said, so I will make just a few observations.

Burnham and Anderson (2014) are mistaken when they say that the relationship between *P* values and AIC differences "holds only for the simplest case (i.e., comparison of two nested models differing by only one parameter)." As shown in Murtaugh (2014) Eqs. 5 and 6, the relationship holds for any k, i.e., for nested models differing by any number of parameters. It is also worth pointing out that the relationship holds for not only for nested linear models with Gaussian errors, as stated by Stanton-Geddes et al. (2014), but also for nested models with non-Gaussian errors if n is large (Murtaugh 2014: Eq. 5).

Burnham and Anderson (2014) comment that information-theoretic methods are "free from arbitrary cutoff values," yet they and others have published arbitrary guidelines for deciding how large a value of ΔAIC is needed for one model to be preferred over another (see Table 1). In any case, it is clear that both the *P* value and ΔAIC are continuous metrics, the interpretation of which is necessarily subjective (see my original Figs. 1 and 3).

De Valpine (2013) comments on the oft-repeated criticism that the *P* value is based on unobserved data, because it is the probability of obtaining a statistic at least as extreme as the observed statistic, given that the null hypothesis is true. As he suggests, any statistical method involving likelihoods is grounded in the assumption that a particular statistical distribution underlies both the observed and unobserved, hypothetical data, so that "part and parcel of that model are the probabilities associated with the unobserved data." I would add that Bayesians working with subjective priors also depend quite heavily on unobserved data.

It seems foolish to discard useful statistical tools because they are old (Burnham and Anderson 2014), or because they can only be applied in certain settings. I think it is healthy that the ecologists challenged by Stanton-Geddes et al. (2014) used a variety of methods to do their analyses, although it is disconcerting that the "participants came to surprisingly different conclusions." I wholeheartedly agree with Stanton-Geddes et

Manuscript received 1 October 2013; accepted 3 October 2013. Corresponding Editor: A. M. Ellison. For reprints of this Forum, see footnote 1, p. 609.

[1] E-mail: murtaugh@science.oregonstate.edu