ELSEVIER

The Journal of Socio-Economics 33 (2004) 631–650

The Journal of Socio-Economics

www.elsevier.com/locate/econbase

No-decision classification: an alternative to testing for statistical significance

Nathan Berg[a,b,*]

[a] *School of Social Sciences, University of Texas at Dallas, GR 31 211300, Box 830688, Richardson, TX 75083-0688, USA*

[b] *Center for Adaptive Behavior and Cognition, Max Planck Institute for Human Development, Berlin, Germany*

Abstract

This paper proposes a new statistical technique for deciding which of two theories is better supported by a given set of data while allowing for the possibility of drawing no conclusion at all. Procedurally similar to the classical hypothesis test, the proposed technique features three, as opposed to two, mutually exclusive data classifications: reject the null, reject the alternative, and no decision. Referred to as No-decision classification (NDC), this technique requires users to supply a simple null and a simple alternative hypothesis based on judgments concerning the smallest difference that can be regarded as an economically substantive departure from the null. In contrast to the classical hypothesis test, NDC allows users to control both Type I and Type II errors by specifying desired probabilities for each. Thus, NDC integrates judgments about the economic significance of estimated magnitudes and the shape of the loss function into a familiar procedural form.

JEL classification: C12; C44; B40; A14

Keywords: Significance; Statistical significance; Economic significance; Hypothesis test; Critical region; Type II; Power

* Corresponding author. Tel.: +1 972 883 2088; fax +1 972 883 2735.
E-mail address: nberg@utdallas.edu.

1053-5357/$ – see front matter © 2004 Elsevier Inc. All rights reserved.
doi:10.1016/j.socec.2004.09.036

1. Introduction

A common goal of economic analysis is to determine which of two theories is better supported by available data. In such cases, the significance test is commonly applied—and commonly misused (Ziliak and McCloskey, 2005). Hardly surprising, such misuse of statistical significance follows from well known problems rooted in the construction of the hypothesis test which lead, as a matter of routine, to strong interpretations based on weak evidence (McCloskey, 1998; McAleer, 1995; McCloskey, 1985; Arrow, 1959). [1]

A key problem with the significance test (referred to alternatively as the hypothesis test, or the standard technique[2]) is its *necessity of choice*, whereby a binary decision (reject, not reject) must be taken no matter how weak the evidence or small the sample. A second problem with the standard test is its asymmetric treatment of Type I and Type II errors. As is well known, the significance test fixes the probability of Type I error with virtually no regard for the probability of correct rejection (i.e., the test's power). Fixing the probability of Type I error, and accepting whatever probabilities of Type II error are implied by the sample size and assumed density, is nearly always suboptimal once the costs and benefits of false versus correct rejection are considered. A third issue, perhaps the most serious, is that by focusing on the probability of extreme observations under the null hypothesis (i.e., statistical significance), the thoughtful analysis of magnitudes, for example, whether a regression coefficient is large enough to be considered important, tends to get crowded out of the analysis. This paper attempts to respond constructively to these frequently remarked upon problems by proposing an alternative statistical technique, referred to as no-decision classification (NDC), that deals directly with the standard procedure's limitations and pitfalls.[3]

One may question whether, despite the limitations and pitfalls, a new statistical procedure is really needed, or rather, whether improved training in the use of the standard technique would suffice. Indeed, it should be acknowledged, with appreciation, that sophisticated users of the hypothesis test are keenly sensitive to its power properties, the role sample size plays, and the importance of considering magnitudes—that is, economic, and not merely

[1] The broad philosophical dimensions of misuse of significance testing, stemming from the application of classical hypothesis testing to pursue analytical goals that are fundamentally Bayesian, is not the focus of this paper. Those issues arise whenever economists draw inferences about hypotheses conditional on data using classical statistics, in spite of the classical perspective's insistence that only probabilistic statements about data *given a hypothesis* are meaningful. This paper holds that contemporary statistical practice in the social sciences incorporates those philosophical tensions as a regular feature of its normal science, resulting in a methodological standard that is a de facto Bayesian-classical hybrid (Gigerenzer et al., 1989). Given that hybrid as the contemporary standard of the economics profession, the statistical decision-making procedure proposed in the present paper aims to improve the most glaring weaknesses of the standard approach without requiring, and waiting for, a major shift of methodological paradigm.

[2] See Dudewicz and Mishra (1988) for precise definitions and descriptions of the classical hypothesis test, and Gigerenzer et al. (1989) for the fascinating story of R.A. Fisher's role in the decades old codification of statistical significance.

[3] Wald (1947) proposed statistical decision procedures which allowed for no decision within the context of a sequence of experimental trials. Although Wald's sequential tests contributed substantially to contemporary operations research and the management of production processes, it remains unclear how to apply Wald's ideas to the secondary analysis of data sets with fixed sample sizes, perhaps the most common task undertaken by non-experimental empirical economists.

statistical, significance. However, as Ziliak and McCloskey (2005) show, misuse of statistical significance remained widespread throughout the 1990s, and likely remains so today, despite rigorous technical training and increased recognition of the distinction between economic versus statistical significance introduced decades earlier (e.g., McCloskey (1985).

Furthermore, misuse matters, not only in the abstract, but also in the real world, by way of policy decisions that too often hinge on little else and carry potentially large social costs. Ziliak and McCloskey (2005) provide specific examples (with author names and full citations) of recent policy analyses published in the *American Economic Review* where misuse of statistical significance is implicated in the formulation of errant policy prescriptions. More general patterns are also recognizable. For example, debates in macroeconomics (in which the question of permanent versus transitory shocks plays a role) frequently depend on "confirmation" of the null hypothesis that an autocorrelation coefficient equals 1 to justify the claim that Gross Domestic Product is a unit root process. Similarly, failure to reject (over-identifying) parameter restrictions in estimates of vector-auto-regressive models is frequently put forth to justify the assertion that the predictions of equilibrium theory hold, or that markets are efficient—or both. Misuse in microeconomics is also common, particularly in the interpretation of statistically insignificant regression coefficients. For example, prominent segments of the empirical micro literature rely on small t statistics to argue that class size has nothing to do with academic performance, that race is no longer an important factor in labor markets, and that political movements cannot explain key features of the institutional environment such as the enactment of new laws. With serious policy questions hanging in the balance, Ziliak and McCloskey (2005) deride the widespread practice of "asterisk economics" (i.e., using the magnitude of t or other test statistics as the sole basis for evaluating the importance of relationships among economic variables) and "sign economics" (reporting only the signs of estimated coefficients without analyzing the reasonableness or importance of their magnitudes).

As a remedy, Ziliak and McCloskey (2005) and McCloskey (1998) suggest that economists include context-specific information to help evaluate the importance of estimated magnitudes, explicitly factoring in their own judgments as a routine component of data analysis in economics. Explicit articulation of one's judgment concerning how large estimated relationships must be in order to count as important, the logic goes, is more transparent than behind-the-scenes incorporation of judgment used in choosing among models, among statistical tests with different power properties, or among data sets with different sample sizes. Although not usually emphasized, even standard implementations of the hypothesis test involve, or should involve, judgment—when choosing the level of significance and, depending on the user's degree of sophistication, when choosing which test to use, the modeling strategy, and the analysis of sample size and power. However, the predominant methodological approach taught in textbooks and practiced by economists does not embrace these roles for judgment. Instead, judgment is relegated to secondary status by adherence to the *set significance at five percent* rule and the (sometimes intentionally) opaque connection between statistical modeling choices and their influence on the hypothesis test's ultimate result.

Ziliak and McCloskey's suggestion to rely and report more on judgment points to a potentially difficult trade-off, however, between efficiency in the use of information and efficiency in the communication of statistical decisions. On the one hand, the significance test wastes information by not including expert judgments about the meaning of magnitudes and the shape of the loss function (i.e., the relative importance of Type I versus Type II errors). On the other hand, the procedural simplicity of the significance test facilitates a non-negligible degree of efficiency in communication that also should be acknowledged. For instance, one thinks of the time and concentration required to read several pages of descriptive text interpreting estimated parameters from various regression models compared with the ease of examining a table of starred coefficients. Similarly, one can appreciate the transactions-cost-reducing value of statistical significance as a useful form of jargon, evident, for example, in the simplicity of arguing against a theory by reporting rejection of a parameter restriction at the five percent level.

Even critics acknowledge that statistical significance is not completely without conceptual merit. Rather than an absolute ban, critics usually call for a more richly contextualized and carefully thought out application of statistical significance, balanced by other sources of information and modes of persuasion. But how to include judgment, achieve the desired balance and contextualization, and still communicate efficiently with the scientific community?

The proposal offered here is the intermediate approach of NDC, which aims to achieve better balance between context-specific judgment and procedural efficiency. NDC draws motivation from the observation that the very existence of a methodological canon implies a certain degree of uniformity in analytical technique. That uniformity carries both costs (e.g., rhetorical constraints, unavoidable methodological prescriptions, and reflexive rituals) and benefits (e.g., ease of reporting, ease of interpreting others' reports, and replicability of findings) for participants in a community of science. Thus, the challenge is to salvage value from the procedural simplicity of the hypothesis test while improving upon its deficiencies. NDC attempts to do precisely this—to recover information from the user that otherwise would be lost, while adhering closely to the procedural norm taught in contemporary textbooks as the hypothesis test.

Rather than obfuscating the role of the user's opinions about the relative importance of various patterns in the data, NDC invites the user to make explicit his or her beliefs concerning magnitudes and the relative costs of Type I and Type II errors. NDC not only makes these judgments explicit, it utilizes judgments consistently across users, requiring as user-supplied initial values the acceptable probabilities of Type I and Type II errors, and the minimum difference in parameter values considered to be economically significant. Thus, NDC is, in the terminology of Gigerenzer et al. (1989), a modified statistical ritual. By design, NDC resembles the standard hypothesis test's algorithmic procedure and therefore inherits its value as a facilitator of efficient communication and replicability of results. The differentiating feature of NDC rests in its capacity to transparently map judgments about magnitudes and the loss function into statistical decisions.

2. Background

Criticism of hypothesis testing is not new. McCloskey (1985) and Denton (1988) cite admonitions against misuse of statistical significance from critics writing in the early 20th century. In some ways, criticism was more vigorous then than now. Gigerenzer et al. (1989)

describe bitter controversies among the originators of hypothesis testing, including Karl Pearson, Jerzy Neyman, and R.A. Fisher, over the correct interpretation of the hypothesis test and its inherent drawbacks. Gigerenzer contrasts those pioneers' interest in the nuances of statistical interpretation and their intense methodological disagreements with contemporary textbook treatments, which present hypothesis testing as a unified construct, free of internal logical tension. Commenting on contemporary practice, Gigerenzer complains that the teaching of hypothesis testing seems to encourage an attitude of reliance upon automatic procedures designed to relieve the analyst from the burden of interpretation rather than attune students to its drawbacks and the debates it sparks. A variety of related observations animate Harlow et al. (1997) book length coverage of the debate over significance testing, *What If There Were No Significance Tests*?.

Among the critiques of significance testing, perhaps the best known is that of McCloskey (1985, 1998). She points out that by relying on statistical significance to answer the question of how big is big, researchers abdicate an important scientific responsibility. McCloskey's claim is not merely that the choice of significance level (e.g., $\alpha = .05$) is arbitrary. The real issue is the importance of thinking through the relative costs of being wrong as a function of what is actually true. That means specifying a loss function which reflects the context of each particular problem. In addition to urging more thorough consideration of power and loss functions, McCloskey asks economists to be more bold in ascribing meaning to magnitudes, taking a clear position on which ones deserve to be referred to as "significant" in the substantive sense.

There have been several attempts to incorporate these qualitative criticisms of hypothesis testing into an improved statistical decision-making procedure. Arrow (1959) proposes an "equal probability" test that treats Type I and Type II errors symmetrically. Selecting simple[4] null and alternative hypotheses, and imposing equality of Type I and Type II error probabilities, Arrow's test allows the magnitude of the error probability (equal to statistical significance) to vary freely within the unit interval, serving as an index of quality associated with inferences based upon it. Arrow illustrates the advantages of the equal probability approach by demonstrating undesirable asymmetries that arise in interpreting regression coefficients and their t statistics. In small to moderate sized samples, Arrow (1959, p. 73)shows that the t test's power can be close to zero, making it difficult to interpret "failure to reject the null." On the other hand, in very large samples, power is close to 1 and appreciable reductions in the probability of Type I error are possible with almost no loss of power. Arrow observes that asymmetric treatment of Type I and Type II errors is, in practice, rarely motivated by careful consideration of those errors' relative costs, as would be the case when using an explicit loss function and decision-theoretic methodology.

Jones and Tukey (2000) discuss a modified test of significance designed to free the analyst from the necessity of choice by allowing for an "indeterminacy" outcome similar in spirit to NDC's possibility of no decision. Wu (1985) proposes a "modified significance

[4] *Simple* hypotheses completely determine the distribution of a random variable, whereas *composite* hypotheses refer to a subset of a family of distributions, allowing for a multiplicity of possible theories about the data generating process. For example, "*X* is normally distributed with mean 7 and standard deviation 12" is simple, whereas "The mean of *X* is greater than 7" is composite, because the statement leaves open many values for the mean, and because its standard deviation is not specified.

test" using a loss-function framework to derive a three-region partition of the space of the test statistic that allows for a no-decision outcome, again, similar to NDC. In fact, both these techniques turn out to be special cases of the procedure proposed in this paper.

Among theoretical econometricians working on hypothesis testing, research priorities appear to be focused largely on extending technical aspects of the standard test rather than modifying its binary decision outcome or the asymmetry-inducing convention of fixing the level of significance. For example, Horowitz (2001) and Godfrey and Orme (2000) propose techniques for adjusting significance levels to reflect differences between finite-sample distributions and their asymptotic approximations. Andrews (1998, 1994) analyzes large sample properties of hypothesis tests with different weights placed on nearby alternatives. Another area of theoretical research on hypothesis testing takes up the question of how to rank competing tests with power functions that are difficult to compare (Terasvirta, 1996; Christian et al., 1993). King (1988) and Elliott et al. (1988) attempt to improve the shape of the power function by picking out simple alternative hypotheses that are more important than others.

At first, these lines of theoretical research seem potentially related to the critiques of standard hypothesis testing discussed in this paper. However, the connections turn out to be rather remote. The binary nature of the classical hypothesis test remains unchallenged. Unmotivated lexicographic prioritization of statistical significance over power also remains. Those papers that do attempt to deal with the power properties of the hypothesis test set out to make the selection of tests automatic and applicable across all problems and contexts, thus missing the essence of Ziliak and McCloskey's critique. Similarly, the goal of studies that advocate consideration of specific simple alternative hypotheses have more to do with defending the selection procedure against the charge of being *ad hoc* than with encouraging economists to think about the choice and justify it in the context of a particular economic problem.

Another area within the econometrics literature with potential links to NDC is non-nested hypothesis testing (Pesaran, 1974; Ericsson, 1986; Godfrey, 1998; Coulibaly and Brorsen, 1999). An embarrassing manifestation of the standard procedure's asymmetry is the intransitive sequences of inferences that arise from pairwise tests involving three or more hypotheses. The problem is acute in empirical studies attempting to use a single data set to falsify one or more theories from a list of several (e.g., Smith and Smyth (1991). When hypothesis testing is called upon to distinguish which economic theory is most consistent with specific data, its inherently asymmetric treatment of Type I and Type II errors winds up privileging one theory over another, stacking the odds in favor of particular conclusions without good justification.

The illogic of the hypothesis test's inflexible prioritization of statistical significance over the test's power is especially obvious in the context of audit tasks in accounting (Srivastava, 1997), where Type II errors (undetected cheats) are typically much more serious than Type I errors (false alarms). In the analysis of data collected from psychological experiments, Hertwig and Todd (2000) argue that the standard test's inherent asymmetry permits researchers to escape from having to fully specify alternative theories. Their analysis describes an unfortunate symbiosis between the hypothesis test's asymmetry and researchers' failure to elaborate precise alternative explanations of why null hypotheses may be inadequate. Gigerenzer (2000) provides numerous examples in which the asymmetry of hypothesis testing, rather than any particular characteristic of the data, virtually determines the conclusions that are drawn.

3. The no-decision classification (NDC) procedure

Let $\boldsymbol{X}$ represent a vector of continuously valued data with a known joint distribution.[5] It is natural to think of the vector $\boldsymbol{X}$ as a random sample, although its elements need not be independent nor identically distributed. In case the observations are themselves vectors, $\boldsymbol{X}$ should be thought of as the design matrix stacked into a single vector.

Let t denote the test statistic, a mapping of $\boldsymbol{X}$ into Ω_t:

$$t : \Omega_X \to \Omega_t. \tag{1}$$

The familiar case is when $\boldsymbol{X}$ is a random sample of length n, and the statistic t is a scalar (e.g., the sample mean). Denote the pdf of $t(\boldsymbol{X})$ as $f_\theta(t)$, a member of a parametric family of distributions indexed by θ.[6]

The no-decision classification procedure is specified below in terms of critical regions, which correspond to the simple hypotheses:[7]

$$H_0 : \theta = \theta_0,\;\; H_1 : \theta = \theta_1, (\theta_0 \neq \theta_1). \tag{2}$$

Before specifying the NDC procedure, several auxiliary definitions are required. The two sets, (C_0, C_1), $C_i \subset \Omega_t$, are said to *overlap* if their intersection is a positive probability event under any value of θ. In other words, if $\max_\theta P_\theta(C_0 \cap C_1) > 0$, then (C_0, C_1) are overlapping, and *non-overlapping* otherwise. Also, denote the complement of a set S with respect to Ω_t as $\overline{S}$. And denote the probability of the event $\cdot$ when the distribution of t is θ as $P_\theta[\cdot]$.

Definition 1 (No-decision classification). Given two simple hypotheses H_0 and H_1, and a test statistic t, the NDC procedure (C_0, C_1) is a pair of non-overlapping sets said to reject H_0, reject H_0, or make no decision, according the following rule:

- reject H_0 when $t \in C_0$
- reject H_1 when $t \in C_1$
- make no decision when $t \in \overline{C_0 \cup C_1}$.

[5] The assumption that the components of $\boldsymbol{X}$ are continuous variables helps avoid inconvenient details later on when expressing equations involving the probability that the test statistic lands in the critical region. In the discrete case, of course, those probabilities jump discontinuously, implying that solutions to equations in which they appear may fail to exist. Generalization to the non-continuous case is straightforward, completely analogous to handling the non-existence of an exact 95for a binomial variable, either with discrete approximation or randomization.

[6] The parametric formulation of hypotheses about the distribution of t can be relaxed to include the non-parametric case in the usual way (see Pagan and Ullah, 1999, for details).

[7] In the spirit of McCloskey-inspired specificity regarding the description of hypotheses, the proposed classification procedure is specified in terms of simple hypotheses, completely determining the distribution of t in each case. However, the logic of the theorems that follow is compatible with composite hypotheses as well, requiring only minor modifications.

Because there are two critical regions, the desired probability of Type II error β can be built into the construction of C_1 without changing the desired level of statistical significance α built into C_0. In other words, given the density of t and the four user-provided inputs θ_0, θ_1, α and β, NDC can be constructed to satisfy the constraints:

$$\alpha = P_{\theta_0}[t(\boldsymbol{X}) \in C_0], \tag{3}$$

$$\beta = P_{\theta_1}[t(\boldsymbol{X}) \in C_1]. \tag{4}$$

The probabilities of Type I and Type II errors, α and β, are referred to as false-rejection probabilities. Although the notation for statistical significance, α, is conventional, the notation here for the probability of rejecting the alternative hypothesis when the alternative is true, β, is not. Unlike the conventional hypothesis test, the power of NDC (i.e., the probability of rejecting H_0 when H_1 is true, $P_{\theta_1}[t(\boldsymbol{X}) \in C_0]$) is not equal to $1 - \beta$. Instead, after accounting for the probability of the no-decision outcome $(\overline{C_0 \cup C_1})$, NDC's power is given by:

$$P_{\theta_1}[t(\boldsymbol{X}) \in C_0] = 1 - \beta - P_{\theta_1}[t(x) \in \overline{C_0 \cup C_1}]. \tag{5}$$

Critics of the standard hypothesis test cite its automatic or "ritualized" implementation as a core methodological weakness. In contrast, this paper argues in favor of procedural automaticity and its transactions-cost-reducing benefits, provided that key analytical judgments are elicited and incorporated into the process. With NDC, the user provides a simple null, a simple alternative, and desired false-rejection probabilities α and β. With an established technique for constructing critical regions given these user-provided values, the procedure becomes automatic once those values are selected. This raises the question of how to construct the critical regions, since there is in general an infinite number of pairs of sets (C_0, C_1) satisfying the constraints (3) and (4). Fortunately, the question of how to construct critical regions has a straightforward answer described in the next section.

3.1. Neyman–Pearson construction of critical regions

The need for a method of constructing critical regions arises because, in general, choices of α and β do not uniquely determine C_0 and C_1. There are many ways of choosing C_0 and C_1 to satisfy (3) and (4).[8] The Neyman–Pearson construction defined below pins down the definitions of the critical regions and provides two key advantages. First, it greatly simplifies the description of the NDC procedure by mapping user-supplied values of α and β into unambiguous definitions of the sets C_0 and C_1. Therefore, rather than describing the desired NDC procedure as two sets, the pre-established method of construction allows the user to describe it with two numbers. Second, the Neyman–Pearson Construction extracts maximal decisiveness from the data by minimizing the chance of no decision.

Definition 2 (Neyman–Pearson construction of critical regions). Given a test statistic density function f in the monotone-likelihood-ratio class of densities,[9] false-rejection proba-

[8] To deal with composite hypotheses, the probabilities on the right hand sides of (3) and (4) would be replaced with the suprema of those probabilities taken with respect to values of θ contained, respectively, in H_0 and H_1.

[9] See Lehmann (1959) for a definition of the "monotone likelihood ratio" class of distributions, and a statement of the Neyman–Pearson Theorem. Lehman provides examples which show that many common distributions, including normal, chi-square and exponential, are included in the monotone-likelihood-ratio class.

bilities α and β, and simple hypotheses θ_0 and θ_1, *the Neyman–Pearson construction of* (C_0, C_1) is defined as:

$$C_0(d_0) = \{t \in \Omega_t | f_{\theta_0}(t) \le d_0 f_{\theta_1}(t)\}, \tag{6}$$

$$C_1(d_1) = \{t \in \Omega_t | d_1 f_{\theta_1}(t) \le f_{\theta_0}(t)\}, \tag{7}$$

where d_0 and d_1 are chosen to satisfy the constraints

$$\int_{C_0(d_0)} f_{\theta_0}(t)\mathrm{d}t = \alpha, \tag{8}$$

$$\int_{C_1(d_1)} f_{\theta_1}(t)\mathrm{d}t = \beta. \tag{9}$$

When t is scalar-valued and f is in the monotone-likelihood-ratio class of distributions, there exists a unique pair of numbers (d_0, d_1) satisfying (6)–(9). This follows from the Neyman–Pearson Theorem. Thus, the Neyman–Pearson construction provides an unambiguous mapping from the four user-provided values (α, β, θ_0 and θ_1) to the NDC procedure (C_0, C_1).

Theorem 1 (*Most decisive NDC*). *Given a density f in the monotone-likelihood-ratio class of distributions, two simple hypotheses* θ_0 *and* θ_1*, and false-rejection probabilities* α *and* β*, the NDC procedure defined by* Eqs. (6)–(9) *maximizes the chance of rejecting either* θ_0 *or* θ_1 *among all pairs of critical regions satisfying the false-rejection requirements* (3) and (4).

Proof. The Neyman–Pearson Theorem implies that:

$$P_{\theta_1}(C_0) \ge P_{\theta_1}(C'_0)\forall C'_0 \subset \Omega_t \text{such that} P_{\theta_0}(C'_0) = \alpha, \tag{10}$$

$$P_{\theta_0}(C_1) \ge P_{\theta_0}(C'_1)\forall C'_1 \subset \Omega_t \text{such that} P_{\theta_1}(C'_1) = \beta. \tag{11}$$

Because the critical regions are (by definition of NDC) non-overlapping, the probability of their union is the sum of their probabilities:

$$P_{\theta_0}(C_0 \cup C_1) = \alpha + P_{\theta_0}(C_1), \tag{12}$$

$$P_{\theta_1}(C_0 \cup C_1) = P_{\theta_1}(C_0) + \beta. \tag{13}$$

Eqs. (10) and (12) imply that, when θ_0 is true, no other critical regions satisfying the false-rejection probability requirements (3) and (4) lead to a larger probability of decision, $P_{\theta_0}(C_0 \cup C_1)$. Similarly, Eqs. (11) and (13) show that, when θ_1 is true, the Neyman–Pearson NDC again maximizes the probability of decision. Thus, regardless of the truth, NDC with critical regions constructed according to the Neyman–Pearson technique is maximally decisive. This completes the proof. □

3.2. *NDC leads to binary classification when critical regions overlap*

By definition, NDC critical regions do not overlap. However, when the user selects hypotheses that are relatively easy for the data to distinguish (while satisfying the false-rejection requirements), the critical regions may at first overlap. In this case, adjustments must be made before proceeding with NDC. Happily, the adjustments wind up working in the user's favor. Both false-rejection probability requirements are to be made more stringent without increasing the chance of no decision. An algorithm for implementing NDC when Neyman–Pearson critical regions initially overlap is presented below. The result of the algorithm is binary rather than three-outcome, no-decision classification. The underlying principle is that when the data are sufficiently decisive (e.g., large sample sizes or other conditions favoring low variance of the test statistic) there is no need for the no-decision region at all.

To illustrate, consider the problem of deciding which of the following two hypotheses is true:

$$X{\sim}N(0,1) \quad \text{versus} \quad X{\sim}N(10,1). \tag{14}$$

A single draw from X can distinguish which hypothesis is true with almost zero probability of either Type I or Type II error using the decision rule, "Take H_1 if $x \geq 5$, and H_0 otherwise." In contrast, the standard hypothesis test at the 5region $[1.645, \infty)$. The problem is that holding the probability of Type I error constant makes little sense. The critical point defining the endpoint of the critical region can be shifted to the right, reducing the chance of Type I error without noticeably sacrificing power. Unless Type I error is costless, the standard approach cannot be optimal.

Suppose instead the user chooses to implement NDC, attempting to distinguish the hypotheses above with false-rejection probabilities fixed at $\alpha = \beta = 0.05$. In this case, the critical regions are $C_0 = [1.645, \infty)$, and $C_1 = (-\infty, 8.335]$. These sets obviously overlap. By holding the ratio $\frac{\alpha}{\beta}$ constant while reducing α and β toward zero, the two critical regions shrink. Eventually when α and β are very close to zero (with $\alpha = \beta$ because their ratio is held constant at 1), the two critical regions become $C_0 = [5, \infty)$ and $C_1 = (-\infty, 5]$. This is the decision rule one derives using the algorithm below.

Theorem 2 (*Algorithm for binary classification when NDC critical regions overlap*). *Assume the test statistic t is a continuous random variable in the monotone-likelihood-ratio class with cdf F_θ. Suppose, too, that the user-provided simple hypotheses $\theta = \theta_0$ and $\theta = \theta_1$ and false-rejection probabilities α_0* and *β_0 lead to Neyman–Pearson critical regions that overlap. Then the following algorithm leads to an NDC with lower than required probabilities of false rejection and zero probability of the no-decision outcome (i.e., NDC becomes binary classification):*

- *Fix the ratio of initially-chosen false rejection probabilities at $\frac{\beta_0}{\alpha_0}$.*
- *Solve $F_{\theta_1}^{-1}(\frac{\beta_0}{\alpha_0}x) = F_{\theta_0}^{-1}(1-x)$ in x and denote the solution x^*.*
- *Set $\alpha = x^*$, $\beta = \frac{\beta_0}{\alpha_0}x^*$, and $c^* = F_{\theta_0}^{-1}(1-x^*)$. Then classify the data as "reject θ_0" if $t > c^*$ and "reject θ_1" otherwise.*

Proof. Given that the critical regions are the Neyman–Pearson type, and that *f* belongs to the monotone-likelihood-ratio class of distributions, critical regions are connected intervals which can, without loss of generality, be written:

$$C_0 = [u, \infty) \quad \text{and} \quad C_1 = (-\infty, l]. \tag{15}$$

□

Thus, critical regions overlap only if $u < l$. Implicit differentiation of the false-rejection probability requirements $1 - F_{\theta_0}(u) = \alpha$, and $F_{\theta_1}(l) = \beta$ shows that u is decreasing in α, and l is increasing in β:

$$\frac{du}{d\alpha} = -\frac{1}{f_{\theta_0}(u)} < 0 \quad \text{and} \quad \frac{\mathrm{d}l}{\mathrm{d}\beta} = \frac{1}{f_{\theta_1}(u)} > 0. \tag{16}$$

The goal, then, is to reduce α and β, keeping the ratio $\frac{\beta}{\alpha}$ fixed at $\frac{\beta_0}{\alpha_0}$ until the two critical regions are separated by a single point $u = l \equiv c^*$. Given that F is continuous, the critical point defining the boundary of the new critical regions exists and is given by the formula:

$$c^* \equiv F_{\theta_0}^{-1}(1 - x^*), \tag{17}$$

where x^* is the solution to

$$F_{\theta_1}^{-1}\left(\frac{\beta_0}{\alpha_0}x\right) = F_{\theta_0}^{-1}(1 - x), \tag{18}$$

which completes the proof.

Four simple examples of NDC are presented below. Example 1 is the standard case with non-overlapping critical regions. Example 2 demonstrates the algorithm from Theorem 2 for the overlapping case. Examples 3 and 4 are non-overlapping, providing formulas for NDC critical regions in the respective cases where the test statistic is normal and exponential.

3.3. Example 1

Suppose the data consist of a single draw from a unit-variance normal distribution X, and that the statistic t is identically $t \equiv X$. NDC is applied to determine which of two simple hypotheses regarding the mean (μ) of X is better supported by the data. Setting $\alpha = \beta = 0.05$, the goal is to classify X as either

$$H_0 : \mu = -1 \text{ or } H_0 : \mu = 1. \tag{19}$$

According to the Neyman–Pearson construction, the numbers d_0 and d_1 which define the critical regions are chosen so that the following two statements hold:

$$P_{\theta_0}\left[\frac{1}{(2\pi)^{0.5}}\mathrm{e}^{1/2(X+1)^2} \leq d_0\frac{1}{(2\pi)^{0.5}}\mathrm{e}^{-1/2(X-1)^2}\right] = 0.05, \tag{20}$$

$$P_{\theta_1}\left[d_1\frac{1}{(2\pi)^{0.5}}\mathrm{e}^{-1/2(X-1)^2} \leq \frac{1}{(2\pi)^{0.5}}\mathrm{e}^{-1/2(X+1)^2}\right] = 0.05. \tag{21}$$

These two equations are equivalent to choosing l and u to satisfy:

$$P_{\theta_0}(X > u) = 0.05 \quad \text{and} \quad P_{\theta_1}(X < l) = 0.05, \tag{22}$$

which leads to the critical regions

$$C_0 = [0.645, \infty) \quad \text{and} \quad C_1 = (-\infty, -0.645], \tag{23}$$

with no-decision region $(-0.645, 0.645)$. The two critical regions do not overlap and, thus, (C_0, C_1) satisfies the definition of an NDC procedure.

3.4. Example 2

Maintaining all other definitions from Example 1, Example 2 applies NDC to distinguish the following pair of hypotheses, which are farther apart and therefore easier to discriminate:

$$H_0 : \mu = -2 \quad \text{versus} \quad H_1 : \mu = 2. \tag{24}$$

In this case, the critical regions overlap:

$$C_0 = [-0.355, \infty), \quad C_1 = (-\infty, 0.355]. \tag{25}$$

The overlap means that X is sufficiently informative to make an unambiguous classification without any need for the no-decision outcome given the required false-rejection probabilities. The algorithm in Theorem 2 is therefore applied with $\frac{\beta_0}{\alpha_0} = \frac{0.05}{0.05}$. Imposing $P_{\mu=-2}(t > c) = P_{\mu=2}(t < c)$, and denoting the standard normal cdf $\Phi(.)$, one solves

$$1 - \Phi(c + 2) = \Phi(c - 2), \tag{26}$$

which has solution $c^* = 0$. The probability of error, whether $\theta = \theta_0$ or $\theta = \theta_1$, is

$$\alpha^* = 1 - F_{\mu=-2}(c^*) = F_{\mu=2}(c^*) = 0.0228. \tag{27}$$

Thus, after applying the algorithm in Theorem 2, NDC specializes to binary classification according to the decision rule, "$\mu = -2$ is rejected if $x > 0$, and $\mu = 2$ is rejected if $x < 0$," achieving lower false-rejection probabilities than required.

3.5. Example 3

Example 3 returns to the non-overlapping case described in Example 1, this time, doing away with the assumption of unit variance. Instead, mean μ and standard deviation σ (of the single observation X) are both unknown, equal to one of two possible values:

$$H_0 : (\mu, \sigma) = (\mu_0, \sigma_0) \quad \text{or} \quad H_1 : (\mu, \sigma) = (\mu_1, \sigma_1). \tag{28}$$

Without loss of generality, assume $\mu_0 < \mu_1$. Using the Neyman–Pearson construction, the critical regions are defined by a pair of (lower and upper) interval endpoints l and u such that

$$C_0 = [u, \infty) \quad \text{and} \quad C_1 = (-\infty, l]. \tag{29}$$

Given α and β, one solves for l and u as solutions to the equations:

$$1 - \Phi\left(\frac{u - \mu_0}{\sigma_0}\right) = \alpha, \tag{30}$$

$$\Phi\left(\frac{l - \mu_1}{\sigma_1}\right) = \beta. \tag{31}$$

NDC critical regions are thus described by the formulas:

$$l = \mu_1 + \sigma_1 \Phi^{-1}(\beta) \quad \text{and} \quad u = \mu_0 + \sigma_0 \Phi^{-1}(1 - \alpha). \tag{32}$$

3.6. Example 4

Example 4 is similar to Example 3, except that X is exponential rather than normal. As before, $t \equiv X$ is a single draw from an exponential distribution with unknown parameter θ. The classification problem is to determine which of two simple hypotheses is best supported by the data X:

$$H_0 : \theta = \theta_0 \quad \text{or} \quad H_1 : \theta = \theta_1. \tag{33}$$

Without loss of generality, assume $\theta_0 < \theta_1$. The Neyman–Pearson construction leads to critical regions of the form

$$C_0 = [u, \infty) \quad \text{and} \quad C_1 = (0, l]. \tag{34}$$

Using the exponential cdf $1 - \mathrm{e}^{-x/\theta}$, l and u are computed by imposing the equations:

$$P_{\theta_0}(X \geq u) = \mathrm{e}^{-u/\theta_0} = \alpha, \tag{35}$$

$$P_{\theta_1}(X \leq l) = 1 - \mathrm{e}^{-l/\theta_1} = \beta, \tag{36}$$

where α, $\beta \in (0, 1)$. Finally,

$$l = \theta_1 \log\left(\frac{1}{1-\beta}\right) \text{and} u = \theta_0 \log\left(\frac{1}{\alpha}\right). \tag{37}$$

3.7. No decision regions and sample size

Examples 3 and 4 provide explicit formulas for NDC critical regions as functions of the simple hypotheses and false-rejection probabilities. When the test statistic t is based on n observations instead of the single observation considered in the previous examples, the size of the critical regions also depends on n. This is particularly easy to see when t is the sample mean from a normally distributed population: $t = \sum_{i=1}^{n} X_i/n$. In this case, the no-decision region for t is:

$$\left[\mu_1 + \frac{\sigma_1}{n}\Phi^{-1}(\beta), \mu_0 + \frac{\sigma_0}{n}\Phi^{-1}(1-\alpha)\right]. \tag{38}$$

The no-decision region shrinks to an empty set and eventually becomes an improper interval (i.e., $u < l$) for large n because $\mu_0 < \mu_1$.

Similarly in the exponential case, the requirement $P_{\theta_0}(\sum_{i=1}^{n} X_j/n > u) = \alpha$ implicitly defines u by the equation $F_{\chi^2(2n)}(2u/\theta_0) = 1 - \alpha$, where $F_{\chi^2(2N)}$ is the chi-square cdf with $2N$ degrees of freedom. These formulas reveal that large sample sizes shrink the no-decision region, eventually leading to improper no-decision intervals unless α and β are made to depend on n.

4. Application 1: Do non-white workers earn less than similarly qualified whites?

Among the most common applications of significance testing is the comparison of expected earnings as a function of demographic traits such as race and gender. Denoting as

y_i the natural logarithm of individual i's annual income, a standard wage regression model can be written as:

$$y_i = \lambda' x_i + \delta d_i + \epsilon_i, \tag{39}$$

where x_i (including a constant) is a vector of i's productivity-related personal characteristics, and d_i is a binary measure of racial/ethnic status:

$$d_i = \begin{cases} 1 & \text{if } i \text{ is white} \\ 0 & \text{otherwise.} \end{cases} \tag{40}$$

Given this specification, the hypothesis of no race-based earnings differential is equivalent to $\mathrm{H}_0 : \delta = 0$.

Using a sample of 2473 full-time workers from the General Social Survey (GSS), Eq. (39) is estimated by OLS (with t statistics appearing below each estimated coefficient) as:

$$\begin{array}{lllll} \hat{y} = 9.54\,\text{CONSTANT} & +\ 0.38\,\text{MALE} & +\ 0.09\,\text{MARRIED} & +\ 0.46\,\text{COLLEGE} \\ \quad (310.42) & \quad 16.87 & \quad 4.02 & \quad 19.36 \\ & +\ 0.90\,\text{AGE} & + -0.74\,\text{AGE}^2 & +\ 0.05\,\text{WHITE} \\ & \quad 14.94 & \quad -12.23 & \quad 1.62 \end{array}$$

How should the estimated coefficient on the variable WHITE be interpreted? Its t statistic is strictly less than the one-sided 95% normal ordinate 1.645, but not by much.

In this case, the standard technique dictates that one report failure to reject the null (regardless of the margin between the test statistic and the critical value) and conclude that there is no evidence of a racial/ethnic earnings differential. Some may additionally, or instead, report the P-value $1 - \Phi(1.62) = 0.0521$, perhaps inviting the reader to give special consideration to the variable WHITE because it is "almost" significant, or because it is "significant at the 90demonstrates the problems described in earlier sections. Within the interpretive boundaries of the standard technique, either the null is rejected or not (necessity of choice). The findings have meaning only under the null (asymmetry with respect to Type I and Type II errors). And instead of context-driven consideration of the economic importance of magnitudes, the analysis centers on less important questions concerning the chance that sampling error could have generated the data were the null true (statistical significance trumps substantive significance).

NDC achieves improvements over the standard procedure with respect to each of these problems. With NDC, necessity of choice is no longer a necessity, because *no decision* is a valid outcome. With regard to Type I and Type II errors, symmetry is restored in the sense that there is no trade-off required between error probabilities—any degree of relative importance can be implemented through the selection of the false rejection probabilities α and β. Finally, the user-supplied simple hypotheses serve to elicit scientific judgment, making explicit users' beliefs about the size of departures from the null that matter. The cost of these improvements is having to defend one's specification of the simple alternative against which NDC is to have the desired power.

Continuing with analysis of the wage regression above, consider the economic significance of a hypothetical race-based earnings differential of \$500. I claim that, over the course of a year, an extra $500 will improve the economic well being of a typical worker in important ways that $50 cannot. $500 can make affordable a short vacation, a noticeably more stylish wardrobe, higher quality groceries, and other amenities that one may reasonably argue improve the economic well being of a worker. Upping the magnitude by a factor of 10 to $5000 clearly reaches the realm of economic significance, in the sense that the possibility of avoiding such a discrepancy would likely induce behavioral change (e.g., changing residences or switching professions) and rises to the level that many lawmakers would consider it a national policy priority. It stretches one's imagination to make similar claims for a differential of $50. Thus, $50 is too small, and the $500 differential is still conservative. Selecting $500 instead of $5000 makes it relatively more difficult for the data to be decisive, illustrating that the smaller the minimally significant departure from the null, the greater the chance of no decision.

To translate the $500 differential into a simple alternative hypothesis, the following equation must be solved for δ:

$$E[e^y|d=1] - E[e^y|d=0] = 500. \tag{41}$$

Solving (41) for δ, together with the assumption that the regression error ϵ_i is normally distributed with variance σ^2, leads to the simple alternative hypothesis $\mathrm{H}_1 : \delta = \delta_1$, where:

$$\delta_1 = \log\left(1 + \frac{500}{\mathrm{e}}^{\lambda'\bar{x}+\frac{1}{2}\sigma^2}\right) \tag{42}$$

Denoting the OLS estimator of δ and its standard error as $\hat{\delta}$ and $\mathrm{S.E.}_{\hat{\delta}}$, respectively, the ratio $\frac{\hat{\delta}-\delta_1}{\mathrm{S.E.}{>}_{\hat{\delta}}}$ has a standard normal asymptotic distribution under the alternative hypothesis, implying:

$$P_{\delta_1}\left(\frac{\hat{\delta}-\delta_1}{\mathrm{S.E.}_{\hat{\delta}}} < -1.645\right) = P_{\delta_1}\left(\frac{\hat{\delta}}{\mathrm{S.E.}_{\hat{\delta}}} < -1.645 + \frac{\delta_1}{\mathrm{S.E.}_{\hat{\delta}}}\right) = 0.05. \tag{43}$$

Thus, under the alternative hypothesis, the expression $-1.645 + \frac{\delta_1}{\mathrm{S.E.}_{\hat{\delta}}}$ gives a lower cutoff point for the t statistic $t \equiv \frac{\hat{\delta}}{\mathrm{S.E.}_{\hat{\delta}}}$.

Finally, NDC provides statistical decisions according to the following formula:

$$\text{reject } \delta = 0 \text{ if } t \in [1.645, \infty), \tag{44}$$

$$\text{reject } \delta = \delta_1 \text{ if } t \in \left(-\infty, -1.645 + \frac{\delta_1}{\mathrm{S.E.}_{\hat{\delta}}}\right] = (\infty, -1.008], \tag{45}$$

$$\text{and take no decision if } t \in (-1.008, 1.645). \tag{46}$$

Because the observed value $t = 1.62$ is in the interval $(-1.008, 1.645)$, NDC makes no decision in this application. The rather large size of the no-decision region is due to the large value of $\mathrm{S.E.}_{\hat{\delta}}$ and the small minimum significant difference $500.

One may use NDC to investigate the related question of how large a minimum significant difference would be needed to reach a statistical decision (i.e., rejection of one of the simple hypotheses). Straightforward algebraic calculations reveal that a minimum significant difference of \$2675 (or larger) would have led to rejection of H_0, since the following upper bound just exceeds the observed value of t:

$$-1.645+\log\left(1+\frac{2675}{e^{\hat{\delta}\bar{z}+\frac{1}{2}\hat{\sigma}^2}}\right)/\mathrm{S.E.}_{\hat{\delta}}=1.630. \tag{47}$$

Another variant of NDC yields Arrow's equal probability test as a special case. Were one to eschew the possibility of no decision and implement NDC with symmetric yet unspecified power (i.e., $\alpha=\beta$), the single critical point that partitions the real line into two critical regions would be:

$$c^*=\frac{\delta_1}{2\mathrm{S.E.}_{\hat{\delta}}}=0.3184. \tag{48}$$

This binary version of NDC, which features no possibility of no decision, rejects $\delta=0$ at the $1-\Phi(0.3184)=0.3751$ level.

5. Application 2: Is U.S. real GDP trend stationary?

Consider the question of permanent versus temporary fluctuations in macroeconomic variables. If output is non-stationary, then policy intervention today can have permanent benefits. But if recessions and rallies are driven entirely by temporary fluctuations about a stable long-run growth path, then policies aimed at controlling output may have a weaker rationale.

There are a variety of tests available for determining whether or not a time series is stationary or not (Leybourne and Newbold, 2000; Shively, 1988) from which conflicting conclusions have been drawn (Shively, 2001). Most tests for stationarity are conducted under the null hypothesis of non-stationarity and do not have sufficient power to detect nearby alternatives (i.e., variables that exhibit a high degree of persistence but are nevertheless stationary). The KPSS test is an exception in this regard Kwiatkowski et al. (1992), Charemza and Syczewska (1998). However, given the frequency with which "null-confirmation" methodology is used, the critiques of standard hypothesis testing still apply.

After removing a linear trend from the natural log of output, y, the autocorrelation coefficient ρ is estimated in the following model:

$$y_t=\rho y_{t-1}+u_t. \tag{49}$$

Next consider the one-sided test:

$$H_0:\rho=1 \quad \text{versus} \quad H_1\rho<1. \tag{50}$$

Using quarterly U.S. GDP data from 1947:1 through 2000:4, the least-squares estimate $\hat{\rho}$ is

$$\hat{\rho}=0.9939\,(0.0078), \tag{51}$$

with the standard error in parentheses. The corresponding t statistic is

$$\frac{\hat{\rho}-1}{\text{S.E.}_{\hat{\rho}}} = -0.7834. \tag{52}$$

The 0.05 level left-tail critical point for a sample of approximately this size is −1.95. Thus, according to the conventional procedure, the null hypothesis of non-stationarity is not rejected. However, this finding reveals little about the stationarity of the process due to the test's low power and highly asymmetric probabilities of Type I and Type II error.

To implement NDC, one must address the question of which simple alternative in the vicinity of non-stationarity should the data be required to distinguish. One approach is to use the time frame of 50 quarters (12.5 years) as a proxy for the long run. The assertion is that if the effect of a shock retains at least half its "oomph" (i.e., magnitude) 50 quarters hence, then it should be regarded as permanent for many practical applications. Other time horizons are, of course, possible and the issue deserves to be debated further. Adopting the 50-quarter half-life as one reasonable boundary between stationarity and non-stationarity, the simple alternative hypothesis becomes $\rho_1 = 0.9862$ (because $0.9862^{50} = 0.50$). Using the ratio $t \equiv \frac{\hat{\rho}-1}{\text{S.E.}_{\hat{\rho}}}$ as the test statistic, lower and upper bounds l and u which define the no-decision region are computed by imposing the conditions:

$$P_{\rho=1}(t < l) = \alpha \quad \text{and} \quad P_{\rho=0.9862}(t > u) = \beta. \tag{53}$$

At $\alpha = \beta = 0.05$, this yields

$$l = F^{-1}_{\rho=1}(0.05) = -1.95, \tag{54}$$

and

$$u = \Phi^{-1}(1-0.05) - \frac{1-0.9862}{\text{S.E.}_{\hat{\rho}}} = -0.1355. \tag{55}$$

(These computations use the approximation $F_{\rho=0.9862} \equiv \Phi(\frac{\hat{\rho}-0.9862}{\text{S.E.}_{\hat{\rho}}})$, since the ratio $\frac{\hat{\rho}-0.9862}{\text{S.E.}_{\hat{\rho}}}$ is asymptotically normal under $\rho = 0.9862$). Thus, the no-decision region is $(-1.95, -0.1355)$, which contains the realized test statistic $t = -0.7834$ and therefore indicates that no decision is to be taken.

If the binary-classification variant of NDC is desired, then the false-rejection probabilities are set equal to one another without specifying their exact values. This leads to a partition of the space of the test statistic with two critical regions and an empty no-decision set. The critical value c^* separating the critical regions satisfies:

$$F_{\rho=1}(c) = \Phi\left(c - \frac{0.9862-1}{\text{S.E.}_{\hat{\rho}}}\right). \tag{56}$$

Interpolating the appropriate tables for the Dickey–Fuller cdf $F_{\rho=1}$ in Hamilton (1994), one finds an approximate solution $c^* = -0.95$ with approximate level $\alpha = 0.20$. In other words, when the no-decision outcome is unacceptable yet one is committed to the symmetric treatment of Type I and Type II errors, stationarity can be rejected with 80% confidence, since $t = -.7834 > c^*$.

6. Conclusion

NDC is a quantitative tool for classifying data into one of three categories: reject the null hypothesis, reject the alternative hypothesis, or no decision, given the user-specified probabilities of Type I and Type II error. In contrast to the standard hypothesis test, NDC does not force an inference in favor of one hypothesis or the other, instead providing a neutral description of data that contain too little information for distinguishing between two theories. Another advantage is that NDC allows users to control the probabilities of both Type I and Type II error, unlike the standard hypothesis test which allows only for control over the probability of Type I error, with sample size and the shape of the data's distribution determining the test's power. Control over the probabilities of Type I and Type II error provides a means of comparing theories against data in accordance with the desired weights or a context-appropriate loss function. Perhaps most importantly, NDC incorporates users' judgments about the meaning of magnitudes—the question of how big is big. By inviting users to provide a simple alternative hypothesis representing the minimum departure from the null that is to be regarded as economically significant, NDC embeds economic significance into the procedural formalism of the standard hypothesis test. Thus, NDC represents a middle of the road trade-off between the transactions-cost-reducing benefits of standard statistical decision-making procedures and the judgment-intensive, context-specific analysis called for by critics of statistical significance.

Theorem 1 shows that the standard Neyman–Pearson construction of critical regions adapted for use with NDC leads to a statistical decision-making procedure which minimizes the chance of no decision. Theorem 2 provides an algorithm for dealing with data and hypotheses that give rise to overlapping critical regions. Overlapping critical regions require modifications to the user-supplied starting values before NDC can be implemented. Rather than a disadvantage, the overlapping case turns out to be beneficial, resulting from abundantly decisive data. Theorem 2 shows that in the overlapping case there exists an NDC procedure which rejects one of the hypotheses (i.e., arrives at a decision) with probability one while achieving lower than required probabilities of false rejection.

Applications of NDC demonstrate its capacity to reverse statistical conclusions regarding important empirical relationships derived from standard hypothesis testing. Thus, NDC's features amount to more than a mere extension of the standard procedure. Rather NDC generates distinct conclusions about the economy and, by extension, the desirability of different economic policies.

The applications and examples in this paper all involve two simple hypotheses supplied by the user. It was argued that this feature is a virtue because it requires users to reflect on and defend claims about the economic significance of magnitudes, not merely the signs of estimated parameters and the size of their t statistics. However, the simple hypothesis structure of NDC is not actually required in order to implement NDC. Settling the issue of whether simple versus composite specifications of the data distribution are more a virtue than a limitation will require further empirical applications of NDC. Based on these, the persuasiveness and replicability of various specifications of hypotheses can be evaluated. Convincingly defending the importance of particular pairs of simple hypotheses will depend crucially on insights that are specific to the economic meaning of the units of measurement in a given context. In general, what can be said is that there is a trade-off between context specificity and standardization in data analysis, and that NDC occupies an intermediate position along a spectrum defined by two poles—ritualized use of standard hypothesis testing on the one hand, and more informative but difficult-to-replicate descriptive analysis on the other. The argument for NDC is that it enjoys the virtues and avoids the drawbacks of both.

Acknowledgements

Thanks to Jim Church and the Department of Mathematics at the University of Kansas for advice in developing quantitative approaches to qualitative criticisms of the classical hypothesis test. The author also thanks Ulrich Hoffrage, Wim Vijverberg, and three anonymous referees for helpful criticism and feedback.

References

Andrews, D.W.K., 1994. The large sample correspondence between classical hypothesis tests and Bayesian posterior odds tests. Econometrica 62, 1207–1232.

Andrews, D.W.K., 1998. Hypothesis testing with a restricted parameter space. Journal of Econometrics 84, 155–199.

Arrow, K.J., 1959. Decision theory and the choice of a level of significance for the *t*-test. In: Olkin, I. (Ed.), Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling. Stanford University Press, Stanford 70–78.

Charemza, W.W., Syczewska, E.M., 1998. Joint application of the Dickey–Fuller and KPSS tests. Economics Letters 61, 17–21.

Christian, R.P., Hwang, G.J.T., Strawderman, W., 1993. Is Pitman closeness a reasonable criterion. Journal of the American Statistical Association 88, 57–63.

Coulibaly, N.B., Brorsen, W., 1999. Monte Carlo sampling approach to testing nonnested hypotheses: Monte Carlo results. Econometric Reviews 18, 195–209.

Denton, F.T., 1988. The significance of significance: rhetorical aspects of statistical hypothesis testing in economics. In: Klamer, A., McCloskey, D.N., Solow, R.M. (Eds.), The Consequences of Economic Rhetoric. Cambridge University Press, New York 163–183.

Dudewicz, E.J., Mishra, S.N., 1988. Modern Mathematical Statistics. Wiley, New York.

Elliott, G., Rothenberg, J., Stock, J.H., 1988. Efficient tests for an autoregressive unit root. Econometrica 64, 813–836.

Ericsson, N.R., 1974. Post-simulation analysis of Monte Carlo experiments: interpreting Pesaran's (1974) study of non-nested hypothesis test statistics. Review of Economic Studies 53 4, 691–707.

Gigerenzer, G., 2000. Adaptive Thinking: Rationality in the Real World. Oxford University Press, New York.

Gigerenzer, G., Swijtink, A., Porter, T., Daston, L., Beatty, J., Kruger, L., 1989. Empire of Chance. Cambridge University Press, Cambridge.

Godfrey, L.G., 1998. Tests of non-nested regression models: some results on small sample behavior and the bootstrap. Journal of Econometrics 84, 59–74.

Godfrey, L.G., Orme, C.D., 2000. Controlling the significance levels of prediction error tests for linear regression models. Econometrics Journal 3, 66–83.

Hamilton, J.D., 1994. Time Series Analysis. Princeton University Press, Princeton.

Harlow, L., Mulaik, S.A., Steiger, J.H. (Eds.), 1997. What If There Were No Significance Tests?. Erlbaum, Mahwah, NJ.

Hertwig, R., Todd, P.M., 2000. Biases to the left, fallacies to the right: stuck in the middle with null hypothesis significance testing. Psycoloquy 11, 11–28.

Horowitz, J.L., 2001. The bootstrap and hypothesis tests in econometrics. Journal of Econometrics 100, 37–40.

Jones, L.V., Tukey, J.W., 2000. A sensible formulation of the significance test. Psychological Methods 5, 411–414.

King, M.L., 1988. Towards a theory of point optimal testing. Econometric Reviews 6, 169–218.

Kwiatkowski, D., Phillips, P.C.B., Schmidt, P., Shin, Y., 1992. Testing the null hypothesis of stationarity against the alternative of a unit root. Journal of Econometrics 54, 159–178.

Lehmann, E.L., 1959. Testing Statistical Hypotheses. John Wiley, New York.

Leybourne, S.J., Newbold, P., 2000. Behaviour of the standard and symmetric Dickey–Fuller-type tests when there is a break under the null hypothesis. Econometrics Journal 3, 1–15.

McAleer, M., 1995. The significance of testing empirical non-nested models. Journal of Econometrics 67, 149–171.

McCloskey, D.N., 1985. The loss function has been mislaid: the rhetoric of significance tests. American Economic Review Papers and Proceedings 75, 201–205.

McCloskey, D.N., 1998. The Rhetoric of Economics, 2nd ed. University of Wisconsin Press, Madison.

Pagan, A., Ullah, A., 1999. Nonparametric Econometrics. Cambridge University Press, Cambridge.

Pesaran, M.H., 1974. On the general problem of model selection. Review of Economic Studies 41, 153–171.

Shively, P.A., 2001. Trend-stationary GNP: evidence from a new exact pointwise most powerful invariant unit root test. Journal of Applied Econometrics 16, 537–552.

Shively, T.S., 1988. An analysis of tests for regression coefficient stability. Journal of Econometrics 39, 367–386.

Smith, M.A., Smyth, D.J., 1991. Multiple and pairwise non-nested tests of the influence of taxes on money demand. Journal of Applied Econometrics 6, 17–30.

Srivastava, R.P., 1997. Analytical Modeling of Multiple Hypotheses Evaluation in Auditing. Working Paper, University of Kansas.

Terasvirta, T., 1996. Power properties of linearity tests for time series. Studies in Nonlinear Dynamics and Econometrics 1, 3–10.

Wald, A., 1947. Sequential Analysis. Wiley, New York.

Wu, D.M., 1985. The Modified Significance Test. Working Paper, University of Kansas.

Ziliak, S.T., McCloskey, D.N., 2005. Size matters: the standard error of regressions in the American Economic Review. Journal of Socio-Economics (this issue).