



ELSEVIER

The Journal of Socio-Economics xxx (2004) xxx–xxx

**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.

© 2004 Published by Elsevier Inc.

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.

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 technique²) 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.

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

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

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

96 Ziliak and McCloskey’s suggestion to rely and report more on judgment points to a
97 potentially difficult trade-off, however, between efficiency in the use of information and
98 efficiency in the communication of statistical decisions. On the one hand, the significance

99 test wastes information by not including expert judgments about the meaning of magnitudes
100 and the shape of the loss function (i.e., the relative importance of Type I versus Type II
101 errors). On the other hand, the procedural simplicity of the significance test facilitates a
102 non-negligible degree of efficiency in communication that also should be acknowledged.
103 For instance, one thinks of the time and concentration required to read several pages of
104 descriptive text interpreting estimated parameters from various regression models compared
105 with the ease of examining a table of starred coefficients. Similarly, one can appreciate
106 the transactions-cost-reducing value of statistical significance as a useful form of jargon,
107 evident, for example, in the simplicity of arguing against a theory by reporting rejection of
108 a parameter restriction at the five percent level.

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

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

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

137 2. Background

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

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

152 Among the critiques of significance testing, perhaps the best known is that of McCloskey
153 (1985, 1998). She points out that by relying on statistical significance to answer the question
154 of how big is big, researchers abdicate an important scientific responsibility. McCloskey's
155 claim is not merely that the choice of significance level (e.g., $\alpha = .05$) is arbitrary. The real
156 issue is the importance of thinking through the relative costs of being wrong as a function
157 of what is actually true. That means specifying a loss function which reflects the context
158 of each particular problem. In addition to urging more thorough consideration of power
159 and loss functions, McCloskey asks economists to be more bold in ascribing meaning to
160 magnitudes, taking a clear position on which ones deserve to be referred to as "significant"
161 in the substantive sense.

162 There have been several attempts to incorporate these qualitative criticisms of hypothesis
163 testing into an improved statistical decision-making procedure. Arrow (1959) proposes
164 an "equal probability" test that treats Type I and Type II errors symmetrically. Selecting
165 simple⁴ null and alternative hypotheses, and imposing equality of Type I and Type II error
166 probabilities, Arrow's test allows the magnitude of the error probability (equal to statistical
167 significance) to vary freely within the unit interval, serving as an index of quality associated
168 with inferences based upon it. Arrow illustrates the advantages of the equal probability
169 approach by demonstrating undesirable asymmetries that arise in interpreting regression
170 coefficients and their t statistics. In small to moderate sized samples, Arrow (1959, p.
171 73) shows that the t test's power can be close to zero, making it difficult to interpret "failure to
172 reject the null." On the other hand, in very large samples, power is close to 1 and appreciable
173 reductions in the probability of Type I error are possible with almost no loss of power.
174 Arrow observes that asymmetric treatment of Type I and Type II errors is, in practice, rarely
175 motivated by careful consideration of those errors' relative costs, as would be the case when
176 using an explicit loss function and decision-theoretic methodology.

177 Jones and Tukey (2000) discuss a modified test of significance designed to free the
178 analyst from the necessity of choice by allowing for an "indeterminacy" outcome similar
179 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.

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

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

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

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

216 The illogic of the hypothesis test’s inflexible prioritization of statistical significance over
217 the test’s power is especially obvious in the context of audit tasks in accounting (Srivastava,
218 1997), where Type II errors (undetected cheats) are typically much more serious than Type
219 I errors (false alarms). In the analysis of data collected from psychological experiments,
220 Hertwig and Todd (2000) argue that the standard test’s inherent asymmetry permits re-
221 searchers to escape from having to fully specify alternative theories. Their analysis describes
222 an unfortunate symbiosis between the hypothesis test’s asymmetry and researchers’ fail-
223 ure to elaborate precise alternative explanations of why null hypotheses may be inadequate.
224 Gigerenzer (2000) provides numerous examples in which the asymmetry of hypothesis test-

ing, rather than any particular characteristic of the data, virtually determines the conclusions that are drawn.

3. The no-decision classification (NDC) procedure

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

Let t denote the test statistic, a mapping of X into Ω_t :

$$t : \Omega_X \rightarrow \Omega_t. \quad (1)$$

The familiar case is when X is a random sample of length n , and the statistic t is a scalar (e.g., the sample mean). Denote the pdf of $t(X)$ 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

$$H_0 : \theta = \theta_0, \quad H_1 : \theta = \theta_1, \quad (\theta_0 \neq \theta_1). \quad (2)$$

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 \bar{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 to 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 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 95% for 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.

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

$$256 \quad \alpha = P_{\theta_0}[t(\mathbf{X}) \in C_0], \quad (3)$$

$$257 \quad \beta = P_{\theta_1}[t(\mathbf{X}) \in C_1]. \quad (4)$$

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

$$264 \quad P_{\theta_1}[t(\mathbf{X}) \in C_0] = 1 - \beta - P_{\theta_1}[t(x) \in \overline{C_0 \cup C_1}]. \quad (5)$$

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

275 3.1. Neyman–Pearson construction of critical regions

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

285 **Definition 2** (Neyman–Pearson construction of critical regions). Given a test statistic den-
 286 sity 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.

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

289
$$C_0(d_0) = \{t \in \Omega_t | f_{\theta_0}(t) \leq d_0 f_{\theta_1}(t)\}, \tag{6}$$

290
$$C_1(d_1) = \{t \in \Omega_t | d_1 f_{\theta_1}(t) \leq f_{\theta_0}(t)\}, \tag{7}$$

291 where d_0 and d_1 are chosen to satisfy the constraints

292
$$\int_{C_0(d_0)} f_{\theta_0}(t) dt = \alpha, \tag{8}$$

293
$$\int_{C_1(d_1)} f_{\theta_1}(t) dt = \beta. \tag{9}$$

294 When t is scalar-valued and f is in the monotone-likelihood-ratio class of distributions,
 295 there exists a unique pair of numbers (d_0, d_1) satisfying (6)–(9). This follows from the
 296 Neyman–Pearson Theorem. Thus, the Neyman–Pearson construction provides an unam-
 297 biguous mapping from the four user-provided values $(\alpha, \beta, \theta_0$ and $\theta_1)$ to the NDC procedure
 298 (C_0, C_1) .

299 **Theorem 1** (*Most decisive NDC*). *Given a density f in the monotone-likelihood-ratio class*
 300 *of distributions, two simple hypotheses θ_0 and θ_1 , and false-rejection probabilities α and*
 301 *β , the NDC procedure defined by Eqs. (6)–(9) maximizes the chance of rejecting either θ_0*
 302 *or θ_1 among all pairs of critical regions satisfying the false-rejection requirements (3) and*
 303 *(4).*

304 **Proof.** The Neyman–Pearson Theorem implies that:

305
$$P_{\theta_1}(C_0) \geq P_{\theta_1}(C'_0) \forall C'_0 \subset \Omega_t \text{ such that } P_{\theta_0}(C'_0) = \alpha, \tag{10}$$

306
$$P_{\theta_0}(C_1) \geq P_{\theta_0}(C'_1) \forall C'_1 \subset \Omega_t \text{ such that } P_{\theta_1}(C'_1) = \beta. \tag{11}$$

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

309
$$P_{\theta_0}(C_0 \cup C_1) = \alpha + P_{\theta_0}(C_1), \tag{12}$$

310
$$P_{\theta_1}(C_0 \cup C_1) = P_{\theta_1}(C_0) + \beta. \tag{13}$$

311 Eqs. (10) and (12) imply that, when θ_0 is true, no other critical regions satisfying the
 312 false-rejection probability requirements (3) and (4) lead to a larger probability of decision,
 313 $P_{\theta_0}(C_0 \cup C_1)$. Similarly, Eqs. (11) and (13) show that, when θ_1 is true, the Neyman–Pearson
 314 NDC again maximizes the probability of decision. Thus, regardless of the truth, NDC
 315 with critical regions constructed according to the Neyman–Pearson technique is maximally
 316 decisive. This completes the proof. \square

317 **3.2. NDC leads to binary classification when critical regions overlap**

318 By definition, NDC critical regions do not overlap. However, when the user selects
 319 hypotheses that are relatively easy for the data to distinguish (while satisfying the false-
 320 rejection requirements), the critical regions may at first overlap. In this case, adjustments

321 must be made before proceeding with NDC. Happily, the adjustments wind up working in
 322 the user's favor. Both false-rejection probability requirements are to be made more stringent
 323 without increasing the chance of no decision. An algorithm for implementing NDC when
 324 Neyman–Pearson critical regions initially overlap is presented below. The result of the algo-
 325 rithm is binary rather than three-outcome, no-decision classification. The underlying princi-
 326 ple is that when the data are sufficiently decisive (e.g., large sample sizes or other conditions
 327 favoring low variance of the test statistic) there is no need for the no-decision region at all.

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

$$330 \quad X \sim N(0, 1) \quad \text{versus} \quad X \sim N(10, 1). \quad (14)$$

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

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

345 **Theorem 2** (*Algorithm for binary classification when NDC critical regions overlap*).
 346 Assume the test statistic t is a continuous random variable in the monotone-likelihood-
 347 ratio class with cdf F_θ . Suppose, too, that the user-provided simple hypotheses $\theta = \theta_0$ and
 348 $\theta = \theta_1$ and false-rejection probabilities α_0 and β_0 lead to Neyman–Pearson critical regions
 349 that overlap. Then the following algorithm leads to an NDC with lower than required
 350 probabilities of false rejection and zero probability of the no-decision outcome (i.e., NDC
 351 becomes binary classification):

- 352 • Fix the ratio of initially-chosen false rejection probabilities at $\frac{\beta_0}{\alpha_0}$.
- 353 • Solve $F_{\theta_1}^{-1}(\frac{\beta_0}{\alpha_0}x) = F_{\theta_0}^{-1}(1 - x)$ in x and denote the solution x^* .
- 354 • 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
 355 $t > c^*$ and “reject θ_1 ” otherwise.

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

$$359 \quad C_0 = [u, \infty) \quad \text{and} \quad C_1 = (-\infty, l]. \quad (15)$$

360

□

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

$$364 \quad \frac{du}{d\alpha} = -\frac{1}{f_{\theta_0}(u)} < 0 \quad \text{and} \quad \frac{dl}{d\beta} = \frac{1}{f_{\theta_1}(u)} > 0. \quad (16)$$

365 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
 366 regions are separated by a single point $u = l \equiv c^*$. Given that F is continuous, the critical
 367 point defining the boundary of the new critical regions exists and is given by the formula:

$$368 \quad c^* \equiv F_{\theta_0}^{-1}(1 - x^*), \quad (17)$$

369 where x^* is the solution to

$$370 \quad F_{\theta_1}^{-1}\left(\frac{\beta_0}{\alpha_0}x\right) = F_{\theta_0}^{-1}(1 - x), \quad (18)$$

371 which completes the proof.

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

376 3.3. Example 1

377 Suppose the data consist of a single draw from a unit-variance normal distribution X ,
 378 and that the statistic t is identically $t \equiv X$. NDC is applied to determine which of two
 379 simple hypotheses regarding the mean (μ) of X is better supported by the data. Setting
 380 $\alpha = \beta = 0.05$, the goal is to classify X as either

$$381 \quad H_0 : \mu = -1 \text{ or } H_0 : \mu = 1. \quad (19)$$

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

$$384 \quad P_{\theta_0} \left[\frac{1}{(2\pi)^{0.5}} e^{1/2(X+1)^2} \leq d_0 \frac{1}{(2\pi)^{0.5}} e^{-1/2(X-1)^2} \right] = 0.05, \quad (20)$$

$$385 \quad P_{\theta_1} \left[d_1 \frac{1}{(2\pi)^{0.5}} e^{-1/2(X-1)^2} \leq \frac{1}{(2\pi)^{0.5}} e^{-1/2(X+1)^2} \right] = 0.05. \quad (21)$$

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

$$387 \quad P_{\theta_0}(X > u) = 0.05 \quad \text{and} \quad P_{\theta_1}(X < l) = 0.05, \quad (22)$$

388 which leads to the critical regions

$$389 \quad C_0 = [0.645, \infty) \quad \text{and} \quad C_1 = (-\infty, -0.645], \quad (23)$$

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

392 3.4. Example 2

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

$$395 H_0 : \mu = -2 \quad \text{versus} \quad H_1 : \mu = 2. \quad (24)$$

396 In this case, the critical regions overlap:

$$397 C_0 = [-0.355, \infty), \quad C_1 = (-\infty, 0.355]. \quad (25)$$

398 The overlap means that X is sufficiently informative to make an unambiguous classi-
 399 fication without any need for the no-decision outcome given the required false-rejection
 400 probabilities. The algorithm in Theorem 2 is therefore applied with $\frac{\beta_0}{\alpha_0} = \frac{0.05}{0.05}$. Imposing
 401 $P_{\mu=-2}(t > c) = P_{\mu=2}(t < c)$, and denoting the standard normal cdf $\Phi(\cdot)$, one solves

$$402 1 - \Phi(c + 2) = \Phi(c - 2), \quad (26)$$

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

$$404 \alpha^* = 1 - F_{\mu=-2}(c^*) = F_{\mu=2}(c^*) = 0.0228. \quad (27)$$

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

408 3.5. Example 3

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

$$412 H_0 : (\mu, \sigma) = (\mu_0, \sigma_0) \quad \text{or} \quad H_1 : (\mu, \sigma) = (\mu_1, \sigma_1). \quad (28)$$

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

$$416 C_0 = [u, \infty) \quad \text{and} \quad C_1 = (-\infty, l]. \quad (29)$$

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

$$418 1 - \Phi\left(\frac{u - \mu_0}{\sigma_0}\right) = \alpha, \quad (30)$$

$$419 \Phi\left(\frac{l - \mu_1}{\sigma_1}\right) = \beta. \quad (31)$$

420 NDC critical regions are thus described by the formulas:

$$421 l = \mu_1 + \sigma_1 \Phi^{-1}(\beta) \quad \text{and} \quad u = \mu_0 + \sigma_0 \Phi^{-1}(1 - \alpha). \quad (32)$$

422 **3.6. Example 4**

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

427
$$H_0 : \theta = \theta_0 \quad \text{or} \quad H_1 : \theta = \theta_1. \tag{33}$$

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

430
$$C_0 = [u, \infty) \quad \text{and} \quad C_1 = (0, l]. \tag{34}$$

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

432
$$P_{\theta_0}(X \geq u) = e^{-u/\theta_0} = \alpha, \tag{35}$$

433
$$P_{\theta_1}(X \leq l) = 1 - e^{-l/\theta_1} = \beta, \tag{36}$$

434 where $\alpha, \beta \in (0, 1)$. Finally,

435
$$l = \theta_1 \log \left(\frac{1}{1 - \beta} \right) \quad \text{and} \quad u = \theta_0 \log \left(\frac{1}{\alpha} \right). \tag{37}$$

436 **3.7. No decision regions and sample size**

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

443
$$\left[\mu_1 + \frac{\sigma_1}{n} \Phi^{-1}(\beta), \mu_0 + \frac{\sigma_0}{n} \Phi^{-1}(1 - \alpha) \right]. \tag{38}$$

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

446 Similarly in the exponential case, the requirement $P_{\theta_0}(\sum_{i=1}^n X_j/n > u) = \alpha$ implicitly
 447 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
 448 $2N$ degrees of freedom. These formulas reveal that large sample sizes shrink the no-decision
 449 region, eventually leading to improper no-decision intervals unless α and β are made to
 450 depend on n .

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

452 Among the most common applications of significance testing is the comparison of ex-
 453 pected earnings as a function of demographic traits such as race and gender. Denoting as

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

$$456 \quad y_i = \lambda' x_i + \delta d_i + \epsilon_i, \tag{39}$$

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

$$459 \quad d_i = \begin{cases} 1 & \text{if } i \text{ is white} \\ 0 & \text{otherwise.} \end{cases} \tag{40}$$

460 Given this specification, the hypothesis of no race-based earnings differential is equivalent
 461 to $H_0 : \delta = 0$.

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

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

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

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

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

488 Continuing with analysis of the wage regression above, consider the economic
 489 significance of a hypothetical race-based earnings differential of \$500. I claim that, over
 490

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

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

$$505 \quad E[e^y | d = 1] - E[e^y | d = 0] = 500. \tag{41}$$

506 Solving (41) for δ , together with the assumption that the regression error ϵ_i is normally
 507 distributed with variance σ^2 , leads to the simple alternative hypothesis $H_1 : \delta = \delta_1$, where:

$$508 \quad \delta_1 = \log \left(1 + \frac{500^{\lambda' \bar{x} + \frac{1}{2} \sigma^2}}{e} \right) \tag{42}$$

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

$$512 \quad P_{\delta_1} \left(\frac{\hat{\delta} - \delta_1}{S.E._{\hat{\delta}}} < -1.645 \right) = P_{\delta_1} \left(\frac{\hat{\delta}}{S.E._{\hat{\delta}}} < -1.645 + \frac{\delta_1}{S.E._{\hat{\delta}}} \right) = 0.05. \tag{43}$$

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

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

$$516 \quad \text{reject } \delta = 0 \text{ if } t \in [1.645, \infty), \tag{44}$$

$$517 \quad \text{reject } \delta = \delta_1 \text{ if } t \in \left(-\infty, -1.645 + \frac{\delta_1}{S.E._{\hat{\delta}}} \right] = (\infty, -1.008], \tag{45}$$

$$518 \quad \text{and take no decision if } t \in (-1.008, 1.645). \tag{46}$$

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

522 One may use NDC to investigate the related question of how large a minimum significant
 523 difference would be needed to reach a statistical decision (i.e., rejection of one of the simple
 524 hypotheses). Straightforward algebraic calculations reveal that a minimum significant

525 difference of \$2675 (or larger) would have led to rejection of H_0 , since the following upper
526 bound just exceeds the observed value of t :

$$527 \quad -1.645 + \log \left(1 + \frac{2675}{e^{\hat{\delta}\bar{z} + \frac{1}{2}\sigma^2}} \right) / \text{S.E.}_{\hat{\delta}} = 1.630. \quad (47)$$

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

$$532 \quad c^* = \frac{\delta_1}{2\text{S.E.}_{\hat{\delta}}} = 0.3184. \quad (48)$$

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

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

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

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

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

$$551 \quad y_t = \rho y_{t-1} + u_t. \quad (49)$$

552 Next consider the one-sided test:

$$553 \quad H_0 : \rho = 1 \quad \text{versus} \quad H_1 \rho < 1. \quad (50)$$

554 Using quarterly U.S. GDP data from 1947:1 through 2000:4, the least-squares estimate
555 $\hat{\rho}$ is

$$556 \quad \hat{\rho} = 0.9939 (0.0078), \quad (51)$$

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

$$558 \quad \frac{\hat{\rho} - 1}{\text{S.E.}_{\hat{\rho}}} = -0.7834. \quad (52)$$

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

563 To implement NDC, one must address the question of which simple alternative in the
564 vicinity of non-stationarity should the data be required to distinguish. One approach is to
565 use the time frame of 50 quarters (12.5 years) as a proxy for the long run. The assertion is
566 that if the effect of a shock retains at least half its "oomph" (i.e., magnitude) 50 quarters
567 hence, then it should be regarded as permanent for many practical applications. Other time
568 horizons are, of course, possible and the issue deserves to be debated further. Adopting the
569 50-quarter half-life as one reasonable boundary between stationarity and non-stationarity,
570 the simple alternative hypothesis becomes $\rho_1 = 0.9862$ (because $0.9862^{50} = 0.50$). Using
571 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
572 no-decision region are computed by imposing the conditions:

$$573 \quad P_{\rho=1}(t < l) = \alpha \quad \text{and} \quad P_{\rho=0.9862}(t > u) = \beta. \quad (53)$$

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

$$575 \quad l = F_{\rho=1}^{-1}(0.05) = -1.95, \quad (54)$$

576 and

$$577 \quad u = \Phi^{-1}(1 - 0.05) - \frac{1 - 0.9862}{\text{S.E.}_{\hat{\rho}}} = -0.1355. \quad (55)$$

578 (These computations use the approximation $F_{\rho=0.9862} \equiv \Phi\left(\frac{\hat{\rho}-0.9862}{\text{S.E.}_{\hat{\rho}}}\right)$, since the ratio
579 $\frac{\hat{\rho}-0.9862}{\text{S.E.}_{\hat{\rho}}}$ is asymptotically normal under $\rho = 0.9862$). Thus, the no-decision region is
580 $(-1.95, -0.1355)$, which contains the realized test statistic $t = -0.7834$ and therefore
581 indicates that no decision is to be taken.

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

$$586 \quad F_{\rho=1}(c) = \Phi\left(c - \frac{0.9862 - 1}{\text{S.E.}_{\hat{\rho}}}\right). \quad (56)$$

587 Interpolating the appropriate tables for the Dickey–Fuller cdf $F_{\rho=1}$ in [Hamilton \(1994\)](#),
588 one finds an approximate solution $c^* = -0.95$ with approximate level $\alpha = 0.20$. In other
589 words, when the no-decision outcome is unacceptable yet one is committed to the symmetric
590 treatment of Type I and Type II errors, stationarity can be rejected with 80% confidence,
591 since $t = -0.7834 > c^*$.

592 6. Conclusion

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

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

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

625 The applications and examples in this paper all involve two simple hypotheses supplied
626 by the user. It was argued that this feature is a virtue because it requires users to reflect
627 on and defend claims about the economic significance of magnitudes, not merely the signs
628 of estimated parameters and the size of their t statistics. However, the simple hypothesis
629 structure of NDC is not actually required in order to implement NDC. Settling the issue of
630 whether simple versus composite specifications of the data distribution are more a virtue
631 than a limitation will require further empirical applications of NDC. Based on these, the
632 persuasiveness and replicability of various specifications of hypotheses can be evaluated.
633 Convincingly defending the importance of particular pairs of simple hypotheses will depend
634 crucially on insights that are specific to the economic meaning of the units of measurement
635 in a given context. In general, what can be said is that there is a trade-off between context

636 specificity and standardization in data analysis, and that NDC occupies an intermediate
637 position along a spectrum defined by two poles—ritualized use of standard hypothesis
638 testing on the one hand, and more informative but difficult-to-replicate descriptive analysis
639 on the other. The argument for NDC is that it enjoys the virtues and avoids the drawbacks
640 of both.

641 Acknowledgements

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

646 References

- 647 Andrews, D.W.K., 1994. The large sample correspondence between classical hypothesis tests and Bayesian pos-
648 terior odds tests. *Econometrica* 62, 1207–1232.
- 649 Andrews, D.W.K., 1998. Hypothesis testing with a restricted parameter space. *Journal of Econometrics* 84, 155–
650 199.
- 651 Arrow, K.J., 1959. Decision theory and the choice of a level of significance for the *t*-test. In: Olkin, I. (Ed.),
652 Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling. Stanford University Press,
653 Stanford 70–78.
- 654 Charemza, W.W., Syczewska, E.M., 1998. Joint application of the Dickey–Fuller and KPSS tests. *Economics*
655 Letters 61, 17–21.
- 656 Christian, R.P., Hwang, G.J.T., Strawderman, W., 1993. Is Pitman closeness a reasonable criterion. *Journal of the*
657 American Statistical Association 88, 57–63.
- 658 Coulibaly, N.B., Brorsen, W., 1999. Monte Carlo sampling approach to testing nonnested hypotheses: Monte Carlo
659 results. *Econometric Reviews* 18, 195–209.
- 660 Denton, F.T., 1988. The significance of significance: rhetorical aspects of statistical hypothesis testing in economics.
661 In: Klamer, A., McCloskey, D.N., Solow, R.M. (Eds.), *The Consequences of Economic Rhetoric*. Cambridge
662 University Press, New York 163–183.
- 663 Dudewicz, E.J., Mishra, S.N., 1988. *Modern Mathematical Statistics*. Wiley, New York.
- 664 Elliott, G., Rothenberg, J., Stock, J.H., 1988. Efficient tests for an autoregressive unit root. *Econometrica* 64,
665 813–836.
- 666 Ericsson, N.R., 1974. Post-simulation analysis of Monte Carlo experiments: interpreting Pesaran's (1974) study
667 of non-nested hypothesis test statistics. *Review of Economic Studies* 53 4, 691–707.
- 668 Gigerenzer, G., 2000. *Adaptive Thinking: Rationality in the Real World*. Oxford University Press, New York.
- 669 Gigerenzer, G., Swijtink, A., Porter, T., Daston, L., Beatty, J., Kruger, L., 1989. *Empire of Chance*. Cambridge
670 University Press, Cambridge.
- 671 Godfrey, L.G., 1998. Tests of non-nested regression models: some results on small sample behavior and the
672 bootstrap. *Journal of Econometrics* 84, 59–74.
- 673 Godfrey, L.G., Orme, C.D., 2000. Controlling the significance levels of prediction error tests for linear regression
674 models. *Econometrics Journal* 3, 66–83.
- 675 Hamilton, J.D., 1994. *Time Series Analysis*. Princeton University Press, Princeton.
- 676 Harlow, L., Mulaik, S.A., Steiger, J.H. (Eds.), 1997. *What If There Were No Significance Tests?*. Erlbaum, Mahwah,
677 NJ.
- 678 Hertwig, R., Todd, P.M., 2000. Biases to the left, fallacies to the right: stuck in the middle with null hypothesis
679 significance testing. *Psychology* 11, 11–28.

- 680 Horowitz, J.L., 2001. The bootstrap and hypothesis tests in econometrics. *Journal of Econometrics* 100, 37–40.
- 681 Jones, L.V., Tukey, J.W., 2000. A sensible formulation of the significance test. *Psychological Methods* 5, 411–414.
- 682 King, M.L., 1988. Towards a theory of point optimal testing. *Econometric Reviews* 6, 169–218.
- 683 Kwiatkowski, D., Phillips, P.C.B., Schmidt, P., Shin, Y., 1992. Testing the null hypothesis of stationarity against
684 the alternative of a unit root. *Journal of Econometrics* 54, 159–178.
- 685 Lehmann, E.L., 1959. *Testing Statistical Hypotheses*. John Wiley, New York.
- 686 Leybourne, S.J., Newbold, P., 2000. Behaviour of the standard and symmetric Dickey–Fuller-type tests when there
687 is a break under the null hypothesis. *Econometrics Journal* 3, 1–15.
- 688 McAleer, M., 1995. The significance of testing empirical non-nested models. *Journal of Econometrics* 67, 149–171.
- 689 McCloskey, D.N., 1985. The loss function has been misled: the rhetoric of significance tests. *American Economic*
690 *Review Papers and Proceedings* 75, 201–205.
- 691 McCloskey, D.N., 1998. *The Rhetoric of Economics*, 2nd ed. University of Wisconsin Press, Madison.
- 692 Pagan, A., Ullah, A., 1999. *Nonparametric Econometrics*. Cambridge University Press, Cambridge.
- 693 Pesaran, M.H., 1974. On the general problem of model selection. *Review of Economic Studies* 41, 153–171.
- 694 Shively, P.A., 2001. Trend-stationary GNP: evidence from a new exact pointwise most powerful invariant unit root
695 test. *Journal of Applied Econometrics* 16, 537–552.
- 696 Shively, T.S., 1988. An analysis of tests for regression coefficient stability. *Journal of Econometrics* 39, 367–386.
- 697 Smith, M.A., Smyth, D.J., 1991. Multiple and pairwise non-nested tests of the influence of taxes on money demand.
698 *Journal of Applied Econometrics* 6, 17–30.
- 699 Srivastava, R.P., 1997. *Analytical Modeling of Multiple Hypotheses Evaluation in Auditing*. Working Paper,
700 University of Kansas.
- 701 Terasvirta, T., 1996. Power properties of linearity tests for time series. *Studies in Nonlinear Dynamics and Econo-*
702 *metrics* 1, 3–10.
- 703 Wald, A., 1947. *Sequential Analysis*. Wiley, New York.
- 704 Wu, D.M., 1985. *The Modified Significance Test*. Working Paper, University of Kansas.
- 705 Ziliak, S.T., McCloskey, D.N., 2005. Size matters: the standard error of regressions in the American Economic
706 *Review*. *Journal of Socio-Economics* (this issue).