Nonparametric Combination (NPC): A Framework for Testing Elaborate Theories

Devin Caughey
MIT

Allan Dafoe
Yale

Jason Seawright
Northwestern

February 27, 2016

Abstract

Social scientists are commonly advised to deduce and test all observable implications of their theories. We describe a principled framework for testing such "elaborate" theories: nonparametric combination. NPC assesses the joint probability of observing the theoretically predicted pattern of results under the sharp null of no effects. NPC accounts for the dependence among the component tests without relying on modeling assumptions or asymptotic approximations. Multiple-testing corrections are also easily implemented with NPC. As we demonstrate with four applications, NPC leverages theoretical knowledge into greater statistical power, which is particularly valuable for studies with strong research designs but small sample sizes. We implement these methods in a new R package, `NPC`.

The authors can be reached at `www.devincaughey.com`, `www.allandafoe.com`, `sites.google.com/site/jnsacademic/`. We thank Jake Bowers, Bear Braumoeller, Thad Dunning, Danny Hidalgo, Adrienne Hosek, Scott Gates, Sara Newland, Kristopher Ramsay, Luigi Salmaso, Jas Sekhon, Guadalupe Tuñón, and Teppei Yamamoto for comments and input. Devin Caughey is an Assistant Professor at MIT, Cambridge, MA, 02142. Allan Dafoe is an Assistant Professor at Yale University, New Haven, CT, 06520. Jason Seawright is an Associate Professor at Northwestern University, Evanston, IL 60208.

Contents

1 Introduction

"Make your theories elaborate."

R. A. Fisher

Social scientists are often advised to test all the observable implications of a given theory, on the logic that doing so provides more opportunities to distinguish rival theories (Cochran 1965; King, Keohane, and Verba 1994; Shadish, Cook, and Campbell 2002; Rosenbaum 2010). This advice is particularly relevant to studies where the research design provides a strong basis for causal inference, but a small sample size limits statistical power. These conditions are especially common in field experiments, where the expense of adding experimental units is often prohibitive, and in natural experiments, where a naturally occurring "as if random" treatment process may affect only a few units. Intuition suggests that researchers should be able to leverage a detailed theory into stronger inferences, but social-science methodologists have provided little statistical guidance on how exactly to do so. As a result, applied researchers have tended either to combine multiple test informally—which can be highly misleading if the tests are correlated, as they usually are—or, more likely, to leave otherwise excellent studies in a file drawer for want of "statistical significance" (Rosenthal 1979).

In this paper, we propose a remedy to this methodological gap based on nonparametric combination (NPC), a simple and generally applicable framework for drawing an overall inference from multiple hypothesis tests.[1] NPC combines the results of multiple hypothesis tests into a single global p-value that takes into account the dependence among the component tests. Being based on permutation inference, NPC does not require modeling assumptions or asymptotic justifications, only that observations be exchangeable (e.g., randomly assigned) under the global null hypothesis that treatment has no effect. Because the

1. Nonparametric combination was first developed by the statistician Fortunato Pesarin (2001). For recent theoretical and applied work on NPC, see Salmaso and Solari (2005), Pesarin and Salmaso (2010, 2012), Brombin, Midena, and Salmaso (2013), Salmaso (2014), Corain and Salmaso (2015), and Pesarin et al. (2015).

relationships among dependent variables need not be modeled, the component tests may be based on different statistical families and thus easily customized to fit different aspects of a theory. Finally, NPC provides a natural framework for implementing multiple-testing corrections, allowing researchers to assess the component hypotheses along with the global one while controlling the familywise error rate (FWER; Finos, Pesarin, and Salmaso 2003).

This paper makes several contributions. Conceptually, it clarifies the problem of testing a global hypothesis derived from an elaborate theory and its relationship to the so-called "multiple testing problem." Methodologically, it shows how NPC, in conjunction with other recent advances in nonparametric statistics, provides a broadly applicable and highly customizable framework for testing global hypotheses. It also sheds light on the often-misunderstood issue of interpreting tests of the "sharp" null hypothesis of no effects. Practically, it demonstrates the advantages of NPC and makes them accessible to applied researchers via the R package `NPC` (Caughey 2015).

We begin with a discussion of the value of elaborate theories and a survey of existing practice. Next, we provide a general review of permutation inference and then a detailed description of the theory and implementation of NPC. The penultimate section applies NPC to four examples drawn from political science: a randomized experiment, two matched observational studies, and a regression-discontinuity design. The final section concludes.

2 Testing Elaborate Theories: Theory and Practice

Theory testing is one of the central concerns of science (Popper 1962). Unlike theories in the physical sciences, however, social-scientific theories typically do not make point predictions but rather directional ones, which are rarely precise enough to rule out other plausible theories (Meehl 1967). Methodologists have thus long advised social scientists to "make [their] theories elaborate"—that is, to "search for additional implications of a hypothesis" so as to test for a precise *pattern* of theoretical predictions across multiple variables (Cochran

1965, 252; King, Keohane, and Verba 1994, 29). "The more specific, the more numerous, and the more varied are the causal implications of a treatment," the more credible are claims to causal inference, especially in observational studies (Shadish, Cook, and Campbell 2002, 485–6; see also Rosenbaum 1994). In different fields, this advice goes by the name of "pattern matching" (Campbell 1966, 1975; Trochim 1985), "process tracing" (Gerring 2004; George and Bennett 2005), and testing "coherent predictions" (Rosenbaum 1997). Elaborate theories take many forms, but in this paper we focus on those that predict "many effects of the same cause" (King, Keohane, and Verba 1994, 223–4), including effects on multiple outcomes as well as alternative measures of the same concept, outcomes measured at different points in time, and mediators of the main outcome of interest.

In the abstract, the analytical leverage derived from testing elaborate theories is clear, but complications quickly arise when it comes to actually testing multiple predictions. Obviously, the evidence for an elaborate theory is strongest when *all* of its implications are confirmed. Unless statistical power is close to 1, however, it is unlikely that every false null hypothesis will be rejected. If β_j is the false-negative rate for test $j \in \{1, \ldots, J\}$ under a given alternative, the probability of rejecting all of J independent tests is $\Pi_j^J(1 - \beta_j)$, which approaches 0 as J increases. Because each additional test increases the probability that one of the tests will fail to reject, such a strict standard has the perverse effect of punishing theories that are more explicit about all of their implications, relative to less explicit ones. When power is particularly low, all predicted effects may be correctly signed but statistically insignificant. Under conventional standards, the null hypotheses would remain standing, even though common sense suggests that such a consistent pattern of results matching a theory's predictions should provide some corroboration for that theory (Westfall 2005).

The foregoing considerations suggest that, especially in low-power situations, we may wish to combine the evidence from multiple tests into an overall conclusion about the theory that generated them. In doing so, however, it is crucial to take into account the dependence among the respective tests (Salmaso 2014). Two closely related variables provide little

information beyond either alone; if one variable happens by chance to be higher in the treatment group, the other is likely to be so as well. By contrast, joint tests of predictions that run "against the grain" of the covariance in the data can be much more powerful than either alone because observing both predicted effects is much less likely to occur by chance. In fact, it is possible for the total evidence for the theory to be quite strong even if neither hypothesis test on its own is statistically significant.

There is thus broad agreement that scholars should elaborate and test all the observable implications of their theories because multiple predictions offer more opportunities to discriminate between theories. To determine the frequency with which social scientists actually test elaborate theories, we surveyed quantitative articles published in the leading political science journals. For each article, we coded the number of distinct hypotheses, all derived from the same theory, that involved the same independent variable in the same sample of data. We found that a substantial minority of articles (nearly 40%) tested at least two predictions. Further, testing multiple predictions has become more prevalent over time: the average number of hypotheses increased from 1.4 predictions in 1960–1989 to 2.2 in 1990–2010. We also found, however, that political scientists—and presumably other social scientists as well—almost never combine multiple inferences using a formal procedure. Rather, their overwhelmingly dominant approach is to first evaluate each hypothesis separately and then combine them using an informal evaluation of the overall evidence for the theory. Because the strength of the joint evidence hinges on the degree of dependence among the tests, informal combination can be highly misleading.

Drawing a global conclusion regarding a theory requires combining multiple pieces of evidence into a single summary measure of support for the theory. This can be done formally in many ways, including creating an index of multiple measures, modeling dependent variables as indicators of a single latent construct, or conducting an F-test or other omnibus test. But the most versatile and general metric on which to combine multiple hypotheses is the p-values of the corresponding tests.

Social-science methodologists have rightly warned against overreliance on significance testing (for an overview, see Harlow, Mulaik, and Steiger 1997). We agree with such critiques insofar as they advocate examination of other statistical quantities alongside p-values, such as point and interval estimates. But we believe that p-values will continue to have a central place in statistical analyses because they summarize an important quantity: how strongly the data deviate from the null in the direction of the alternative hypothesis (Lehmann and Romano 2005, 63–4). Given that social-scientific theories rarely yield precise point predictions, directional tests are often the most appropriate way for social scientists to corroborate theories.

An important advantage of combining inferences on the metric of the p-values is that it allows analysts to tailor a test to each component hypothesis—Fisher's exact test for one, Wilcoxon's rank-sum for another—without having to worry about standardizing variables, creating comparable test statistics, or stretching a method to accommodate outcomes with different levels of measurement. In contrast to many off-the-shelf methods, several of which cannot even accommodate one-sided predictions, combining p-values provides a straightforward way to optimize a global test for either a very general or a highly specific pattern of anticipated results (cf. Rosenthal and Rosnow 1985). Within this range, specificity leads to tests that discriminate more clearly among rival theories and to greater statistical power. It is possible to combine p-values parametrically, typically under the assumption that the component tests are independent, but nonparametric combination provides a much more general approach that is valid under arbitrary dependence structures.

3 Nonparametric Combination

NPC involves three basic steps:

1. Determine which observations are *exchangeable* under the null hypothesis.

2. Test each theoretical implication using a *test statistic* that is sensitive to the corre-

sponding empirical prediction.

3. *Combine* the p-values of the component tests into a single global p-value.

Each of these steps corresponds to an analytic decision that applied researchers can generally make on the basis of their theoretical and empirical knowledge. Thus, although the theory and implementation of NPC (detailed below) are quite technical, the method makes realistic demands of the typical user.

Determining which responses are exchangeable requires knowledge of the treatment assignment mechanism. Typically, this means identifying sets of units within which treatment was (as-if) randomly assigned with equal probability. To be sure, exchangeability is typically a strong assumption outside of experimental settings (see, e.g., Sekhon 2009), but most other methods for estimating causal effects and hypothesis tests rely on similarly strong assumptions of unconfounded, ignorable, or exogenous treatment assignment (Freedman 2010, §3.6; Imbens and Rubin 2015, §3). As long as exchangeability is correctly determined, the resulting permutation test will have correct size under the null hypothesis. No additional assumptions are required.

While only Step 1 is required for valid hypothesis test, Step 2 affects the statistical power of the test—its probability of rejecting false nulls. Unlike its size, the power of a permutation test depends on the distribution of the data under the alternative, so the test statistic in Step 2 should be tailored to the specific predictions of the theory in question. Step 3, the choice of function to combine the p-values, also affects the power (but not the size) of the joint test. Section 3.2 discusses which combining functions are most appropriate for NPC.

3.1 Permutation Tests

Since NPC is implemented within a permutation framework, we first review the theory and practice of permutation tests before describing NPC itself. Historically, permutation methods have been little-used in political science compared to parametric methods. However,

as recent articles such as Keele, McConnaughy, and White (2012), Bowers, Fredrickson, and Panagopoulos (2013), and Glynn and Ichino (2014) indicate, the discipline's interest in permutation inference has been growing, due in part to the increased availability of the computer resources needed to enumerate permutation distributions and in part to increased interest in experiments.

In permutation inference, the decision of whether to reject a null hypothesis is based on a comparison between the observed value of a test statistic (e.g., the difference of means) and its permutation distribution under the null. In some cases, the null distribution can be calculated analytically, but in general it can be simulated with arbitrary accuracy by shuffling the group labels (e.g., "treated" and "control") of units many times and calculating the value of the test statistic in each permutation. Permutations are only permitted among units that are exchangeable under the null hypothesis (see below for a discussion of exchangeability). Assuming without loss of generality that test statistics are expected to be large in the alternative, the permutation p-value is the probability across permutations of observing a value of the test statistic at least as extreme as the one actually observed. A distinguishing characteristic of permutation tests is that they are *exact*—that is, their probability of rejecting a true null hypothesis is no greater than the p-value indicates—regardless of the probability distribution that generated the data.

3.1.1 The Assumption of Exchangeability

As noted above, the key assumption of permutation inference is that the responses of units in different groups are exchangeable under the null hypothesis. A set of observations is said to be *exchangeable* if their joint distribution is invariant under permutation of the order of the observations.[2] Independent and identical distribution is a sufficient but not a necessary condition for exchangeability (Greenland and Draper 2011).

In the context of permutation tests, exchangeability is typically justified under either a

2. We will sometimes say that certain units are exchangeable, by which we mean that the random variables associated with these units, such as their responses, are exchangeable.

"population model" or a "randomization model" (Lehmann 2006, 64–5). Under the population model, observations are considered random samples from one or more populations. Under the null hypothesis that the (unknown) population distributions are equal, observations in different groups are exchangeable, and thus permutation tests may be used to test null hypotheses of distributional equality between groups (Pitman 1937).

Permutation tests are more commonly motivated under a randomization model, in which exchangeability is justified by random assignment of treatment rather than random sampling (Fisher 1935; Rosenbaum 2002, 27–40).[3] The randomization model is most obviously applicable to randomized controlled experiments (see Keele, McConnaughy, and White 2012). But it also encompasses "natural experiments" in which the randomization is not controlled by the researcher, as well as observational studies in which treatment can be considered "as if" randomly assigned (e.g., Ho and Imai 2006; see also Dunning 2012). In many observational studies, stratification or matching can also be used to create subsets of exchangeable observations. To respect the restricted nature of the putative randomization, permutation inference for these studies, as well as for their experimental counterparts (e.g., block-randomized experiments), must be based on permutations within exchangeable subsets (on matching and permutation inference, see Rosenbaum 2002).

3.1.2 Strong Null Hypotheses

Under both the population model and the randomization model, permutation tests are typically conducted under the null hypothesis that the probability distributions of (possibly multivariate) responses $\mathbf{Y}$ are identical across the G groups being compared:

$$H_0 : \ \mathbf{Y}_g \stackrel{d}{=} \mathbf{Y}_h, \quad \forall \ g, h \in \{1, \ldots, G\}. \tag{1}$$

3. The association between permutation tests and randomization is so close that they are often referred to as *randomization tests*. We prefer the term *permutation tests* because random treatment assignment is not a necessary condition for the use of permutation tests.

In the context of causal inference, where groups correspond to $G = 2$ levels of treatment, Equation 1 is often referred to as Fisher's "sharp" null hypothesis of no effects. In the notation of potential outcomes (Rubin 1974), where $Y_i(0)$ and $Y_i(1)$ respectively indicate unit i's hypothetical responses under control and treatment, the sharp null is written as

$$H_0 : Y_i(1) = Y_i(0), \quad \forall\ i \in \{1, \ldots, N\}. \tag{2}$$

We use the term *strong null* to refer to null hypotheses under which the responses $\mathbf{Y}$ can be transformed so as to satisfy Equation 1. Any null hypothesis that meets this requirement can be tested exactly using permutation inference. A simple example is the null of a constant additive treatment effect, which can be transformed into the sharp null by subtracting the stipulated effect from the responses of the treated units. More theoretically sophisticated null hypotheses, including spillover effects, multiplicative effects, and a variety of non-constant effect models, can be transformed into Equation 1 via similar procedures (Bowers, Fredrickson, and Panagopoulos 2013; Rosenbaum 2003; Rosenbaum 2010, 40–56).

By contrast, most non-permutation tests are conducted under a so-called *weak null* hypothesis. A weak null states that some aspect of the probability distributions that generated the data is equal across treatment groups (e.g., $\mathbb{E}[Y_1] = \mathbb{E}[Y_2]$), whereas a strong null states that the distributions of the possibly transformed responses are equal in all respects ($Y_1 \stackrel{d}{=} Y_2$). Like null hypothesis testing generally, the strong and weak nulls have been the subject of considerable debate. To some, strong nulls are superior because they yield exact tests without relying on often-dubious parametric assumptions or large-sample approximations (e.g., Fisher 1935; Rosenbaum 2010). Others find strong nulls so restrictive as to be trivially false and thus almost useless for scientific inference (Neyman 1935; Gelman 2013).

Our position accords most closely with that of Imbens and Rubin (2015, ch. 5–6), who argue that the strong and weak nulls have different strengths and serve different inferential purposes, and that it often makes sense to use both frameworks to analyze the same study.

The weak null is most appropriate when the relevant scientific question is best answered by estimating a specific parameter, such as the population average treatment effect, and when parametric assumptions or asymptotic approximations are tolerably accurate. By contrast, tests of the strong null are exact and "distribution-free," but these virtues come at the cost of a more restrictive null hypothesis, the rejection of which is often but not always scientifically interesting.

3.1.3 Test Statistics, Alternative Hypotheses, and Statistical Power

The substantive meaning of rejecting a strong null hypothesis hinges on the test statistic used in the corresponding permutation test. Some test statistics, such as the Kolmogorov-Smirnov statistic, are sensitive to generic distributional differences, and thus rejecting a strong null with this statistic says little about *how* responses differ across groups. By contrast, the difference-of-means statistic is most sensitive to differences in location (i.e., the center of the distribution). Difference-of-means permutation tests also have the useful property that if the strong null $Y_i(1) - Y_i(0) = \delta \; \forall i$ can be rejected, so can any null under which $Y_i(1) - Y_i(0) \leq \delta \; \forall i$ (Caughey, Dafoe, and Miratix 2015).[4] In other words, if the observed data provide clear evidence against a given strong null, then they provide even clearer evidence against null hypotheses farther from the alternative hypothesis. This property is particularly useful for testing competing theories that predict effects in opposite directions.

A limitation of the difference-of-means permutation test is that it may reject the strong null even if the weak (average) null is true—for example, if the variances differ (Romano 1990). This limitation can be mitigated, however, by using Student's t as the permutation test statistic, which provides an asymptotically valid test of the weak null as well as an exact test of the sharp null. The same property holds for a wide variety of other test statistics (e.g.,

4. This property, which is shared by the rank-sum test and a large class of other test statistics, follows from the fact that the rejection probability of the difference-of-means permutation test is non-decreasing in the unit-level treatment effects. Thus, the rejection probability under the sharp null (i.e., the size of the test, α) is always at least as great as under any null that stipulates non-positive effects (or non-negative, if negative effects are expected).

the difference of medians) once they have been "studentized" using a consistent estimate of their standard error (Janssen 1997; Chung and Romano 2013). Permutation tests with studentized test statistics are thus especially appealing because they can be simultaneously interpreted as approximations of the corresponding parametric test. On the other hand, since studentized statistics are not weakly increasing in $\delta_i = Y_i(1) - Y_i(0)$, studentized permutation tests (like parametric tests) cannot be interpreted as tests of every null for which $\delta_i \leq \delta \ \forall i$.

In short, while any permutation test provides an exact test of the sharp null, the choice of test statistic is still consequential, substantively as well as statistically. A good rule of thumb is to choose a statistic that is highly sensitive to (i.e., powerful under) the expected alternative while remaining robust to deviations from the hypothesized model (Lehmann 2009). The difference of means is a good default choice that is optimal for light-tailed symmetric distributions, but rank tests such as the Wilcoxon rank-sum can be much more powerful when the responses contain outliers. It may also be wise to select a specialized test statistic for such problems as skewed distributions, covariate-adjusted responses, censored data, large-but-rare effects, and stratified and/or clustered designs (Hogg, Fisher, and Randles 1975; Hothorn et al. 2006; Hansen and Bowers 2008; Small, Ten Have, and Rosenbaum 2008; Rosenbaum 2010). Moreover, if it is a priority to reject only when the mean (or other parameter) differs across groups, use a studentized statistic.

3.2 NPC in Theory

The basic insight underlying NPC is that since p-values are functions of the observed data, any scalar function of multiple p-values is a valid test statistic (compare Imbens and Rubin 2015, 70). The joint evidence provided by multiple tests can thus be evaluated by comparing the observed value of the combined statistic with its distribution under the null hypothesis. Combining p-values can be done analytically when tests are independent, but determining the joint distribution of dependent p-values generally requires resampling methods, such as permutation tests (Westfall 2005).

Nonparametric combination is a general permutation-based framework for decomposing a complex hypothesis, testing its constituent sub-hypotheses, and combining the resulting p-values in a way that accounts for the dependence among the tests. In the NPC framework, the *global* null hypothesis consists of the intersection of $J > 1$ *partial* sub-hypotheses: $\bigcap_{j=1}^{J} H_{0j}$, where each H_{0j} is a strong null amenable to permutation inference. In essence, the global null states that all of its constituent sub-hypotheses are true. The global alternative hypothesis is the union of J sub-alternatives, $\bigcup_{j=1}^{J} H_{1j}$, so the global null is false if any sub-alternative is true. As long as every partial test is exact, the global test will be exact as well.

3.2.1 Combining Functions

Although any NPC global test will be exact, its power will depend both on the power of the partial tests and on the function used to combine their p-values. As a default choice of combining function, we recommend the *product* function (also known as Fisher's chi-square combination). The product function is appealing because it is relatively powerful when all sub-alternatives are true, but it does not over-punish weak evidence for single sub-alternative. To limit the risk of "p-hacking," we recommend that if researchers use a different combining function, they offer a principled justification for doing so and also report results using the product function. Below, we discuss alternative combining functions and describe their relative advantages and disadvantages.

One desirable property in a combining function is that it leads to a global test that is consistent, which means that the probability of rejecting a false global null approaches 1 as the sample size goes to infinity. The value of consistency is illustrated by the intuitively appealing, but generally undesirable, combining function of the average of the p-values. Because it is not consistent, the average combining function may not reject if one p-value is above a certain threshold, regardless of how large the sample size is or how small the other p-values are (Loughin 2004, 470). For example, in the case of two independent test, the average will not reject at level α unless *both* p-values are less than $\sqrt{2\alpha}$. This is because the average

gives insufficient weight to partial tests that are extremely significant. For instance, the average combining function treats the p-value pair $\mathbf{p}_1 = (0.0000001, 0.3)$ as providing weaker evidence against the global null than the pair $\mathbf{p}_2 = (0.1, 0.2)$, which seems inappropriate given that under the null a p-value smaller than 0.0000001 is one-millionth as likely as one smaller than 0.1.

Pesarin and Salmaso (2010, 128–35) identify a class of consistent combining functions (i.e., functions that lead to consistent global tests).[5] The first function we consider is Liptak's *normal* combining function,

$$\psi_{\Phi} = -\sum_{j=1}^{J} \Phi^{-1}(p_j), \tag{3}$$

which is equivalent to converting p-values to z-scores and taking their average. Among the consistent functions that we consider, the normal has the greatest relative power when all sub-alternatives are true. Conversely, the normal combination is relatively unlikely to reject when the evidence is imbalanced across partial tests. Consider, for example, the normal combination of three independent tests (see Westfall 2005). If all three p-values are 0.10—that is, if there is moderate evidence for all three alternative hypotheses—the global p-value will be 0.01. If, by contrast, one of the p-values is 0.10 but two are 0.50 (the expected value under the null hypothesis), the p-value of the global test will be 0.23. In order to reject the global null at 0.01, the evidence for the first hypothesis would have to be extremely powerful, with a p-value less than 0.00003. Due to this behavior, the normal combining function should be used only when researchers are confident that their theory strongly implies every one of the alternative hypotheses being tested.

A second consistent function is the *minimum* combining function,

$$\psi_{min} = -\min_{1 \leq j \leq J}(p_j), \tag{4}$$

5. We have modified the notation and names of the following combining functions, but our modified versions are permutationally equivalent to the original sources. To avoid infinite combined statistics, we recommend that p-values be mapped to the open $(0, 1)$ interval using the transformation $p_{(0,1)} = (p + (2B)^{-1})/(1 + B^{-1})$, where B is the number of samples from the permutation space.

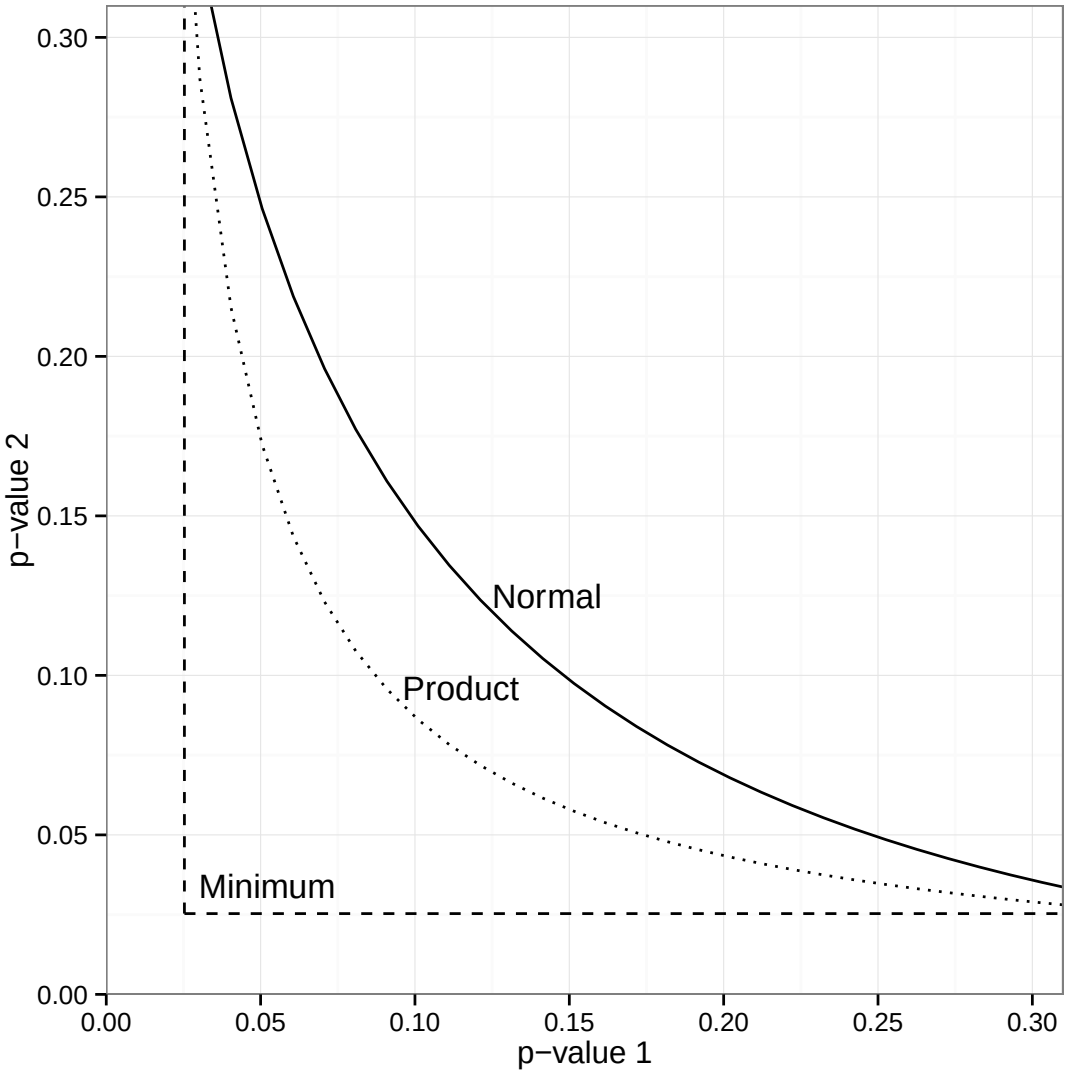


Figure 1: Comparison of the rejection regions (lower-left) of the minimum, product, and normal combining functions in the special case of two independent tests. Any pair of p-values left of and below a given curve will be rejected at $\alpha = 0.05$ by the corresponding combining function. Note that the plot region does not cover the entire $(0, 1)$ p-value space.

which because it ignores all but the smallest p-value has greatest relative power when only a single sub-alternative is true. While useful in some contexts, this property is generally a poor fit for testing elaborate theories. The main advantage of the minimum is that it allows for computationally efficient corrections for multiple tests, which can be valuable when testing many hypotheses (see Section 3.4 for further discussion).

Finally, our recommended choice of Fisher's *product* function,

$$\psi_{\Pi} = -2 \sum_{j=1}^{J} \log(p_j), \tag{5}$$

is a compromise between the minimum and normal combining functions. (We label it the "product" function because it is permutationally equivalent to $-\Pi_{j=1}^{J} p_j$.) Unlike the minimum, the product function rewards a test for finding support for multiple predictions. And unlike the normal, it avoids over-punishing the test if evidence for one prediction is weak. Figure 1 compares the rejection behavior of the minimum, product, and normal combining functions in the special case of two independent tests (also see Loughin 2004). In our experience, we have found that under most circumstances the product and normal combining functions yield very similar results.

3.3 NPC in Practice

Since NPC works on the metric of the p-value, it requires calculating J p-values for the observed data as well as J "pseudo p-values" for each of the B permutations. Roughly speaking, this means that each permutation is ranked on each sub-hypothesis according to its support for the corresponding sub-alternative. Then, the p-values in each permutation are combined across tests using a suitable function, producing a global test statistic for each permutation. The global p-value is the proportion of permutations with global test statistics at least as large as the one observed. Note that all variables associated with a given unit are permuted together, so NPC automatically accounts for the dependence across tests.

NPC can be carried out using the following algorithm (adapted from Pesarin and Salmaso 2010, 125–7):

1. Calculate the vector $\mathbf{T}^{\text{obs}} = (T_1^{\text{obs}}, \dots T_j^{\text{obs}}, \dots T_J^{\text{obs}})$ of observed test statistics corresponding to the J partial tests. For example, if the test statistic is the difference of means and the groups are treated and control, calculate $T_j^{\text{obs}} = \bar{Y}_j^T - \bar{Y}_j^C$ for each response variable j.

2. Repeat the following B times:

   (a) Randomly permute the group labels of units that are exchangeable under the global null hypothesis (e.g., within blocks in the case of block-randomized experiment).

   (b) In each permutation $b \in \{1, \dots, B\}$, calculate the vector $\mathbf{T}_b^* = (T_{1b}^*, \dots T_{jb}^*, \dots T_{Jb}^*)$ of J test statistics.

3. Presuming that the partial test statistics are expected to be large in the alternative, let $\hat{L}_j(t) = B^{-1} \sum_{m=1}^{B} \mathbb{I}(T_{jm}^* \geq t)$ be the estimated significance level for any test statistic value $t \in \mathbb{R}^1$ corresponding to partial test j. Calculate the vector of estimated significance levels for the observed data: $\hat{\mathbf{p}} = (\hat{p}_1, \dots \hat{p}_j, \dots \hat{p}_J)$, where $\hat{p}_j = \hat{L}_j(T_j^{\text{obs}})$. Like any p-value, $\hat{p}_j$ indicates the probability under the null hypothesis of a test statistic at least as extreme as the one observed. Then, for each permutation b, calculate the vector of pseudo p-values $\hat{\mathbf{L}}_b^* = (\hat{L}_{1b}^*, \dots \hat{L}_{jb}^*, \dots \hat{L}_{Jb}^*)$, where $\hat{L}_{jb}^* = \hat{L}_j(T_{jb}^*)$.

4. Using combining function ψ, combine the vector of J estimated significance levels into a global test statistic $T''^{\text{obs}} = \psi(\hat{\mathbf{p}})$, which captures the observed divergence from the null across all partial tests. Then calculate the analogous statistic $T_b''^* = \psi(\hat{\mathbf{L}}_b^*)$ for each permutation b.

5. Estimate the combined significance level (p-value) of the global test as

$$\hat{p}''_{\psi} = B^{-1} \sum_{b=1}^{B} \mathbb{I}(T''^{*}_{b} \geq T''^{\text{obs}}). \tag{6}$$

If all permutations are enumerated, the global significance level $\hat{p}''_{\psi}$ in (6) is exact, as are the significance levels $\hat{\mathbf{p}}$ of the partial tests. In practice, permutation significance levels can be estimated to an arbitrary degree of accuracy by randomly sampling a large number of permutations from the permutation space. Since the global p-value is calculated directly from the permutation distribution of partial p-values, this algorithm is not computationally demanding. For example, it takes about 5 seconds to run the algorithm on a 100-observation dataset with 10 response variables, using $B = 10{,}000$ permutations and the mean difference as the test statistic.[6]

3.4 NPC and the Multiple Testing Problem

Before demonstrating how NPC can be applied to real examples, we first clarify the subtle relationship between global tests of complex hypotheses and the "multiple testing problem." The multiple testing problem describes the inflation in false-positive error rates that occurs when multiple hypotheses are tested in the same study. Most multiple-testing procedures aim to control the familywise error rate: the probability of rejecting at least one true null hypothesis in a "family" of tests. Thus in contrast to NPC's focus on the *global test*, multiple-testing adjustments are concerned with the validity of the *partial tests*. In fact, if only the global p-value is considered in an NPC analysis, the multiple testing problem does not arise.

Nevertheless, NPC and multiplicity control have an important point of connection in that methods of FWER adjustment typically involve testing intersection hypotheses. Marcus, Peritz, and Gabriel (1976) show that it is possible to control the FWER for a set of elemental hypotheses by testing all possible intersections of the hypotheses and rejecting a hypothesis

6. This example was run on a 2014 MacBook Air with a 1.7 GHz Intel Core i7 processor and 8GB of RAM.

only if all intersection hypotheses that include it are statistically significant. If they account for the dependence among tests, such "closed testing" procedures are generally much more powerful than simple methods such as the Bonferroni correction. As Pesarin and Salmaso (2010, 177–96) note, NPC provides a natural method for conducting the intersection tests required by closed-testing procedures. FWER control can thus be achieved by applying any NPC combining function to every intersection hypothesis and adjusting the elemental p-values up to the maximum of all intersection tests that include it. While appealing, closed testing is computationally demanding because the number of intersection hypotheses increases exponentially in the number of elemental tests. If the minimum combining function is used, however, it is possible to use a short-cut procedure described by Westfall and Young (1993, 66–7), "step-down MinP," for which computation time increases only linearly. The `R` package `NPC` implements both step-down MinP and general closed testing adjustment, using the latter only if there are fewer than 15 elemental hypotheses. As we illustrate in several examples, it is often useful to apply two rounds of NPC: first to test the global null, and second to adjust the partial p-values so that they can be individually assessed without inflating the FWER.

4 Applications

NPC is applicable whenever one has an elaborate theory (a theory that makes multiple predictions against a strong null) and exchangeable treatment assignment. We group the many possible applications into six classes, which we describe briefly below. We then discuss four specific applications.

1. **Multiple experiments or subgroups.** Hypothesis tests of predictions for multiple experiments or subgroups of the same experiment can be combined with NPC by pooling the data, permuting treatment in accord with the original treatment assignment process, and calculating the partial test statistics and p-values on data from the

corresponding experiment/subgroup of interest.

2. **Multiple Treatment Levels and Dose Response.** Treatment often has multiple levels (different "dosages"), and theories typically posit specific responses to dosage (e.g., increasing effects, or an "umbrella" pattern). Such patterns can be tested by decomposing the hypothesized pattern into a series of comparisons between treatment levels, and combining the tests with NPC. We illustrate a simple example in Section 4.1 (see also Pesarin and Salmaso 2010, §6.5).

3. **Causal Mechanisms.** Many social scientific theories posit causal mechanisms linking causal factors to outcomes. Thus, like process tracing in qualitative case studies, tests of treatment effects on hypothesized mediators of the causal process offer additional opportunities to corroborate such theories. These tests can then be combined using NPC. Doing so can be an appropriate first step in mediation analysis, for it tests a necessary but insufficient condition of causal mediation—that treatment has an effect on each mediator—under less stringent assumptions than full mediation analysis (cf. Green, Ha, and Bullock 2010; Imai et al. 2011).

4. **Multiple Outcomes.** Whether or not we are able to observe mediators of the causal process, we are often able to deduce multiple consequences of the theorized causal process. These outcomes may each be of inherent substantive interest, or some may be side effects of a motivating main relationship; the inferential logic is the same in either setup. Section 4.2 provides an example, which was what motivated our initial interest in NPC. Multiple outcomes is the canonical application of NPC.

5. **Multiple Outcome Measures or Time-Periods.** Two specific kinds of multiple outcomes arises when researchers have (a) several noisy measures of an imperfectly observed outcome or (b) repeated measures over time of the outcome. Typical solutions are to use just one measure, or to combine the measures into an index by, for example, taking the average or by extracting the first factor/component. NPC provides a non-

paramtric solution to this problem, which we illustrate with an application to repeated measures in Section 4.3.

6. **Placebo Tests.** Experimental and natural experimental designs imply a large number of known zero associations. Testing for these is a good practice for evaluating the credibility of one's identifying assumptions: that treatment is truly (as if) random. NPC provides a principled and flexible way of combining many placebo tests into a single hypothesis test, as we illustrate in Section 4.4.

4.1 Cluster-Randomized Experiment with Dose Response

We begin with a simple re-analysis of Wantchekon's (2003) well-known field experiment on clientelistic campaign appeals. Wantchekon convinced presidential candidates in Benin to randomly vary their campaign appeals across 24 villages stratified by electoral district. Villages were assigned to one of three "doses" of clientelism: a purely clientelistic campaign, a mix of clientelistic and policy-based appeals, or a purely policy-based campaign. Wantchekon's main finding is that vote share increased with the dose of clientelism: candidates' average vote share was lowest in villages where they ran a policy-based campaign (69%), better where they ran a mixed campaign (74%), and best where they made exclusively clientelistic appeals (84%). Wantchekon reports that the increase at each dosage level is highly statistically significant, but Green and Vavreck (2008, 139–40) suggest that their significance may be "grossly exaggerated" because Wantchekon did not "account for the fact that [voters] were embedded in village-level clusters."

With this critique in mind, we re-analyze Wantchekon's data at the level of the village ($n = 8$ per treatment condition), the unit at which treatment was assigned. To operationalize Wantchekon's dose-response hypothesis, we decompose it into two sub-hypotheses:

H1: $\mathbb{E}(\textit{Vote Share} \mid \textit{Campaign} = \text{Policy}) < \mathbb{E}(\textit{Vote Share} \mid \textit{Campaign} = \text{Mixed})$

H2: $\mathbb{E}(\textit{Vote Share} \mid \textit{Campaign} = \text{Mixed}) < \mathbb{E}(\textit{Vote Share} \mid \textit{Campaign} = \text{Clientelism})$

As a baseline for comparison, we follow standard econometric practice and regress vote share on dummies for treatment categories, basing inference on the larger of conventional and robust standard errors (in this case, conventional).[7] Consistent with Green and Vavreck's skepticism, both H1 and H2 fall short of statistical significance, even with one-sided tests ($p_{\text{H1}} = 0.23$, $p_{\text{H2}} = 0.11$), as does an F-test of the global null that both differences are zero ($p_F = 0.15$).

An analysis using difference-of-means permutation tests suggests slightly stronger support for the sub-hypotheses ($p_{\text{H1}} = 0.22$ and $p_{\text{H2}} = 0.03$). The real advantage of permutation inference, however, comes from combining the tests with NPC, which allows us to test a more focused alternative than the non-directional F-test. Following Finos, Salmaso, and Solari (2007), we do so by conducting one-sided tests of the mean difference at each treatment threshold and combining them using Fisher's product function, which yields $p_{\text{NPC}} < 0.01$.[8] The NPC analysis thus decisively rejects the global null in favor of Wantchekon's original hypothesis that clientelistic campaigns are more effective. Crucially, it does so while respecting the clustered nature of the experimental randomization and without relying on parametric assumptions or asymptotic approximations.

4.2 Cluster-Matched Observational Study with Multiple Outcomes

Our second example application is the one that originally motivated our interest in NPC: a matched observational study of conflict differences between Southern and non-Southern U.S. presidents (Dafoe and Caughey, Forthcoming). We hypothesize that because Southern presidents were raised in a "culture of honor," they were socialized to display intense concern

7. Angrist and Pischke (2009, 302–04) recommend this strategy as "the best of both worlds" in terms of the trade-off between bias and variance. They further note that in balanced experiments like Wantchekon's, in which sample sizes across treatment conditions are equal, conventional and robust variance estimators differ very little. In this application, however, the standard error estimates are quite sensitive to which heteroskedasticity-consistent formula is used to calculate them, suggesting that they may be unreliable in this small sample.

8. The normal combining function produces identical results. The minimum function yields a slightly higher p-value (0.02), as one would expect given that its under the alternative that both H1 and H2 are true is lower than the other combining functions. The results for Student's t are also extremely similar.

of their reputation for resolve, leading in turn to systematically different behavior in militarized interstate disputes (MIDs). With the aid of the formal model of conflict escalation, we derived three observable implications of our theory. MID that occurred while a Southerner was president should be:

H1: More likely to involve the use of force by the United States (dichotomous)

H2: Longer in duration (measured in days, censored at the end of presidential terms)

H3: Resolved more favorably for the United States (defeat $= -1$, draw $= 0$, victory $= +1$).

We test these three sub-hypotheses on a dataset of 11 matched pairs of Southern and non-Southern presidents. As with Wantchekon's cluster-randomized experiment, the analysis must respect the fact that treatment was assigned to presidents rather than MID. The conventional way to do so would be to estimate the average treatment effect on the treated (ATT) using OLS with standard errors clustered by president. According to this method, none of the ATT estimates is significant at the 5% level (one-sided). Given that there are only 22 clusters, however, the standard error estimates are of uncertain reliability.

Table 4.2 reports the results for an analysis using permutation tests and NPC with Liptak's normal combining function. In the first row, the variable-specific p-values correspond to permutation tests using ATT estimates listed above them as the permutation test statistic. The second row highlights the flexibility and generality of NPC by reporting the results of using a different test statistic for each sub-hypothesis:

H1: Weighted mean of the pair-specific mean differences (Hansen and Bowers 2008)

H2: Log-rank statistic for hazard-rate differences in censored duration data

H3: Kolmogorov-Smirnov statistic for stochastic dominance

The dependence among these three tests would be difficult if not impossible to derive analytically, but is trivial to account for using NPC. The partial p-values differ somewhat depending

Table 1: Permutation p-values of differences in conflict behavior between Southern and non-Southern presidents.

	Use of Force ATT = +0.17	**Duration** ATT = +46	**Outcome** ATT = +0.18	**NPC**
Difference of Means	0.10	0.02	0.04	0.01
Wtd. Mean, Log-Rank, K-S	0.07	0.10	0.05	0.01

on which test statistic is used, but both NPC p-values indicate strong evidence against the global null. This analysis thus illustrates both NPC's flexibility and the potential gain in statistical power that it can bring when used to test an elaborate theory.

4.3 Matched Panel Analysis with Multiple Endpoints

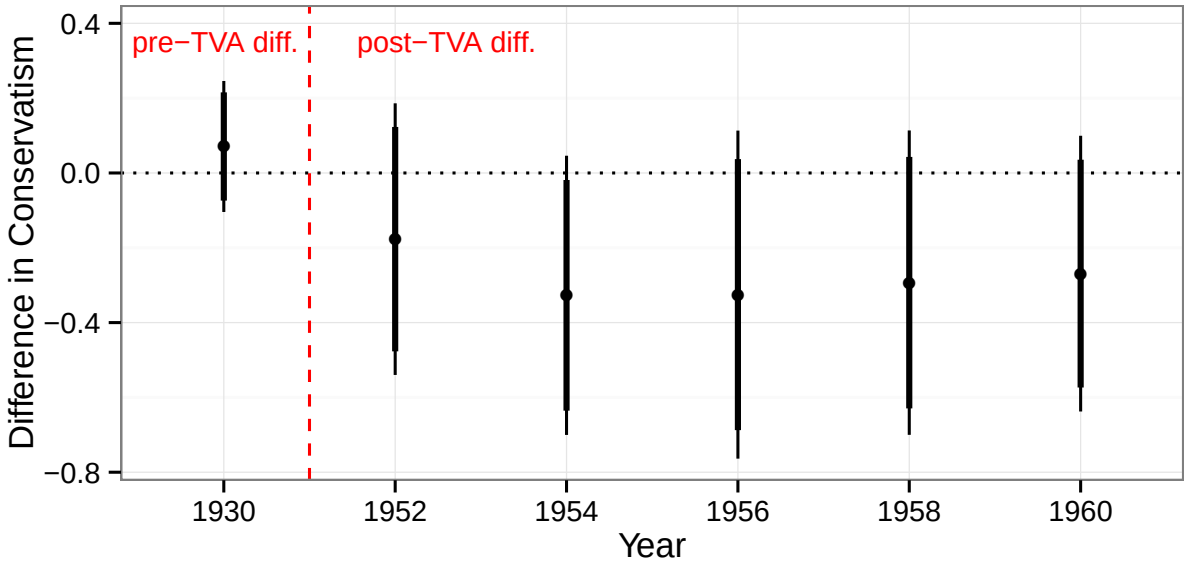
In this example, we apply NPC to panel data, a special case of multiple dependent variables where the responses for each unit are measured at multiple points in time. Specifically, we use a matched panel design (Heckman, Ichimura, and Todd 1997) to estimate the political effects of a major policy intervention, the Tennessee Valley Authority (TVA).[9] We use this example to compare the performance of mean-based and rank-based studentized test statistics, as well as to illustrate how testing and estimation can serve complementary purposes in the same study.

Historians have argued that the TVA, which originated in 1933 as part of President Franklin Roosevelt's New Deal, fostered support for New Deal liberalism in the areas of the South it affected (e.g., Schulman 1994). We test this theory by comparing the estimated conservatism of U.S. House members from matched TVA and non-TVA districts in five consecutive congresses. Twenty matched pairs of districts were created, based on pre-intervention demographic and political covariates. The ATT estimates plotted in Figure 2 suggested that TVA representatives were less conservative in years after 1952, but note that all of the 95% confidence intervals cover zero.

Table 2 presents analogous results for two sets of permutation tests, one using the studen-

9. An earlier analysis of these data appeared in Caughey (2012).

Figure 2: Year-specific estimated average effects of the TVA on representatives' conservatism (with two-sided 90% and 95% confidence intervals). Standard errors were calculated following Abadie and Imbens (2006) and compared to t distribution with 19 degrees of freedom (one fewer than the number of pairs).



tized difference of means and the other the studentized Wilcoxon rank-sum statistic.[10] The first and third rows contain the raw p-values of the partial tests. Below them, in parentheses, are closed testing–adjusted p-values, treating each set of five tests as a family. The mean-based p-values are insignificant in every year except 1954, which too becomes insignificant once adjusted for multiplicity. Given the presence of several outliers in the data, however, rank-based tests may be more powerful. Consistent with this supposition, the rank-based partial p-values are uniformly smaller, and the global p-value is about half as large as the mean-based one.

This example illustrates how Neyman-style estimation and NPC-based testing can serve complementary purposes in the same analysis. The ATT estimates in Figure 2 provide a sense of the magnitude of the average effect in each year and its associated uncertainty. As one would expect given their asymptotic equivalence, the confidence intervals and the

10. The studentized Wilcoxon statistic is asymptotically valid as a test of $H_0 : \Pr(Y_1 < Y_2) = \Pr(Y_1 > Y_2)$ against the alternative that observations tend to be larger in one group (Chung and Romano 2013, 492).

	1952	1954	1956	1958	1960	NPC
Mean:	0.15 (0.15)	0.03 (0.07)	0.06 (0.09)	0.06 (0.09)	0.06 (0.09)	0.06
Rank:	0.03 (0.04)	0.02 (0.04)	0.06 (0.06)	0.05 (0.06)	0.03 (0.05)	0.03

Table 2: One-sided permutation p-values of the null hypothesis that the TVA had no effects on representives' conservatism, based on studentized difference-of-means and rank-sum statistics. Parenthesis-enclosed p-values have been adjusted for multiple tests using closed testing. The product combining function was used to combine the partial tests.

unadjusted p-values for the studentized mean differences correspond very closely, whereas the rank-based results are somewhat stronger due to their resistance to outliers. The added value of NPC in this analysis comes from the multiplicity-adjusted p-values, which control the FWER, and especially from the global p-values, which provide a convenient and principled summary of the overall evidence. Even though the null of zero average effect cannot be rejected for any year without inflating the type-I error rate, the totality of the evidence indicates solid support for the theory that the TVA had a liberalizing effect on congressional representation.

4.4 Covariate Balance in a Regression-Discontinuity Design

Our final example illustrates the application of NPC to the problem of testing covariate balance in a natural experiment, a purpose for which it is particularly well suited. In an influential article, Lee (2008) argues that very close elections can be considered natural experiments that randomly assign one candidate into office, and that regression-discontinuity (RD) designs can thus be used to identify the causal effects of elections. Caughey and Sekhon (2011) critique Lee's application of RD to postwar U.S. House elections, claiming that important pretreatment covariates are not balanced between close Democratic and Republican victories. Recently, Eggers et al. (2015) have challenged these conclusions, arguing *inter alia* that given the large number of balance tests in Caughey and Sekhon (2011), the covariate imbalance they report could easily have arisen by random chance.

NPC provides a natural way to adjudicate this debate because an implication of as-if randomized treatment assignment is equality of the multivariate distribution of all pretreatment covariates (Hansen and Bowers 2008). In this spirit, we use the replication data from Caughey and Sekhon (2011) to reanalyze their Figure 2 (p. 394), which reports balance statistics for 25 covariates in 85 House elections decided by less than 0.5%. To deal with the fact that many of the covariates contain missing data, we use specialized test statistics of the form

$$T = S_g\sqrt{\nu_h/\nu_g} - S_h\sqrt{\nu_g/\nu_h}, \tag{7}$$

where g and h index treatment groups, S_g and S_h are the sum of the (possibly transformed) observed responses in each group, and ν_g and ν_h are the number of non-missing responses in each group. Under the assumption that the data are missing completely at random, these statistics provide nearly exact tests of the null that the observed data are equal in distribution (the test is "nearly" exact because only the mean and variance of the test statistic are invariant under permutation; Pesarin and Salmaso 2010, 234–44). For consistency with Caughey and Sekhon (2011), we use $S = \sum y_i$ for dichotomous covariates and $S = \sum \text{rank}(y_i)$ for all others, with all 25 tests made two-sided by taking the absolute value of the statistic. We combine the tests with the product function and adjust for multiplicity with step-down MinP.

Even after correcting for multiple testing, nearly half of the 25 covariates in Figure 3 are significantly imbalanced at the 10% level, and seven of the adjusted p-values can be rejected at the 5% level as well. Most importantly, the global p-value is 0.0002, indicating dramatic departure from the null of distributional equality. The advantages of NPC in this application are obvious. NPC directly answers the question of interest—*Is the null of as-if random assignment plausible in close House elections?*—with a clear *No*. It does so without parametric assumptions or asymptotic approximations, all while controlling the FWER on the partial tests in a way that maximizes statistical power.

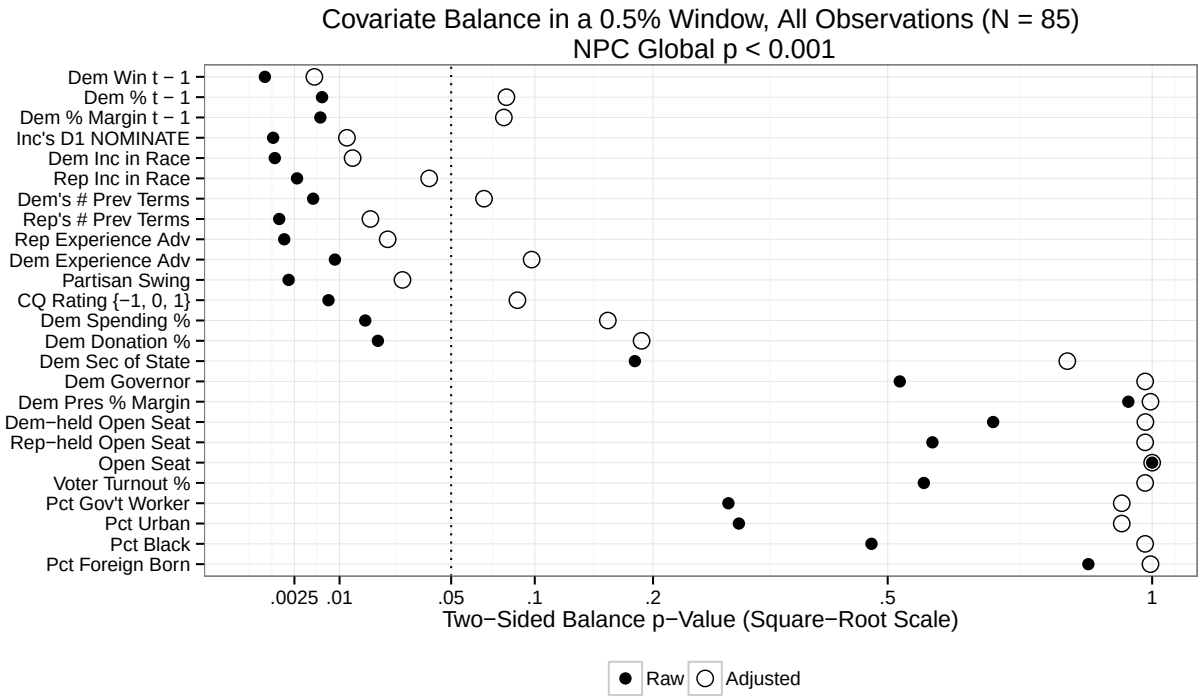


Figure 3: Covariate balance in close U.S. House elections. Hollow points represent p-values that have been adjusted using permutation step-down MinP. All p-values are two-sided.

5 Conclusion

Fisher's advice to elaborate and test all of a theory's empirical implications often goes unheeded. This need not be the case. In this paper, we have described a principled, transparent, and easily tailored method for testing elaborate theories: nonparametric combination. NPC can be used in any context where observations are exchangeable under the null hypothesis and an elaborate pattern of outcome differences is anticipated. As we have demonstrated, such contexts include both observational and experimental designs, as well as stratified or clustered treatment assignments, dose-response relationships, missing data, and multiplicity control.

Although NPC is a powerful framework that can be applied in many contexts, it is only one tool among many, and whether NPC should be used in a given study depends on the goals of the analysis. Obviously, NPC is not useful for testing only a single prediction, or for testing multiple predictions derived from contradictory theories. If the dependence across multiple tests can plausibly be modeled, then a parametric approach may be superior to one based on permutation inference. In addition, since NPC is designed for hypothesis testing, other methods should typically be employed when estimation is the primary goal, especially if treatment effects are thought to be heterogeneous.

Nevertheless, NPC and other approaches should not be considered mutually exclusive. In most applications, it makes sense to evaluate the partial tests individually as well as the overall evidence provided by the NPC global test. It is also often profitable to use permutation hypothesis tests and NPC in conjunction with estimation, for the two approaches convey different kinds of information and require different assumptions. What matters most is not which specific methods researchers use, but that they routinely elaborate and test, in a principled way, all the implications of their theories.

References

Abadie, Alberto, and Guido W. Imbens. 2006. "Large Sample Properties of Matching Estimators for Average Treatment Effects." *Econometrica* 74 (1): 235–267.

Angrist, Joshua David, and Jörn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion.* Princeton, NJ: Princeton University Press.

Bowers, Jake, Mark M. Fredrickson, and Costas Panagopoulos. 2013. "Reasoning about Interference Between Units: A General Framework." *Political Analysis* 21 (1): 97–124.

Brombin, Chiara, Edoardo Midena, and Luigi Salmaso. 2013. "Robust Non-Parametric Tests for Complex-Repeated Measures Problems in Ophthalmology." *Statistical Methods in Medical Research* 22 (6): 643–660.

Campbell, Donald T. 1966. "Pattern Matching as an Essential in Distal Knowing." In *The Psychology of Egon Brunswik,* 81–106. New York: Rinehart & Winston.

———. 1975. "'Degrees of Freedom' and the Case Study." *Comparative Political Studies* 8 (2): 178–193.

Caughey, Devin. 2012. "Congress, Public Opinion, and Representation in the One-Party South, 1930s–1960s." PhD diss., University of California, Berkeley.

———. 2015. *NPC: Nonparametric Combination of Hypothesis Tests.* R package version 1.0.2. `http://CRAN.R-project.org/package=NPC`.

Caughey, Devin, Allan Dafoe, and Luke Miratix. 2015. "Beyond the Sharp Null: Permutation Tests Actually Test Heterogeneous Effects." Paper presented at MacMillan-CSAP Workshop on Quantitative Research Methods, Yale University, November 17.

Caughey, Devin, and Jasjeet S. Sekhon. 2011. "Elections and the Regression Discontinuity Design: Lessons from Close U.S. House Races, 1942–2008." *Political Analysis* 19 (4): 385–408.

Chung, EunYi, and Joseph P. Romano. 2013. "Exact and Asymptotically Robust Permutation Tests." *Annals of Statistics* 41 (2): 484–507.

Cochran, W. G. 1965. "The Planning of Observational Studies of Human Populations." *Journal of the Royal Statistical Society (Series A)* 128 (2): 234–66.

Corain, Livio, and Luigi Salmaso. 2015. "ImprovIng Power of Multivariate Combination-Based Permutation Tests." *Statistics and Computing* 15 (2): 203–214.

Dafoe, Allan, and Devin Caughey. Forthcoming. "Honor and War: Southern U.S. Presidents and the Effects of Concern for Reputation." *World Politics* 68 (2).

Dunning, Thad. 2012. *Natural Experiments in the Social Sciences: A Design-Based Approach.* New York: Cambridge UP.

Eggers, Andrew C., Anthony Fowler, Jens Hainmueller, Andrew B. Hall, and James M. Snyder Jr. 2015. "On the Validity of the Regression Discontinuity Design for Estimating Electoral Effects: New Evidence from Over 40,000 Close Races." *American Journal of Political Science* 59 (1): 259–274.

Finos, Livio, Fortunato Pesarin, and Luigi Salmaso. 2003. "Test combinati per il controllo della molteplicità mediante procedure di closed testing [Combined tests for controlling multiplicity in closed testing procedures]." *Statistica Applicata* 15 (2): 301–329.

Finos, Livio, Luigi Salmaso, and Aldo Solari. 2007. "Conditional Inference under Simultaneous Stochastic Ordering Constraints." *Journal of Statistical Planning and Inference* 137 (8): 2633–2641.

Fisher, Ronald A. 1935. *Design of Experiments.* Edinburgh: Oliver and Boyd.

Freedman, David A. 2010. "Statistical Models and Shoe Leather." In *Statistical Models and Causal Inference: A Dialogue with the Social Sciences,* edited by David Collier, Jasjeet S. Sekhon, and Philip B. Stark, 45–64. New York: Cambridge UP.

Gelman, Andrew. 2013. "In Which I Side with Neyman over Fisher." *Statistical Modeling, Causal Inference, and Social Science* (blog), May 24. `http://andrewgelman.com/2013/05/24/in-which-i-side-with-neyman-over-fisher`.

George, Alexander L., and Andrew Bennett. 2005. *Case Studies and Theory Development in the Social Sciences.* Cambridge, MA: MIT Press.

Gerring, John. 2004. "What Is a Case Study and What Is It Good for?" *American Political Science Review* 98 (2): 341–354.

Glynn, Adam N., and Nahomi Ichino. 2014. "Using Qualitative Information to Improve Causal Inference." *American Journal of Political Science* 59 (4): 1055–1071.

Green, Donald P., Shang E. Ha, and John G. Bullock. 2010. "Enough Already about 'Black Box' Experiments: Studying Mediation Is More Difficult than Most Scholars Suppose." *ANNALS of the American Academy of Political and Social Science* 628 (1): 200–208.

Green, Donald P., and Lynn Vavreck. 2008. "Analysis of Cluster-Randomized Experiments: A Comparison of Alternative Estimation Approaches." *Political Analysis* 16 (2): 138–152.

Greenland, Sander, and David Draper. 2011. "Exchangeability." In *International Encyclopedia of Statistical Science,* edited by Miodrag Lovric, 474–476. London: Springer.

Hansen, Ben B., and Jake Bowers. 2008. "Covariate Balance in Simple, Stratified and Clustered Comparative Studies." *Statistical Science* 23 (2): 219–236.

Harlow, Lisa L., Stanley A. Mulaik, and James H. Steiger, eds. 1997. *What If There Were No Significance Tests?* Mahwaw, NJ: Lawrence Erlbaum Associates.

Heckman, James J., Hidehiko Ichimura, and Petra E. Todd. 1997. "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme." *Review of Economic Studies* 64 (4): 605–654.

Ho, Daniel E., and Kosuke Imai. 2006. "Randomization Inference With Natural Experiments." *Journal of the American Statistical Association* 101 (475): 888–900.

Hogg, Robert V., Doris M. Fisher, and Ronald H. Randles. 1975. "A Two-Sample Adaptive Distribution-Free Test." *Journal of the American Statistical Association* 70 (351): 656–661.

Hothorn, Torsten, Kurt Hornik, Mark A. van de Wiel, and Achim Zeileis. 2006. "A Lego System for Conditional Inference." *The American Statistician* 60 (3): 257–263.

Imai, Kosuke, Luke Keele, Dustin Tingley, and Teppei Yamamoto. 2011. "Unpacking the Black Box of Causality: Learning about Causal Mechanisms from Experimental and Observational Studies." *American Political Science Review* 105 (4): 765–789.

Imbens, Guido W., and Donald B. Rubin. 2015. *Causal Inference in Statistics, Social, and Biomedical Sciences: An Introduction.* New York: Cambridge UP.

Janssen, Arnold. 1997. "Studentized Permutation Tests for Non-i.i.d. Hypotheses and the Generalized Behrens-Fisher Problem." *Statistics & Probability Letters* 36 (1): 9–21.

Keele, Luke, Corrine McConnaughy, and Ismail White. 2012. "Strengthening the Experimenter's Toolbox: Statistical Estimation of Internal Validity." *American Journal of Political Science* 56 (2): 484–499.

King, Gary, Robert Keohane, and Sidney Verba. 1994. *Designing Social Inquiry.* Princeton, NJ: Princeton UP.

Lee, David S. 2008. "Randomized Experiments from Non-Random Selection in U.S. House Elections." *Journal of Econometrics* 142:675–697.

Lehmann, E. L., and Joseph P. Romano. 2005. *Testing Statistical Hypotheses.* 3rd. London: Springer.

Lehmann, Erich L. 2006. *Nonparametrics: Statistical Methods Based on Ranks.* Revised 1st. New York: Springer.

———. 2009. "Parametric Versus Nonparametrics: Two Alternative Methodologies." *Journal of Nonparametric Statistics* 21 (4): 397–405.

Loughin, Thomas M. 2004. "A Systematic Comparison of Methods for Combining P-Values from Independent Tests." *Computational Statistics & Data Analysis* 47 (3): 467–485.

Marcus, Ruth, Eric Peritz, and K. R. Gabriel. 1976. "On Closed Testing Procedures with Special Reference to Ordered Analysis of Variance." *Biometrika* 63 (3): 655–660.

Meehl, Paul E. 1967. "Theory-Testing in Psychology and Physics: A Methodological Paradox." *Philosophy of Science* 34 (2): 103–115.

Neyman, J. 1935. "Statistical Problems in Agricultural Experimentation." *Supplement to the Journal of the Royal Statistical Society* 2 (2): 107–180.

Pesarin, Fortunato. 2001. *Multivariate Permutation Tests: With Applications in Biostatistics.* New York: Wiley.

Pesarin, Fortunato, and Luigi Salmaso. 2010. *Permutation Tests for Complex Data.* Chichester, UK: Wiley.

———. 2012. "A Review and Some New Results on Permutation Testing for Multivariate Problems." *Statistics and Computing* 22:639–646.

Pesarin, Fortunato, Luigi Salmaso, Eleonora Carrozzo, and Rosa Arboretti. 2015. "Union-Intersection Permutation Solution for Two-Sample Equivalence Testing." *Statistics and Computing.*

Pitman, E. J. G. 1937. "Significance Tests Which May Be Applied to Samples from Any Populations." *Supplement to the Journal of the Royal Statistical Society* 4 (1): 119–130.

Popper, Karl R. 1962. *Conjectures and Refutations: The Growth of Scientific Knowledge.* New York: Basic Books.

Romano, Joseph P. 1990. "On the Behavior of Randomization Tests Without a Group Invariance Assumption." *Journal of the American Statistical Association* 85 (411): 686–692.

Rosenbaum, Paul R. 1994. "Coherence in Observational Studies." *Biometrics* 50 (2): 368–374.

———. 1997. "Signed Rank Statistics for Coherent Predictions." *Biometrics* 53 (2): 556–566.

———. 2002. *Observational Studies.* 2nd. New York: Springer.

———. 2003. "Exact Confidence Intervals for Nonconstant Effects by Inverting the Signed Rank Test." *The American Statistician* 57 (2): 132–138.

———. 2010. *Design of Observational Studies.* New York: Springer.

Rosenthal, Robert. 1979. "The 'File Drawer Problem' and Tolerance for Null Results." *Psychological Bulletin* 86 (3): 638–641.

Rosenthal, Robert, and Ralph L. Rosnow. 1985. *Contrast Analysis: Focused Comparisons in the Analysis of Variance.* New York: Cambridge UP.

Rubin, Donald B. 1974. "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies." *Journal of Educational Psychology* 66 (5): 688–701.

Salmaso, Luigi. 2014. "Combination-Based Permutation Tests: Equipower Property and Power Behaviour in Presence of Correlation." *Communications in Statistics - Theory and Methods.* doi:10.1080/03610926.2013.810270.

Salmaso, Luigi, and Aldo Solari. 2005. "Multiple Aspect Testing for Case-Control Designs." *Metrika* 62 (2-3): 331–340.

Schulman, Bruce J. 1994. *From Cotton Belt to Sunbelt: Federal Policy, Economic Development, and the Transformation of the South, 1938–1980.* Durham, NC: Duke University Press.

Sekhon, Jasjeet S. 2009. "Opiates for the Matches: Matching Methods for Causal Inference." *Annual Review of Political Science* 12 (1): 487–508.

Shadish, William R., Thomas D. Cook, and Donald T. Campbell. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference.* Boston: Houghton-Mifflin.

Small, Dylan S., Thomas R. Ten Have, and Paul R. Rosenbaum. 2008. "Randomization Inference in a Group-Randomized Trial of Treatments for Depression." *Journal of the American Statistical Association* 103 (481): 271–279.

Trochim, William M. K. 1985. "Pattern Matching, Validity, and Conceptualization in Program Evaluation." *Evaluation Review* 9 (5): 575–604.

Wantchekon, Leonard. 2003. "Clientelism and Voting Behavior: Evidence from a Field Experiment in Benin." *World Politics* 55 (3): 399–422.

Westfall, P. H., and S. S. Young. 1993. *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment.* New York: Wiley.

Westfall, Peter H. 2005. "Combining P Values." In *Encyclopedia of Biostatistics.* Chichester, UK: John Wiley & Sons, Ltd.