



ANNUAL REVIEWS **Further**

Click [here](#) for quick links to Annual Reviews content online, including:

- Other articles in this volume
- Top cited articles
- Top downloaded articles
- Our comprehensive search

Hypothesis Testing in Econometrics

Joseph P. Romano,¹ Azeem M. Shaikh,² and Michael Wolf³

¹Departments of Economics and Statistics, Stanford University, Stanford, California 94305; email: romano@stanford.edu

²Department of Economics, University of Chicago, Chicago, Illinois 60637

³Institute for Empirical Research in Economics, University of Zürich, CH-8006 Zürich, Switzerland

Annu. Rev. Econ. 2010.2:75-104

First published online as a Review in Advance on February 9, 2010

The *Annual Review of Economics* is online at econ.annualreviews.org

This article's doi:
10.1146/annurev.economics.102308.124342

Copyright © 2010 by Annual Reviews.
All rights reserved

1941-1383/10/0904-0075\$20.00

Key Words

asymptotics, multiple testing, optimality, resampling

Abstract

This article reviews important concepts and methods that are useful for hypothesis testing. First, we discuss the Neyman-Pearson framework. Various approaches to optimality are presented, including finite-sample and large-sample optimality. Then, we summarize some of the most important methods, as well as resampling methodology, which is useful to set critical values. Finally, we consider the problem of multiple testing, which has witnessed a burgeoning literature in recent years. Along the way, we incorporate some examples that are current in the econometrics literature. While many problems with well-known successful solutions are included, we also address open problems that are not easily handled with current technology, stemming from such issues as lack of optimality or poor asymptotic approximations.

1. INTRODUCTION

This review highlights many current approaches to hypothesis testing in the econometrics literature. We consider the general problem of testing in the classical Neyman-Pearson framework, reviewing the key concepts in Section 2. As such, optimality is defined via the power function. Section 3 briefly addresses control of the size of a test. Because the ideal goal of the construction of uniformly most powerful tests (defined below) cannot usually be realized, several general approaches to optimality are reviewed in Section 4, which attempt to bring about a simplification of the problem. First, we consider restricting tests by the concepts of unbiasedness, conditioning, monotonicity, and invariance. We also discuss notions of optimality that do not place any such restrictions, namely maximin tests, tests maximizing average power, and locally most powerful tests. Large-sample approaches to optimality are reviewed in Section 5. All these approaches, and some in combination, have been successfully used in econometric problems.

Next, various methods used to construct hypothesis tests are discussed. The generalized likelihood ratio test and the tests of Wald and Rao are first introduced in Section 6 in the context of parametric models. We then describe how these tests extend to the extremum estimation framework, which encompasses a wide variety of semiparametric and nonparametric models used in econometrics. Afterwards, we discuss in Section 7 the use of resampling methods for the construction of critical values, including randomization methods, the bootstrap, and subsampling.

Finally, Section 8 expands the discussion from tests of a single null hypothesis to the simultaneous testing of multiple null hypotheses. This scenario occurs whenever more than one hypothesis of interest is tested at the same time, and therefore it is common in applied economic research. The easiest, and most common, approach to deal with the problem of multiple tests is simply to ignore it and test each individual hypothesis at the usual nominal level. However, such an approach is problematic because the probability of rejecting at least one true null hypothesis increases with the number of tests, and can even become very close to one. The procedures presented in Section 8 are designed to account for the multiplicity of tests so that the probability of rejecting any true null hypothesis is controlled. Other measures of error control are considered as well. Special emphasis is given to the construction of procedures based on resampling techniques.

2. THE NEYMAN-PEARSON PARADIGM

Suppose data X is generated from some unknown probability distribution P in a sample space \mathcal{X} . In anticipation of asymptotic results, we may write $X = X^{(n)}$, where n typically refers to the sample size. A model assumes that P belongs to a certain family of probability distributions $\{P_\theta, \theta \in \Omega\}$, although we make no rigid requirements for Ω ; it may be a parametric, semiparametric, or nonparametric model. A general hypothesis about the underlying model can be specified by a subset of Ω .

In the classical Neyman-Pearson setup that we consider, the problem is to test the null hypothesis $H_0 : \theta \in \Omega_0$ against the alternative hypothesis $H_1 : \theta \in \Omega_1$. Here, Ω_0 and Ω_1 are disjoint subsets of Ω with union Ω . A hypothesis is called *simple* if it completely specifies the distribution of X , or equivalently a particular θ ; otherwise, a hypothesis is called *composite*. The goal is to decide whether to reject H_0 (and thereby decide that H_1 is true) or accept H_0 . A *nonrandomized test* assigns to each possible value $X \in \mathcal{X}$ one of these two

decisions, thus dividing the sample space \mathcal{X} into two complementary regions, S_0 (the region of acceptance of H_0) and S_1 (the rejection or critical region). Declaring H_1 is true when H_0 is true is called a *Type 1 error*, while accepting H_0 when H_1 is true is called a *Type 2 error*. The main problem of constructing hypothesis tests lies in the construction of a decision rule, or equivalently the construction of a critical region S_1 , which keeps the probabilities of these two types of errors to a minimum. Unfortunately, both probabilities cannot be controlled simultaneously (except in a degenerate problem). In the Neyman-Pearson paradigm, a Type 1 error is considered the more serious of the errors. As a consequence, one selects a number $\alpha \in (0,1)$, called the *significance level*, and restricts attention to critical regions S_1 satisfying

$$P_\theta\{S_1\} \leq \alpha \quad \text{for all } \theta \in \Omega_0.$$

Importantly, acceptance of H_0 does not necessarily demonstrate that H_0 is indeed true; there simply may be insufficient data to show inconsistency of the data with the null hypothesis. Therefore, the decision that “accepts” H_0 should be interpreted as a failure to reject H_0 .

More generally, it is convenient for theoretical reasons to allow for the possibility of a *randomized test*. A randomized test is specified by a test (or critical) function $\phi(X)$, taking values in $[0,1]$. If the observed value of X is x , then H_0 is rejected with probability $\phi(x)$. For a nonrandomized test with critical region S_1 , the corresponding test function is just the indicator of S_1 . In general, the *power function*, $\beta_\phi(\cdot)$, of a particular test $\phi(X)$ is given by $\beta_\phi(\theta) = E_\theta[\phi(X)] = \int \phi(x) dP_\theta(x)$.

Thus, $\beta_\phi(\theta)$ is the probability of rejecting H_0 if θ is true. The *level constraint* of a test ϕ is expressed as

$$E_\theta[\phi(X)] \leq \alpha \quad \text{for all } \theta \in \Omega_0. \tag{1}$$

A test satisfying Equation 1 is said to be *level α* . The supremum over $\theta \in \Omega_0$ of the left-hand side of Equation 1 is the *size* of the test ϕ .

3. CONTROL OF THE SIZE OF A TEST

Typically, a test procedure is specified by a test statistic $T = T(X)$, with the rejection region $S_1 = S_1(\alpha)$ taking the form $T(X) > c$. For a prespecified significance level α , the *critical value* c is chosen (possibly in a data-dependent way and also dependent on α) to control the size of the test, although one often resorts to asymptotic approximations, some of which are described later.

3.1. *p*-Values

Suppose that, for each $\alpha \in (0,1)$, nonrandomized tests are specified with nested rejection regions $S_1(\alpha)$, i.e.,

$$S_1(\alpha) \subseteq S_1(\alpha') \quad \text{if } \alpha < \alpha'.$$

Then, the usual practice is to report the *p*-value, defined as

$$\hat{p} = \inf\{\alpha : X \in S_1(\alpha)\}. \tag{2}$$

If the test with rejection region $S_1(\alpha)$ is level α , then it is easy to see that

$$\theta \in \Omega_0 \Rightarrow P_\theta\{\hat{p} \leq u\} \leq u \quad \text{for all } 0 \leq u \leq 1. \tag{3}$$

3.2. The Bahadur-Savage Result

The problem of constructing a level α test can be nontrivial, in the sense that the level constraint may prohibit the construction of a test that has any power to detect H_1 . In other words, there may exist situations in which it is impossible to construct a level α test that has power larger than α against even one alternative. Bahadur & Savage (1956) provide a classical instance of the nonexistence of any useful level α test. The result is stated in terms of testing the mean of a population in a nonparametric setting. Suppose X_1, \dots, X_n are independent and identically distributed (i.i.d.) with cumulative distribution function F on the real line, where F belongs to some large class of distributions \mathbf{F} . Let $\mu(F)$ denote the mean of F . Here, F (rather than θ) is used to index the model \mathbf{F} . The family \mathbf{F} is assumed to satisfy the following:

1. For every $F \in \mathbf{F}$, $\mu(F)$ exists and is finite.
2. For every real m , there is an $F \in \mathbf{F}$ with $\mu(F) = m$.
3. If $F_1 \in \mathbf{F}$ and $\gamma \in (0, 1)$, then $\gamma F_1 + (1 - \gamma)F_2 \in \mathbf{F}$.

Consider the problem of testing the null hypothesis $H_0 : \mu(F) = 0$ against $H_1 : \mu(F) \neq 0$. Suppose $\phi = \phi(X_1, \dots, X_n)$ is a level α test. Then, for any F with $\mu(F) \neq 0$, $E_F(\phi) \leq \alpha$; that is, the power of the test cannot exceed α for any $F \in \mathbf{F}$.

For example, the result applies when \mathbf{F} is the family of all distributions having infinitely many moments. Unfortunately, the result has consequences for testing any (mean-like) parameter that is influenced by tail behavior. The only remedy is to restrict \mathbf{F} , for example, by assuming the support of F lies in a fixed compact set (see Romano & Wolf 2000; for other nonexistence results, see Dufour 1997 and Romano 2004). In summary, in some problems, there may not exist methods controlling the Type 1 error that are any better than the test that rejects H_0 with probability α , independent of the data, and the only way to avoid this is to reduce the size of the model.

4. OPTIMALITY CONSIDERATIONS

For a given alternative $\theta_1 \in \Omega_1$, the problem of determining ϕ to maximize $\beta_\phi(\theta_1)$ subject to Equation 1 is one of maximizing a real-valued function from the space of test functions satisfying the level constraint; it can be shown that such a test exists under weak conditions. Such a test is then called *most powerful* (MP) level α . Typically, the optimal ϕ will depend on the fixed alternative θ_1 . If a test ϕ exists that maximizes the power for all $\theta_1 \in \Omega_1$, then ϕ is called *uniformly most powerful* (UMP) level α .

In the restricted situation where both hypotheses are simple and specified as $\Omega_i = \{\theta_i\}$, then the Neyman-Pearson Lemma provides necessary and sufficient conditions for a test to be the MP level α test. Specifically, if p_i denotes the density of X under H_i (with respect to any dominating measure), then a sufficient condition for a size α test to be most powerful among α level tests is that, for some constant k , $\phi(X) = 1$ if $p_1(X) > k \cdot p_0(X)$ and $\phi(X) = 0$ if $p_1(X) < k \cdot p_0(X)$. Evidence against H_0 is ordered by the value of the *likelihood ratio* $p_1(X)/p_0(X)$.

For parametric models indexed by a real-valued parameter θ , UMP tests exist for one-sided hypotheses H_0 specified by $\Omega_0 = \{\theta : \theta \leq \theta_0\}$ for some fixed θ_0 , assuming the underlying family has monotone likelihood ratio. For two-sided hypotheses, UMP tests are rare. In multiparameter models in which $\theta \in \mathbb{R}^d$ or Ω is infinite dimensional, UMP tests typically do not exist. The nonparametric sign test is an exception (see Lehmann & Romano 2005b,

example 3.8.1). The following is an example in which a UMP test exists in a multivariate setting. Its importance stems from the fact that, in large samples, many testing problems can be approximated by the one in the example (see Section 5.1 for details).

Example 4.1 (Multivariate normal mean). Suppose X is multivariate normal with unknown mean vector $\theta \in \mathbb{R}^d$ and known covariance matrix Σ . Fix a vector $(a_1, \dots, a_d)^T \in \mathbb{R}^d$ and a number δ . For testing the null hypothesis

$$\Omega_0 = \left\{ \theta : \sum_{i=1}^d a_i \theta_i \leq \delta \right\}$$

against $\Omega_1 = \mathbb{R}^d \setminus \Omega_0$, there exists a UMP level α test that rejects H_0 when $\sum_i a_i X_i > \sigma z_{1-\alpha}$, where $\sigma^2 = a^T \Sigma a$, and $z_{1-\alpha}$ is the $1 - \alpha$ quantile of the standard normal distribution.

The lack of UMP tests in many applications has led to the search for tests under less stringent requirements of optimality. We now review several successful approaches.

4.1. Uniformly Most Powerful Unbiased Tests

A test ϕ is called level α unbiased if

$$\begin{aligned} \beta_\phi(\theta) &\leq \alpha & \text{if } \theta \in \Omega_0, \\ \beta_\phi(\theta) &\geq \alpha & \text{if } \theta \in \Omega_1, \end{aligned} \tag{4}$$

so that the probability of rejecting H_0 if any alternative $\theta \in \Omega_1$ is true is no smaller than the probability of rejecting H_0 when $\theta \in \Omega_0$. A test ϕ is called UMP unbiased (or UMPU) at level α if $\beta_\phi(\theta)$ is maximized uniformly over $\theta \in \Omega_1$ among all level α unbiased tests.

The restriction to unbiasedness is most successful in one- and two-sided testing about a univariate parameter in the presence of a (possibly multivariate) nuisance parameter in a large class of multiparameter models. In particular, many testing problems in multiparameter exponential family models of full rank admit UMPU level α tests. Exponential families are studied in Brown (1986). Some other success stories include the comparison of binomial (or Poisson) parameters, testing independence in a two-by-two contingency table, and inference for the mean and variance from a normal population. The notion of unbiasedness also applies to some nonparametric hypotheses, leading to the class of randomization tests described later. A well-known example in which unbiasedness does not lead to an optimal procedure is the famous Behrens-Fisher problem, which is the testing of equality of means of two normal populations with possibly different unknown variances. We also mention the testing of moment inequalities (in a simplified parametric setting), which has led to a recent burgeoning literature in econometrics (see Example 4.3).

4.2. Conditional Tests

The usual approach to determining a UMPU test is to condition on an appropriate statistic T so that the conditional distribution of X given $T = t$ is free of nuisance parameters. If T has Neyman structure, and other conditions hold, as described in Lehmann & Romano (2005b, chapter 4), UMPU tests can be derived. But even without these assumptions, conditioning is often successful because it reduces the dimension of the problem. However, it reduces the class of tests considered because we now demand that not only Equation 1

holds, but also the stronger conditional level constraint that, for (almost) all outcomes t of a conditioning statistic T ,

$$E_{\theta}[\phi(X)|T=t] \leq \alpha \quad \text{for all } \theta \in \Omega_0. \quad (5)$$

An optimal test may exist within this smaller class of tests, although the reduction to such a class may or may not have any compelling merit to it, as better tests may exist outside the class. The philosophical basis for conditioning is strongest when the statistic T is chosen to be ancillary, i.e., when its distribution does not depend on θ (see Lehmann & Romano 2005b, section 10.3, for some optimal conditional tests in which conditioning is done using an ancillary statistic). We now mention a recent important example where the conditioning statistic is not ancillary, although conditioning does reduce the problem from a curved two-parameter exponential family to a one-parameter exponential family.

Example 4.2 (Unit root testing). The problem of testing for a unit root has received considerable attention by econometricians, dating back to Dickey & Fuller (1979). We discuss some issues with the following simplified version of the problem with an autoregressive process of order one with Gaussian errors. Specifically, let $X_0 = 0$ and

$$X_t = \theta \cdot X_{t-1} + \varepsilon_t \quad t = 1, \dots, n,$$

where $\theta \in (-1, 1]$ and the ε_t are unobserved and i.i.d. Gaussian with mean 0 and known variance σ^2 . Consider the problem of testing $\theta = 1$ against $\theta < 1$. The likelihood function $L_n(\theta)$ is given by

$$L_n(\theta) = \exp \left[n(\theta - 1)U_n - \frac{n^2(\theta - 1)^2}{2} V_n \right] \cdot b(X_1, \dots, X_n),$$

where

$$U_n = \frac{1}{n\sigma^2} \sum_{t=1}^n X_{t-1}(X_t - X_{t-1}) \quad \text{and} \quad V_n = \frac{1}{n^2\sigma^2} \sum_{t=1}^n X_{t-1}^2,$$

and the function b does not depend on θ . This constitutes a curved exponential family. For a fixed alternative θ' , the MP test rejects for large values of

$$(\theta' - 1)U_n - \frac{(\theta' - 1)^2}{2} V_n.$$

As the optimal rejection region is seen to depend on θ' , no UMP test exists. An interesting way to choose θ' is suggested by Elliot et al. (1996), although the use of a particular θ' does not imply any optimality of the power function at another θ' . Crump (2008) constructed optimal tests conditional on U_n because, conditionally, the family of distributions becomes a one-parameter exponential family with monotone likelihood ratio. Although he gives an interesting case for conditioning on U_n , one can instead condition on V_n . To date, no particular test has any strong optimality property, and the problem warrants further study.

4.3. Uniformly Most Powerful Invariant Tests

Some testing problems exhibit symmetries, which lead to natural restrictions on the family of tests considered. Here we describe the mathematical expression of symmetry. Suppose G is a group of one-to-one transformations from the sample space \mathcal{X} onto itself. Suppose that, if $g \in G$ and if X is governed by the parameter θ , then gX also has a distribution in the model; that is, gX has a distribution governed by some $\theta' \in \Omega$. The element θ' obtained in this manner is denoted by $\bar{g}\theta$. In general, we say a parameter set $\omega \subseteq \Omega$ remains *invariant* under g if $\bar{g}\theta \in \omega$ whenever $\theta \in \omega$, and also if, for any $\theta' \in \omega$, there exists $\theta \in \omega$ such that $\bar{g}\theta = \theta'$. We then say the problem of testing Ω_0 against Ω_1 remains invariant under G if both Ω_0 and Ω_1 remain invariant under any $g \in G$. This structure implies that a statistician testing Ω_0 against Ω_1 based on data X is faced with the identical problem based on data $X' = gX$, for any $g \in G$. Therefore, the idea of invariance is that the decision based on X and X' be the same. So, we say a test ϕ is invariant under G if $\phi(gx) = \phi(x)$ for all $x \in \mathcal{X}$ and $g \in G$. A test that uniformly maximizes power among invariant level α tests is called uniformly most powerful invariant (UMPI) at level α .

Invariance considerations apply to some interesting models, such as location and scale models. Perhaps the greatest success is testing parameters in some Gaussian linear models, encompassing applications such as regression and analysis of variance, in which least squares procedures and standard F tests are shown to have optimality properties among invariant procedures. However, UMPI tests may be inadmissible in some problems. We refer the reader to Andrews et al. (2006) for the use of invariance restrictions in instrumental variables regression. Both conditioning and invariance considerations are utilized in Moreira (2003).

4.4. Monotone Tests

In some problems, it may be reasonable to impose monotonicity restrictions on the testing procedure. We illustrate the idea with two examples.

Example 4.3 (Moment inequalities). Suppose $X = (X_1, \dots, X_d)^T$ is multivariate normal with unknown mean vector $\theta = (\theta_1, \dots, \theta_d)^T$ and known nonsingular covariance matrix Σ . The null hypothesis specifies $\Omega_0 = \{\theta : \theta_i \leq 0 \forall i\}$. In fact, the only unbiased test for this problem is the trivial test $\phi \equiv \alpha$ (see Lehmann & Romano 2005b, problem 4.8). Nor do any invariance considerations generally apply. In the special case that Σ exhibits compound symmetry (meaning the diagonal elements are all the same, and the off-diagonal elements are all the same as well), then the problem remains invariant under permutations of the coordinates of X , leading to procedures that are invariant under permutations. Even so, such transformations do not reduce the problem sufficiently to lead to any optimal procedure.

However, a natural monotonicity restriction on a test ϕ is the following. Specifically, if ϕ rejects based on data X , so that $\phi(X) = 1$, and $Y = (Y_1, \dots, Y_d)^T$ with $Y_i \geq X_i$ for all i , then a monotonicity requirement demands that $\phi(Y) = 1$. We return to this example below. We point out now, however, that many currently suggested tests for this problem do not obey such a monotonicity constraint.

Example 4.4 (Testing for superiority or stochastic dominance). Assume the model of Example 4.3, except now the problem is to demonstrate that θ satisfies $\theta_i > 0$ for all i . The null hypothesis parameter space is specified by $\Omega_0 = \{\theta : \text{not all } \theta_i > 0\}$. This is a simplified version of the problem of testing for stochastic dominance (see Davidson & Duclos 2006). Among tests obeying the same monotonicity restriction as in Example 4.3, there exists a UMP level α test (see Lehmann 1952).

The restrictions to unbiased, invariant, conditional, or monotone tests impose certain constraints on the class of available procedures. We now mention some notions of optimality that do not limit the class of available procedures, at the expense of weaker notions of optimality.

4.5. Maximin Tests

For testing Ω_0 against Ω_1 , let $\omega_1 \subseteq \Omega_1$ be a (possibly strict) subset of Ω_1 . A level α test ϕ is *maximin* with respect to ω_1 at level α if it is level α and it maximizes $\inf_{\theta \in \omega_1} E_\theta[\phi(X)]$ among level α tests.

Example 4.5 (Moment inequalities, continued). In Example 4.3, it is possible to combine monotonicity and maximin restrictions to obtain an optimal test. For example, if Σ has all diagonal elements equal, and also all off-diagonal elements equal, then the test that rejects for large $\max X_i$ is optimal (see Lehmann 1952 for the case $d = 2$). The result generalizes, and such a test is admissible among all tests (obeying the level constraint) without any monotonicity restriction (J.P. Romano, A.M. Shaikh & M. Wolf, unpublished work).

4.6. Tests Maximizing Average or Weighted Power

Let Λ_1 be a probability distribution (or generally a nonnegative measure) over Ω_1 . The *average* or *weighted* power of a test ϕ with respect to Λ_1 is given by $\int_{\Omega_1} E_\theta[\phi(X)] d\Lambda_1(\theta)$. A level α test ϕ maximizing this quantity among all level α tests maximizes average power with respect to Λ_1 . The approach to determining such a test is to note that this average power of ϕ can be expressed as the power of ϕ against the mixture distribution M , which assigns to a set E the probability

$$M\{E\} = \int_{\Omega_1} P_\theta\{E\} d\Lambda_1(\theta),$$

and so the problem is reduced to finding the MP level α test of Ω_0 against the simple alternative hypothesis $X \sim M$.

Example 4.6 (Moment inequalities, Example 4.3, continued). In the setup of Example 4.3, Chiburis (2008) considers tests that (approximately) maximize average power (see also Andrews 1998). Such an approach can provide tests with reasonably good power properties, although the choice of the averaging distribution Λ_1 is unclear.

4.7. Locally Most Powerful Tests

Let $d(\theta)$ be a measure of distance of an alternative $\theta \in \Omega_1$ to the given null hypothesis parameter space Ω_0 . A level α test ϕ is said to be locally most powerful (LMP) if, given any other level α test ϕ' , there exists $\Delta > 0$ such that

$$E_\theta[\phi(X)] \geq E_\theta[\phi'(X)] \quad \text{for all } \theta \text{ with } 0 < d(\theta) < \Delta.$$

Example 4.7 (Unit root testing, continued). In the setup of Example 4.2, it is easily seen that the LMP test rejects for small values of U_n .

5. LARGE-SAMPLE CONSIDERATIONS

Outside a narrow class of problems, finite-sample optimality notions do not directly apply. However, an asymptotic approach to optimality applies in a much broader class of models. Furthermore, control of the size of a test is often only approximate, and it is important to distinguish various notions of approximation.

As above, suppose that data $X^{(n)}$ comes from a model indexed by a parameter $\theta \in \Omega$. Consider testing Ω_0 against Ω_1 . Here we study sequences of tests $\phi_n = \phi_n(X^{(n)})$.

For a given level α , a sequence of tests $\{\phi_n\}$ is *pointwise asymptotically level α* if, for any $\theta \in \Omega_0$,

$$\limsup_{n \rightarrow \infty} E_\theta[\phi_n(X^{(n)})] \leq \alpha. \quad (6)$$

Condition 6 does not guarantee the size of ϕ_n is asymptotically no larger than α because the convergence is stated pointwise in θ . For this purpose, uniform convergence is required. The sequence $\{\phi_n\}$ is *uniformly asymptotically level α* if

$$\limsup_{n \rightarrow \infty} \sup_{\theta \in \Omega_0} E_\theta[\phi_n(X^{(n)})] \leq \alpha. \quad (7)$$

If, instead of Equation 7, the sequence $\{\phi_n\}$ satisfies

$$\lim_{n \rightarrow \infty} \sup_{\theta \in \Omega_0} E_\theta[\phi_n(X^{(n)})] = \alpha, \quad (8)$$

then this value of α is called the limiting size of $\{\phi_n\}$. Of course, we also study the approximate behavior of tests under the alternative hypothesis. For example, a sequence $\{\phi_n\}$ is *pointwise consistent in power* if, for any $\theta \in \Omega_1$,

$$\lim_{n \rightarrow \infty} E_\theta[\phi_n(X^{(n)})] = 1. \quad (9)$$

We note that the Bahadur-Savage result is not just a finite-sample phenomenon. In the context of their result, any test sequence whose size tends to α cannot have limiting power against any fixed alternative (or sequence of alternatives) larger than α . Uniformity is particularly important when the test statistic has an asymptotic distribution that is in some sense discontinuous in θ . Some recent papers in which uniformity plays a key role are Mikusheva (2007) and Andrews & Guggenberger (2009). However, the knowledge that ϕ_n is uniformly asymptotically level α does not alone guarantee anything about the size of ϕ_n for a given n ; one would also need to know how large an n is required for the size to be within a given ε of α .

5.1. Asymptotic Optimality

A quite general approach to asymptotic optimality is based on Le Cam's notion of convergence of experiments. The basic idea is that a general statistical problem (not just a testing problem) can often be approximated by a simpler problem (usually in the limit as the sample size tends to infinity). For example, it is a beautiful and astounding finding that the experiment consisting of observing n i.i.d. observations from an appropriately smooth parametric model $\{P_\theta, \theta \in \Omega\}$, where Ω is an open subset of \mathbb{R}^k , can be approximated by the experiment of observing a single multivariate normal vector X in \mathbb{R}^k with unknown mean and known covariance matrix.

The appropriate smoothness conditions are known as *quadratic mean differentiability*, which we now define. The context is that $X^{(n)} = (X_1, \dots, X_n)$ consists of n i.i.d. observations according to F_θ , where $\theta \in \Omega$, an open subset of \mathbb{R}^k . In other words, $P_\theta = F_\theta^n$. Assume F_θ is dominated by a common σ -finite measure μ , and let $f_\theta(x) = dF_\theta(x)/d\mu$. The family $\{F_\theta, \theta \in \Omega\}$ is *quadratic mean differentiable* (q.m.d.) at θ_0 if there exists a vector of real-valued functions $\eta(\cdot, \theta_0) = (\eta_1(\cdot, \theta_0), \dots, \eta_k(\cdot, \theta_0))^T$ such that

$$\int_X \left[\sqrt{f_{\theta_0+h}(x)} - \sqrt{f_{\theta_0}(x)} - \langle \eta(x, \theta_0), h \rangle \right]^2 d\mu(x) = o(|h|^2) \quad (10)$$

as $|h| \rightarrow 0$. For such a model, the Fisher Information matrix is defined to be the matrix $I(\theta)$ with (i, j) entry

$$I_{i,j}(\theta) = 4 \int \eta_i(x, \theta) \eta_j(x, \theta) d\mu(x).$$

The important consequence of q.m.d. models is Le Cam's expansion of the log of the likelihood function, which we now describe. Let $L_n(\theta) = \prod_{i=1}^n f_\theta(X_i)$ denote the likelihood function. Fix θ_0 . Define the normalized score vector Z_n by

$$Z_n = Z_{n,\theta_0} = n^{-1/2} \sum_{i=1}^n \tilde{\eta}(X_i, \theta_0), \quad (11)$$

where $\tilde{\eta}(x, \theta) = \frac{2\eta(x, \theta)}{f_\theta^{1/2}(x)}$. Then, if $I(\theta_0)$ is nonsingular,

$$\log \left[L_n(\theta_0 + hn^{-1/2}) / L_n(\theta_0) \right] = [b^T Z_n - \frac{1}{2} b^T I(\theta_0) b] + o_{P_{\theta_0}}(1). \quad (12)$$

If X is distributed as Q_b , the multivariate normal distribution with mean vector b and covariance matrix $I(\theta_0)$, then the term in brackets on the right-hand side of Equation 12 with $Z = I(\theta_0)X$ in place of Z_n is exactly $\log(dQ_b/dQ_0)$. In this sense, the log of the likelihood ratios approximates those from an experiment consisting of observing X from a multivariate normal distribution with unknown mean b and covariance matrix $I(\theta_0)$.

Such a local asymptotically normal expansion implies, among other things, the following. Suppose $\phi_n = \phi_n(X_1, \dots, X_n)$ is any sequence of tests. For fixed θ_0 , let $\beta_n(b) = E_{\theta_0+bn^{-1/2}}(\phi_n)$ be the *local power function*. Suppose that $\beta_n(b)$ converges to some function $\beta(b)$ for every b . Then, $\beta(b) = E_b(\phi(X))$ is the power function of a test ϕ in the limit experiment consisting of observing X from a multivariate normal distribution with unknown mean b and covariance matrix $I(\theta_0)$. Thus, the best achievable limiting power can be obtained by determining the optimal power in the limiting normal experiment. The

above results are developed in Lehmann & Romano (2005b, chapter 13), including numerous applications. To provide one example, suppose $\theta = (\theta_1, \dots, \theta_k)^T$, and the problem is to test $H_0: \theta_1 \leq 0$ versus $H_1: \theta_1 > 0$. The limit problem corresponds to testing $b_1 \leq 0$ versus $b_1 > 0$ based on X , and a UMP test exists for this problem, as described above in Example 4.1. For a test whose limiting size is no bigger than α , the resulting optimal limiting power against alternatives $\theta_1 = b_1 n^{-1/2}$ with $\theta_2, \dots, \theta_k$ fixed is

$$1 - \Phi\left(z_{1-\alpha} - b_1 \left\{I^{-1}(0, \theta_2, \dots, \theta_k)_{1,1}\right\}^{-1/2}\right). \quad (13)$$

Tests that achieve this limiting power are described later.

Even in nonparametric problems, the above development is useful because one can consider optimal limiting power among all appropriately smooth parametric submodels. The submodel yielding the smallest asymptotic power is then least favorable (see van der Vaart 1998, theorem 25.44).

On the other hand, there are many important nonstandard problems in econometrics in which the local asymptotically normal expansion does not hold. Even so, the idea of approximating by a limit experiment is still quite useful, as in the unit root problem.

Example 5.1 (Unit root testing, continued). In the setup of Example 4.2, the log likelihood ratio is given by

$$\log[L_n(1 + bn^{-1})/L_n(1)] = bU_n - \frac{1}{2}b^2V_n^2. \quad (14)$$

As is well-known, (U_n, V_n) tends to a limit law (under $\theta = \theta_n(b) = 1 + b/n$), which depends on the local parameter b . Even though the right-hand sides of both Equations 12 and 14 are quadratic in b , there are some crucial differences. First, the local parameter is of order n^{-1} from $\theta_0 = 1$, as opposed to the more typical case where it is of order $n^{-1/2}$. More important is that V_n tends to a limit law that is nondegenerate, which prevents the existence of a UMP or even an asymptotically UMP one-sided tests (e.g., see Crump 2008, lemma 1). Nevertheless, the limit experiment approach offers important insight into the behavior of power functions of various tests (see Jansson 2008, who removes the Gaussian assumption, among other things).

6. METHODS FOR HYPOTHESIS TESTING

There is no single method for constructing tests that is desirable for, or even applicable in, all circumstances. We therefore instead present several general principles that have been useful in different situations. We begin by considering parametric models and describe likelihood methods for testing certain hypotheses in such models. Then we introduce a broad class of possibly nonparametric models in which the parameter of interest is defined as the minimizer of a criterion function. This setup is sometimes referred to as the *extremum estimation* framework.

6.1. Testing in Parametric Models Using Likelihood Methods

In this section, assume that Ω is a subset of \mathbb{R}^k . For concreteness, we assume throughout this section that $P_\theta = F_\theta^n$, where each F_θ is absolutely continuous with respect to a common,

σ -finite dominating measure μ . Denote by f_θ the density of F_θ with respect to μ . In this notation,

$$L_n(\theta) = \prod_{1 \leq i \leq n} f_\theta(X_i).$$

6.1.1. Generalized likelihood ratio tests. As mentioned earlier, when both the null and alternative hypotheses are simple and specified as $\Omega_i = \{\theta_i\}$, MP tests are given by the *likelihood ratio test*, which rejects for large values of $L_n(\theta_1)/L_n(\theta_0)$. More generally, when either Ω_0 or Ω_1 is not simple, the *generalized likelihood ratio test* rejects for large values of

$$\frac{\sup_{\theta \in \Omega} L_n(\theta)}{\sup_{\theta \in \Omega_0} L_n(\theta)}.$$

Example 6.1 (Multivariate normal mean). Suppose that F_θ is the multivariate normal distribution with unknown mean vector $\theta \in \mathbb{R}^k$ and known covariance matrix Σ . Consider first testing the null hypothesis

$$\Omega_0 = \{0\}$$

versus the alternative $\Omega_1 = \mathbb{R}^k \setminus \Omega_0$. The generalized likelihood ratio test rejects for large values of

$$n\bar{X}_n^T \Sigma^{-1} \bar{X}_n. \tag{15}$$

If the critical value is chosen to be the $c_{k,1-\alpha}$, the $1 - \alpha$ quantile of the χ_k^2 distribution, then the resulting test has exact level α . Now consider testing the null hypothesis

$$\Omega_0 = \{\theta : \theta_i \leq 0 \ \forall i\}$$

versus the alternative $\Omega_1 = \mathbb{R}^k \setminus \Omega_0$. In this case, the generalized likelihood ratio test rejects for large values of

$$\inf_{\theta \in \Omega_0} n(\bar{X}_n - \theta)^T \Sigma^{-1} (\bar{X}_n - \theta).$$

If the critical value is chosen such that $P_0\{\inf_{\theta \in \Omega_0} n(\bar{X}_n - \theta)^T \Sigma^{-1} (\bar{X}_n - \theta) > c\} = \alpha$, then the resulting test again has exact level α .

6.1.2. Wald tests. Wald tests are based on a suitable estimator of θ . To describe this approach, we specialize to the case in which

$$\Omega_0 = \{\theta_0\},$$

$\Omega_1 = \mathbb{R}^k \setminus \Omega_0$, and the family $\{F_\theta : \theta \in \mathbb{R}^k\}$ is quadratic mean differentiable at θ_0 with nonsingular Fisher Information matrix $I(\theta_0)$ and score function Z_n defined in Equation 11. Assume further that $\hat{\theta}_n$ is an estimator of θ satisfying

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = I^{-1}(\theta_0)Z_n + o_{P_{\theta_0}}(1). \tag{16}$$

In some instances, such an estimator may be given by the *maximum likelihood estimator* (MLE) of θ , defined as

$$\hat{\theta}_n = \arg \max_{\theta \in \mathbb{R}^k} L_n(\theta).$$

Lehmann & Casella (1998), for example, present sufficient conditions for the existence of an estimator $\hat{\theta}_n$ satisfying Equation 16. From Equation 16, it follows that

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, I^{-1}(\theta_0)) \quad \text{under } P_{\theta_0}.$$

An example of a Wald test is the test that rejects for large values of

$$n(\hat{\theta}_n - \theta_0)^T I(\theta_0)(\hat{\theta}_n - \theta_0).$$

If the critical value is chosen to be $c_{k,1-\alpha}$, then the resulting test is pointwise asymptotically level α .

6.1.3. Rao score tests. Consider again the problem of testing the null hypothesis

$$\Omega_0 = \{\theta_0\}$$

versus the alternative $\Omega_1 = \mathbb{R}^k \setminus \Omega_0$. Suppose, as before, that $\{F_\theta : \theta \in \mathbb{R}^k\}$ is differentiable in quadratic mean at θ_0 with nonsingular Fisher Information $I(\theta_0)$ and score function Z_n defined in Equation 11. A disadvantage of Wald tests is that they involve the computation of a suitable estimator satisfying Equation 16. An alternative due to Rao that avoids this difficulty is based directly on Z_n defined in Equation 11. Under these assumptions,

$$Z_n \xrightarrow{d} N(0, I(\theta_0)) \quad \text{under } P_{\theta_0}.$$

An example of a Rao test in this case is the test that rejects for large values of

$$Z_n^T I^{-1}(\theta_0) Z_n.$$

If the critical value is chosen, as before, to be $c_{k,1-\alpha}$, then the resulting test is pointwise asymptotically level α .

Typically, the three preceding tests will behave similarly against alternatives local to the null hypothesis. For example, when testing the null hypothesis

$$\Omega_0 = \{\theta : \theta_1 \leq 0\}$$

versus the alternative $\Omega_1 = \mathbb{R}^k \setminus \Omega_0$, each of these three tests has limiting power given by Equation 13 against alternatives $\theta_1 = b/\sqrt{n}$, with $\theta_2, \dots, \theta_k$ fixed. On the other hand, there may still be important differences in the behavior of the three tests at nonlocal alternatives. A classical instance is provided by the Cauchy location model (see Lehmann & Romano 2005b, example 13.3.3).

6.2. Testing in the Extremum Estimation Framework

We now introduce a class of models in which Ω is not required to be a subset of \mathbb{R}^k . The extremum estimation framework provides a broad class of models that includes many nonparametric models. For ease of exposition, we also assume throughout this section that $P_\theta = F_\theta^\mu$, where each F_θ is absolutely continuous with respect to a common, σ -finite dominating measure μ . Denote by f_θ the density of F_θ with respect to μ . In this framework, we assume that the parameter of interest, $\gamma(F_\theta)$, may be written as

$$\gamma(F_\theta) = \arg \min_{\gamma \in \Gamma} Q(\gamma, F_\theta),$$

where

$$\Gamma = \{\gamma(F_\theta) : \theta \in \Omega\} \subseteq \mathbb{R}^d$$

and $Q : \Gamma \times \{F_\theta : \theta \in \Omega\} \rightarrow \mathbb{R}$. Denote by $\hat{Q}_n(\gamma)$ an estimate of $Q(\gamma, F_\theta)$ computed from $X^{(n)}$.

The following examples describe some important special cases of this framework that encompass a wide variety of applications in econometrics.

Example 6.2 (M-estimators). In many instances, $Q(\gamma, F_\theta) = E_\theta[q(X_i, \gamma)]$. Here, it is reasonable to choose

$$\hat{Q}_n(\gamma) = \frac{1}{n} \sum_{1 \leq i \leq n} q(X_i, \gamma).$$

The estimator $\hat{\gamma}_n = \arg \min_{\gamma \in \Gamma} \hat{Q}_n(\gamma)$ is referred to as an *M-estimator* in this case.

Example 6.3 (Generalized method of moments). Hansen (1982) considers the choice

$$Q(\gamma, P_\theta) = E_\theta[h(X_i, \gamma)]^T W(F_\theta) E_\theta[h(X_i, \gamma)],$$

where $W(F_\theta)$ is a positive definite matrix. Note that the dimension of h may exceed the dimension of γ . Here, it is reasonable to choose

$$\hat{Q}_n(\gamma) = \left[\frac{1}{n} \sum_{1 \leq i \leq n} h(X_i, \gamma) \right]^T \hat{W}_n \left[\frac{1}{n} \sum_{1 \leq i \leq n} h(X_i, \gamma) \right],$$

where \hat{W}_n is a consistent estimator of $W(F_\theta)$. The estimator $\hat{\gamma}_n = \arg \min_{\gamma \in \Gamma} \hat{Q}_n(\gamma)$ is referred to as the generalized method of moments (GMM) estimator in this case. If one wishes to minimize the asymptotic variance of the GMM estimator, then it is optimal to choose

$$W(F_\theta) = \left\{ E_\theta \left[h(X_i, \gamma(F_\theta)) h(X_i, \gamma(F_\theta))^T \right] \right\}^{-1}. \quad (17)$$

A consistent estimate of Equation 17 can be obtained in two steps, with $\gamma(F_\theta)$ consistently estimated in the first step. The large-sample efficiency of such estimators is studied in Chamberlain (1987).

Remark 6.1. If we take $\Gamma = \Theta$ and $q(X_i, \gamma) = -\log f_\gamma(X_i)$ in Example 6.2, then we see that the MLE is an *M-estimator*. In some cases, the MLE may also be characterized by the system of equations $\frac{1}{n} \sum_{1 \leq i \leq n} \nabla_\gamma \log f_\gamma(X_i) = 0$. When this is true, it can be thought of as a GMM estimator by taking $\Gamma = \Theta$ and $h(X_i, \gamma) = \nabla_\gamma \log f_\gamma(X_i)$ in Example 6.3. But, importantly, the MLE may not always be characterized in this fashion. To see this, we simply consider the example where F_θ is the uniform distribution on $[0, \theta]$.

Remark 6.2. In many applications, the parameter of interest may not be uniquely determined by the distribution of the observed data. We say that the parameter of interest in such models is *partially identified* (see Manski 2003 for numerous examples). For this reason, it is interesting to allow for the possibility of multiple minimizers of $Q(\gamma, F_\theta)$. Inference for such models is an

active area of research (for some recent contributions, see Chernozhukov et al. 2007 and Romano & Shaikh 2006a, 2008).

We consider testing the null hypothesis

$$\Omega_0 = \{\theta \in \Omega : \gamma(F_\theta) \in \Gamma_0\},$$

where Γ_0 is a fixed subset of Γ , versus the alternative

$$\Omega_1 = \{\theta \in \Omega : \gamma(F_\theta) \in \Gamma \setminus \Gamma_0\}.$$

The generalized likelihood ratio, Wald, and Rao tests have natural analogs in the extremum estimation framework. We now briefly describe these tests. The reader is referred to Newey & McFadden (1994) for further details.

6.2.1. Distance tests. By analogy with generalized likelihood ratio tests, distance tests are based on comparisons of $\inf_{\gamma \in \Gamma_0} \hat{Q}_n(\gamma)$ and $\inf_{\gamma \in \Gamma} \hat{Q}_n(\gamma)$. For example, one such test would reject the null hypothesis for large values of

$$n \left(\inf_{\gamma \in \Gamma_0} \hat{Q}_n(\gamma) - \inf_{\gamma \in \Gamma} \hat{Q}_n(\gamma) \right). \quad (18)$$

Example 6.4 (Generalized method of moments, continued). Recall the setup of Example 6.3 and suppose further that

$$\Gamma_0 = \{\theta \in \Omega : a(\gamma(F_\theta)) = 0\},$$

where $a : \Gamma \rightarrow \mathbb{R}^r$ is differentiable, and $\nabla_\gamma a(\gamma(F_\theta))$ has rank r for all $\theta \in \Omega_0$. Newey & West (1987) propose rejecting the null hypothesis for large values of Equation 18. If the critical value is chosen to be $c_{r,1-\alpha}$, then the resulting test is pointwise asymptotically level α under weak assumptions on Ω .

6.2.2. Wald tests. As mentioned above, Wald tests are based on a suitable estimator of $\gamma(F_\theta)$. To describe this approach, we specialize to the case in which

$$\Gamma_0 = \{\gamma_0\}.$$

We assume further that there is an estimator of $\gamma(F_\theta)$ satisfying

$$\sqrt{n}(\hat{\gamma}_n - \gamma(F_\theta)) \xrightarrow{d} N(0, V(F_\theta)),$$

where $V(F_\theta)$ is nonsingular, under P_θ with $\theta \in \Omega_0$. In some instances, such an estimator may be given by

$$\hat{\gamma}_n = \arg \min_{\gamma \in \Gamma} \hat{Q}_n(\gamma).$$

For sufficient conditions for the existence of such an estimator, we refer the reader to Newey & McFadden (1994), for example (see also van der Vaart & Wellner 1996 for empirical process techniques that are especially relevant for M -estimators). An example of a Wald test in this case is the test that rejects for large values of

$$n(\hat{\gamma}_n - \gamma_0)^T \hat{V}_n^{-1} (\hat{\gamma}_n - \gamma_0),$$

where \hat{V}_n is a consistent estimate of $V(F_\theta)$ under P_θ with $\theta \in \Omega_0$. If the critical value is chosen to be $c_{d,1-\alpha}$, then the resulting test is pointwise asymptotically level α .

6.2.3. Lagrange multiplier tests. Consider again the special case in which

$$\Gamma_0 = \{\gamma_0\}.$$

As discussed above, a disadvantage of Wald tests is that they require the computation of a suitable estimator of $\gamma(F_\theta)$. Suppose that $\hat{Q}_n(\gamma)$ is differentiable and that

$$\sqrt{n}\nabla_\gamma\hat{Q}_n(\gamma(F_\theta)) \xrightarrow{d} N(0, V(F_\theta)),$$

where $V(F_\theta)$ is nonsingular, under P_θ with $\theta \in \Omega_0$. In this case, one may overcome this difficulty by considering instead tests based on

$$\nabla_\gamma\hat{Q}_n(\gamma_0).$$

An example of a Lagrange multiplier test in this case is the test that rejects for large values of

$$n\nabla_\gamma\hat{Q}_n(\gamma_0)^T\hat{V}_n^{-1}\nabla_\gamma\hat{Q}_n(\gamma_0),$$

where \hat{V}_n is a consistent estimate of $V(F_\theta)$ under P_θ with $\theta \in \Omega_0$. If the critical value is chosen to be $c_{d,1-\alpha}$, then the resulting test is again pointwise asymptotically level α .

7. CONSTRUCTION OF CRITICAL VALUES

In the preceding section, we described several principles for constructing tests in both parametric and nonparametric models. Critical values were typically chosen by exploiting the fact that the test statistics under consideration were either pivots or asymptotic pivots; that is, their distributions or limiting distributions under P_θ with $\theta \in \Omega_0$ did not depend on P_θ . We now introduce some approaches for constructing critical values that may be applicable even when the test statistics are not so well behaved. In particular, we discuss randomization methods, the bootstrap, and subsampling. Even when the test statistics are pivots or asymptotic pivots, we see that there may be compelling reasons to use these methods instead.

7.1. Randomization Methods

We now introduce a general construction of tests that have exact level α for any sample size n whenever a certain invariance restriction holds. To describe this approach in more detail, let \mathbf{G} be a group of transformations of the data X . We require that $gX \stackrel{d}{=} X$ for any $g \in \mathbf{G}$ and $X \sim P_\theta$ with $\theta \in \Omega_0$. This assumption is sometimes referred to as the *randomization hypothesis*. For an appropriate choice of \mathbf{G} , the randomization hypothesis holds in a variety of commonly encountered testing problems.

Example 7.1 (One sample tests). Let $X^{(n)} = (X_1, \dots, X_n)$ consist of n i.i.d. observations from a distribution F_θ on the real line. Consider testing the null hypothesis

$$\Omega_0 = \{\theta \in \Omega : F_\theta \text{ symmetric about } 0\}.$$

The randomization hypothesis holds in this case with

$$\mathbf{G} = \{-1, 1\}^n$$

and the action of $g = (\varepsilon_1, \dots, \varepsilon_n) \in \mathbf{G}$ on X defined by $gX = (\varepsilon_1 X_1, \dots, \varepsilon_n X_n)$.

Example 7.2 (Two sample tests). Let $X^{(n)} = (X_1, \dots, X_n) = (Y_1, \dots, Y_\ell, Z_1, \dots, Z_m)$ be distributed according to P_θ , where Y_1, \dots, Y_ℓ are i.i.d. with distribution F_θ^Y and Z_1, \dots, Z_m are i.i.d. with distribution F_θ^Z . Consider testing the null hypothesis

$$\Omega_0 = \{\theta \in \Omega : F_\theta^Y = F_\theta^Z\}.$$

The randomization hypothesis holds in this case with \mathbf{G} given by the group of permutations of n elements and the action of g on X defined by $gX = (X_{g(1)}, \dots, X_{g(n)})$.

We now describe the construction. Let $T(X)$ be any real-valued test statistic such that we reject the null hypothesis for large values of $T(X)$. Suppose the group \mathbf{G} has M elements and let

$$T^{(1)}(X) \leq \dots \leq T^{(M)}(X)$$

denote the ordered values of $\{T(gX) : g \in \mathbf{G}\}$. Define $k = \lceil M \cdot (1 - \alpha) \rceil$, where $\lceil \cdot \rceil$ denotes the function that returns the least integer greater than or equal to its argument. Let

$$a(X) = \frac{M\alpha - M^+(X)}{M^0(X)},$$

where

$$M^0(X) = |\{1 \leq j \leq M : T^{(j)}(X) = T^{(k)}(X)\}|$$

$$M^+(X) = |\{1 \leq j \leq M : T^{(j)}(X) > T^{(k)}(X)\}|.$$

The test $\phi(X)$ that equals 1, $a(X)$, or 0 according to whether $T(X) > T^{(k)}(X)$, $T(X) = T^{(k)}(X)$, or $T(X) < T^{(k)}(X)$, respectively, has exact level α whenever the randomization hypothesis holds.

Remark 7.1. Even though it has exact size α , the test constructed above may not be interesting if it has poor power. After all, the test that simply rejects the null hypothesis with probability α also has this feature. It is therefore worth examining the power properties of tests constructed using randomization methods. For example, when testing whether the mean is less than or equal to zero versus greater than zero in a normal location model, the UMP test is, of course, the t -test. One may instead consider using the randomization test based on the group of transformations described in Example 7.1 and the t -statistic for this same problem. The randomization test is not UMP but has the benefit of not requiring the assumption of normality. On the other hand, it is possible to show that the randomization test has the same limiting power against contiguous alternatives, so there is no great loss of power, at least in large samples.

7.2. The Bootstrap

Unfortunately, randomization methods apply only to a restricted class of problems. The bootstrap was introduced in Efron (1979) as a broadly applicable method for approximating the sampling distribution of a statistic or, more generally, a *root*. A root is simply a real-valued function of the parameter of interest and the data. For ease of exposition, we assume that $P_\theta = F_\theta^n$. Denote by $J_n(x, F_\theta)$ the distribution of a root $R_n(X^{(n)}, \gamma(F_\theta))$ under $P_\theta = F_\theta^n$; that is,

$$J_n(x, F_\theta) = P_\theta \left\{ R_n(X^{(n)}, \gamma(F_\theta)) \leq x \right\}.$$

Our goal is to estimate $J_n(x, F_\theta)$ or its appropriate quantiles, which are typically unknown because F_θ is unknown. The bootstrap estimate of $J_n(x, F_\theta)$ is simply the plug-in estimate given by $J_n(x, \hat{F}_n)$, where \hat{F}_n is an estimate of F_θ . Because the data $X^{(n)} = (X_1, \dots, X_n)$ consists of n i.i.d. observations, one can use Efron's (1979) bootstrap (i.e., nonparametric bootstrap) or a suitable model-based bootstrap (i.e., parametric bootstrap) (e.g., see Davison & Hinkley 1997).

Sufficient conditions required for the validity of the bootstrap can be described succinctly in terms of a metric $d(\cdot, \cdot)$ on the space of distributions. In this notation, if we assume that (a) $J_n(x, F_n)$ converges weakly to a continuous limiting distribution $J(x, F_\theta)$ whenever $d(F_n, F_\theta) \rightarrow 0$ and $\theta \in \Omega_0$ and (b) $d(\hat{F}_n, F_\theta) \xrightarrow{P_\theta} 0$ whenever $\theta \in \Omega_0$, then

$$P_\theta \left\{ R_n(X^{(n)}, \gamma(F_\theta)) > J_n^{-1}(1 - \alpha, \hat{F}_n) \right\} \rightarrow \alpha$$

for all $\theta \in \Omega_0$. Here,

$$J_n^{-1}(1 - \alpha, \hat{F}_n) = \inf \left\{ x \in \mathbb{R} : J_n(x, \hat{F}_n) \geq 1 - \alpha \right\}.$$

In other words, we require that $J_n(x, F_\theta)$ must be sufficiently smooth in F_θ for the bootstrap to succeed.

There are often benefits to using the bootstrap even in simple problems. To illustrate this feature, suppose F_θ is a distribution on the real line with finite, nonzero variance for all $\theta \in \Omega$. Consider testing the null hypothesis

$$\Omega_0 = \{ \theta \in \Omega : \mu(F_\theta) \leq 0 \}$$

versus the alternative $\Omega_1 = \Omega \setminus \Omega_0$. For this problem, one possible test rejects when

$$\frac{\sqrt{n}\bar{X}_n}{\hat{\sigma}_n} > z_{1-\alpha}. \tag{19}$$

Instead of using $z_{1-\alpha}$, one could use $J_n^{-1}(1 - \alpha, \hat{F}_n)$, where \hat{F}_n is the empirical distribution of X_1, \dots, X_n and

$$R_n(X^{(n)}, \gamma(F_\theta)) = \frac{\sqrt{n}(\bar{X}_n - \mu(F_\theta))}{\hat{\sigma}_n}.$$

Both these tests are pointwise asymptotically level α , but, under further technical conditions ensuring the validity of Edgeworth expansions, it is possible to show that for any F_θ with $\theta \in \Omega_0$, the difference between the rejection probability and the nominal level is of order $O(n^{-1/2})$ for the first test and of order $O(n^{-1})$ for the second

test. Informally, the reason for this phenomenon is that the bootstrap approximation to the distribution of left-hand side of Equation 19, unlike the standard normal approximation, does not assume that the skewness of the finite-sample distribution of the t -statistic is zero (see Hall & Horowitz 1996 for related results in the context of GMM and Horowitz 2001 and MacKinnon 2007 for other applications of the bootstrap in econometrics).

In the above example, one could also use the bootstrap to approximate the distribution of the left-hand side of Equation 19 directly. In that case, one should use an estimate of F_θ that satisfies the constraints of the null hypothesis because critical values should be determined as if the null hypothesis were true. Such an approach is most useful for problems in which the hypotheses cannot be framed nicely in terms of parameters, such as testing for goodness of fit or for independence.

Unfortunately, there are many instances in which the required smoothness of $J_n(x, F_\theta)$ for the validity of the bootstrap does not hold. Examples include extreme order statistics, Hodges' superefficient estimator, and situations where the parameter lies on the boundary (see Beran 1984; Politis et al. 1999, chapter 1; and Andrews 2000 for further details).

7.3. Subsampling

Although the bootstrap is not universally applicable, an approach based on subsamples is often valid, at least in the sense that the probability of rejection tends to α under every P_θ with $\theta \in \Omega_0$, under very weak assumptions. To describe this approach, we also assume that $P_\theta = F_\theta^n$, but the approach can be easily modified for dependent data (see Politis et al. 1999, chapter 3). The key insight underlying this approach is that each subset of size b from these n observations constitutes b i.i.d. observations from F_θ . This suggests that the empirical distribution of the statistic of interest computed over these $\binom{n}{b}$ subsets of data should provide a reasonable estimate of the unknown distribution of the statistic.

More formally, let $\tilde{J}_n(x, F_\theta)$ be the distribution of a statistic T_n under P_θ . Index by $i = 1, \dots, \binom{n}{b}$ the subsets of data of size b and denote by $T_{n,b,i}$ the statistic T_n computed using the i -th subset of data of size b . Define

$$L_{n,b}(x) = \frac{1}{\binom{n}{b}} \sum_{1 \leq i \leq \binom{n}{b}} I\{T_{n,b,i} \leq x\}.$$

For the validity of this approach, we require that $b \rightarrow \infty$ so that $b/n \rightarrow 0$ and that $\tilde{J}_n(x, F_\theta)$ converge weakly to a continuous limiting distribution $\tilde{J}(x, F_\theta)$ whenever $\theta \in \Omega_0$. Under these assumptions,

$$P_\theta\{T_n > L_{n,b}^{-1}(1 - \alpha)\} \rightarrow \alpha \tag{20}$$

for all $\theta \in \Omega_0$. Here,

$$L_{n,b}^{-1}(1 - \alpha) = \inf\{x \in \mathbb{R} : L_{n,b}(x) \geq 1 - \alpha\}.$$

Remarkably, it is possible to show that

$$\sup_{\theta \in \Omega} P_{\theta} \{ \sup_{x \in \mathbb{R}} |L_{n,b}(x) - \tilde{J}_b(x, F_{\theta})| > \varepsilon \} \rightarrow 0$$

for any $\varepsilon > 0$ regardless of Ω . This suggests that whenever $\tilde{J}_b(x, F_{\theta})$ is suitably close to $\tilde{J}_n(x, F_{\theta})$, then subsampling may yield tests controlling the probability of a false rejection more strictly than Equation 20. For example, if

$$\lim_{n \rightarrow \infty} \sup_{\theta \in \Omega} \sup_{x \in \mathbb{R}} \{ \tilde{J}_b(x, F_{\theta}) - \tilde{J}_n(x, F_{\theta}) \} \leq \alpha,$$

then one has in fact

$$\lim_{n \rightarrow \infty} \sup_{\theta \in \Omega} P_{\theta} \{ T_n > L_{n,b}^{-1}(1 - \alpha) \} \leq \alpha \quad (21)$$

(see Romano & Shaikh 2008 for further details). Related results have also been obtained independently by Andrews & Guggenberger (2009), who go on to establish formulae for the left-hand side of Equation 21. Using these formulae, they establish in a variety of problems that the left-hand side of Equation 21 exceeds α , sometimes by a large margin. This problem may occur when the limiting distribution of the test statistic is discontinuous in F_{θ} . On the other hand, even when this is the case, subsampling may yield tests satisfying Equation 21, as shown by the following example.

Example 7.3 (Moment inequalities). The recent literature on partially identified models has focused considerable attention on testing the null hypothesis

$$\Omega_0 = \{ \theta \in \Omega : E_{\theta}[b(X_i, \gamma_0)] \leq 0 \}$$

for some fixed $\gamma_0 \in \Gamma$ versus the alternative $\Omega_1 = \Omega \setminus \Omega_0$. Note here that the dimension of $b(\cdot, \cdot)$ is allowed to be greater than one. This problem is closely related to the parametric problem described in Example 4.3. For this problem, subsampling leads to tests satisfying Equation 21 under very weak assumptions on Ω (for details, see Romano & Shaikh 2008 and Andrews & Guggenberger 2009).

8. MULTIPLE TESTING

8.1. Motivation

Much empirical research in economics involves the simultaneous testing of several hypotheses. For example, (a) one fits a multiple regression model and wishes to decide which coefficients are different from zero; (b) one compares several investment strategies to a benchmark and wishes to decide which strategies are outperforming the benchmark; and (c) one studies a number of active labor market programs and wishes to decide which programs are successful at bringing back the unemployed to the active labor force.

If one does not take the multiplicity of tests into account, there typically results a large probability that some of the true hypotheses will get rejected by chance alone. Take the case of $S = 100$ hypotheses being tested at the same time, all of them being true, with the size and level of each test exactly equal to α . For $\alpha = 0.05$, one expects five true hypotheses to be rejected. Furthermore, if all tests are mutually independent, then the probability that at least one true null hypothesis will be rejected is given by $1 - 0.95^{100} = 0.994$.

Of course, there is no problem if one focuses on a particular hypothesis, and only one of them, a priori. The decision can still be based on the corresponding individual p -value. The problem only arises if one searches the list of p -values for significant results a posteriori. Unfortunately, the latter case is much more common.

8.2. Notation and Various Error Rates

Again, we assume that data $X = X^{(n)}$ is generated from some probability distribution P_θ , with $\theta \in \Omega$. The problem is to simultaneously test the S null hypotheses $H_{0,s} : \theta \in \Omega_{0,s}$, with $H_{0,s}$ being tested against $H_{1,s} : \theta \in \Omega_{1,s}$. We also assume that a test of the individual hypothesis $H_{0,s}$ is based on a test statistic $T_{n,s}$, with large values indicating evidence against $H_{0,s}$. An individual p -value for testing $H_{0,s}$ is denoted by $\hat{p}_{n,s}$.

The term *false discovery* refers to the rejection of a true null hypothesis.¹ Also, let $\mathcal{I}(\theta)$ denote the set of true null hypotheses if θ is true; that is, $s \in \mathcal{I}(\theta)$ if and only if (iff) $\theta \in \Omega_{0,s}$.

Accounting for the multiplicity of individual tests can be achieved by controlling an appropriate *error rate*. The traditional *familywise error rate* (FWE) is the probability of one or more false discoveries:

$$\text{FWE}_\theta = P_\theta\{\text{reject at least one hypothesis } H_{0,s} : s \in \mathcal{I}(\theta)\}.$$

Of course, this criterion is strict; not even a single true hypothesis is allowed to be rejected. When S is very large, the corresponding multiple testing procedure (MTP) might result in low power, where we loosely define power as the ability to reject false null hypotheses.² Therefore, it can be beneficial to relax the criterion in return for higher power. There exist several possibilities to this end.

The *generalized familywise error rate* (k -FWE) is concerned with the probability of k or more false discoveries, where k is some positive integer:

$$k\text{-FWE}_\theta = P_\theta\{\text{reject at least } k \text{ hypotheses } H_{0,s} : s \in \mathcal{I}(\theta)\}.$$

Obviously, the special case $k = 1$ simplifies to the traditional FWE.

A related measure of error control is the average number of false discoveries, also known as the *per-family error rate* (PFER). To this end, let F denote the number of false rejections made by an MTP. Then, $\text{PFER}_\theta = E_\theta(F)$, where the concern now is to ensure $\text{PFER}_\theta \leq \lambda$ for some $\lambda \in [0, \infty)$.

Instead of error rates based only on the number of false discoveries, one can consider error rates based on the fraction of false discoveries (among all discoveries). Let R denote the total number of rejections. Then the *false discovery proportion* (FDP) is defined as $\text{FDP} = (F/R) \cdot 1\{R > 0\}$, where $1\{\cdot\}$ denotes the indicator function. One then is concerned with the probability of the FDP exceeding a small, prespecified proportion: $P_\theta\{\text{FDP} > \gamma\}$, for some $\gamma \in [0, 1)$. The special choice of $\gamma = 0$ simplifies to the traditional FWE.

Finally, the *false discovery rate* (FDR) is given by the expected value of the FDP. Namely, $\text{FDR}_\theta = E_\theta(\text{FDP})$, where the concern now is to ensure $\text{FDR}_\theta \leq \gamma$ for some $\gamma \in [0, 1)$.

¹Analogously, the term *discovery* refers to the rejection of any null hypothesis, and the term *true discovery* refers to the rejection of a false null hypothesis.

²If there is more than one null hypothesis under test, there no longer exists a unique definition of power. Some reasonable definitions include (a) the probability of rejecting at least one false null hypothesis, (b) the probability of rejecting all false null hypotheses, and (c) the average probability of rejection over the set of false null hypotheses.

The k -FWE, PFER, FDP, and FDR can all be coined *generalized error rates* in the sense that they relax or generalize the FWE. Although they are distinct, they share a common philosophy: By relaxing the FWE criterion and allowing for a small number (k -FWE), a small expected number (PFER), a small proportion (FDP), or a small expected proportion (FDR) of false discoveries, one is afforded greater power in return.

Having defined the various error rates, we next discuss what is meant by control of these error rates and what sort of conclusions one is afforded when applying corresponding MTPs to a set of data.

Control of the k -FWE means that, for a given significance level α ,

$$k\text{-FWE}_\theta \leq \alpha \quad \text{for any } \theta, \tag{22}$$

whereas control of the PFER means that, for a given integer k , $\text{PFER}_\theta \leq k$ for any θ . Moreover, control of the FDP means that, for a given significance level α and for a given proportion $\gamma \in [0, 1)$, $P_\theta \{\text{FDP} > \gamma\} \leq \alpha$ for any θ . Finally, control of the FDR means that, for a given proportion $\gamma \in [0, 1)$, $\text{FDR}_\theta \leq \gamma$ for any θ .

Which conclusions can be drawn when the various error rates are controlled? Control of the k -FWE allows one to be $1 - \alpha$ confident that there are at most $k - 1$ false discoveries among the rejected hypotheses. In particular, for $k = 0$, one can be $1 - \alpha$ confident that there are no false discoveries at all.

On the other hand, control of the PFER does not really allow one to draw any meaningful conclusion about the realized value of F (except for some crude bounds, based on Markov's inequality). The general reason is that by controlling an expected value, one can conclude little about the realization of the underlying random variable.

Control of the FDP allows one to be $1 - \alpha$ confident that the proportion of false discoveries among all rejected hypotheses is at most γ or, in other words, that the realized FDP is at most γ .

On the other hand, control of the FDR does not really allow one to draw any meaningful conclusion about the realized FDP. The general reason is, again, that by controlling an expected value, one can conclude little about the realization of the underlying random variable. Unfortunately, this important point is not always appreciated by researchers applying MTPs that control the FDR. Instead, by a law of large numbers, one might conclude that the average realized FDP—when FDR control is repeatedly applied to large number of data sets—will be at most γ (plus some small ε).

Remark 8.1 (Finite-sample versus asymptotic control). For this remark, we restrict attention to the FWE. The issues are completely analogous for the other error rates. Control of the FWE is equated with finite-sample control: Equation 22, with $k = 1$, is required to hold for any given sample size n . However, such a requirement can sometimes only be achieved under strict parametric assumptions (such as multivariate normality with known covariance matrix when testing a collection of individual means) or for special permutation setups. Instead, one settles for (pointwise) *asymptotic* control of the FWE:

$$\limsup_{n \rightarrow \infty} \text{FWE}_\theta \leq \alpha \quad \text{for any } \theta. \tag{23}$$

(In this section, all asymptotic considerations are restricted to pointwise asymptotics.)

Next, we discuss MTPs that (asymptotically) control the aforementioned error rates. Such procedures can roughly be classified according to two criteria. The first criterion is whether the method is based only on the individual p -values $\hat{p}_{n,s}$ or whether it is something more complex, trying to account for the dependence structure between the individual test statistics $T_{n,s}$. In general, methods of the latter type are more powerful. The second criterion is whether the method is a single-step method or a stepwise method. In general, methods of the latter type are more powerful. We begin by discussing the second criterion.

8.3. Single-Step Versus Stepwise Methods

In single-step methods, individual test statistics are compared to their critical values simultaneously, and after this simultaneous joint comparison, the multiple testing method stops. Often there is only one common critical value, but this need not be the case. More generally, the critical value for the s -th test statistic may depend on s . An example is the weighted Bonferroni method discussed below.

Often, single-step methods can be improved in terms of power via stepwise methods, while nevertheless maintaining control of the desired error rate. Stepdown methods start with a single-step method but then continue by possibly rejecting further hypotheses in subsequent steps. This is achieved by decreasing the critical values for the remaining hypotheses depending on the hypotheses already rejected in previous steps. As soon as no further hypotheses are rejected, the method stops. An example is given by the Holm (1979) method discussed below.

Such stepwise methods that improve upon single-step methods by possibly rejecting less significant hypotheses in subsequent steps are called stepdown methods. Intuitively, this is because such methods start with the most significant hypotheses, having the largest test statistics, and then step down to further examine the remaining hypotheses corresponding to smaller test statistics.

In contrast, there also exist stepup methods that start with the least significant hypotheses, having the smallest test statistics, and then step up to further examine the remaining hypotheses having larger test statistics. The crucial difference is that, at any given step, the question is whether to reject all remaining hypotheses or not. Therefore, the hypotheses sorted out in previous steps correspond to not-rejected hypotheses rather than rejected hypotheses, as in stepdown methods. A prominent example is the FDR controlling method of Benjamini & Hochberg (1995) discussed below.

8.4. Methods Based on Individual p -Values

MTPs falling in this category only work with the list of the individual p -values. They do not attempt to incorporate any dependence structure between these p -values. There are two advantages to such methods. First, one might have access only to the list of p -values from a past study, but not to the underlying complete data set. Second, such methods can be quickly implemented on the computer or even be carried out with paper and pencil. On the other hand, as we will see later, such methods are generally suboptimal in terms of power.

To show that such methods control the desired error rate, one needs a condition on the p -values corresponding to the true null hypotheses:

$$\theta \in \Omega_{0,s} \Rightarrow P_{\theta}\{\hat{p}_{n,s} \leq u\} \leq u \quad \text{for any } u \in (0, 1). \quad (24)$$

The classical method to control the FWE is the Bonferroni method. It is a single-step method providing control of the FWE. Specifically, it rejects $H_{0,s}$ iff $\hat{p}_{n,s} \leq \alpha/S$. More generally, the weighted Bonferroni method is a single-step method with the s -th cutoff value given by $w_s \cdot \alpha/S$, where the constants w_s reflect the importance of the individual hypotheses, satisfying $w_s \geq 0$ and $\sum w_s = 1$.

A stepdown improvement is obtained by the method of Holm (1979). The individual p -values are ordered from smallest to largest, $\hat{p}_{n,(1)} \leq \hat{p}_{n,(2)} \leq \dots \leq \hat{p}_{n,(s)}$, with their corresponding null hypotheses labeled accordingly. $H_{0,(1)}, H_{0,(2)}, \dots, H_{0,(s)}$. Then, $H_{0,(s)}$ is rejected iff $\hat{p}_{n,(j)} \leq \alpha/(S - j + 1)$ for $j = 1, \dots, s$. In other words, the method starts with testing the most significant hypothesis by comparing its p -value to α/S , just as in the Bonferroni method. If the hypothesis is rejected, the method moves on to the second most significant hypothesis by comparing its p -value to $\alpha/(S - 1)$, and so on, until the procedure comes to a stop. Necessarily, all hypotheses rejected by the Bonferroni method will also be rejected by the Holm method, but potentially the Holm method will reject a few more. So, trivially, the method is more powerful. But it still controls the FWE under Equation 24.

Both the Bonferroni and Holm methods can be easily generalized to control the k -FWE; these generalizations are due to Hommel & Hoffman (1988) and Lehmann & Romano (2005a). For the Bonferroni method, simply change the cutoff value from α/S to $k \cdot \alpha/S$. For the Holm method, change the cutoff values for the k most significant hypotheses to also $k \cdot \alpha/S$ and only then start subtracting one from the denominator in each subsequent step: Therefore, for $j > k$, the cutoff value in the j -th step is given by $k \cdot \alpha/(S - j + k)$. It becomes quite clear that even for a small value of $k > 1$, potentially many more hypotheses can be rejected as compared to FWE control.

The (generalized) Bonferroni and Holm methods are robust against the dependence structure of the p -values. They only need Equation 24 to provide control of the FWE and the k -FWE, respectively. Intuitively, they achieve this by ensuring control under a worst-case dependence structure.³ In contrast, the most widely known p -value-based methods to control the FDP and the FDR assume certain restrictions on the dependence structure.

Lehmann & Romano (2005a) develop a stepdown method to control the FDP. The individual p -values are ordered from smallest to largest again, like for the Holm method. Then, $H_{0,(s)}$ is rejected if $\hat{p}_{n,(j)} \leq \alpha_j$ for $j = 1, \dots, s$, with

$$\alpha_j = \frac{(\lfloor \gamma j \rfloor + 1)\alpha}{S + \lfloor \gamma j \rfloor + 1 - j},$$

where $\lfloor \cdot \rfloor$ denotes the integer part. This method provides control of the FDP under Equation 24 and the additional assumption that the p -values are mutually independent, or at least positively dependent in a certain sense (see Lehmann & Romano 2005a).

Benjamini & Hochberg (1995) propose a stepup method to control the FDR based on the ordered p -values. Define

$$j^* = \max\{j : \hat{p}_{n,(j)} \leq \gamma_j\}, \quad \text{where } \gamma_j = \frac{j}{S} \gamma,$$

³For example, as far as the Bonferroni method is concerned, this worst-case dependence structure is close to independence. Under independence, the cutoff value could be chosen as $1 - (1 - \alpha)^{1/S}$, which tends to be only slightly larger than α/S for nonextreme values of α and S ; e.g., for $\alpha = 0.05$ and $S = 100$, one obtains 0.000513 instead of 0.0005.

and then reject $H_{0,(1)}, \dots, H_{0,(j^*)}$. If no such j^* exists, reject no hypothesis. Unlike the stepdown methods discussed above, this MTP starts by examining the least significant hypothesis. If $\hat{p}_{n,(S)} \leq \gamma$, then all hypotheses are rejected. If not, $\hat{p}_{n,(S-1)}$ is compared to $(S-1)/S \cdot \gamma$, and so on. Benjamini & Hochberg (1995) prove control of this method under the assumption of independence. Benjamini & Yekutieli (2001) extend the validity of the method to a more general positive regression dependency.

Both the Lehmann & Romano (2005a) method to control the FDP and the Benjamini & Hochberg (1995) method to control the FDR can be modified to provide control under any dependence structure of the p -values. To this end, the cutoff values need to be suitably enlarged. However, the modified methods then turn quite conservative, so some users might shy away from them (for details, see Benjamini & Yekutieli 2001, as well as Lehmann & Romano 2005a and Romano & Shaikh 2006b). Stepup methods based on individual p -values to control the FWER, k -FWER, and FDP are discussed by Romano & Shaikh (2006c).

Remark 8.2 (Adaptive Benjamini and Hochberg method). Under conditions that ensure finite-sample control of the Benjamini & Hochberg (1995) method, it can be shown that $\text{FDR}_\theta = (S_0/S) \cdot \gamma$, where $S_0 = |\mathcal{I}(\theta)|$. So the method will generally be conservative, unless all null hypotheses are true. Therefore, power could be improved, while maintaining control of the FDR, by replacing the cutoff values by $\gamma_j = (j/S_0) \cdot \gamma$. Of course, S_0 is unknown in practice. But there exist several strategies to first come up with a (conservative) estimator of S_0 , denoted by \hat{S}_0 , and to then apply the method with cutoff values $\gamma_j = (j/\hat{S}_0) \cdot \gamma$. The literature in this field is quite extensive, and we refer the reader to Storey et al. (2004), Benjamini et al. (2006), Gavrilov et al. (2009), and the references therein.

Remark 8.3 (Finite-sample versus asymptotic control). So far, this subsection has assumed finite-sample validity of the null p -values expressed by Equation 24. However, often p -values are derived by asymptotic approximations or resampling methods, only guaranteeing asymptotic validity: For any (fixed) θ ,

$$\theta \in \Omega_{0,s} \Rightarrow \limsup_{n \rightarrow \infty} P_\theta \{ \hat{p}_{n,s} \leq u \} \leq u \quad \text{for any } u \in (0, 1). \quad (25)$$

Under this more realistic condition, the MTPs presented in this subsection only provide asymptotic control of their target error rates.

8.5. Resampling Methods Accounting for Dependence

As discussed above, p -value-based methods often achieve (asymptotic) control of their target error rates by assuming (a) a worst-case dependence structure or (b) a convenient dependence structure (such as mutual independence). This has two potential disadvantages. In case (a), the method can be quite suboptimal in terms of power if the true dependence structure is quite far away from the worst-case scenario. In case (b), asymptotic control may fail if the dependence structure does not hold.

As an example of case (a), consider the Bonferroni method. If there were perfect dependence between the p -values, the cut-off value could be changed from α/S to α . Perfect

dependence rarely happens in practice, of course. But this example is just to make a point. In the realistic setup of strong cross dependence, the cut-off value could be changed to something a lot larger than α/S while still maintaining control of the FWE. As an example of case (b), consider the adaptive method of Storey et al. (2004) to control the FDR. It assumes (near) mutual independence of the individual p -value. If this assumption is violated, the method can turn quite anticonservative, failing to control the FDR (see Romano et al. 2008a). Hence, both in terms of power and controlling an error rate, it is desirable to account for the underlying dependence structure.

Of course, this dependence structure is unknown and must be (implicitly) estimated from the available data. Consistent estimation, in general, requires that the sample size grow to infinity. Therefore, in this subsection, we settle for asymptotic control of the various error rates. In addition, we specialize to making simultaneous inference on the elements of a parameter vector $\theta = (\theta_1, \dots, \theta_S)^T$. The individual hypotheses can be all one-sided of the form

$$H_{0,s} : \theta_s \leq 0 \quad \text{vs.} \quad H_{1,s} : \theta_s > 0 \tag{26}$$

or they can be all two-sided of the form

$$H_{0,s} : \theta_s = 0 \quad \text{vs.} \quad H_{1,s} : \theta_s \neq 0. \tag{27}$$

Here we treat the one-sided case Equation (26); the necessary modifications for the two-sided case Equation (27) are given below.

The test statistics are of the form $T_{n,s} = \hat{\theta}_{n,s} / \hat{\sigma}_{n,s}$. Here, $\hat{\theta}_{n,s}$ is an estimator of θ_s computed from $X^{(n)}$. Furthermore, $\hat{\sigma}_{n,s}$ is either a standard error for $\hat{\theta}_{n,s}$ or simply equal to $1/\sqrt{n}$ in case such a standard error is not available or is quite difficult to obtain.

We start by discussing a single-step method for asymptotic control of the k -FWE. An idealized method would reject all $H_{0,s}$ for which $T_{n,s} \geq d_1$, where d_1 is the $1 - \alpha$ quantile under P_θ of the random variable $k\text{-max}_s (\hat{\theta}_{n,s} - \theta_s) / \hat{\sigma}_{n,s}$. Here, the k -max function selects the k -th largest element of an input vector. Naturally, the quantile d_1 depends not only on the marginal distributions of the centered statistics $(\hat{\theta}_{n,s} - \theta_s) / \hat{\sigma}_{n,s}$ but, crucially, also on their dependence structure.

As P_θ is unknown, the idealized critical value d_1 is not available. But it can be estimated consistently, under weak regularity conditions, as follows. Take \hat{d}_1 as the $1 - \alpha$ quantile under \hat{P}_n of $k\text{-max}_s (\hat{\theta}_{n,s}^* - \theta_{n,s}) / \hat{\sigma}_{n,s}^*$. Here, \hat{P}_n is an *unrestricted* estimate of P_θ . Furthermore, $\hat{\theta}_{n,s}^*$ and $\hat{\sigma}_{n,s}^*$ are the estimator of θ_s and its standard error (or simply $1/\sqrt{n}$), respectively, computed from $X^{(n),*}$, where $X^{(n),*} \sim \hat{P}_n$. In other words, we use the bootstrap to estimate d_1 . The particular choice of \hat{P}_n depends on the situation. If the data are i.i.d., one can use Efron's (1979) bootstrap (i.e., nonparametric bootstrap) or a suitable model-based bootstrap (i.e., parametric bootstrap) (e.g., see Davison & Hinkley 1997). If the data are dependent over time, one must use a suitable time-series bootstrap (e.g., see Lahiri 2003).

We have thus described a single-step MTP. However, a stepdown improvement is possible. Unfortunately, it is rather complex for general k ; the reader is referred to Romano et al. (2008b) for the details. However, it is straightforward for the special case of $k = 1$. In any given step j , one simply discards the hypotheses that have been rejected so far and applies the single-step MTP to the remaining family of nonrejected hypotheses. The resulting critical value \hat{d}_j necessarily satisfies $\hat{d}_j \leq \hat{d}_{j-1}$ so that new rejections may result; otherwise the method stops.

The modifications to the two-sided case Equation (27) are straightforward. First, the individual test statistics are now given by $z_{n,s} = |\hat{\theta}_{n,s}|/\hat{\sigma}_{n,s}$. Second, the idealized critical constants are now given by quantiles under P_θ of the random variable $k\text{-max}_s|\hat{\theta}_{n,s} - \theta_s|/\hat{\sigma}_{n,s}$, with the obvious implication for their estimation via the bootstrap.

Being able to control the k -FWE for any k enables us to easily control the FDP, accounting for the dependence structure. Set $k = 1$ and apply k -FWE control. If the number of rejections is less than $k/\gamma - 1$, stop. If not, let $k = k + 1$ and continue. In other words, one applies successive control of the k -FWE, with increasing k , until a stopping rule dictates termination.

Remark 8.4 (Asymptotic validity). The MTPs presented so far provide asymptotic control of their target error rates, namely k -FWE and FDP under remarkably weak regularity conditions. Mainly, it is assumed that $\sqrt{n}(\hat{\theta} - \theta)$ converges in distribution to a (multivariate) continuous limit distribution and that the bootstrap consistently estimates this limit distribution. In addition, if standard errors are employed for $\hat{\sigma}_{n,ss}$ as opposed to simply using $1/\sqrt{n}$, it is assumed that they converge to the same finite, nonzero limiting values in probability, both in the real world and in the bootstrap world. Under even weaker regularity conditions, a subsampling approach could be used instead. Furthermore, when a randomization setup applies, randomization methods can be used as an alternative (see Romano & Wolf 2005, 2007 for details).

Remark 8.5 (Alternative methods). Related bootstrap methods are developed in White (2000) and Hansen (2005). However, both works only treat the special case $k = 1$ and are restricted to single-step methods. In addition, White (2000) does not consider studentized test statistics.

Stepwise bootstrap methods to control the FWE have been already proposed in Westfall & Young (1993). An important difference in their approach is that they bootstrap under the joint null; that is, they use a *restricted* estimate of P_θ in which the constraints of all null hypotheses jointly hold. This approach requires the so-called *subset pivotality* condition and is generally less valid than the approaches discussed above based on an unrestricted estimate of P_θ (e.g., see Romano & Wolf 2005, example 4.1).

There exist alternative MTPs to control the k -FWE and the FDP: namely, the *augmentation* procedures of van der Laan et al. (2004) and *empirical Bayes* procedures of van der Laan et al. (2005). However, the former are suboptimal in terms of power, whereas the latter do not always provide asymptotic control and can be quite anticonservative (see Romano & Wolf 2007).

We finally turn to FDR control. As these methods are lengthy to describe, we restrict ourselves to a brief listing. Yekutieli & Benjamini (1999) propose a bootstrap method without discussing asymptotic properties. Dudoit et al. (2008) propose an empirical Bayes method that does not always provide asymptotic control and can be quite anticonservative. Romano et al. (2008a) propose a bootstrap method and prove asymptotic control under suitable regularity conditions. Also, in the simulations they consider, their method is more powerful than the Benjamini & Hochberg (1995) method and its adaptive versions which also are robust to a general dependence structure.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

We give special thanks to Tim Bresnahan for helpful comments. Our research was supported by the National Science Foundation grant DMS-0707085 (J.P.R.), by the National Science Foundation grant DMS-0820310 (A.M.S.), and by the Swiss National Science Foundation (NCCR FINRISK, Module A3) (M.W.).

LITERATURE CITED

- Andrews DWK. 1998. Hypothesis testing with a restricted parameter space. *J. Econom.* 84:155–99
- Andrews DWK. 2000. Inconsistency of the bootstrap when a parameter is on the boundary of the parameter space. *Econometrica* 68:399–405
- Andrews DWK, Guggenberger P. 2009. Validity of subsampling and ‘plug-in asymptotic’ inference for parameters defined by moment inequalities. *Econom. Theory* 25(3):669–709
- Andrews DWK, Moreira MJ, Stock JH. 2006. Optimal two-sided invariant similar tests for instrumental variables regression. *Econometrica* 74:715–52
- Bahadur RR, Savage LJ. 1956. The nonexistence of certain statistical procedures in nonparametric problems. *Ann. Math. Stat.* 27(4):1115–22
- Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* 57(1):289–300
- Benjamini Y, Krieger AM, Yekutieli D. 2006. Adaptive linear step-up procedures that control the false discovery rate. *Biometrika* 93(3):491–507
- Benjamini Y, Yekutieli D. 2001. The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.* 29(4):1165–88
- Beran R. 1984. Bootstrap methods in statistics. *Jahresber. Dtsch. Math. Ver.* 86:14–30
- Brown LD. 1986. *Fundamentals of Statistical Exponential Families*. Vol. 9. Lect. Notes Monogr. Ser. Hayward, CA: Inst. Math. Stat.
- Chamberlain G. 1987. Asymptotic efficiency in estimation with conditional moment restrictions. *J. Econom.* 34(3):305–34
- Chernozhukov V, Hong H, Tamer E. 2007. Estimation and confidence regions for parameter sets in econometric models. *Econometrica* 75(5):1243–84
- Chiburis R. 2008. *Approximately most powerful tests for moment inequalities*. Job market paper, Dep. Econ., Princeton Univ.
- Crump R. 2008. *Optimal conditional inference in nearly-integrated autoregressive processes*. Job market paper, Dep. Econ., Univ. Calif., Berkeley
- Davidson R, Duclos J-Y. 2006. *Testing for restricted stochastic dominance*. Discuss. Pap. 2047, IZA. <http://ssrn.com/abstract=894061>
- Davison AC, Hinkley DV. 1997. *Bootstrap Methods and Their Application*. Cambridge, UK: Cambridge Univ. Press
- Dickey DA, Fuller WA. 1979. Distribution of the estimators for autoregressive time series with a unit root. *J. Am. Stat. Assoc.* 74:427–31
- Dudoit S, Gilbert H, van der Laan MJ. 2008. Resampling-based empirical Bayes multiple testing procedures for controlling generalized tail probability and expected value error rates: focus on the false discovery rate and simulation study. *Biom. J.* 50(5):716–44

- Dufour J-M. 1997. Some impossibility theorems in econometrics with applications to structural and dynamic models. *Econometrica* 65(6):1365–87
- Efron B. 1979. Bootstrap methods: another look at the jackknife. *Ann. Stat.* 7:1–26
- Elliot G, Rothenberg TJ, Stock JH. 1996. Efficient tests for an autoregressive unit root. *Econometrica* 64:813–36
- Gavrilov Y, Benjamini Y, Sarkar SK. 2009. An adaptive step-down procedure with proven FDR control. *Ann. Stat.* 37(2):619–29
- Hall P, Horowitz J. 1996. Bootstrap critical values for tests based on generalized-method-of-moments estimators. *Econometrica* 64(4):891–916
- Hansen L. 1982. Large sample properties of generalized method of moments estimators. *Econometrica* 50:1029–54
- Hansen PR. 2005. A test for superior predictive ability. *J. Bus. Econ. Stat.* 23:365–80
- Holm S. 1979. A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* 6:65–70
- Hommel G, Hoffman T. 1988. Controlled uncertainty. In *Multiple Hypothesis Testing*, ed. P Bauer, G Hommel, E Sonnemann, pp. 154–61. Heidelberg: Springer
- Horowitz J. 2001. The bootstrap. In *Handbook of Econometrics*, Vol. 5, ed. Z Griliches, JJ Heckman, MD Intriligator, EE Leamer, pp. 3159–228. Amsterdam: North Holland
- Jansson M. 2008. Semiparametric power envelopes for tests of the unit root hypothesis. *Econometrica* 76:1103–42
- Lahiri SN. 2003. *Resampling Methods for Dependent Data*. New York: Springer
- Lehmann EL. 1952. Testing multiparameter hypotheses. *Ann. Math. Stat.* 23:541–52
- Lehmann EL, Casella G. 1998. *Theory of Point Estimation*. New York: Springer
- Lehmann EL, Romano JP. 2005a. Generalizations of the familywise error rate. *Ann. Stat.* 33(3):1138–54
- Lehmann EL, Romano JP. 2005b. *Testing Statistical Hypotheses*. New York: Springer. 3rd ed.
- MacKinnon J. 2007. *Bootstrap hypothesis testing*. Work. Pap. 1127, Dep. Econ., Queen's Univ.
- Manski CF. 2003. *Partial Identification of Probability Distributions*. New York: Springer
- Mikusheva A. 2007. Uniform inference in autoregressive models. *Econometrica* 75(5):1411–52
- Moreira MJ. 2003. A conditional likelihood ratio test for structural models. *Econometrica* 71(4):1027–48
- Newey WK, McFadden DL. 1994. Large-sample estimation and hypothesis testing. In *Handbook of Econometrics*, Vol. 4, ed. RF Engle, DL McFadden, pp. 2111–245. Amsterdam: Elsevier
- Newey WK, West KD. 1987. A simple positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica* 55:703–8
- Politis DN, Romano JP, Wolf M. 1999. *Subsampling*. New York: Springer
- Romano JP. 2004. On nonparametric testing, the uniform behavior of the t -test, and related problems. *Scand. J. Stat.* 31(4):567–84
- Romano JP, Shaikh AM. 2006a. Inference for the identified set in partially identified econometric models. *Tech. Rep. 2006-10*, Dep. Statistics, Stanford Univ.
- Romano JP, Shaikh AM. 2006b. On stepdown control of the false discovery proportion. In *Optimality: The Second Erich L. Lehmann Symposium*, ed. J Rojo, pp. 33–50. Beachwood, OH: Inst. Math. Stat.
- Romano JP, Shaikh AM. 2006c. Stepup procedures for control of generalizations of the familywise error rate. *Ann. Stat.* 34(4):1850–73
- Romano JP, Shaikh AM. 2008. Inference for identifiable parameters in partially identified econometric models. *J. Stat. Plan. Inference* 138(9):2786–807
- Romano JP, Shaikh AM, Wolf M. 2008a. Control of the false discovery rate under dependence using the bootstrap and subsampling (with discussion). *Test* 17(3):417–42
- Romano JP, Shaikh AM, Wolf M. 2008b. Formalized data snooping based on generalized error rates. *Econom. Theory* 24(2):404–47

- Romano JP, Wolf M. 2000. Finite sample nonparametric inference and large sample efficiency. *Ann. Stat.* 28(3):756–78
- Romano JP, Wolf M. 2005. Exact and approximate stepdown methods for multiple hypothesis testing. *J. Am. Stat. Assoc.* 100:94–108
- Romano JP, Wolf M. 2007. Control of generalized error rates in multiple testing. *Ann. Stat.* 35(4):1378–408
- Storey JD, Taylor JE, Siegmund D. 2004. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *J. R. Stat. Soc. B* 66(1):187–205
- van der Laan MJ, Birkner MD, Hubbard AE. 2005. Empirical bayes and resampling based multiple testing procedure controlling tail probability of the proportion of false positives. *Stat. Appl. Genet. Mol. Biol.* 4(1):29
- van der Laan MJ, Dudoit S, Pollard KS. 2004. Augmentation procedures for control of the generalized family-wise error rate and tail probabilities for the proportion of false positives. *Stat. Appl. Genet. Mol. Biol.* 3(1):15
- van der Vaart AW. 1998. *Asymptotic Statistics*. Cambridge, UK: Cambridge University Press
- van der Vaart AW, Wellner JA. 1996. *Weak Convergence and Empirical Processes*. New York: Springer
- Westfall PH, Young SS. 1993. *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment*. New York: John Wiley
- White HL. 2000. A reality check for data snooping. *Econometrica* 68(5):1097–126
- Yekutieli D, Benjamini Y. 1999. Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. *J. Stat. Plann. Inference* 82:171–96



Contents

Questions in Decision Theory <i>Itzhak Gilboa</i>	1
Structural Estimation and Policy Evaluation in Developing Countries <i>Petra E. Todd and Kenneth I. Wolpin</i>	21
Currency Unions in Prospect and Retrospect <i>J.M.C. Santos Silva and Silvana Tenreyro</i>	51
Hypothesis Testing in Econometrics <i>Joseph P. Romano, Azeem M. Shaikh, and Michael Wolf</i>	75
Recent Advances in the Empirics of Organizational Economics <i>Nicholas Bloom, Raffaella Sadun, and John Van Reenen</i>	105
Regional Trade Agreements <i>Caroline Freund and Emanuel Ornelas</i>	139
Partial Identification in Econometrics <i>Elie Tamer</i>	167
Intergenerational Equity <i>Geir B. Asheim</i>	197
The World Trade Organization: Theory and Practice <i>Kyle Bagwell and Robert W. Staiger</i>	223
How (Not) to Do Decision Theory <i>Eddie Dekel and Barton L. Lipman</i>	257
Health, Human Capital, and Development <i>Hoyt Bleakley</i>	283
Beyond Testing: Empirical Models of Insurance Markets <i>Liran Einav, Amy Finkelstein, and Jonathan Levin</i>	311
Inside Organizations: Pricing, Politics, and Path Dependence <i>Robert Gibbons</i>	337

Identification of Dynamic Discrete Choice Models <i>Jaap H. Abbring</i>	367
Microeconomics of Technological Adoption <i>Andrew D. Foster and Mark R. Rosenzweig</i>	395
Heterogeneity, Selection, and Wealth Dynamics <i>Lawrence Blume and David Easley</i>	425
Social Interactions <i>Steven N. Durlauf and Yannis M. Ioannides</i>	451
The Consumption Response to Income Changes <i>Tullio Jappelli and Luigi Pistaferri</i>	479
Financial Structure and Economic Welfare: Applied General Equilibrium Development Economics <i>Robert Townsend</i>	507
Models of Growth and Firm Heterogeneity <i>Erzo G.J. Luttmer</i>	547
Labor Market Models of Worker and Firm Heterogeneity <i>Rasmus Lentz and Dale T. Mortensen</i>	577
The Changing Nature of Financial Intermediation and the Financial Crisis of 2007–2009 <i>Tobias Adrian and Hyun Song Shin</i>	603
Competition and Productivity: A Review of Evidence <i>Thomas J. Holmes and James A. Schmitz, Jr.</i>	619
Persuasion: Empirical Evidence <i>Stefano DellaVigna and Matthew Gentzkow</i>	643
Commitment Devices <i>Gharad Bryan, Dean Karlan, and Scott Nelson</i>	671

Errata

An online log of corrections to *Annual Review of Economics* articles may be found at <http://econ.annualreviews.org>