Institute for Empirical Research in Economics
University of Zurich

Working Paper Series
ISSN 1424-0459

Working Paper No. 444

# Hypothesis Testing in Econometrics

Joseph P. Romano, Azeem M. Shaikh and Michael Wolf

September 2009

# Hypothesis Testing in Econometrics

Joseph P. Romano
Departments of Econmics and Statistics
Stanford University
romano@stanford.edu

Azeem M. Shaikh
Department of Economics
University of Chicago
amshaikh@uchicago.edu

Michael Wolf
Institute for Empirical Research in Economics
University of Zurich
mwolf@iew.uzh.ch

September 2009

**Abstract**

This paper 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, some of the most important methods are summarized, 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 we include many problems with well-known successful solutions, we also include open problems that are not easily handled with current technology, stemming from issues like lack of optimality or poor asymptotic approximations.

KEY WORDS: Asymptotics, multiple testing, optimality, resampling.

JEL CLASSIFICATION NOS: C12.

ACKNOWLEDGMENTS: Special thanks to Tim Bresnahan for helpful comments. Research of Romano supported by the National Science Foundation grant DMS-0707085. Research of Shaikh supported by the National Science Foundation grant DMS-0820310. Research of Wolf supported by the Swiss National Science Foundation (NCCR FINRISK, Module A3).

# 1 INTRODUCTION

This paper highlights many of the 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 which 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 of these approaches, and sometimes in combination, have been successfully used in econometric problems.

Next, various methods which are 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 constructing 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 is very 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 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 $\{P_\theta,\ \theta \in \Omega\}$, though we make no rigid requirements for $\Omega$; it may be a parametric, semiparametric or nonparametric model. A general hypothesis about the underlying model can be specified by a subset of $\Omega$.

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, $\Omega_0$ and $\Omega_1$ are disjoint subsets of $\Omega$ with union $\Omega$. A hypothesis is called *simple* if it completely specifies the distribution of $X$, or equivalently a particular $\theta$; 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 can be described as constructing 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

$$P_\theta\{S_1\} \le \alpha \qquad \text{for all } \theta \in \Omega_0 \ .$$

It is important to note that acceptance of $H_0$ does not necessarily show $H_0$ is indeed true; there simply may be insufficient data to show inconsistency of the data with the null hypothesis. So, the decision which “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

$$\beta_\phi(\theta) = E_\theta\big[\phi(X)\big] = \int \phi(x)dP_\theta(x) \ .$$

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

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

A test satisfying (1) is said to be *level* $\alpha$. The supremum over $\theta \in \Omega_0$ of the left side of (1) is the *size* of the test $\phi$.

# 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 pre-specified significance level $\alpha$, the *critical value* $c$ is chosen (possibly in a data-dependent way and on $\alpha$) to control the size of the test, though 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 $\alpha$, then it is easy to see that,

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

## 3.2 The Bahadur-Savage Result

The problem of constructing a level $\alpha$ test can be nontrivial, in the sense that the level constraint may prohibit the construction of a test that has any power to detect $H_1$. To put it another way, there may exist situations where it is impossible to construct a level $\alpha$ test that has power bigger than $\alpha$ against even one alternative. A classical instance of the nonexistence of any useful level $\alpha$ test was provided by Bahadur and Savage (1956). The result is stated in terms of testing the mean of a population in a nonparametric setting. Suppose $X_1, \ldots, X_n$ are i.i.d. with c.d.f. $F$ 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 $\theta$) is used to index the model $\mathbf{F}$. The family $\mathbf{F}$ is assumed to satisfy the following:

(i) For every $F \in \mathbf{F}$, $\mu(F)$ exists and is finite.

(ii) For every real $m$, there is an $F \in \mathbf{F}$ with $\mu(F) = m$.

(iii) If $F_i \in \mathbf{F}$ 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 $\phi = \phi(X_1, \ldots, X_n)$ is a level $\alpha$ test. Then, for any $F$ with $\mu(F) \neq 0$, $E_F(\phi) \leq \alpha$; that is, the power of the test cannot exceed $\alpha$ 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 which 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 and Wolf (2000). For other nonexistence results, see Dufour (1997) and Romano (2004). In summary, the main point of the example is that, 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 $\alpha$, independent of the data, and the only way to avoid this is "reduce" the size of the model.

## 4 OPTIMALITY CONSIDERATIONS

For a given alternative $\theta_1 \in \Omega_1$, the problem of determining $\phi$ to maximize $\beta_\phi(\theta_1)$ subject to (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 $\alpha$. Typically, the optimal $\phi$ will depend on the fixed alternative $\theta_1$. If a test $\phi$ exists that maximizes the power for all $\theta_1 \in \Omega_1$, then $\phi$ is called *uniformly most powerful* (UMP) level $\alpha$.

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 $\alpha$ 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 level $\alpha$ test to be most powerful 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 $\theta$, UMP tests exist for one-sided hypotheses $H_0$ specified by $\Omega_0 = \{\theta : \ \theta \leq \theta_0\}$ for some fixed $\theta_0$, assuming the underlying family has monotone likelihood ratio. For two-sided hypotheses, UMP tests are rare. In multiparameter models where $\theta \in \mathbb{R}^d$ or where $\Omega$ is infinite-dimensional, UMP tests typically do not exist. The nonparametric sign test is an exception; see Example 3.8.1 of Lehmann and Romano (2005b). The following is an example where 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 $\Sigma$. Fix a vector $(a_1, \ldots, a_d)^T \in \mathbb{R}^d$ and a number $\delta$. For testing the null hypothesis

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

against $\Omega_1 = \mathbb{R}^d \setminus \Omega_0$, there exists a UMP level $\alpha$ test which 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 UMPU Tests

A test $\phi$ is called level $\alpha$ *unbiased* if

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

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 $\phi$ is called UMP unbiased (or UMPU) at level $\alpha$ if $\beta_\phi(\theta)$ is maximized uniformly over $\theta \in \Omega_1$ among all level $\alpha$ 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 $\alpha$ 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; 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 where 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 Chapter 4 of Lehmann and Romano (2005b), 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 (1) hold, but also the stronger conditional level constraint that, for (almost) all outcomes $t$ of a conditioning statistic $T$,

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

An optimal test may exist within this smaller class of tests, though the reduction to such a class may or may not have any compelling merit to it, since 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 $\theta$. See Section 10.3 of Lehmann and Romano (2005b) for some optimal conditional tests, where conditioning is done using an ancillary statistic. We now mention a recent important example where the conditioning statistic is not ancillary, though 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 and Fuller (1979). We discuss some of the 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

$$X_t = \theta X_{t-1} + \epsilon_t \quad t = 1, \ldots, n \ ,$$

where $\theta \in (-1, 1]$ and the $\epsilon_t$ are unobserved and i.i.d. Gaussian with mean 0 and known variance $\sigma^2$. Consider the problem of testing $\theta = 1$ against $\theta < 1$. The likelihood function

$L_n(\theta)$ is given by

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

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 $h$ does not depend on $\theta$. This constitutes a curved exponential family. For a fixed alternative $\theta'$, the MP test rejects for large values of

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

Since the optimal rejection region is seen to depend on $\theta'$, no UMP test exists. An interesting way to choose $\theta'$ is suggested in Elliot et al. (1996), though using a particular $\theta'$ does not imply any optimality of the power function at another $\theta$. In Crump (2008), optimal tests are constructed conditional on $U_n$, since 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 UMPI Tests

Some testing problems exhibit symmetries, which lead to natural restrictions on the family of tests considered. The mathematical expression of symmetry is now described. 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 $\theta$, then $gX$ also has a distribution in the model; that is, $gX$ has distribution governed by some $\theta' \in \Omega$. The element $\theta'$ 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 $\Omega_0$ against $\Omega_1$ remains invariant under $G$ if both $\Omega_0$ and $\Omega_1$ remain invariant under any $g \in G$. This structure implies that a statistician testing $\Omega_0$ against $\Omega_1$ 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 $\phi$ 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 $\alpha$ tests is called uniformly most powerful invariant (UMPI) at level $\alpha$.

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 like regression and analysis of variance, where least squares procedures and standard $F$ tests are shown to have optimality properties among invariant procedures. Note, however, UMPI tests may be inadmissible in some problems. See 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 $X = (X_1, \ldots, X_d)^T$ is multivariate normal with unknown mean vector $\theta = (\theta_1, \ldots, \theta_d)^T$ and known nonsingular covariance matrix $\Sigma$. 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 Problem 4.8 in Lehmann and Romano (2005b). Nor do any invariance considerations generally apply. In the special case that $\Sigma$ 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 which are invariant under permutations. Even so, such transformations do not reduce the problem sufficiently far to lead to any optimal procedure.

However, a natural monotonicity restriction on a test $\phi$ is the following. Specifically, if $\phi$ rejects based on data $X$, so that $\phi(X) = 1$, and $Y = (Y_1, \ldots, Y_d)^T$ with $Y_i \geq X_i$ for all $i$, then a monotonicity requirement demands that $\phi(Y) = 1$. We will return to this example later. 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 $\theta$ 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 problem is a simplified version of the problem of testing for stochastic dominance; see Davidson and Duclos (2006). Among tests obeying the same monotonicity restriction as in Example 4.3, there exists a UMP level $\alpha$ test; see Lehmann (1952). ■

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

### 4.5 Maximin Tests

For testing $\Omega_0$ against $\Omega_1$, let $\omega_1 \subseteq \Omega_1$ be a (possibly strict) subset of $\Omega_1$. A level $\alpha$ test $\phi$ is *maximin* with respect to $\omega_1$ at level $\alpha$ if it is level $\alpha$ and it maximizes $\inf_{\theta \in \omega_1} E_\theta\big[\phi(X)\big]$ among level $\alpha$ 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 $\Sigma$ 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) in the case $d = 2$. The result generalizes, and it can also be shown (in unpublished work) that such a test is admissible among all tests (obeying the level constraint) without any monotonicity restriction. ■

### 4.6 Tests Maximizing Average or Weighted Power

Let $\Lambda_1$ be a probability distribution (or generally a nonnegative measure) over $\Omega_1$. The *average* or *weighted power* of a test $\phi$ with respect to $\Lambda_1$ is given by $\int_{\Omega_1} E_\theta\big[\phi(X)\big] d\Lambda_1(\theta)$. A level $\alpha$ test $\phi$ maximizing this quantity among all level $\alpha$ tests maximizes average power with respect to $\Lambda_1$. The approach to determining such a test is to note that this average power of $\phi$ can be expressed as the power of $\phi$ 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 most powerful level $\alpha$ test of $\Omega_0$ 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 which (approximately) maximize average power. Also, see Andrews (1998). Such an approach can provide tests with reasonably good power properties, though the choice of the averaging distribution $\Lambda$ 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 $\Omega_0$. A level $\alpha$ test $\phi$ is said to be locally most powerful (LMP) if, given any other level $\alpha$ test $\phi'$, there exists $\Delta > 0$ such that

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

**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 approximated, and it is important to distinguish various notions of approximation.

As before, suppose that data $X^{(n)}$ comes from a model indexed by a parameter $\theta \in \Omega$. Consider testing $\Omega_0$ against $\Omega_1$. We will be studying sequences of tests $\phi_n = \phi_n(X^{(n)})$.

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

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

Condition (6) does not guarantee the size of $\phi_n$ is asymptotically no bigger than $\alpha$ since the convergence is stated pointwise in $\theta$. For this purpose, uniform convergence is required.

The sequence $\{\phi_n\}$ is *uniformly asymptotically level* $\alpha$ if

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

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

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

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

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

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 $\alpha$ cannot have limiting power against any fixed alternative (or sequence of alternatives) bigger than $\alpha$. Uniformity is particularly important when the test statistic has an asymptotic distribution which is in some sense discontinuous in $\theta$. Some recent papers where uniformity plays a key role are Mikusheva (2007), and Andrews and Guggenberger (2009). Note, however, knowing that $\phi_n$ is uniformly asymptotically level $\alpha$ does not alone guarantee anything about the size of $\phi_n$ for a given $n$; one would also need to know how large an $n$ is required for the size to be within a given $\epsilon$ of $\alpha$.

### 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 $\Omega$ 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 $X^{(n)} = (X_1, \ldots, X_n)$ consists of $n$ i.i.d. observations according to $F_\theta$, where $\theta \in \Omega$, an open subset of $\mathbb{R}^k$. In other words, $P_\theta = F_\theta^n$. Assume $F_\theta$ is dominated by a common $\sigma$-finite measure $\mu$, and let $f_\theta(x) = dF_\theta(x)/d\mu$. The family $\{F_\theta, \theta \in \Omega\}$ is *quadratic mean differentiable* (abbreviated q.m.d.) at $\theta_0$ if there exists a vector of real-valued functions $\eta(\cdot, \theta_0) = \big(\eta_1(\cdot, \theta_0), \ldots, \eta_k(\cdot, \theta_0)\big)^T$ such that

$$\int_{\mathcal{X}} \left[ \sqrt{f_{\theta_0+h}(x)} - \sqrt{f_{\theta_0}(x)} - < \eta(x, \theta_0), h > \right]^2 d\mu(x) = o\big(|h|^2\big) \tag{10}$$

as $|h| \to 0$. 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 $\theta_0$. Define the normalized score vector $Z_n$ by

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

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

$$\log\bigl[L_n(\theta_0 + hn^{-1/2})/L_n(\theta_0)\bigr] = [h^T Z_n - \frac{1}{2}h^T I(\theta_0)h] + o_{P_{\theta_0}}(1). \tag{12}$$

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

Such a local asymptotically normal (LAN) expansion implies, among other things, the following. Suppose $\phi_n = \phi_n(X_1, \ldots, X_n)$ is any sequence of tests. For fixed $\theta_0$, let $\beta_n(h) = E_{\theta_0 + hn^{-1/2}}(\phi_n)$ be the *local power function*. Suppose that $\beta_n(h)$ converges to some function $\beta(h)$ for every $h$. Then, $\beta(h) = E_h\bigl(\phi(X)\bigr)$ is the power function of a test $\phi$ in the limit experiment consisting of observing $X$ from a multivariate normal distribution with unknown mean $h$ 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 Chapter 13 of Lehmann and Romano (2005b), including numerous applications. To provide one example, suppose $\theta = (\theta_1, \ldots, \theta_k)^T$ and the problem is to test $H_0 : \theta_1 \le 0$ versus $H_1 : \theta_1 > 0$. The limit problem corresponds to testing $h_1 \le 0$ versus $h_1 > 0$ 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 $\alpha$, the resulting optimal limiting power against alternatives $\theta_1 = h_1 n^{-1/2}$ with $\theta_2, \ldots, \theta_k$ fixed, is

$$1 - \Phi\Bigl(z_{1-\alpha} - h_1\bigl\{I^{-1}(0, \theta_2, \ldots, \theta_k)_{1,1}\bigr\}^{-1/2}\Bigr) \ . \tag{13}$$

Tests that achieve this limiting power will be 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 Theorem 25.44 of van der Vaart (1998).

On the other hand, there are many important nonstandard problems in econometrics, where the LAN 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

$$\log\bigl[L_n(1 + hn^{-1})/L_n(1)\bigr] = hU_n - \frac{1}{2}h^2 V_n^2 \ . \tag{14}$$

As is well-known, $(U_n, V_n)$ tends to a limit law (under $\theta = \theta_n(h) = 1 + h/n$), which depends on the local parameter $h$. Even though the right sides of both (14) and (12) are quadratic in $h$, note 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 which is nondegenerate, which prevents the existence of a UMP or even an asymptotically UMP one-sided tests; e.g., see Lemma 1 of Crump (2008). 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 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, a broad class of possibly nonparametric models is introduced 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 $\Omega$ is a subset of $\mathbb{R}^k$. For concreteness, we assume throughout this section that $P_\theta = F_\theta^n$, where each $F_\theta$ is absolutely continuous with respect to a common, $\sigma$-finite dominating measure $\mu$. Denote by $f_\theta$ the density of $F_\theta$ with respect to $\mu$. 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 $\Omega_0$ or $\Omega_1$ 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 $F_\theta$ is the multivariate normal distribution with unknown mean vector $\theta \in \mathbb{R}^k$ and known covariance matrix $\Sigma$. 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 $\chi_k^2$ distribution, then the resulting test has exact level $\alpha$. 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 $P_0\big\{\inf_{\theta\in\Omega_0} n(\bar{X}_n - \theta)^T \Sigma^{-1}(\bar{X}_n - \theta) > c\big\} = \alpha$, then the resulting test again has exact level $\alpha$. ■

### 6.1.2 Wald Tests

Wald tests are based on a suitable estimator of $\theta$. In order to describe this approach, we will 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 $\theta_0$ with nonsingular Fisher Information matrix $I(\theta_0)$ and score function $Z_n$ defined in (11). Assume further that $\hat{\theta}_n$ is an estimator of $\theta$ 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 $\theta$, defined as

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

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

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

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 $\alpha$.

### 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 $\theta_0$ with nonsingular Fisher Information $I(\theta_0)$ and score function $Z_n$ defined in (11). A disadvantage of Wald tests is that it involves the computation of a suitable estimator satisfying (16). An alternative due to Rao that avoids this difficulty is based directly on $Z_n$ defined in (11). Under these assumptions,

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

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 $\alpha$.

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 (13) against alternatives $\theta_1 = h/\sqrt{n}$ with $\theta_2, \ldots, \theta_k$ 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 Example 13.3.3 of Lehmann and Romano (2005b).

## 6.2 Testing in the Extremum Estimation Framework

We now introduce a class of models in which $\Omega$ 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 $P_\theta = F_\theta^n$, where each $F_\theta$ is absolutely continuous with respect to a common, $\sigma$-finite dominating measure $\mu$. Denote by $f_\theta$ the density of $F_\theta$ with respect to $\mu$. 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

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

and $Q : \Gamma \times \{F_\theta : \theta \in \Omega\} \to \mathbb{R}$. 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, $Q(\gamma, F_\theta) = E_\theta\big[q(X_i, \gamma)\big]$. Here, it is reasonable to choose

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

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 (GMM))** Hansen (1982) consider the choice

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

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

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

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 GMM estimator in this case. If one wishes to minimize the asymptotic variance of the GMM estimator, then it is optimal to choose

$$W(F_\theta) = E_\theta\Big[h\big(X_i, \gamma(F_\theta)\big) h\big(X_i, \gamma(F_\theta)\big)^T\Big]^{-1} \ . \tag{17}$$

A consistent estimate of (17) can be obtained in two steps, where in the first step $\gamma(F_\theta)$ is consistently estimated. 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 it is important to note that the MLE may not always be characterized in this fashion. To see this, simply consider the example where $F_\theta$ 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. See Chernozhukov et al. (2007) and Romano and Shaikh (2006a, 2008) for some recent contributions. ■

We consider testing the null hypothesis

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

where $\Gamma_0$ is a fixed subset of $\Gamma$, versus the alternative

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

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 and 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

$$n\big(\inf_{\gamma \in \Gamma_0} \hat{Q}_n(\gamma) - \inf_{\gamma \in \Gamma} \hat{Q}_n(\gamma)\big) . \tag{18}$$

**Example 6.4 (GMM, Continued)** Recall the setup of Example 6.3 and suppose further that

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

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

### 6.2.2 Wald Tests

As before, Wald tests are based on a suitable estimator of $\gamma(F_\theta)$. In order 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

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

where $V(F_\theta)$ is nonsingular, under $P_\theta$ 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, see, for example, Newey and McFadden (1994). See also van der Vaart and 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 $P_\theta$ 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 $\alpha$.

### 6.2.3 Lagrange Multiplier Tests

Consider again the special case in which

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

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

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

where $V(F_\theta)$ is nonsingular, under $P_\theta$ 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 $P_\theta$ 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 $\alpha$.

# 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 $P_\theta$ with $\theta \in \Omega_0$ did not depend on $P_\theta$. 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 will discuss randomization methods, bootstrap, and subsampling. Even when the test statistics are pivots or asymptotic pivots, we will 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 $\alpha$ for any sample size $n$ whenever a certain invariance restriction holds. In order to describe this approach in more detail, let $\mathbf{G}$ be a group of transformations of the data $X$. We require that $gX \overset{d}{=} X$ 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 $F_\theta$ 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 $g = (\epsilon_1, \dots, \epsilon_n) \in \mathbf{G}$ on $X$ defined by $gX = (\epsilon_1 X_1, \dots, \epsilon_n X_n)$. ■

**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 $P_\theta$, where $Y_1, \dots, Y_\ell$ are i.i.d. with distribution $F_\theta^Y$ and $Z_1, \dots, Z_m$ are i.i.d. with distribution $F_\theta^Z$. 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

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

denote the ordered values of $\{T(gX) : g \in \mathbf{G}\}$. Define $k = \lceil M(1-\alpha) \rceil$, 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

$$\begin{aligned} M^0(X) &= \left|\{1 \leq j \leq M : T^{(j)}(X) = T^{(k)}(X)\}\right| \\ M^+(X) &= \left|\{1 \leq j \leq M : T^{(j)}(X) > T^{(k)}(X)\}\right| \ . \end{aligned}$$

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 $\alpha$ whenever the randomization hypothesis holds.

**Remark 7.1** Even though it has exact size $\alpha$, the test constructed above may not be very interesting if it has poor power. After all, the test that simply rejects the null hypothesis with probability $\alpha$ also has this feature. It is therefore interesting to examine 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 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,

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

Our goal is to estimate $J_n(x, F_\theta)$ or its appropriate quantiles, which are typically unknown because $F_\theta$ 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 $F_\theta$. Since the data $X^{(n)} = (X_1, \ldots, X_n)$ consists of $n$ i.i.d. observations, one can use Efron's (1979) bootstrap (i.e., non-parametric bootstrap) or a suitable model-based bootstrap (i.e., parametric bootstrap); e.g., see Davison and 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 (i) $J_n(x, F_n)$ converges weakly to a continuous limiting distribution $J(x, F_\theta)$ whenever $d(F_n, F_\theta) \to 0$ and $\theta \in \Omega_0$ and (ii) $d(\hat{F}_n, F_\theta) \overset{F_\theta}{\to} 0$ whenever $\theta \in \Omega_0$, then

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

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

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

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

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

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

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 $X_1, \ldots, X_n$ and

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

Both of these tests are pointwise asymptotically level $\alpha$, but, under further technical conditions ensuring the validity of Edgeworth expansions, it is possible to show that for any $F_\theta$ 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 $O(n^{-1})$ for the second test. Informally, the reason for this phenomenon is that the bootstrap approximation to the distribution of left-hand side of (19), unlike the standard normal approximation, does not assume the skewness of the finite-sample distribution of the $t$-statistic is zero. See Hall and 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 (19) directly. In that case, one should use an estimate of $F_\theta$ that satisfies the constraints of the null hypothesis since critical values should be determined as if the null hypothesis were true. Such an approach is most useful for problems in which the hypotheses can not 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), Chapter 1 of Politis et al. (1999), and Andrews (2000) for further details.

## 7.3 Subsampling

While 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 $\alpha$ under every $P_\theta$ with

$\theta \in \Omega_0$, under very weak assumptions. In order to describe this approach, we also assume that $P_\theta = F_\theta^n$, but the approach can be easily modified for dependent data; see Chapter 3 of Politis et al. (1999). The key insight underlying this approach is that each subset of size $b$ from these $n$ observations constitutes $b$ i.i.d. observations from $F_\theta$. 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 $P_\theta$. Index by $i = 1, \ldots, \binom{n}{b}$ 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 $b \to \infty$ so that $b/n \to 0$ and $\tilde{J}_n(x, F_\theta)$ converges weakly to a continuous limiting distribution $\tilde{J}(x, F_\theta)$ whenever $\theta \in \Omega_0$. Under these assumptions,

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

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

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

Remarkably, it is possible to show that

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

for any $\epsilon > 0$ regardless of $\Omega$. 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 (20). For example, if

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

then one has in fact

$$\limsup_{n \to \infty} \sup_{\theta \in \Omega} P_\theta\big\{T_n > L_{n,b}^{-1}(1-\alpha)\big\} \leq \alpha \ . \tag{21}$$

See Romano and Shaikh (2008) for further details. Related results have also been obtained independently by Andrews and Guggenberger (2009), who go on to establish formulae for the left-hand side of (21). Using these formulae, they establish in a variety of problems that the left-hand side of (21) exceeds $\alpha$, sometimes by a large margin. This problem may occur when the limiting distribution of the test statistic is discontinuous in $F_\theta$. On the other hand, even when this is the case, subsampling may yield tests satisfying (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

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

for some fixed $\gamma_0 \in \Gamma$ versus the alternative $\Omega_1 = \Omega \backslash \Omega_0$. Note here that the dimension of $h(\cdot, \cdot)$ 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 (21) under very weak assumptions on $\Omega$. For details, see Romano and Shaikh (2008) and Andrews and Guggenberger (2009). ■

# 8 MULTIPLE TESTING

## 8.1 Motivation

Much empirical research in economics involves simultaneous testing of several hypotheses. To list just three examples: (i) one fits a multiple regression model and wishes to decide which coefficients are different from zero; (ii) one compares several investment strategies to a benchmark and wishes to decide which strategies are outperforming the benchmark; (iii) 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 $\alpha$. For $\alpha = 0.05$, one expects five true hypotheses to be rejected. Further, 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

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 $\theta$ is true; that is, $s \in \mathcal{I}(\theta)$ if and only if (iff) $\theta \in \Omega_{0,s}$.

Again, we assume that data $X = X^{(n)}$ is generated from some probability distribution $P_\theta$, 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 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 $\hat{p}_{n,s}$.

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:

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

Of course, this criterion is very 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:

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

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 a MTP. Then, $\text{PFER}_\theta = E_\theta(F)$, where the concern now is to ensure $\text{PFER}_\theta \leq \lambda$ for some $\lambda \in [0, \infty)$.

[1] Analogously, 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.

[2] If there is more than one null hypothesis under test, there no longer exists a unique definition of power. Some reasonable definitions include: (i) the probability of rejecting at least one false null hypothesis; (ii) the probability of rejecting all false null hypotheses; (iii) the average probability of rejection over the set of false null hypotheses.

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 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, pre-specified 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 $\gamma \in [0, 1)$.

The $k$-FWE, PFER, FDP, and FDR can all be coined *generalized error rates* in the sense that they relax and generalize the FWE. While 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 $\alpha$,

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

Control of the PFER means that, for a given integer $k$, $\text{PFER}_\theta \leq k$ for any $\theta$.

Control of the FDP means that, for a given significance level $\alpha$ and for a given proportion $\gamma \in [0, 1)$, $P_\theta\{\text{FDP} > \gamma\} \leq \alpha$ for any $\theta$.

Finally, control of the FDR means that, for a given proportion $\gamma \in [0, 1)$, $\text{FDR}_\theta \leq \gamma$ for any $\theta$.

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 $F$ at all (except for some very 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 $\gamma$. Or, in other words, that the realized FDP is at most $\gamma$.

On the other hand, control of the FDR does not really allow one to draw any meaningful conclusion about the realized FDP at all. 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 which 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 $\gamma$ (plus some small $\varepsilon$).

**Remark 8.1 (Finite-sample vs. 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: (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 set-ups. Instead, one settles for (pointwise) *asymptotic* control of the FWE:

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

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

Next, we discuss MTPs that (asymptotically) control these error rates. Such procedures can roughly be classified according to two criteria. The first criterion is whether the method is based on the individual $p$-values $\hat{p}_{n,s}$ only 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 vs. 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 anymore, the method stops. An example is given by the Holm (1979) method discussed below.

Such stepwise methods which improve upon single-step methods by possible 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. And so 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 and 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 only have access to the list of $p$-values from a past study, but not to the underlying complete data set. Second, such methods can be very 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 sub-optimal 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:

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

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: $\hat{p}_{n,(1)} \leq \hat{p}_{n,(2)} \leq \ldots \leq \hat{p}_{n,(S)}$ with their corresponding null hypotheses labeled accordingly: $H_{0,(1)}, H_{0,(2)}, \ldots, H_{0,(S)}$. Then, $H_{0,(s)}$ is rejected iff $\hat{p}_{n,(j)} \leq \alpha/(S - j + 1)$ for $j = 1, \ldots, s$. In other words, the method starts with testing the most significant hypothesis by comparing its $p$-value to $\alpha/S$, 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 Bonferroni will also be rejected by Holm, but potentially a few more. So, trivially, the method is more powerful. But it still controls the FWE under (24).

Both the Bonferroni and Holm methods can be easily generalized to control the $k$-FWE; these generalizations are due to Hommel and Hoffman (1988) and Lehmann and Romano (2005a). For Bonferroni, simply change the cutoff value from $\alpha/S$ to $k \cdot \alpha/S$. For Holm, 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: so 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 (24) in order 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 and 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,

[3] For 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 slighter larger than $\alpha/S$ for 'non-extreme' values of $\alpha$ and $S$; e.g., for $\alpha = 0.05$ and $S = 100$, one obtains 0.000513 instead of 0.0005.

$H_{0,(s)}$ is rejected if $\hat{p}_{n,j} \le \alpha_j$ for $j = 1, \ldots, s$, with:

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

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

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

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

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

Both the Lehmann and Romano (2005a) method to control the FDP and the Benjamini and 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 the details, see Benjamini and Yekutieli (2001) as well as Lehmann and Romano (2005a) and Romano and Shaikh (2006b), respectively.

Stepup methods based on individual $p$-values to control the FWER, $k$-FWER, and FDP are discussed by Romano and Shaikh (2006c).

**Remark 8.2 (Adaptive Benjamini and Hochberg Method)** Under conditions which ensure finite-sample control of the Benjamini and Hochberg (1995) method, it can be shown that $\text{FDR}_\theta = (S_0/S) \cdot \gamma$, where $S_0 = \big|\mathcal{I}(\theta)\big|$. 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 vs. Asymptotic Control)** So far, this subsection has assumed 'finite-sample validity' of the null $p$-values expressed by (24). However, often $p$-values are derived by asymptotic approximations or resampling methods, only guaranteeing 'asymptotic validity' instead: for any (fixed) $\theta$,

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

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

As an example of case (i), consider the Bonferroni method. If there were perfect dependence between the $p$-values, the cut-off value could be changed from $\alpha/S$ to $\alpha$. Perfect dependence rarely happens in practice, of course. But this example is just to make a point. In the realistic set-up of 'strong cross dependence', the cut-off value could be changed to something a lot larger than $\alpha/S$ while still maintaining control of the FWE. As an example of case (ii), 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 anti-conservative, 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 will settle for asymptotic control of the various error rates. In addition, we will specialize to making simultaneous inference on the elements of a parameter vector $\theta = (\theta_1, \ldots, \theta_S)^T$. 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:

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

For the time being, we will treat the one-sided case (26); the necessary modifications for the two-sided case (27) will be given later.

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 $\theta_s$ computed from $X^{(n)}$. Further, $\hat{\sigma}_{n,s}$ 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 only very 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 $P_\theta$ 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 $d_1$ does not only depend 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.

Since $P_\theta$ 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 $k\text{-max}_s(\hat{\theta}^*_{n,s} - \hat{\theta}_{n,s})/\hat{\sigma}^*_{n,s}$. Here, $\hat{P}_n$ is an *unrestricted* estimate of $P_\theta$. Further $\hat{\theta}^*_{n,s}$ and $\hat{\sigma}^*_{n,s}$ are the estimator of $\theta_s$ 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., non-parametric bootstrap) or a suitable model-based bootstrap (i.e., parametric bootstrap); e.g., see Davison and 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 non-rejected 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 (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 $P_\theta$ 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 $\hat{\sigma}_{n,s}$, as opposed to simply using $1/\sqrt{n}$, it is assumed that they converge to the same non-zero 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 and 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 are already proposed in Westfall and Young (1993). An important difference in their approach is that they bootstrap under the joint null, that is, they use a *restricted* estimate of $P_\theta$ where the contraints of all null hypotheses jointly hold. This approach requires the so-called *subset pivotality* condition and is generally less valid than the approaches discussed so far based on an unrestricted estimate of $P_\theta$; e.g., see Example 4.1 of Romano and Wolf (2005).

There exist alternative MTPs to control the $k$-FWE and the FDP. Namely, *augmentation* procedures of van der Laan et al. (2004) and *empirical Bayes* procedures of van der Laan et al. (2005). However, the former are sub-optimal in terms of power while the latter do not always provide asymptotic control and can be quite anti-conservative; see Romano and Wolf (2007). ■

We finally turn to FDR control. Since these methods are very lengthy to describe, we restrict ourselves to a brief listing. Yekutieli and Benjamini (1999) propose a bootstrap method without discussing asymptotic properties. Dudoit et al. (2008) propose an empirical Bayes method which does not always provide asymptotic control and can be quite anti-conservative.

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 and Hochberg (1995) method and its adaptive versions which also are robust to a general dependence structure.

## References

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

Andrews, D. W. K. (2000). Inconsistency of the bootstrap when a parameter is on the boundary of the parameter space. *Econometrica*, 68:399–405.

Andrews, D. W. K. and Guggenberger, P. (2009). Validity of subsampling and 'plug-in asymptotic' inference for parameters defined by moment inequalities. *Econometric Theory*, 25(3):669–709.

Andrews, D. W. K., Moreira, M. J., and Stock, J. H. (2006). Optimal two-sided invariant similar tests for instrumental variables regression. *Econometrica*, 74:715–752.

Bahadur, R. R. and Savage, L. J. (1956). The nonexistence of certain statistical procedures in nonparametric problems. *Annals of Mathematical Statistics*, 27(4):1115–1122.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, 57(1):289–300.

Benjamini, Y., Krieger, A. M., and Yekutieli, D. (2006). Adaptive linear step-up procedures that control the false discovery rate. *Biometrika*, 93(3):491–507.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, 29(4):1165–1188.

Beran, R. (1984). Bootstrap methods in statistics. *Jahresberichte des Deutschen Mathematischen Vereins*, 86:14–30.

Brown, L. D. (1986). *Fundamentals of Statistical Exponential Families (With Application to Statistical Decision Theory)*, volume 9. Institute of Mathematical Statistics Lecture Notes Monograph Series, Hayward, CA.

Chamberlain, G. (1987). Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of Econometrics*, 34(3):305–334.

Chernozhukov, V., Hong, H., and Tamer, E. (2007). Estimation and confidence regions for parameter sets in econometric models. *Econometrica*, 75(5):1243–1284.

Chiburis, R. (2008). Approximately most powerful tests for moment inequalities. Job market paper, Dept. of Economics, Princeton University.

Crump, R. (2008). Optimal conditional inference in nearly-integrated autoregressive processes. Job market paper, Dept. of Economics, University of Berkeley.

Davidson, R. and Duclos, J.-Y. (2006). Testing for restricted stochastic dominance. Discussion paper 2047, IZA. Available at `http://ssrn.com/abstract=894061`.

Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap Methods and their Application.* Cambridge University Press, Cambridge.

Dickey, D. A. and Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, 74:427–431.

Dudoit, S., Gilbert, H., and van der Laan, M. J. (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. *Biometrical Journal*, 50(5):716–744.

Dufour, J.-M. (1997). Some impossibility theorems in econometrics with applications to structural and dynamic models. *Econometrica*, 65(6):1365–1387.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, 7:1–26.

Elliot, G., Rothenberg, T. J., and Stock, J. H. (1996). Efficient tests for an autoregressive unit root. *Econometrica*, 64:813–836.

Gavrilov, Y., Benjamini, Y., and Sarkar, S. K. (2009). An adaptive step-down procedure with proven FDR control. *Annals of Statistics*, 37(2):619–629.

Hall, P. and 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*, 50:1029–1054.

Hansen, P. R. (2005). A test for superior predictive ability. *Journal of Business and Economics Statistics*, 23:365–380.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6:65–70.

Hommel, G. and Hoffman, T. (1988). Controlled uncertainty. In Bauer, P., Hommel, G., and Sonnemann, E., editors, *Multiple Hypotheses Testing*, pages 154–161. Springer, Heidelberg.

Horowitz, J. (2001). The bootstrap. In Griliches, Z., Heckman, J. J., Intriligator, M. D., and Leamer, E. E., editors, *Handbook of Econometrics*, volume 5, chapter 52, pages 3159–3228. North Holland, Amsterdam.

Jansson, M. (2008). Semiparametric power envelopes for tests of the unit root hypothesis. *Econometrica*, 76:1103–1142.

Lahiri, S. N. (2003). *Resampling Methods for Dependent Data.* Springer, New York.

Lehmann, E. L. (1952). Testing multiparameter hypotheses. *Annals of Mathematical Statistics*, 23:541–552.

Lehmann, E. L. and Casella, G. (1998). *Theory of Point Estimation.* Springer, Ney York.

Lehmann, E. L. and Romano, J. P. (2005a). Generalizations of the familywise error rate. *Annals of Statistics*, 33(3):1138–1154.

Lehmann, E. L. and Romano, J. P. (2005b). *Testing Statistical Hypotheses.* Springer, New York, third edition.

MacKinnon, J. (2007). Bootstrap hypothesis testing. Working Paper 1127, Dept. of Economics, Queen's University.

Manski, C. F. (2003). *Partial Identification of Probability Distributions.* Springer, New York.

Mikusheva, A. (2007). Uniform inference in autoregressive models. *Econometrica*, 75(5):1411–1452.

Moreira, M. J. (2003). A conditional likelihood ratio test for structural models. *Econometrica*, 71(4):1027–1048.

Newey, W. K. and McFadden, D. L. (1994). Large-sample estimation and hypothesis testing. In Engle, R. F. and McFadden, D. L., editors, *Handbook of Econometrics*, volume IV. Elsevier, Amsterdam.

Newey, W. K. and West, K. D. (1987). A simple positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55:703–708.

Politis, D. N., Romano, J. P., and Wolf, M. (1999). *Subsampling.* Springer, New York.

Romano, J. P. (2004). On nonparametric testing, the uniform behavior of the $t$-test, and related problems. *Scandinavian Journal of Statistics*, 31(4):567–584.

Romano, J. P. and Shaikh, A. M. (2006a). Inference for the identified set in partially identified econometric models. Technical Report 2006–10, Department of Statistics, Stanford University.

Romano, J. P. and Shaikh, A. M. (2006b). On stepdown control of the false discovery proportion. In Rojo, J., editor, *IMS Lecture Notes—Monograph Series, 2nd Lehmann Symposium—Optimality*, pages 33–50.

Romano, J. P. and Shaikh, A. M. (2006c). Stepup procedures for control of generalizations of the familywise error rate. *Annals of Statistics*, 34(4):1850–1873.

Romano, J. P. and Shaikh, A. M. (2008). Inference for identifiable parameters in partially identified econometric models. *Journal of Statistical Planning and Inference – Special Issue in Honor of Ted Anderson*, 138(9):2786–2807.

Romano, J. P., Shaikh, A. M., and Wolf, M. (2008a). Control of the false discovery rate under dependence using the bootstrap and subsampling (with discussion). *TEST*, 17(3):417–442.

Romano, J. P., Shaikh, A. M., and Wolf, M. (2008b). Formalized data snooping based on generalized error rates. *Econometric Theory*, 24(2):404–447.

Romano, J. P. and Wolf, M. (2000). Finite sample nonparametric inference and large sample efficiency. *Annals of Statistics*, 28(3):756–778.

Romano, J. P. and Wolf, M. (2005). Exact and approximate stepdown methods for multiple hypothesis testing. *Journal of the American Statistical Association*, 100(469):94–108.

Romano, J. P. and Wolf, M. (2007). Control of generalized error rates in multiple testing. *Annals of Statistics*, 35(4):1378–1408.

Storey, J. D., Taylor, J. E., and Siegmund, D. (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: A unified approach. *Journal of the Royal Statistical Society, Series B*, 66(1):187–205.

van der Laan, M. J., Birkner, M. D., and Hubbard, A. E. (2005). Empirical bayes and resampling based multiple testing procedure controlling tail probability of the proportion of false positives. *Statistical Applications in Genetics and Molecular Biology*, 4(1):Article 29. Available at `http://www.bepress.com/sagmb/vol4/iss1/art29/`.

van der Laan, M. J., Dudoit, S., and Pollard, K. S. (2004). Augmentation procedures for control of the generalized family-wise error rate and tail probabilities for the proportion of false positives. *Statistical Applications in Genetics and Molecular Biology*, 3(1):Article 15. Available at `http://www.bepress.com/sagmb/vol3/iss1/art15/`.

van der Vaart, A. W. (1998). *Asymptotic Statistics.* Cambridge University Press, Cambridge.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes: with Applications to Statistics.* Springer, New York.

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

White, H. L. (2000). A reality check for data snooping. *Econometrica*, 68(5):1097–1126.

Yekutieli, D. and Benjamini, Y. (1999). Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. *Journal of Statistical Planning and Inference*, 82:171–196.