Data-Snooping Biases in Tests of Financial Asset Pricing Models

Andrew W. Lo; A. Craig MacKinlay

The Review of Financial Studies, Volume 3, Issue 3 (1990), 431-467.

Stable URL:
http://links.jstor.org/sici?sici=0893-9454%281990%293%3A3%3C431%3ADBITOF%3E2.0.CO%3B2-9

Data-Snooping Biases in Tests of Financial Asset Pricing Models

Andrew W. Lo
Sloan School of Management
Massachusetts Institute of Technology

A. Craig MacKinlay
Wharton School
University of Pennsylvania

Tests of financial asset pricing models may yield misleading inferences when properties of the data are used to construct the test statistics. In particular, such tests are often based on returns to portfolios of common stock, where portfolios are constructed by sorting on some empirically motivated characteristic of the securities such as market value of equity. Analytical calculations, Monte Carlo simulations, and two empirical examples show that the effects of this type of data snooping can be substantial.

The reliance of economic science upon nonexperimental inference is, at once, one of the most challenging and most nettlesome aspects of the discipline. Because of the virtual impossibility of controlled experimentation in economics, the importance of sta-

Research support from the Batterymarch Fellowship (Lo), the Geewax-Terker Research Fund (MacKinlay), the John M. Olin Fellowship at the National Bureau of Economic Research (Lo), and the National Science Foundation (SES-8821583) is gratefully acknowledged. We thank David Aldous, Cliff Ball, Michael Brennan, Herbert David, Mike Gibbons, Jay Shanken, a referee, and seminar participants at the Board of Governors of the Federal Reserve, Boston College, Columbia, Dartmouth, Harvard, M.I.T., Northwestern, Princeton, Stanford, University of Chicago, University of Michigan, University of Wisconsin at Madison, Washington University, and Wharton for useful comments and suggestions. Address reprint requests to Andrew Lo, Sloan School of Management, M.I.T., 50 Memorial Drive, Cambridge, MA 02139.

The Review of Financial Studies 1990 Volume 3, number 3, pp. 431–467

tistical data analysis is now well-established. However, there is a growing concern that the procedures under which formal statistical inference have been developed may not correspond to those followed in practice.[1] For example, the classical statistical approach to selecting a method of estimation generally involves minimizing an expected loss function, irrespective of the actual data. Yet in practice the properties of the realized data almost always influence the choice of estimator.

Of course, ignoring obvious features of the data can lead to nonsensical inferences even when the estimation procedures are optimal in some metric. But the way we incorporate those features into our estimation and testing procedures can affect subsequent inferences considerably. Indeed, by the very nature of empirical innovation in economics, the axioms of classical statistical analysis are violated routinely: future research is often motivated by the successes and failures of past investigations. Consequently, few empirical studies are free of the kind of data-instigated pretest biases discussed in Leamer (1978). Moreover, we can expect the degree of such biases to increase with the number of published studies performed on any single data set—the more scrutiny a collection of data is subjected to, the more likely will interesting (spurious) patterns emerge. Since stock market prices are perhaps the most studied economic quantities to date, tests of financial asset pricing models seem especially susceptible.

In this paper, we attempt to quantify the inferential biases associated with one particular method of testing financial asset pricing models such as the capital asset pricing model (CAPM) and the arbitrage pricing theory (APT). Because there are often many more securities than there are time series observations of stock returns, asset pricing tests are generally performed on the returns of *portfolios* of securities. Besides reducing the cross-sectional dimension of the joint distribution of returns, grouping into portfolios has also been advanced as a method of reducing the impact of measurement error. However, the selection of securities to be included in a given portfolio is almost never at random, but is often based on some of the stocks' empirical characteristics. The formation of size-sorted portfolios, portfolios based on the market value of the companies' equity, is but one example. Conducting classical statistical tests on portfolios formed this way creates potentially significant biases in the test statistics. These are

[1] Perhaps the most complete analysis of such issues in economic applications is by Leamer (1978). Recent papers by Lakonishok and Smidt (1988), Merton (1987), and Ross (1987) address data snooping in financial economics. Of course, data snooping has been a concern among probabilists and statisticians for quite some time, and is at least as old as the controversy between Bayesian and classical statisticians. Interested readers should consult Berger and Wolpert (1984, chapter 4.2) and Leamer (1978, chapter 9) for further discussion.

examples of "data-snooping statistics," a term used by Aldous (1989, p. 252) to describe the situation "where you have a family of test statistics $T(a)$ whose null distribution is known for fixed a, but where you use the test statistic $T = T(a)$ for some a chosen using the data." In our application the quantity a may be viewed as a vector of zeros and ones that indicates which securities are to be included in or omitted from a given portfolio. If the choice of a is based on the data, then the sampling distribution of the resulting test statistic is generally not the same as the null distribution with a fixed a; hence, the actual size of the test may differ substantially from its nominal value under the null. Under plausible assumptions our calculations show that this kind of data snooping can lead to rejections of the null hypothesis with probability 1 even when the null hypothesis is true!

Although the term "data snooping" may have an unsavory connotation, our usage neither implies nor infers any sort of intentional misrepresentation or dishonesty. That prior empirical research may influence the way current investigations are conducted is often unavoidable, and this very fact results in what we have called data snooping. Moreover, it is not at all apparent that this phenomenon necessarily imparts a "bias" in the sense that it affects inferences in an undesirable way. After all, the primary reason for publishing scientific discoveries is to add to a store of common knowledge on which future research may build.

But when scientific discovery is statistical in nature, we must weigh the significance of newly discovered relations in view of past inferences. This is recognized implicitly in many formal statistical circumstances, as in the theory of sequential hypothesis testing. But it is considerably more difficult to correct for the effects of specification searches in practice since such searches often consist of *sequences* of empirical studies undertaken by many individuals over many years.[2] For example, as a consequence of the many investigations relating the behavior of stock returns to size, Chen, Roll, and Ross (1986, p. 394) write: "It has been facetiously noted that size may be the best theory we now have of expected returns. Unfortunately, this is less of a theory than an empirical observation." Then, as Merton (1987, p. 107) asks in a related context: "Is it reasonable to use the standard t-statistic as a valid measure of significance when the test is conducted on the same data used by many earlier studies whose results influenced the choice of theory to be tested?" We rephrase this question

[2] Statisticians have considered a closely related problem, known as the "file drawer problem," in which the overall significance of several published studies must be assessed while accounting for the possibility of unreported insignificant studies languishing in various investigators' file drawers. An excellent review of the file drawer problem and its remedies, which has come to be known as "meta-analysis," is provided by Iyengar and Greenhouse (1988).

in the following way: Are standard tests of significance valid when the construction of the test statistics is influenced by empirical relations derived from the very same data to be used in the test? Our results show that using prior information only marginally correlated with statistics of interest can distort inferences dramatically.

In Section 1, we quantify the data-snooping biases associated with testing financial asset pricing models with portfolios formed by sorting on some empirically motivated characteristic. Using the theory of induced order statistics, we derive in closed form the asymptotic distribution of a commonly used test statistic before and after sorting. This not only yields a measure of the effect of data snooping, but also provides the appropriate sampling theory when snooping is unavoidable. In Section 2, we report the results of Monte Carlo experiments designed to gauge the accuracy of the asymptotic approximations used in Section 1. In Section 3, two empirical examples are provided that illustrate the potential importance of data-snooping biases in existing tests of asset pricing models, and, in Section 4, we show how these biases can arise naturally from our tendency to focus on the unusual. We conclude in Section 5.

1. Quantifying Data-Snooping Biases With Induced Order Statistics

Many tests of the CAPM and APT have been conducted on returns of groups of securities rather than on individual security returns, where the grouping is often according to some empirical characteristic of the securities. Perhaps the most common attribute by which securities are grouped is market value of equity or "size." The prevalence of size-sorted portfolios in recent tests of asset pricing models has not been precipitated by any economic theory linking size to asset prices. It is a consequence of a series of empirical studies demonstrating the statistical relation between size and the stochastic behavior of stock returns.[3] Therefore, we must allow for our foreknowledge of size-related phenomena in evaluating the actual significance of tests performed on size-sorted portfolios. More generally, grouping securities by some characteristic that is empirically motivated may affect the size of the usual significance tests,[4] particularly when the empirical motivation is derived from the very data set on which the test is based.

[3] See Banz (1978, 1981), Brown, Kleidon, and Marsh (1983), and Chan, Chen, and Hsieh (1985), for example. Although Banz's (1978) original investigation may have been motivated by theoretical considerations, virtually all subsequent empirical studies exploiting the size effect do so because of Banz's empirical findings, and not his theory.

[4] Unfortunately the use of "size" to mean both market value of equity and type I error is unavoidable. Readers beware.

We quantify these effects in the following sections by appealing to asymptotic results for induced order statistics, and show that even mild forms of data snooping can change inferences substantially. In Section 1.1, a brief summary of the asymptotic properties of induced order statistics is provided. In Section 1.2, results for tests based on individual securities are presented, and in Section 1.3, corresponding results for portfolios are reported. We provide a more positive interpretation of data-snooping biases as power against deviations from the null hypothesis in Section 1.4.

1.1. Asymptotic properties of induced order statistics

Since the particular form of data snooping we are investigating is most common in empirical tests of financial asset pricing models, our exposition will lie in that context. Suppose for each of N securities we have some consistent estimator $\hat{\alpha}_i$ of a parameter α_i which is to be used in the construction of an aggregate test statistic. For example, in the Sharpe–Lintner CAPM, $\hat{\alpha}_i$ would be the estimated intercept from the following regression:

$$R_{it} - R_{ft} = \hat{\alpha}_i + (R_{mt} - R_{ft})\beta_i + \epsilon_{it} \tag{1}$$

where R_{it}, R_{mt}, and R_{ft} are the period-t returns on security i, the market portfolio, and a risk-free asset, respectively. A test of the null hypothesis that $\alpha_i = 0$ would then be a proper test of the Sharpe–Lintner version of the CAPM; thus, $\hat{\alpha}_i$ may serve as a test statistic itself. However, more powerful tests may be obtained by combining the $\hat{\alpha}_i$'s for many securities. But how should we combine them?

Suppose for each security i we observe some characteristic X_i, such as its out-of-sample market value of equity or average annual earnings, and we learn that X_i is correlated empirically with $\hat{\alpha}_i$. By this we mean that the relation between X_i and $\hat{\alpha}_i$ is an empirical fact uncovered by "searching" through the data, and not motivated by any a priori theoretical considerations. This search need not be a systematic sifting of the data, but may be interpreted as any one of Leamer's (1978) six specification searches, which even the most meticulous of classical statisticians has conducted at some point. The key feature is that our interest in characteristic X_i is derived from a look at the data, the same data to be used in performing our test. Common intuition suggests that using information contained in the X_i's can yield a more powerful test of economic restrictions on the $\hat{\alpha}_i$'s. But if this characteristic is not a part of the original null hypothesis, and only catches our attention after a look at the data (or after a look at another's look at the data), using it to form our test statistics may lead us to reject those economic restrictions even when they obtain. More formally, if we write $\hat{\alpha}_i$ as

$$\hat{\alpha}_i = \alpha_i + \zeta_i, \tag{2}$$

then it is evident that under the null hypothesis where $\alpha_i = 0$, any correlation between X_i and $\hat{\alpha}_i$ must be due to correlation between the characteristic and estimation or measurement error ζ_i. Although measurement error is usually assumed to be independent of all other relevant economic variables, the very process by which the characteristic comes to our attention may induce spurious correlation between X_i and ζ_i. We formalize this intuition in Section 4 and proceed now to show that such spurious correlation has important implications for testing the null hypothesis.

This is most evident in the extreme case where the null hypothesis $\alpha_i = 0$ is tested by performing a standard t-test on the largest of the $\hat{\alpha}_i$'s. Clearly such a test is biased toward rejection unless we account for the fact that the largest $\hat{\alpha}_i$ has been drawn from the set $\{\hat{\alpha}_j\}$. Otherwise, extreme realizations of estimation error will be confused with a violation of the null hypothesis. If, instead of choosing $\hat{\alpha}_i$ by its value relative to other $\hat{\alpha}_j$'s, our choice is based on some characteristic X_i correlated with the estimation errors of $\hat{\alpha}_{i,}$ a similar bias might arise, albeit to a lesser degree.

To formalize the preceding intuition, suppose that only a subset of n securities is used to form the test statistic and these n are chosen by sorting the X_i's. That is, let us reorder the bivariate vectors $[X_i \, \hat{\alpha}_i]'$ according to their first components, yielding the sequence

$$\begin{pmatrix} X_{1:N} \\ \hat{\alpha}_{[1:N]} \end{pmatrix}, \begin{pmatrix} X_{2:N} \\ \hat{\alpha}_{[2:N]} \end{pmatrix}, \ldots, \begin{pmatrix} X_{N:N} \\ \hat{\alpha}_{[N:N]} \end{pmatrix}, \tag{3}$$

where $X_{1:N} < X_{2:N} < \cdots < X_{N:N}$ and the notation $X_{i:N}$ follows that of the statistics literature in denoting the ith order statistic from the sample of N observations $\{X_i\}$.[5] The notation $\hat{\alpha}_{[i:N]}$ denotes the ith ***induced order statistic*** corresponding to $X_{i:N}$, or the ith ***concomitant*** of the order statistic $X_{i:N}$.[6] That is, if the bivariate vectors $[X_i \, \hat{\alpha}_i]'$ are ordered according to the X_i entries, $\hat{\alpha}_{[i:N]}$ is defined to be the second component of the ith ordered vector. The $\hat{\alpha}_{[i:N]}$'s are not themselves

[5] It is implicitly assumed throughout that both $\hat{\alpha}_i$ and X_i have continuous joint and marginal cumulative distribution functions; hence, strict inequalities suffice.

[6] The term ***concomitant*** of an order statistic was introduced by David (1973), who was perhaps the first to systematically investigate its properties and applications. The term ***induced*** order statistic was coined by Bhattacharya (1974) at about the same time. Although the former term seems to be more common usage, we use the latter in the interest of brevity. See Bhattacharya (1984) for an excellent review.

ordered but correspond to the ordering of the $X_{i:N}$'s.[7] For example, if X_i is firm size and $\hat{\alpha}_i$ is the intercept from a market-model regression of firm i's excess return on the excess market return, then $\hat{\alpha}_{[j:N]}$ is the $\hat{\alpha}$ of the jth smallest of the N firms. We call this procedure *induced ordering* of the $\hat{\alpha}_i$'s.

It is apparent that if we construct a test statistic by choosing n securities according to the ordering (3), the sampling theory cannot be the same as that of n securities selected independently of the data. From the following remarkably simple result by Yang (1977), an asymptotic sampling theory for test statistics based on induced order statistics may be derived analytically:[8]

Theorem 1.1. *Let the vectors $[X_i\ \hat{\alpha}_i]'$, $i = 1, \ldots, N$, be independently and identically distributed and let $1 < i_1 < i_2 < \cdots < i_n < N$ be sequences of integers such that, as $N \to \infty$, $i_k/N \to \xi_k \in (0, 1)$ $(k = 1, 2, \ldots, n)$. Then*

$$\lim_{N\to\infty} \Pr(\hat{\alpha}_{[i_1:N]} < a_1, \ldots, \hat{\alpha}_{[i_n:N]} < a_n) = \prod_{k=1}^{n} \Pr(\hat{\alpha}_k < a_k \mid F_x(X_k) = \xi_k), \tag{4}$$

where $F_x(\cdot)$ is the marginal cumulative distribution function of X_i.

Proof. See Yang (1977). ■

This result gives the large-sample joint distribution of a finite subset of induced order statistics whose identities are determined solely by their relative rankings ξ_k (as ranked according to the order statistics $X_{i:N}$). From (4) it is evident that the $\hat{\alpha}_{[i_k:N]}$'s are mutually independent in large samples. If X_i were the market value of equity of the ith company, Theorem 1.1 shows that the $\hat{\alpha}_i$ of the security with size at, for example, the 27th percentile is asymptotically independent of the $\hat{\alpha}_j$ of the security with size at the 45th percentile.[9] If the characteristics $\{X_i\}$ and $\{\hat{\alpha}_i\}$ are statistically independent, the joint distribution of

[7] If the vectors are independently and identically distributed and X_i is perfectly correlated with $\hat{\alpha}_i$, then $\hat{\alpha}_{[i:N]}$ are also order statistics. But as long as the correlation coefficient ρ is strictly between -1 and 1, then, for example, $\hat{\alpha}_{[N:N]}$ will generally not be the largest $\hat{\alpha}_i$.

[8] See also David and Galambos (1974) and Watterson (1959). In fact, Yang (1977) provides the exact finite-sample distribution of any finite collection of induced order statistics, but even assuming bivariate normality does not yield a tractable form of this distribution.

[9] This is a limiting result and implies that the identities of the stocks with 27th and 45th percentile sizes will generally change as N increases.

the latter clearly cannot be influenced by ordering according to the former. It is tempting to conclude that as long as the correlation between X_i and $\hat{\alpha}_i$ is economically small, induced ordering cannot greatly affect inferences. Using Yang's result we show the fallacy of this argument in Sections 1.2 and 1.3.

1.2 Biases of tests based on individual securities

We evaluate the bias of induced ordering under the following assumption:

(A) The vectors $[X_i \ \hat{\alpha}_i]'$ $(i = 1, 2, \ldots, N)$ are independently and identically distributed bivariate normal random vectors with mean $[\mu_x \ \alpha]'$, variance $[\sigma_x^2 \ \ \sigma_{\hat{\alpha}}^2]'$, and correlation $\rho \in (-1, 1)$.

The null hypothesis H is then

$$H: \alpha = 0.$$

Examples of asset pricing models that yield restrictions of this form are the Sharpe–Lintner CAPM and the exact factor pricing version of Ross's APT.[10] Under this null hypothesis, the $\hat{\alpha}_i$'s deviate from zero solely through estimation error.

Since the sampling theory provided by Theorem 1.1 is asymptotic, we construct our test statistics using a finite subset of n securities where it is assumed that $n \ll N$. If these securities are selected without the prior use of data, then we have the following well-known result:

$$\theta \equiv \frac{1}{\hat{\sigma}_{\alpha}^2} \sum_{i=1}^{n} \hat{\alpha}_i^2 \overset{a}{\sim} \chi_n^2, \tag{5}$$

where $\hat{\sigma}_{\alpha}^2$ is any consistent estimator of $\sigma_{\hat{\alpha}}^2$.[11] Therefore, a 5 percent test of H may be performed by checking whether θ is greater or less than $C_{.05}^n$, where $C_{.05}^n$ is defined by

$$F_{\chi_n^2}(C_{.05}^n) = .95 \tag{6}$$

and $F_{\chi_n^2}(\cdot)$ is the cumulative distribution function of a χ_n^2 variate.

Now suppose we construct θ from the induced order statistics

[10] See Chamberlain (1983), Huberman and Kandel (1987), Lehmann and Modest (1988), and Wang (1988) for further discussion of exact factor pricing models. Examples of tests that fit into the framework of H are those in Campbell (1987), Connor and Korajczyk (1988), Gibbons, Ross, and Shanken (1989), Huberman and Kandel (1987), Lehmann and Modest (1988), and MacKinlay (1987).

[11] In most contexts the consistency of $\hat{\sigma}_{\alpha}^2$ is with respect to the number of time series observations T. In that case something must be said of the relative rates at which T and N increase without bound so as to guarantee convergence of θ. However, under H the parameter $\sigma_{\hat{\alpha}}^2$ may be estimated cross-sectionally; hence, the relation $\overset{a}{\sim}$ in (5) need only represent N-asymptotics.

$\hat{\alpha}_{[i_k:N]}$, $k = 1, \ldots, n$, instead of the $\hat{\alpha}_i$'s. Specifically, define the following test statistic:

$$\tilde{\theta} \equiv \frac{1}{\hat{\sigma}_\alpha^2} \sum_{k=1}^{n} \hat{\alpha}_{[i_k:N]}^2. \tag{7}$$

Using Theorem 1.1, the following proposition is easily established:

Proposition 1.1. *Under the null hypothesis H and assumption (A), as N increases without bound the induced order statistics $\hat{\alpha}_{[i_k:N]}$ ($k = 1, \ldots, n$) converge in distribution to independent gaussian random variables with mean μ_k and variance σ_k^2, where*

$$\mu_k \equiv \rho(\sigma_\alpha/\sigma_x)[F_x^{-1}(\xi_k) - \mu_x] = \rho\sigma_\alpha\Phi^{-1}(\xi_k), \tag{8}$$

$$\sigma_k^2 \equiv \sigma_\alpha^2(1 - \rho^2), \tag{9}$$

which implies

$$\tilde{\theta} \overset{a}{\sim} (1 - \rho^2)\cdot\chi_n^2(\lambda), \tag{10}$$

with noncentrality parameter

$$\lambda = \sum_{k=1}^{n} \left(\frac{\mu_k}{\sigma_k}\right)^2 = \frac{\rho^2}{1 - \rho^2} \sum_{k=1}^{n} [\Phi^{-1}(\xi_k)]^2, \tag{11}$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function.

Proof. This follows directly from the definition of a noncentral chi-squared variate. The second equality in (8) follows from the fact that $\Phi(\xi_k) = F_x(\xi_k\sigma_x + \mu_x)$. ∎

Proposition 1.1 shows that the null hypothesis H is violated by induced ordering since the means of the ordered $\hat{\alpha}_i$'s are no longer zero. Indeed, the mean of $\hat{\alpha}_{[i_k:N]}$ may be positive or negative depending on ρ and the (limiting) relative rank ξ_k. For example, if $\rho = .10$ and $\sigma_\alpha = 1$, the mean of the induced order statistic in the 95th percentile is 0.164.

The simplicity of $\tilde{\theta}$'s asymptotic distribution follows from the fact that the $\hat{\alpha}_{[i_k:N]}$'s become independent as N increases without bound. It follows from the fact that induced order statistics are conditionally independent when conditioned on the order statistics that determine the induced ordering. This seemingly counterintuitive result is easy to see when $[X_i\ \hat{\alpha}_i]$ is bivariate normal, since, in this case

$$\hat{\alpha}_i = \alpha + \rho(\sigma_\alpha/\sigma_x)[X_i - \mu_x] + Z_i,$$

$$Z_i \quad \text{i.i.d.} \quad N(0, \sigma_\alpha^2(1 - \rho^2)), \tag{12}$$

where X_i and Z_i are independent. Therefore, the induced order statistics may be represented as

$$\hat{\alpha}_{[i_k:N]} = \alpha + \rho(\sigma_\alpha/\sigma_x)[X_{i_k:N} - \mu_x] + Z_{[i_k]},$$
$$Z_{[i_k]} \quad \text{i.i.d.} \quad \mathrm{N}(0, \sigma_\alpha^2(1 - \rho^2)), \tag{13}$$

where the $Z_{[i_k]}$ are independent of the (order) statistics $X_{i_k:N}$. But since $X_{i_k:N}$ is an order statistic, and since the sequence i_k/N converges to ξ_k, $X_{i_k:N}$ converges to the ξ_kth quantile, $F^{-1}(\xi_k)$. Using (13) then shows that $\hat{\alpha}_{[i_k:N]}$ is gaussian, with mean and variance given by (8) and (9), and independent of the other induced order statistics.[12]

To evaluate the size of a 5 percent test based on the statistic $\tilde{\theta}$, we need only evaluate the cumulative distribution function of the noncentral $\chi_n^2(\lambda)$ at the point $C_{.05}^n/(1 - \rho^2)$, where $C_{.05}^n$ is given in (6). Observe that the noncentrality parameter λ is an increasing function of ρ^2. If $\rho^2 = 0$ then the distribution of $\tilde{\theta}$ reduces to a central χ_n^2 which is identical to the distribution of θ in (5)—sorting on a characteristic that is statistically independent of the $\hat{\alpha}_i$'s cannot affect the null distribution of θ. As $\hat{\alpha}_i$ and X_i become more highly correlated, the noncentral χ^2 distribution shifts to the right. However, this does not imply that the actual size of a 5 percent test necessarily increases since the relevant critical value for $\tilde{\theta}$, $C_{.05}^n/(1 - \rho^2)$, also grows with ρ^2.[13]

Numerical values for the size of a 5 percent test based on $\tilde{\theta}$ may be obtained by first specifying choices for the relative ranks $\{\xi_k\}$ of the n securities. We choose three sets of $\{\xi_k\}$, yielding three distinct test statistics $\tilde{\theta}_1$, $\tilde{\theta}_2$, and $\tilde{\theta}_3$:

$$\tilde{\theta}_1 \Leftrightarrow \xi_k = \frac{k}{n+1}, \qquad k = 1, 2, \ldots, n; \tag{14}$$

[12] In fact, this shows how our parametric specification may be relaxed. If we replace normality by the assumption that $\hat{\alpha}_i$ and X_i satisfy the linear regression equation,

$$\hat{\alpha}_i = \mu_\alpha + \beta_i(X_i - \mu_x) + Z_i,$$

where Z_i is independent of X_i, then our results remain unchanged. Moreover, this specification may allow us to relax the rather strong i.i.d. assumption since David (1981, chapters 2.8 and 5.6) does present some results for order statistics in the nonidentically distributed and the dependent cases separately. However, combining and applying them to the above linear regression relation is a formidable task which we leave to the more industrious.

[13] In fact, if $\rho^2 = 1$, the limiting distribution of $\tilde{\theta}$ is degenerate since the test statistic converges in probability to the following limit:

$$\sum_{k=1}^{n} [\Phi^{-1}(\xi_k)]^2.$$

This limit may be greater or less than C_{05}^n depending on the values of ξ_k; hence, the size of the test in this case may be either zero or unity.

$$\tilde{\theta}_2 \Leftrightarrow \xi_k = \begin{cases} \dfrac{k}{(m+1)(n_o+1)}, & \text{for } k=1,2,\ldots,n_o, \\[2ex] \dfrac{k+m(n_o+1)-n_o}{(m+1)(n_o+1)}, & \text{for } k=n_o+1,\ldots,2n_o; \end{cases} \tag{15}$$

$$\tilde{\theta}_3 \Leftrightarrow \xi_k = \begin{cases} \dfrac{k+n_o+1}{(m+1)(n_o+1)}, & \text{for } k=1,2,\ldots,n_o, \\[2ex] \dfrac{k+(m-1)(n_o+1)-n_o}{(m+1)(n_o+1)}, & \text{for } k=n_o+1,\ldots,2n_o; \end{cases} \tag{16}$$

where $n \equiv 2n_o$ and n_o is an arbitrary positive integer. The first method (14) simply sets the ξ_k's so that they divide the unit interval into n equally spaced increments. The second procedure (15) first divides the unit interval into $m + 1$ equally spaced increments, sets the first half of the ξ_k's to divide the *first* such increment into equally spaced intervals each of width $1/(m + 1)(n_o + 1)$, and then sets the remaining half so as to divide the *last* increment into equally spaced intervals also of width $1/(m + 1)(n_o + 1)$ each. The third procedure is similar to the second, except that the ξ_k's are chosen to divide the second smallest and second largest $m + 1$ increments into equally spaced intervals of width $1/(m + 1)(n_o + 1)$.

These three ways of choosing n securities allow us to see how an attempt to create (or remove) dispersion—as measured by the characteristic X_i—affects the null distribution of the statistics. The first choice for the relative ranks is the most disperse, being evenly distributed on (0, 1). The second yields the opposite extreme: the $\hat{\alpha}_{[i_k:N]}$'s selected are those with characteristics in the lowest and highest $100/(m + 1)$-percentiles. As the parameter m is increased, more extreme outliers are used to compute $\tilde{\theta}_2$. This is also true for $\tilde{\theta}_3$, but to a lesser extent since the statistic is based on $\hat{\alpha}_{[i_k:N]}$'s in the second lowest and second highest $100/(m + 1)$-percentiles.

Table 1 shows the size of the 5 percent test using $\tilde{\theta}_1$, $\tilde{\theta}_2$, and $\tilde{\theta}_3$ for various values of n, ρ^2, and m. For concreteness, observe that ρ^2 is simply the R^2 of the cross-sectional regression of $\hat{\alpha}_i$ on X_i, so that $\rho = \pm.10$ implies that only 1 percent of the variation in $\hat{\alpha}_i$ is explained by X_i. For this value of R^2, the entries in the second panel of Table 1 show that the size of a 5 percent test using $\tilde{\theta}_1$ is 4.9 percent for samples of 10 to 100 securities. However, using securities with extreme characteristics does affect the size, as the entries in the "$\tilde{\theta}_2$-test" and "$\tilde{\theta}_3$-test" columns indicate. Nevertheless the largest deviation is only 8.1 percent. As expected, the size is larger for the test based on $\tilde{\theta}_2$ than for that of $\tilde{\theta}_3$ since the former statistic is based on more extreme induced order statistics than the latter.

Table 1
Theoretical sizes of nominal 5 percent χ^2_n-tests of H: $\alpha_i = 0$ ($i = 1, \ldots, n$) using the test statistics $\tilde{\theta}_j$

n	$\tilde{\theta}_1$-test	$\tilde{\theta}_2$-test ($m = 4$)	$\tilde{\theta}_3$-test ($m = 4$)	$\tilde{\theta}_2$-test ($m = 9$)	$\tilde{\theta}_3$-test ($m = 9$)	$\tilde{\theta}_2$-test ($m = 19$)	$\tilde{\theta}_3$-test ($m = 19$)
$R^2 = .005$							
10	0.049	0.051	0.049	0.053	0.050	0.054	0.052
20	0.050	0.052	0.049	0.054	0.050	0.056	0.052
50	0.050	0.053	0.048	0.056	0.050	0.060	0.053
100	0.050	0.054	0.047	0.059	0.050	0.064	0.054
$R^2 = .01$							
10	0.049	0.053	0.048	0.056	0.050	0.059	0.053
20	0.049	0.054	0.047	0.058	0.050	0.063	0.054
50	0.049	0.056	0.046	0.063	0.051	0.071	0.057
100	0.049	0.059	0.045	0.069	0.051	0.081	0.059
$R^2 = .05$							
10	0.045	0.063	0.041	0.080	0.051	0.101	0.066
20	0.045	0.070	0.038	0.096	0.052	0.130	0.073
50	0.046	0.086	0.033	0.135	0.053	0.201	0.087
100	0.047	0.107	0.028	0.190	0.054	0.304	0.106
$R^2 = .10$							
10	0.040	0.076	0.032	0.116	0.052	0.166	0.083
20	0.041	0.093	0.028	0.158	0.053	0.244	0.099
50	0.042	0.133	0.020	0.267	0.055	0.442	0.137
100	0.043	0.192	0.014	0.423	0.058	0.680	0.191
$R^2 = .20$							
10	0.030	0.104	0.019	0.202	0.052	0.330	0.121
20	0.032	0.146	0.013	0.318	0.054	0.528	0.163
50	0.034	0.262	0.006	0.599	0.059	0.862	0.272
100	0.036	0.432	0.002	0.857	0.064	0.987	0.429

$\tilde{\theta}_j \equiv \Sigma_{k=1}^{2n} \hat{\alpha}^2_{[i_k(j)\ N]}/\hat{\sigma}^2_\alpha$, $j = 1, 2, 3$, for various sample sizes n. The statistic $\tilde{\theta}_1$ is based on induced order statistics with relative ranks evenly spaced in (0,1); $\tilde{\theta}_2$ is constructed from induced order statistics ranked in the lowest and highest $100/(m + 1)$-percent fractiles; and $\tilde{\theta}_3$ is constructed from those ranked in the second lowest and second highest $100/(m + 1)$-percent fractiles. The R^2 is the square of the correlation between $\hat{\alpha}_i$ and the sorting characteristics.

When the R^2 increases to 10 percent the bias becomes more important. Although tests based on a set of securities with evenly spaced characteristics still have sizes approximately equal to their nominal 5 percent value, the size deviates more substantially when securities with extreme characteristics are used. For example, the size of the $\tilde{\theta}_2$ test that uses the 100 securities in the lowest and highest characteristic decile is 42.3 percent! In comparison, the 5 percent test based on the second lowest and second highest deciles exhibits only a 5.8 percent rejection rate. These patterns become even more pronounced for R^2's higher than 10 percent.

The intuition for these results may be found in (8)—the more extreme induced order statistics have means farther away from zero; hence, a statistic based on evenly distributed $\hat{\alpha}_{[i_k:N]}$'s will not provide evidence against the null hypothesis $\alpha = 0$. If the relative ranks are extreme, as is the case for $\tilde{\theta}_2$ and $\tilde{\theta}_3$, the resulting $\hat{\alpha}_{[i_k:N]}$'s may appear to be statistically incompatible with the null.

1.3 Biases of tests based on portfolios of securities

The entries in Table 1 show that as long as the n securities chosen have characteristics evenly distributed in relative rankings, test statistics based on individual securities yield little inferential bias. However, in practice the ordering by characteristics such as market value of equity is used to group securities into *portfolios*, and the portfolio returns are used to construct test statistics. For example, let $n \equiv n_o q$, where n_o and q are arbitrary positive integers, and consider forming q portfolios with n_o securities in each portfolio, where the portfolios are formed randomly. Under the null hypothesis H we have the following:

$$\phi_k \equiv \frac{1}{n_o} \sum_{j=(k-1)n_o+1}^{kn_o} \hat{\alpha}_j \sim N\left(0, \frac{\sigma_\alpha^2}{n_o}\right), \quad k = 1, 2, \ldots, q, \tag{17}$$

$$\theta_p \equiv \frac{n_o}{\hat{\sigma}_\alpha^2} \sum_{k=1}^{q} \phi_k^2 \overset{a}{\sim} \chi_q^2, \tag{18}$$

where ϕ_k is the estimated alpha of portfolio k and θ_p is the aggregate test statistic for the q portfolios. To perform a 5 percent test of H using θ_p, we simply compare it with the critical value $C^q_{.05}$ defined by

$$F_{\chi_q^2}(C^q_{.05}) = .95. \tag{19}$$

Suppose, however, we compute this test statistic using the induced order statistics $\{\hat{\alpha}_{[i_k:N]}\}$ instead of randomly chosen $\{\hat{\alpha}_i\}$. From Theorem 1.1 we have:

Proposition 1.2. *Under the null hypothesis H and assumption (A), as N increases without bound, the statistics $\tilde{\phi}_k$ ($k = 1, 2, \ldots, q$) and $\tilde{\theta}_p$ converge in distribution to the following:*

$$\tilde{\phi}_k \equiv \frac{1}{n_o} \sum_{j=(k-1)n_o+1}^{kn_o} \hat{\alpha}_{[i_j:N]} \overset{a}{\sim} \mathrm{N}\left(\sum_{j=(k-1)n_o+1}^{kn_o} \frac{\mu_j}{n_o}, \frac{\sigma_\alpha^2(1-\rho^2)}{n_o} \right), \tag{20}$$

$$\tilde{\theta}_p \equiv \frac{n_o}{\hat{\sigma}_\alpha^2} \sum_{k=1}^{q} \tilde{\phi}_k^2 \overset{a}{\sim} (1-\rho^2)\cdot\chi_q^2(\lambda), \tag{21}$$

with noncentrality parameter

$$\lambda = \frac{n_o \rho^2}{1-\rho^2} \sum_{k=1}^{q} \left(\frac{1}{n_o} \sum_{j=(k-1)n_o+1}^{kn_o} [\Phi^{-1}(\xi_j)] \right)^2. \tag{22}$$

Proof. Again, this follows directly from the definition of a noncentral chi-squared variate and the asymptotic independence of the induced order statistics. ■

The noncentrality parameter (22) is similar to that of the statistic based on individual securities—it is increasing in ρ^2 and equals zero when $\rho = 0$. However, it differs in one respect: because of portfolio aggregation, each term of the outer sum (the sum with respect to k) is the average of $\Phi^{-1}(\xi_j)$ over all securities in the kth portfolio. To see the importance of this, consider the case where the relative ranks ξ_j are chosen to be evenly spaced in (0, 1), that is,

$$\xi_j = j/(n_o q + 1). \tag{23}$$

Recall from Table 1 that for individual securities the size of 5 percent tests based on *evenly spaced* ξ_j's was not significantly biased. Table 2 reports the size of 5 percent tests based on the portfolio statistic $\tilde{\theta}_p$, also using evenly spaced relative rankings. The contrast is striking—even for as low an R^2 as 1 percent, which implies a correlation of only ±10 percent between $\hat{\alpha}_i$ and X_i, a 5 percent test based on 50 portfolios with 50 securities in each rejects 67 percent of the time! We can also see how portfolio grouping affects the size of the test for a fixed number of securities by comparing the ($q = i$, $n_o = j$) entry with the ($q = j$, $n_o = i$) entry. For example, in a sample of 250 securities a test based on 5 portfolios of 50 securities has size 16.5 percent, whereas a test based on 50 portfolios of 5 securities has only a 7.5 percent rejection rate. Grouping securities into portfolios increases the size considerably. The entries in Table 2 are also monotonically increasing across rows and across columns, implying that the test size increases with the number of securities, regardless of whether the number of portfolios or the number of securities per portfolio is held fixed.

To understand why forming portfolios yields much higher rejection rates than using individual securities, recall from (8) and (9) that the mean of $\hat{\alpha}_{[i_k:N]}$ is a function of its relative rank i_k/N (in the limit), whereas its variance $\sigma_\alpha^2(1 - \rho^2)$ is fixed. Forming a portfolio of the induced order statistics within a characteristic-fractile amounts to averaging a collection of n_o approximately independent random variables with similar means and identical variances. The result is a statistic $\tilde{\phi}_k$ with a comparable mean but with a variance n_o times smaller than each of the $\hat{\alpha}_{[i_k:N]}$'s. This variance reduction amplifies the importance of the deviation of the $\tilde{\phi}_k$ mean from zero, and is ultimately reflected in the entries of Table 2. A more dramatic illustration is provided in Table 3, which reports the appropriate 5 percent critical values for the tests in Table 2—when $R^2 = .05$, the 5 percent critical

Table 2
Theoretical sizes of nominal 5 percent χ^2_q-tests of H: $\alpha_i = 0$ ($i = 1, \ldots, n$) using the test statistic $\tilde{\theta}_p$

q	$n_o = 5$	$n_o = 10$	$n_o = 20$	$n_o = 25$	$n_o = 50$
$R^2 = .005$					
5	0.053	0.058	0.068	0.073	0.102
10	0.055	0.062	0.077	0.086	0.134
20	0.057	0.067	0.091	0.105	0.185
25	0.058	0.070	0.097	0.113	0.208
50	0.062	0.079	0.123	0.148	0.311
$R^2 = .01$					
5	0.056	0.066	0.087	0.099	0.165
10	0.060	0.075	0.110	0.130	0.247
20	0.065	0.088	0.146	0.179	0.382
25	0.067	0.093	0.161	0.202	0.440
50	0.075	0.117	0.232	0.302	0.669
$R^2 = .05$					
5	0.080	0.140	0.288	0.368	0.716
10	0.104	0.212	0.477	0.602	0.941
20	0.142	0.333	0.728	0.854	0.998
25	0.159	0.387	0.808	0.914	1.000
50	0.235	0.607	0.971	0.995	1.000
$R^2 = .10$					
5	0.114	0.255	0.568	0.697	0.971
10	0.174	0.434	0.847	0.935	1.000
20	0.276	0.688	0.985	0.998	1.000
25	0.323	0.773	0.996	1.000	1.000
50	0.523	0.960	1.000	1.000	1.000
$R^2 = .20$					
5	0.193	0.514	0.913	0.971	1.000
10	0.348	0.816	0.997	1.000	1.000
20	0.596	0.980	1.000	1.000	1.000
25	0.688	0.994	1.000	1.000	1.000
50	0.926	1.000	1.000	1.000	1.000

$\tilde{\theta}_p \equiv n_o \Sigma_{k=1}^{q} \tilde{\phi}_k^2/\sigma_\alpha^2$, and $\tilde{\phi}_k \equiv (1/n_o)\Sigma_{j=(k-1)q+1}^{kq} \hat{\alpha}_{[t_j:N]}$ is constructed from portfolio k, with portfolios formed by sorting on some characteristic correlated with estimates $\hat{\alpha}_i$. This induced ordering alters the null distribution of $\hat{\theta}_p$ from χ_q^2 to $(1 - R^2)/\chi_q^2(\lambda)$, where the noncentrality parameter λ is a function of the number q of portfolios, the number n_o of securities in each portfolio, and the squared correlation coefficient R^2 between $\hat{\alpha}_i$ and the sorting characteristic.

value for the χ^2 test with 50 securities in each of 50 portfolios is 211.67. If induced ordering is unavoidable, these critical values may serve as a method for bounding the effects of data snooping on inferences.

When the R^2 increases to 10 percent, implying a cross-sectional correlation of about ±32 percent between $\hat{\alpha}_i$ and X_i, the size approaches unity for tests based on 20 or more portfolios with 20 or more securities in each portfolio. These results are especially surprising in view of the sizes reported in Table 1, since the portfolio test statistic is based on evenly spaced induced order statistics

Table 3

Critical values C_{05} for 5 percent χ^2-tests of H: $\alpha_i = 0$ ($i = 1, \ldots, n$) using the test statistic $\hat{\theta}_p$

q	C_{05}-χ^2_q	C_{05}-$\chi^2_q(\lambda)$ ($n_o = 5$)	C_{05}-$\chi^2_q(\lambda)$ ($n_o = 10$)	C_{05}-$\chi^2_q(\lambda)$ ($n_o = 20$)	C_{05}-$\chi^2_q(\lambda)$ ($n_o = 25$)	C_{05}-$\chi^2_q(\lambda)$ ($n_o = 50$)
$R^2 = .005$						
5	11.07	11.22	11.45	11.93	12.16	13.29
10	18.31	18.60	19.03	19.87	20.28	22.31
20	31.41	31.97	32.72	34.22	34.96	38.58
25	37.65	38.33	39.24	41.05	41.94	46.33
50	67.50	68.78	70.44	73.72	75.35	83.39
$R^2 = .01$						
5	11.07	11.36	11.83	12.74	13.19	15.31
10	18.31	18.89	19.73	21.36	22.16	26.00
20	31.41	32.52	34.01	36.93	38.36	45.31
25	37.65	39.01	40.81	44.34	46.08	54.52
50	67.50	70.05	73.33	79.79	82.98	98.60
$R^2 = .05$						
5	11.07	12.45	14.53	18.39	20.21	28.68
10	18.31	21.09	24.88	32.00	35.41	51.54
20	31.41	36.72	43.62	56.75	63.09	93.59
25	37.65	44.18	52.56	68.59	76.35	113.82
50	67.50	79.85	95.41	125.47	140.16	211.67
$R^2 = .10$						
5	11.07	13.65	17.45	24.37	27.63	42.96
10	18.31	23.58	30.62	43.74	50.02	79.98
20	31.41	41.60	54.63	79.32	91.27	148.98
25	37.65	50.21	66.13	96.44	111.15	182.43
50	67.50	91.49	121.42	179.11	207.33	345.24
$R^2 = .20$						
5	11.07	15.70	22.44	34.82	40.71	68.73
10	18.31	27.98	40.86	65.01	76.65	132.76
20	31.41	50.51	74.89	121.32	143.91	253.93
25	37.65	61.32	91.29	148.61	176.58	313.10
50	67.50	113.43	170.67	281.43	335.83	603.10

$\tilde{\theta}_p \equiv n_o \Sigma^q_{k=1} \tilde{\phi}^2_k/\sigma^2_\alpha$, and $\tilde{\phi}_k \equiv (1/n_o)\Sigma^{kq}_{j=(k-1)q+1}\, \hat{\alpha}_{[i_j\ N]}$ is constructed from portfolio k, with portfolios formed by sorting on some characteristic correlated with estimates $\hat{\alpha}_i$. This induced ordering alters the null distribution of $\hat{\theta}_p$ from χ^2_q to $(1 - R^2)/\chi^2_q(\lambda)$, where the noncentrality parameter λ is a function of the number q of portfolios, the number n_o of securities in each portfolio, and the squared correlation coefficient R^2 between $\hat{\alpha}_i$ and the sorting characteristic. $C_{.05}$ is defined implicitly by the relation $\Pr(\tilde{\theta}_p > C_{05}) = 1 - F_{\chi^2_q(\lambda)}\,(C_{.05}/(1 - R^2)) = .05$. For comparison, we also report the 5 percent critical value of the central χ^2_q distribution in the second column.

$\hat{\alpha}_{[i_k:N]}$. Using 100 securities, Table 1 shows a size of 4.3 percent with evenly spaced $\hat{\alpha}_{[i_k:N]}$'s; Table 2 shows that placing those 100 securities into 5 portfolios with 20 securities in each increases the size to 56.8 percent. Computing $\tilde{\theta}_p$ with extreme $\hat{\alpha}_{[i_k:N]}$ would presumably yield even higher rejection rates. The biases reported in Tables 2 and 3 are even more surprising in view of the limited use we have made of the data. The only data-related information impounded in the induced order statistics is the rankings of the characteristics $\{X_i\}$. Nowhere have we exploited the actual values of the X_i's, which contain considerably more precise information about the $\hat{\alpha}_i$'s.

1.4 Interpreting data-snooping bias as power

We have so far examined the effects of data snooping under the null hypothesis that $\alpha_i = 0$, for all i. Therefore, the degree to which induced ordering increases the probability of rejecting this null is implicitly assumed to be a bias, an increase in type I error. However, the results of the previous sections may be reinterpreted as describing the power of tests based on induced ordering against certain alternative hypotheses.

Recall from (2) that $\hat{\alpha}_i$ is the sum of α_i and estimation error ζ_i. Since all α_i's are zero under H, the induced ordering of the estimates $\hat{\alpha}_i$ creates a spurious incompatibility with the null arising solely from the sorting of the estimation errors ζ_i. But if the α_i's are nonzero and vary across i, then sorting by some characteristic X_i related to α_i and forming portfolios does yield a more powerful test. Forming portfolios reduces the estimation error through diversification (or the law of large numbers), and grouping by X_i maintains the dispersion of the α_i's across portfolios. Therefore what were called biases in Sections 1.1–1.3 may also be viewed as measures of the power of induced ordering against alternatives in which the α_i's differ from zero and vary cross-sectionally with X_i. The values in Table 2 show that grouping on a marginally correlated characteristic can increase the power substantially.[14]

To formalize the above intuition within our framework, suppose that the α_i's were i.i.d. random variables independent of ζ_i and have mean μ_α and variance σ_α^2. Then the $\hat{\alpha}_i$'s are still independently and identically distributed, but the null hypothesis that $\alpha_i = 0$ is now violated. Suppose the estimation error ζ_i were identically zero, so that all variation in $\hat{\alpha}_i$ was due to variations in α_i. Then the values in Table 2 would represent the *power* of our test against this alternative, where the squared correlation is now given by

$$\rho_p^2 = \frac{\text{Cov}^2[X_i, \alpha_i]}{\text{Var}[X_i]\cdot\text{Var}[\alpha_i]} . \tag{24}$$

If, as under our null hypothesis, all α_i's were identically zero, then

[14] However, implicit in Table 2 is the assumption that the $\hat{\alpha}_i$'s are cross-sectionally independent, which may be too restrictive a requirement for interesting alternative hypotheses. For example, if the null hypothesis $\alpha_i = 0$ corresponds to the Sharpe–Lintner CAPM, then one natural alternative might be a two-factor APT. In that case, the $\hat{\alpha}_i$'s of assets with similar factor loadings would tend to be positively cross-sectionally correlated as a result of the omitted factor. This positive correlation reduces the benefits of grouping. Grouping by induced ordering does tend to cluster $\hat{\alpha}_i$'s with similar (nonzero) means together, but correlation works against the variance reduction that gives portfolio-based tests their power. The importance of cross-sectional dependence is evident in MacKinlay's (1987) power calculations. We provide further discussion in Section 2.3.

the values in Table 2 must be interpreted as the *size* of our test, where the squared correlation reduces to

$$\rho_s^2 = \frac{\mathrm{Cov}^2[X_i, \zeta_i]}{\mathrm{Var}[X_i]\cdot\mathrm{Var}[\zeta_i]} . \tag{25}$$

More generally, the squared correlation ρ^2 is related to ρ_s^2 and ρ_p^2 in the following way:

$$\rho^2 = \frac{\mathrm{Cov}^2[X_i, \hat{\alpha}_i]}{\mathrm{Var}[X_i]\cdot\mathrm{Var}[\hat{\alpha}_i]} = \frac{(\mathrm{Cov}[X_i, \alpha_i] + \mathrm{Cov}[X_i, \zeta_i])^2}{\mathrm{Var}[X_i]\cdot(\mathrm{Var}[\alpha_i] + \mathrm{Var}[\zeta_i])} \tag{26}$$

$$= \left(\rho_s\sqrt{\pi} + \rho_p\sqrt{1-\pi}\right)^2, \quad \pi \equiv \frac{\mathrm{Var}[\zeta_i]}{\mathrm{Var}[\hat{\alpha}_i]} . \tag{27}$$

Holding the correlations ρ_s and ρ_p fixed, the importance of the spurious portion of ρ^2, given by ρ_s, increases with π, the fraction of variability in $\hat{\alpha}_i$ due to estimation error. Conversely, if the variability of $\hat{\alpha}_i$ is largely due to fluctuations in α_i, then ρ^2 will reflect mostly ρ_p^2.

Of course, the essence of the problem lies in our inability to identify π except in very special cases. We observe an empirical relation between X_i and $\hat{\alpha}_i$, but we do not know whether the characteristic varies with α_i or with estimation error ζ_i. It is a type of identification problem that is unlikely to be settled by data analysis alone, but must be resolved by providing theoretical motivation for a relation, or no relation, between X_i and α_i. That is, economic considerations must play a dominant role in determining π. We shall return to this issue in the empirical examples of Section 3.

2. Monte Carlo Results

Although the values in Tables 1–3 quantify the magnitude of the biases associated with induced ordering, their practical relevance may be limited in at least three respects. First, the test statistics we have considered are similar in spirit to those used in empirical tests of asset pricing models, but implicitly use the assumption of cross-sectional independence. The more common practice is to estimate the covariance matrix of the N asset returns using a finite number T of time series observations, from which an F-distributed quadratic form may be constructed. Both sampling error from the covariance matrix estimator and cross-sectional dependence will affect the null distribution of $\tilde{\theta}$ in finite samples.

Second, the sampling theory of Section 1 is based on asymptotic approximations, and few results on rates of convergence for Theorem

1.1 are available.[15] How accurate are such approximations for empirically realistic sample sizes?

Finally, the form of the asymptotics does not correspond exactly to procedures followed in practice. Recall that the limiting result involves a finite number n of securities with relative ranks that converge to fixed constants ξ_i as the number of securities N increases without bound. This implies that as N increases, the number of securities in between any two of our chosen n must also grow without bound. However, in practice characteristic-sorted portfolios are constructed from *all* securities within a fractile, not just from those with particular relative ranks. Although intuition suggests that this may be less problematic when n is large (so that within any given fractile there will be many securities), it is surprisingly difficult to verify.[16]

In this section we report results from Monte Carlo experiments that show the asymptotic approximations of Section 1 to be quite accurate in practice despite these three reservations. In Section 2.1, we evaluate the quality of the asymptotic approximations for the $\tilde{\theta}_p$ test used in calculating Tables 2 and 3. In Section 2.2, we consider the effects of induced ordering on F-tests with fixed N and T when the covariance matrix is estimated and the data-generating process is cross-sectionally independent. In Section 2.3, we consider the effects of relaxing the independence assumption.

2.1 Simulation results for $\tilde{\theta}_p$

The $\chi^2_q(\lambda)$ limiting distribution of $\tilde{\theta}_p$ obtains because any finite collection of induced order statistics, each with a fixed distinct limiting relative rank ξ_i in (0, 1), becomes mutually independent as the total number N of securities increases without bound. This asymptotic approximation implies that between any two of the n chosen securities there will be an increasing number of securities omitted from all portfolios as N increases. In practice, all securities within a particular characteristic fractile are included in the sorted portfolios; hence, the theoretical sizes of Table 2 may not be an adequate approximation to this more empirically relevant situation. To explore this possibility we simulate bivariate normal vectors $(\hat{\alpha}_i, X_i)$ with squared correlation R^2, form portfolios using the induced ordering by the X_i's, compute $\tilde{\theta}_p$ using *all* the $\hat{\alpha}_{[i:N]}$'s (in contrast to the asymptotic

[15] However, see Bhattacharya (1984) and Sen (1981).

[16] When n is large relative to a finite N, the asymptotic approximation breaks down. In particular, the dependence between adjacent induced order statistics becomes important for nontrivial n/N. A few elegant asymptotic approximations for sums of induced order statistics are available using functional central limit theory and may allow us to generalize our results to the more empirically relevant case. See, for example, Bhattacharya (1974), Nagaraja (1982a, 1982b, 1984), Sandström (1987), Sen (1976, 1981), and Yang (1981a, 1981b). However, our Monte Carlo results suggest that this generalization may be unnecessary.

Table 4
Empirical sizes of nominal 5 percent χ^2_q-tests of H: $\alpha_i = 0$ ($i = 1, \ldots, n$) using the test statistic $\tilde{\theta}_p$

q	$n_o = 5$	$n_o = 10$	$n_o = 20$	$n_o = 25$	$n_o = 50$
$R^2 = .005$					
5	0.055	0.057	0.067	0.075	0.108
10	0.054	0.063	0.080	0.084	0.139
20	0.056	0.068	0.086	0.106	0.182
25	0.062	0.070	0.104	0.112	0.209
50	0.059	0.077	0.119	0.146	0.314
$R^2 = .01$					
5	0.058	0.064	0.093	0.105	0.174
10	0.059	0.076	0.119	0.130	0.257
20	0.057	0.083	0.140	0.188	0.385
25	0.069	0.100	0.170	0.206	0.445
50	0.083	0.118	0.244	0.300	0.679
$R^2 = .05$					
5	0.091	0.149	0.310	0.392	0.723
10	0.117	0.227	0.493	0.611	0.943
20	0.156	0.351	0.744	0.854	0.999
25	0.163	0.401	0.818	0.916	1.000
50	0.249	0.616	0.971	0.997	1.000
$R^2 = .10$					
5	0.141	0.285	0.601	0.721	0.973
10	0.197	0.473	0.854	0.937	1.000
20	0.308	0.709	0.985	0.998	1.000
25	0.338	0.789	0.995	1.000	1.000
50	0.545	0.961	1.000	1.000	1.000
$R^2 = .20$					
5	0.267	0.577	0.922	0.974	1.000
10	0.405	0.833	0.997	1.000	1.000
20	0.635	0.982	1.000	1.000	1.000
25	0.728	0.996	1.000	1.000	1.000
50	0.933	1.000	1.000	1.000	1.000

$\tilde{\theta}_p \equiv n_o \Sigma_{k=1}^q \tilde{\phi}_k^2/\sigma_\alpha^2$, and $\tilde{\phi}_k \equiv (1/n_o)\Sigma_{j=(k-1)q+1}^{kq} \hat{\alpha}_{[i_j:N]}$ is constructed from portfolio k, with portfolios formed by sorting on some characteristic correlated with estimates $\hat{\alpha}_i$. This induced ordering alters the null distribution of $\hat{\theta}_p$ from χ_q^2 to $(1 - R^2) \cdot \chi_q^2(\lambda)$, where the noncentrality parameter λ is a function of the number q of portfolios, the number n_o of securities in each portfolio, and the squared correlation coefficient R^2 between $\hat{\alpha}_i$ and the sorting characteristic. Each simulation is based on 5000 replications; asymptotic standard errors for the size estimates may be obtained from the usual binomial approximation, and is 3.08×10^{-3} for the 5 percent test.

experiment where only those induced order statistics of given relative ranks are used), and then repeat this procedure 5,000 times to obtain the finite sample distribution.

Table 3 reports the results of these simulations for the same values of R^2, n_o, and q as in Table 2. Except when both n_o and q are small, the empirical sizes of Table 4 match their asymptotic counterparts in Table 2 closely. Consider, for example, the $R^2 = .05$ panel; with five portfolios each with five securities, the difference between the theoretical and empirical size is 1.1 percentage points, whereas this difference is only 0.2 percentage points for 25 portfolios each with 25 securities. When n_o and q are both small, the theoretical and empirical sizes differ more for larger R^2, by as much as 7.4 percent when $R^2 = .20$. However, for the more relevant values of R^2, the empirical and theoretical sizes of the $\tilde{\theta}_p$ test are virtually identical.

2.2 Effects of induced ordering on *F*-tests

Although the results of Section 2.1 support the accuracy of our asymptotic approximation to the sampling distribution of $\tilde{\theta}_p$, the closely related F-statistic is used more frequently in practice. In this section we consider the finite-sample distribution of the F-statistic after induced ordering. We perform Monte Carlo experiments under the now standard multivariate data-generating process common to virtually all static financial asset pricing models. Let r_{it} denote the return of asset i between dates $t - 1$ and t, where $i = 1, 2, \ldots, N$ and $t = 1, 2, \ldots, T$. We assume that for all assets i and dates t the following obtains:

$$r_{it} = \alpha_i + \sum_{j=1}^{k} \beta_{ij} r_t^j + \epsilon_{it}, \tag{28}$$

where α_i and β_{ij} are fixed parameters, r_t^j is the return on some portfolio j (systematic risk), and ϵ_{it} is mean-zero (idiosyncratic) noise. Depending on the particular application, r_{it} may be taken to be nominal, real, or excess asset returns. The process (28) may be viewed as a factor model where the factors correspond to particular portfolios of traded assets, often called the "mimicking portfolios" of an exact factor pricing model. In matrix notation, we have

$$r_t = \alpha + Br_t^p + \epsilon_t, \quad E[\epsilon_t \mid r_t^p] = 0, \quad E[r_t^p] = \mu_p; \tag{29}$$

$$E[\epsilon_s \epsilon_t'] = \begin{cases} \Sigma, & \text{for } s = t, \\ 0, & \text{otherwise;} \end{cases} \tag{30}$$

$$E[(r_s^p - \mu_p)(r_t^p - \mu_p)'] = \begin{cases} \Omega, & \text{for } s = t, \\ 0, & \text{otherwise.} \end{cases} \tag{31}$$

Here, r_t is the $N \times 1$ vector of asset returns at a time t, B is the $N \times k$ matrix of factor loadings, r_t^p is the $k \times 1$ vector of time-t spanning portfolio returns, and α and ϵ_t are $N \times 1$ vectors of asset return intercepts and disturbances, respectively.

This data-generating process is the starting point of the two most popular static models of asset pricing, the CAPM and the APT. Further restrictions are usually imposed by the specific model under consideration, often reducing to the following null hypothesis:

$$H\colon g(\alpha, B) = 0,$$

where the function g is model dependent.[17] Many tests simply set $g(\alpha, B) = \alpha$ and define r_t as excess returns, such as those of the Sharpe–Lintner CAPM and the exact factor-pricing APT. With the added assumption that r_t and r_t^p are jointly normally distributed, the finite-sample distribution of the following test statistic is well known:

$$\psi = \kappa \cdot \frac{\hat{\alpha}'\hat{\Sigma}^{-1}\hat{\alpha}}{1 + \bar{r}^p\hat{\Omega}^{-1}\bar{r}^p} \sim F_{N,T-k-N}, \quad \kappa \equiv \frac{T-k-N}{N}, \tag{32}$$

where $\hat{\Sigma}$ and $\hat{\Omega}$ are the maximum likelihood estimators of the covariance matrices of the disturbances $\hat{\epsilon}_t$ and the spanning portfolio returns r_t^p, respectively, and $\bar{r}^p$ is the vector of sample means of r_t^p. If the number of available securities N is greater than the number of time series observations T less $k + 1$, the estimator $\hat{\Sigma}$ is singular and the test statistic (32) cannot be computed without additional structure. This problem is most often circumvented in practice by forming portfolios. That is, let r_t be a $q \times 1$ vector of returns of q portfolios of securities where $q \ll N$. Since the return-generating process is linear for each security i, a linear relation also obtains for portfolio returns. However, as the analysis of Section 1 foreshadows, if the portfolios are constructed by sorting on some characteristic correlated with $\hat{\alpha}$ then the null distribution of ψ is altered.

To evaluate the null distribution of ψ under characteristic-sorting data snooping, we design our simulation experiments in the following way. The number of time series observations T is set to 60 for all simulations. With little loss in generality, we set the number of spanning portfolios k to zero so that $\hat{\alpha}_i = \Sigma_{t=1}^{T} r_{it}/T$. To separate the effects of *estimating* the covariance matrix from the effects of cross-sectional dependence, we first assume that the covariance matrix Σ of ϵ_t is equal to the identity matrix I—this assumption is relaxed in Section 2.3. We simulate T observations of the $N \times 1$ gaussian vector r_t (where N takes the values 200, 500, and 1000), and compute $\hat{\alpha}$. We then form q portfolios (where q takes the values 10 and 20) by constructing a characteristic X_i that has correlation ρ with $\hat{\alpha}_i$ (where ρ^2 takes the

[17] Examples of tests that fit into this framework are those in Campbell (1987), Connor and Korajczyk (1988), Gibbons (1982), Gibbons and Ferson (1985), Gibbons, Ross, and Shanken (1989), Huberman and Kandel (1987), MacKinlay (1987), Lehmann and Modest (1988), Stambaugh (1982), and Shanken (1985).

values .005, .01, .05, .10, and .20), and then sorting the $\hat{\alpha}_i$'s by this characteristic. To do this, we define

$$X_i \equiv \hat{\alpha}_i + \eta_i, \qquad \eta_i \text{ i.i.d. } \mathrm{N}(0, \sigma_\eta^2), \qquad \sigma_\eta^2 = \frac{1-\rho^2}{T\rho^2}\,. \tag{33}$$

Having constructed the X_i's, we order $\{\hat{\alpha}_i\}$ to obtain $\{\hat{\alpha}_{[i:N]}\}$, construct portfolio intercept estimates that we call $\hat{\phi}_k$, $k = 1, \ldots, n$,

$$\hat{\phi}_k = \frac{1}{n_o} \sum_{i=(k-1)n_o+1}^{kn_o} \hat{\alpha}_{[i:N]}, \qquad N \equiv n_o q, \tag{34}$$

from which we form the F-statistic

$$\psi = \kappa \cdot \hat{\phi}' \hat{\Sigma}^{-1} \hat{\phi} \sim F_{q,T-q}, \qquad \kappa \equiv (T-q)/q, \tag{35}$$

where $\hat{\phi}$ denotes the $q \times 1$ vector of $\hat{\phi}_k$'s, and $\hat{\Sigma}$ is the maximum likelihood estimator of the $q \times q$ covariance matrix of the q portfolio returns. This procedure is repeated 5000 times, and the mean and standard deviation of the resulting distribution for the statistic ψ are reported in Table 5, as well as the size of 1, 5, and 10 percent F-tests.

Even for as small an R^2 as 1 percent, the empirical size of the 5 percent F-test differs significantly from its nominal value for all values of q and n_o. For the sample of 1000 securities grouped into ten portfolios, the empirical rejection rate of 36.7 percent deviates substantially from 5 percent. When the 1000 securities are grouped into 20 portfolios, the size is somewhat lower—26.8 percent—matching the pattern in Table 2. Also similar is the monotonicity of the size with respect to the number of securities. For 200 securities the empirical size is only 7.1 percent with 10 portfolios, but it is more than quintupled with 1000 securities. When the squared correlation between $\hat{\alpha}_i$ and X_i increases to 10 percent, the size of the F-test is essentially unity for sample sizes of 500 or more. Thus even for finite sample sizes of practical relevance, the importance of data snooping via induced ordering cannot be overemphasized.

2.3 *F*-tests with cross-sectional dependence

The substantial bias that induced ordering imparts on the size of portfolio-based F-tests comes from the fact that the induced order statistics $\{\hat{\alpha}_{[i:N]}\}$ generally have nonzero means;[18] hence, the averages of these statistics within sorted portfolios also have nonzero means but reduced variances about those means. Alternatively, the bias from portfolio formation is a result of the fact that the $\hat{\alpha}_i$'s of the extreme portfolios do not approach zero as more securities are combined,

[18] Only those $\hat{\alpha}_{[i:N]}$ for which $i/N \to \frac{1}{2}$ will have zero expectation under the null hypothesis H.

Table 5
Empirical size of $F_{q,T-q}$ tests based on q portfolios sorted by a random characteristic whose squared correlation with $\hat{\alpha}_i$ is R^2

q	n_o	n	Mean	Std. Dev.	Size 10%	Size 5%	Size 1%
$R^2 = .005$							
10	20	200	1.111	0.542	0.124	0.041	0.014
20	10	200	1.081	0.424	0.107	0.054	0.009
10	50	500	1.238	0.611	0.177	0.070	0.026
20	25	500	1.147	0.462	0.152	0.079	0.018
10	100	1000	1.406	0.679	0.270	0.118	0.046
20	50	1000	1.240	0.500	0.194	0.114	0.033
$R^2 = .01$							
10	20	200	1.225	0.619	0.181	0.071	0.026
20	10	200	1.148	0.460	0.148	0.079	0.018
10	50	500	1.512	0.728	0.318	0.152	0.070
20	25	500	1.301	0.514	0.240	0.143	0.036
10	100	1000	2.030	0.908	0.576	0.367	0.203
20	50	1000	1.554	0.596	0.405	0.268	0.098
$R^2 = .05$							
10	20	200	1.980	0.883	0.549	0.342	0.189
20	10	200	1.505	0.582	0.369	0.241	0.082
10	50	500	3.501	1.335	0.945	0.846	0.700
20	25	500	2.264	0.801	0.798	0.670	0.382
10	100	1000	5.991	1.976	0.999	0.997	0.986
20	50	1000	3.587	1.169	0.992	0.972	0.879
$R^2 = .10$							
10	20	200	2.961	1.196	0.868	0.713	0.538
20	10	200	1.977	0.727	0.658	0.510	0.257
10	50	500	5.939	1.931	0.999	0.997	0.987
20	25	500	3.526	1.128	0.988	0.968	0.868
10	100	1000	10.888	3.050	1.000	1.000	1.000
20	50	1000	6.123	1.811	1.000	1.000	0.999
$R^2 = .20$							
10	20	200	4.831	1.657	0.997	0.982	0.937
20	10	200	2.895	0.992	0.948	0.882	0.667
10	50	500	10.796	3.022	1.000	1.000	1.000
20	25	500	6.005	1.758	1.000	1.000	0.998
10	100	1000	20.695	5.112	1.000	1.000	1.000
20	50	1000	11.194	2.988	1.000	1.000	1.000

n_o is the number of securities in each portfolio and $n \equiv n_o q$ is the total number of securities. The number of time series observations T is set to 60. The mean and standard deviation of the test statistic over the 5000 replications are reported. The population mean and standard deviation of $F_{10,50}$ are 1.042 and 0.523, respectively; those of the $F_{20,40}$ are 1.053 and 0.423, respectively. Asymptotic standard errors for the size estimates may be obtained from the usual binomial approximation; they are 4.24×10^{-3}, 3.08×10^{-3}, and 1.41×10^{-3} for the 10, 5, and 1 percent tests, respectively.

whereas the residual variances of the portfolios (and consequently the variances of the portfolio $\hat{\alpha}_i$'s) do tend to zero. Of course, our assumption that the disturbances ϵ_t of (29) are cross-sectionally independent implies that the portfolio residual variance approaches zero rather quickly (at rate $1/n_o$). But in many applications (such as the CAPM), cross-sectional independence is counterfactual. Firm size and industry membership are but two factors that might induce cross-sectional correlation in return residuals. In particular, when the residuals are positively cross-sectionally correlated, the bias is likely to be smaller since there is less variance reduction in forming portfolios than in the cross-sectionally independent case.

To see how restrictive the independence assumption is, we simulate a data-generating process in which disturbances are cross-sectionally correlated. The design is identical to that of Section 2.2 except that the residual covariance matrix Σ is no longer diagonal. Instead, we set

$$\Sigma = \delta\delta' + I, \tag{36}$$

where δ is an $N \times 1$ vector of parameters and I is the identity matrix. Such a covariance matrix would arise, for example, from a single common factor model for the $N \times 1$ vector of disturbances ϵ_t:

$$\epsilon_t = \delta\Lambda_t + \nu_t, \tag{37}$$

where Λ_t is some i.i.d. zero-mean unit-variance common factor independent of ν_t, and ν_t is N-dimensional vector white noise with covariance matrix I. For our simulations, the parameters δ are chosen to be equally spaced in the interval $[-1, 1]$. With this design the cross-correlation of the disturbances will range from -0.5 to 0.5. The X_i's are constructed as in (33) with

$$\sigma_\eta^2 = \frac{(1 - \rho^2)\sigma^2(\alpha)}{\rho^2}, \qquad \sigma^2(\alpha) \equiv \frac{1}{NT}\sum_{i=1}^{N} (\delta_i^2 + 1), \tag{38}$$

where ρ^2 is fixed at .05.

Under this design, the results of the simulation experiments may be compared to the third panel of Table 5, and are reported in Table 6.[19] Despite the presence of cross-sectional dependence, the impact of induced ordering on the size of the F-test is still significant. For example, with 20 portfolios each containing 25 securities the empirical size of the 5 percent test is 32.3 percent; with 10 portfolios of 50 securities each the empirical size increases to 82.0 percent. As in the cross-sectionally independent case, the bias increases with the number of securities given a fixed number of portfolios, and the bias decreases as the number of portfolios is increased given a fixed number of securities. Not surprisingly, for fixed n_o and q, cross-sectional dependence of the $\hat{\alpha}_i$'s lessens the bias. However, the entries in Table 6 demonstrate that the effects of data snooping may still be substantial even in the presence of cross-sectional dependence.

[19] The correspondence between the two tables is not exact because the dependency introduced in (36) induces cross-sectional heteroskedasticity in the $\hat{\alpha}_i$'s; hence, $\rho^2 = .05$ yields an R^2 of .05 only approximately.

Table 6
Empirical size of $F_{q,T-q}$ tests based on q portfolios sorted by a random characteristic whose squared correlation with $\hat{\alpha}_i$ is approximately .05

q	n_o	n	Mean	Std. Dev.	Size 10%	Size 5%	Size 1%
$R^2 \approx .05$							
10	20	200	1.700	0.763	0.422	0.216	0.100
20	10	200	1.372	0.528	0.270	0.167	0.047
10	50	500	2.520	1.041	0.765	0.565	0.367
20	25	500	1.867	0.693	0.593	0.322	0.205
10	100	1000	3.624	1.605	0.925	0.820	0.682
20	50	1000	2.516	0.966	0.844	0.743	0.501

n_o is the number of securities in each portfolio and $n \equiv n_o q$ is the total number of securities. The $\hat{\alpha}_i$'s of the portfolios are cross-sectionally correlated, where the source of correlation is an i.i.d. zero-mean common factor in the returns. The number of time series observations T is set to 60. The mean and standard deviation of the test statistic over the 5000 replications are reported. The population mean and standard deviation of $F_{10,50}$ are 1.042 and 0.523, respectively; those of the $F_{20,40}$ are 1.053 and 0.423, respectively. Asymptotic standard errors for the size estimates may be obtained from the usual binomial approximation; they are 4.24×10^{-3}, 3.08×10^{-3}, and 1.41×10^{-3} for the 10, 5, and 1 percent tests, respectively.

3. Two Empirical Examples

To illustrate the potential relevance of data-snooping biases associated with induced ordering, we provide two examples drawn from the empirical literature. The first example is taken from the early tests of the Sharpe–Lintner CAPM, where portfolios were formed by sorting on out-of-sample betas. We show that such tests can be biased towards falsely rejecting the CAPM if in-sample betas are used instead, underscoring the importance of the elaborate sorting procedures used by Black, Jensen, and Scholes (1972) and Fama and MacBeth (1973). Our second example concerns tests of the APT that reject the zero-intercept null hypothesis when applied to portfolio returns sorted by market value of equity. We show that data-snooping biases can account for much the same results, and that only additional economic restrictions will determine the ultimate source of the rejections.

3.1 Sorting by beta

Although tests of the Sharpe–Lintner CAPM may be conducted on individual securities, the potential benefits of using multiple securities are well known. One common approach for allocating securities to portfolios has been to rank them by their betas and then group the sorted securities. Beta-sorted portfolios will exhibit more risk dispersion than portfolios of randomly chosen securities, and may therefore yield more information about the CAPM's risk–return relation. Ideally, portfolios would be formed according to their true betas. However, since the population betas are unobservable, in practice portfolios have grouped securities by their estimated betas. For example, both Black, Jensen, and Scholes (1972) and Fama and MacBeth (1973) use portfolios formed by sorting on estimated betas, where the betas are estimated with a *prior* sample of stock returns. Their motivation for this more complicated procedure was to avoid grouping common estimation or measurement error since, within the sample, securities with high estimated betas will tend to have positive realizations of estimation error, and vice versa for securities with low estimated betas.

Suppose, instead, that securities are grouped by betas estimated *in sample.* Can grouping common estimation error change inferences substantially? To answer this question within our framework, suppose the Sharpe–Lintner CAPM obtains so that

$$r_{it} = \beta_i r_{mt} + \epsilon_{it}, \qquad \mathrm{E}[\epsilon_t \mid r_{mt}] = 0, \qquad \mathrm{E}[\epsilon_t \epsilon_t'] = \sigma_\epsilon^2 I, \tag{39}$$

where r_{it} denotes the excess return of security i, r_{mt} is the excess market return, and ϵ_t is the $N \times 1$ vector of disturbances. To assess the impact of sorting on in-sample betas, we require the squared correlation of $\hat{\alpha}_i$ and $\hat{\beta}_i$. However, since our framework requires that both $\hat{\alpha}_i$ and $\hat{\beta}_i$ be independently and identically distributed, and since $\hat{\beta}_i$ is the sum of β_i and estimation error ζ_i, we assume β_i to be random to allow for cross-sectional variation in the betas. Therefore, let

$$\beta_i \text{ i.i.d. } \mathrm{N}(\mu_\beta, \sigma_\beta^2), \qquad i = 1, 2, \ldots, N,$$

where each β_i is independent of all ϵ_{jt} in (39). The squared correlation between $\hat{\alpha}_i$ and $\hat{\beta}_i$ may then be explicitly calculated as

$$\rho^2(\hat{\alpha}_i, \hat{\beta}_i) = \frac{\mathrm{Cov}^2[\hat{\alpha}_i, \hat{\beta}_i]}{\mathrm{Var}[\hat{\alpha}_i]\,\mathrm{Var}[\hat{\beta}_i]} = \frac{\hat{S}_m^2}{1 + \hat{S}_m^2} \cdot \frac{1}{1 + (\sigma_\beta^2 \hat{\sigma}_m^2/\sigma_\epsilon^2)T}, \tag{40}$$

where $\hat{\mu}_m$ and $\hat{\sigma}_m$ are the sample mean and standard deviation of the excess market return, respectively, $S_m \equiv \hat{\mu}_m/\hat{\sigma}_m$ is the ex post Sharpe measure, and T is the number of time series observations used to estimate the α_i's and β_i's.

The term $\sigma_\beta^2 \hat{\sigma}_m^2 T/\sigma_\epsilon^2$ in (40) captures the essence of the errors-in-variables problem for in-sample beta sorting. This is simply the ratio of the cross-sectional variance in betas, σ_β^2, to the variance of the beta estimation error, $\sigma_\epsilon^2/(\hat{\sigma}_m^2 T)$. When the cross-sectional dispersion of the betas is much larger than the variance of the estimation errors, this ratio is large, implying a small value for ρ^2 and little data-snooping bias. In fact, since the estimation error of the betas declines with the number of observations T, as the time period lengthens, in-sample beta sorting becomes less problematic. However, when the variance of the estimation error is large relative to the cross-sectional variance

Table 7
Theoretical sizes of nominal 5 percent χ_q^2-tests under the null hypothesis of the Sharpe–Lintner CAPM using q in-sample beta-sorted portfolios with n_o securities per portfolio

Sample period	$\hat{R}^2$	$q = 10$ $n_o = 250$	$q = 20$ $n_o = 125$	$q = 50$ $n_o = 50$
January 1954–December 1958	.044	1.000	1.000	1.000
January 1959–December 1963	.007	0.790	0.656	0.435
January 1964–December 1968	.048	1.000	1.000	1.000
January 1969–December 1973	.008	0.869	0.756	0.529
January 1974–December 1978	.001	0.183	0.139	0.100
January 1979–December 1983	.023	1.000	1.000	0.991
January 1984–December 1988	.002	0.248	0.183	0.123

$\hat{R}^2$ is the estimated squared correlation between $\hat{\beta}_i$ and $\hat{\alpha}_i$ under the null hypothesis that $\alpha_i = 0$ and that the β_i's are i.i.d. normal random variables with mean and variance μ_β and σ_β^2, respectively. Within each subsample, the estimate $\hat{R}^2$ is based on the first 200 stocks in the CRSP monthly returns files with complete return histories over the five-year subperiod, and the CRSP equal-weighted index. For illustrative purposes, the theoretical size is computed under the assumption that the total number of securities $n \equiv n_o q$ is fixed at 2500.

of the betas, then ρ^2 is large and grouping common estimation errors becomes a more serious problem.

To show just how serious this might be in practice, we report in Table 7 the estimated ρ^2 between $\hat{\alpha}_i$ and $\hat{\beta}_i$ for five-year subperiods from January 1954 to December 1988, where each estimate is based on the first 200 securities listed in the CRSP monthly returns files with complete return histories within the particular five-year subsample, and the CRSP equal-weighted index. Also reported is the probability of rejecting the null hypothesis $\alpha_i = 0$ when it is true using a 5 percent test, assuming a sample of 2500 securities, where the number of portfolios q is 10, 20, or 50 and the number of securities per portfolio n_o is defined accordingly.[20]

The entries in Table 7 show that the null hypothesis is quite likely to be rejected even when it is true. For many of the subperiods, the probability of rejecting the null is unity, and when only 10 beta-sorted portfolios are used, the smallest size of a nominal 5 percent test is still 18.3 percent. We conclude, somewhat belatedly, that the elaborate out-of-sampling sorting procedures used by Black, Jensen, and Scholes (1972) and Fama and MacBeth (1973) were indispensable to the original tests of the Sharpe–Lintner CAPM.

3.2 Sorting by size

As a second example of the practical relevance of data-snooping biases, we consider Lehmann and Modest's (1988) multivariate test of a 15-

[20] Our analysis is limited by the counterfactual assumption that the market model disturbances are cross-sectionally uncorrelated. But the simulation results presented in Section 2.3 indicate that biases are still substantial even in the presence of cross-sectional dependence. A more involved application would require a deeper analysis of cross-sectional dependence in the ϵ_{it}'s.

factor APT model, in which they reject the zero-intercept null hypothesis using five portfolios formed by grouping securities ordered by market value of equity.[21] We focus on this particular study because of the large number of factors employed—our framework requires the disturbances ϵ_t of (29) to be cross-sectionally independent, and since 15 factors are included in Lehmann and Modest's cross-sectional regressions, a diagonal covariance matrix for ϵ_t is not implausible.

It is well-known that the estimated intercept $\hat{\alpha}_i$ from the single-period CAPM regression (excess individual security returns regressed on an intercept and the market risk premium) is negatively cross-sectionally correlated with log size.[22] Since this $\hat{\alpha}_i$ will in general be correlated with the estimated intercept from a 15-factor APT regression, it is likely that the estimated APT intercept and log size will also be empirically correlated.[23] Unfortunately, we do not have a direct measure of the correlation of the APT intercept and log size which is necessary to derive the appropriate null distribution after induced ordering.[24] As an alternative, we estimate the cross-sectional R^2 of the estimated CAPM alpha with the logarithm of size, and we use this R^2 as well as $\frac{1}{2}R^2$ and $\frac{1}{4}R^2$ to estimate the bias attributable to induced ordering.

Following Lehmann and Modest (1988), we consider four five-year time periods from January 1963 to December 1982. X_i is defined to be the logarithm of beginning-of-period market values of equity. The $\hat{\alpha}_i$'s are the intercepts from regressions of excess returns on the market risk premium as measured by the difference between an equal-weighted NYSE index and monthly Treasury bill returns, where the NYSE index is obtained from the Center for Research in Security Prices (CRSP) database. The R^2's of these regressions are reported in the second column of Table 8. One cross-sectional regression of $\hat{\alpha}_i$ on log size X_i is run for each five-year time period using monthly NYSE-AMEX data from CRSP. We run regressions only for those stocks having complete return histories within the relevant five-year period.

Table 8 contains the test statistics for a 15-factor APT framework using five size-sorted portfolios. The first four rows contain results

[21] See Lehmann and Modest (1988, table 1, last row). Connor and Korajczyk (1988) report similar findings.

[22] See, for example, Banz (1981) and Brown, Kleidon, and Marsh (1983).

[23] We recognize that correlation is not transitive, so if X is correlated with Y and Y with Z, X need not be correlated with Z. However, since the intercepts from the two regressions will be functions of some common random variables, situations in which they are independent are the exception rather than the rule.

[24] Nor did Lehmann and Modest prior to their extensive investigations. If they are subject to any data-snooping biases it is only from their awareness of size-related empirical results for the single-period CAPM, and of corresponding results for the APT as in Chan, Chen, and Hsieh (1985).

Table 8
Comparison of p-values for Lehmann and Modest's (1988) tests of the APT with and without correcting for the effects of induced ordering

Sample	N	$\hat{R}^2$	$\tilde{\theta}_p$	χ^2 p-value	$\chi^2(\lambda_1)$ p-value	$\chi^2(\lambda_2)$ p-value	$\chi^2(\lambda_3)$ p-value
6301–6712	1001	0.015	13.70	0.018	0.687	0.315	0.131
6801–7212	1359	0.040	15.50	0.008	1.000	0.919	0.520
7301–7712	1346	0.033	10.20	0.070	1.000	0.963	0.720
7801–8212	1281	0.004	12.05	0.034	0.272	0.134	0.078
Aggregate	—	—	51.45	0.00014	1.000	0.917	0.298

In the absence of data snooping, the appropriate test statistics and their p-values (using the central χ^2 distribution) are given in Lehmann and Modest (1988, table 1) and reported below in columns 4 and 5 (we transform their F-statistics into χ^2 variates for purposes of comparison). Corresponding p-values that account for induced ordering are calculated in columns labeled "$\chi^2(\lambda_i)$ p-value" ($i = 1, 2, 3$) (using the noncentral χ^2 distribution), where λ_1, λ_2, and λ_3 are noncentrality parameters computed with $\hat{R}^2$, $\frac{1}{2}\hat{R}^2$, and $\frac{1}{4}\hat{R}^2$, respectively. In all cases, five portfolios are formed from the total number of securities; this yields five degrees of freedom for the χ^2 statistics in the first four rows, and 20 degrees of freedom for the aggregate χ^2 statistics.

for each of the four subperiods and the last row contains aggregate test statistics. To apply the results of Sections 1 and 2 we transform Lehmann and Modest's (1988) F-statistics into (asymptotic) χ^2 variates.[25] The total number of available securities ranges from a minimum of 1001 for the first five-year subperiod to a maximum of 1359 for the second subperiod. For each test statistic in Table 8 we report four different p-values: the first is with respect to the null distribution that ignores data snooping, and the next three are with respect to null distributions that account for induced ordering to various degrees.

The entries in Table 8 show that the potential biases from sorting by characteristics that have been empirically selected can be immense. The p-values range from 0.008 to 0.070 in the four subperiods according to the standard theoretical null distribution, yielding an aggregate p-value of 0.00014, considerable evidence against the null. When we adjust for the fact that the sorting characteristic is selected empirically (using the $\hat{R}^2$ from the cross-sectional regression of $\hat{\alpha}_i$ on X_i), the p-values for these same four subperiods range from 0.272 to 1.000, yielding an aggregate p-value of 1.000! Therefore, whether or not induced ordering is allowed for can change inferences dramatically.

The appropriate R^2 in the preceding analysis is the squared correlation between log size and the intercept from a 15-factor APT regression, and not the one used in Table 8. To see how this may affect our conclusions, recall from (2) that the cross-sectional correlation between $\hat{\alpha}_i$ and log size can arise from two sources: the

[25] Since Lehmann and Modest (1988) use weekly data, the null distribution of their test statistics is $F_{5,240}$ In practice the inferences are virtually identical using the χ^2_5 distribution after multiplying the test statistic by 5.

estimation error ζ_i in $\hat{\alpha}_i$, and the cross-sectional dispersion in the "true" CAPM α_i (which is zero under the null hypothesis). Correlation between X_i and ζ_i will be partially reflected in correlation between the estimated APT intercept and log size. The second source of correlation will not be relevant ***under the APT null hypothesis*** since under that scenario we assume that the 15-factor APT obtains and therefore the intercept vanishes for all securities. As a conservative estimate for the appropriate R^2 to be used in Table 8, we set the squared correlation equal to $\frac{1}{2}\hat{R}^2$ and $\frac{1}{4}\hat{R}^2$, yielding the p-values reported in the last two columns of Table 8. Even when the squared correlation is only $\frac{1}{4}\hat{R}^2$, the inferences change markedly after induced ordering, with p-values ranging from 0.078 to 0.720 in the four subperiods and 0.298 in the aggregate. This simple example illustrates the severity with which even a mild form of data snooping can bias our inferences in practice.

Nevertheless, it should not be inferred from Table 8 that all size-related phenomena are spurious. After all, the correlation between X_i and $\hat{\alpha}_i$ may be the result of cross-sectional variations in the population α_i's, and not estimation error. Even so, tests using size-sorted portfolios are still biased if based on the same data from which the size effect was previously observed. A procedure that is free from such biases is to decide today that size is an interesting characteristic, collect ten years of new data, and then perform tests on size-sorted portfolios from this fresh sample. Provided that the old and new samples are statistically independent, this will yield a perfectly valid test of the null hypothesis H, since the only possible source of correlation between the X_i's and the $\hat{\alpha}_i$'s in the new sample is from the α_i's (presumably the result of some underlying economic relation between the two), and not from the estimation errors. In such cases, induced ordering cannot affect the distribution of the test statistics under the null hypothesis, and will yield a considerably more powerful test against many alternatives.

4. How the Data Get Snooped

Whether the probabilities of rejection in Table 2 are to be interpreted as size or power depends, of course, on the particular null and alternative hypotheses at hand, the key distinction being the source of correlation between $\hat{\alpha}_i$ and the characteristic X_i. Since our starting point in Section 1 was the assertion that this correlation is "spurious," we view the values of Table 2 as probabilities of falsely rejecting the null hypothesis. We suggested in Section 1 that the source of this spurious correlation is correlation between the characteristic and the estimation errors in $\hat{\alpha}_i$, since such errors are the only source of variation in $\hat{\alpha}_i$ under the null. But how does this correlation arise? One possibility is the very mechanism by which characteristics are selected. Without any economic theories for motivation, a plausible behavioral model of how we determine characteristics to be particularly "interesting" is that we tend to focus on those that have unusually large squared sample correlations or R^2's with the $\hat{\alpha}_i$'s. In the spirit of Ross (1987), economists study "interesting" events, as well as events that are interesting from a theoretical perspective. If so, then even in a collection of K characteristics all of which are independent of the $\hat{\alpha}_i$'s, correlation between the $\hat{\alpha}_i$'s and the most "interesting" characteristic is artificially induced.

More formally, suppose for each of N securities we have a collection of K distinct and mutually independent characteristics Y_{ik}, $k = 1, 2, \ldots, K$, where Y_{ik} is the kth characteristic of the ith security. Let the null hypothesis obtain so that $\alpha_i = 0$, for all i, and assume that all characteristics are independent of $\{\hat{\alpha}_i\}$. This last assumption implies that the distribution of a test statistic based on grouped $\hat{\alpha}_i$'s is unaffected by sorting on any of the characteristics. For simplicity let each of the characteristics and the $\hat{\alpha}_i$'s be normally distributed with zero mean and unit variance, and consider the sample correlation coefficients:

$$\hat{\rho}_k = \frac{\Sigma_{i=1}^{N}(Y_{ik} - \bar{Y}_k)(\hat{\alpha}_i - \bar{\hat{\alpha}})}{\sqrt{\Sigma_{i=1}^{N}(Y_{ik} - \bar{Y}_k)^2} \cdot \sqrt{\Sigma_{i=1}^{N}(\hat{\alpha}_i - \bar{\hat{\alpha}})^2}}, \quad k = 1, 2, \ldots, K, \tag{41}$$

where $\bar{Y}_k$ and $\bar{\hat{\alpha}}$ are the sample means of characteristic k and the $\hat{\alpha}_i$'s, respectively. Suppose we choose as our sorting characteristic the one that has the largest squared correlation with the $\hat{\alpha}_i$'s, and call this characteristic X_i. That is, $X_i \equiv Y_{ik^*}$, where the index k^* is defined by

$$\hat{\rho}_{k^*}^2 = \operatorname*{Max}_{1 \leq k \leq K} \hat{\rho}_k^2 . \tag{42}$$

This X_i is a new characteristic in the statistical sense, in that its distribution is no longer the same as that of the Y_{ik}'s.[26] It is apparent that X_i and $\hat{\alpha}_i$ are not mutually independent since the $\hat{\alpha}_i$'s were used in selecting this characteristic. By construction, extreme realizations of the random variables $\{X_i\}$ tend to occur when extreme realizations of $\{\hat{\alpha}_i\}$ occur.

To estimate the magnitude of correlation spuriously induced between X_i and $\hat{\alpha}_i$, first observe that although the correlation between Y_{ik} and $\hat{\alpha}_i$ is zero for all k, $E[\hat{\rho}_k^2] = 1/(N - 1)$ under our normality assumption. Therefore, $1/(N - 1)$ should be our benchmark in assessing the degree of spurious correlation between X_i and $\hat{\alpha}_i$. Since the $\hat{\rho}_k^2$'s are well-known to be independently and identically distributed

Beta$(\frac{1}{2}, \frac{1}{2}(N-2))$ variates, the distribution and density functions of $\hat{\rho}^2_{k*}$, denoted by $F_*(v)$ and $f_*(v)$, respectively, may be readily derived as[27]

$$F_*(v) = [F_\beta(v)]^K, \quad v \in (0, 1), \tag{43}$$

$$f_*(v) = K[F_\beta(v)]^{K-1} f_\beta(v), \quad v \in (0, 1), \tag{44}$$

where F_β and f_β are the cumulative distribution function and probability density function of the Beta distribution with parameters $\frac{1}{2}$ and $\frac{1}{2}(N-2)$. A measure of that portion of squared correlation between X_i with $\hat{\alpha}_i$ due to sorting on $\hat{\rho}^2_k$ is then given by

$$\gamma \equiv E[\hat{\rho}^2_{k*}] - E[\hat{\rho}^2_k] = \int_0^1 v f_*(v)\, dv - \frac{1}{N-1}. \tag{45}$$

For 25 securities and 50 characteristics, γ is 20.5 percent![28] With 100 securities, γ is still 5.4 percent and only declines to 1.1 percent for 500 securities. With only 25 characteristics, the values of γ for 25, 100, and 500 securities fall to 16.4, 4.2, and 0.8 percent, respectively. However, these smaller values of γ can still yield misleading inferences for tests based on few portfolios, each containing many securities. This is seen in Table 9, in which the theoretical sizes of 5 percent tests with R^2's equal to the appropriate γ for each cell are displayed. For example, the first entry in the first row of Table 9, 0.163, is the size of the 5 percent portfolio-based test with five portfolios and five securities in each, where the R^2 used to perform the calculation is the γ corresponding to 25 securities and 25 characteristics, or 16.4 percent. As the number of securities per portfolio grows, γ declines but the bias worsens—with 50 securities in each of 5 portfolios, γ is only 1.7 percent but the actual size of a 5 percent test is 26.4 percent. Although there is in fact no statistical relation between

[26] In fact, if we denote by Y_k the $N \times 1$ vector containing values of characteristic k for each of the N securities, then the vector most highly correlated with $\hat{\alpha}$ (which we have called X) may be viewed as the concomitant $Y_{[K\ K]}$ of the Kth order statistic $\hat{\rho}^2_{[K\ K]} = \hat{\rho}^{2*}_k$. As in the scalar case, induced ordering does change the distribution of the vector concomitants.

[27] That the squared correlation coefficients are i.i.d. Beta random variables follows from our assumptions of normality and the mutual independence of the characteristics and the $\hat{\alpha}_i$'s [see Stuart and Ord (1987, chapter 16.28) for example]. The distribution and density functions of the maximum follow directly from this.

[28] Note that γ is only an approximation to the squared population correlation:

$$\left[\frac{E(X_i - E[X])(\hat{\alpha}_i - E[\hat{\alpha}])}{\sqrt{E(X_i - E[X])^2} \cdot \sqrt{E(\hat{\alpha}_i - E[\hat{\alpha}])^2}}\right]^2.$$

However, Monte Carlo simulations with 10,000 replications show that this approximation is excellent even for small sample sizes. For example, fixing K at 50, the correlation from the simulations is 22.82 percent for $N = 25$, whereas (45) yields $\gamma = 20.47$ percent; for $N = 100$ the simulations yield a correlation of 6.25 percent, compared to a γ of 5.39 percent.

Table 9
Theoretical sizes of nominal 5 percent χ_q^2-tests of H: $\alpha_i = 0$ ($i = 1, \ldots, n$) using the test statistic $\tilde{\theta}_p$

q	$n_o = 5$	$n_o = 10$	$n_o = 20$	$n_o = 25$	$n_o = 50$
$K = 25$					
5	0.163	0.216	0.246	0.253	0.264
10	0.150	0.182	0.200	0.202	0.210
20	0.125	0.144	0.153	0.155	0.159
25	0.117	0.132	0.140	0.142	0.145
50	0.096	0.104	0.109	0.110	0.112
$K = 50$					
5	0.197	0.270	0.311	0.319	0.337
10	0.183	0.228	0.254	0.259	0.270
20	0.151	0.178	0.192	0.195	0.201
25	0.141	0.163	0.175	0.177	0.182
50	0.112	0.125	0.131	0.133	0.136

$\tilde{\theta}_p \equiv n_o \Sigma_{k=1}^{q} \tilde{\phi}_k^2/\sigma_\alpha^2$, and $\tilde{\phi}_k \equiv (1/n_o)\Sigma_{j=(k-1)q+1}^{kq} \hat{\alpha}_{[ij\ N]}$ is constructed from portfolio k, with portfolios formed by sorting on some characteristic correlated with estimates $\hat{\alpha}_i$. This induced ordering alters the null distribution of $\hat{\theta}_p$ from χ_q^2 to $(1 - R^2)\cdot\chi_q^2(\lambda)$, where the noncentrality parameter λ is a function of the number q of portfolios, the number n_o of securities in each portfolio, and the squared correlation coefficient R^2 between $\hat{\alpha}_i$ and the sorting characteristic. The values of R^2 used for the size calculations vary with the total number of securities $n_o q$ and with K, the total number of independent characteristics from which the most "interesting" is selected.

any of the characteristics and the $\hat{\alpha}_i$'s, a procedure that focuses on the most striking characteristic can *create* spurious statistical dependence.

As the number of securities N increases, this particular source of dependence becomes less important since all the sample correlation coefficients $\hat{\rho}_k$ converge almost surely to zero, as does γ. However, recall from Table 2 that as the sample size grows the bias increases if the number of portfolios is held fixed; hence, as Table 9 illustrates, a larger N and thus a smaller γ need not imply a smaller bias. Moreover, since γ is increasing in the number of characteristics K, we cannot find refuge in the law of large numbers without weighing the number of securities against the number of characteristics and portfolios in some fashion. Table 9 provides one informal measure of this trade-off.

Perhaps even the most unscrupulous investigator might hesitate at the kind of data snooping we have just considered. However, the very review process that published research undergoes can have much the same effect, since competition for limited journal space tilts the balance in favor of the most striking and dissonant of empirical results. Indeed, the "Anomalies" section of the *Journal of Economic Perspectives* is the most obvious example of our deliberate search for the unusual in economics. As a consequence, interest may be created in otherwise theoretically irrelevant characteristics. In the absence of an economic paradigm, such data-snooping biases are not easily distinguishable from violations of the null hypothesis. This inability to separate pretest bias from alternative hypotheses is the most compelling criticism of "measurement without theory."

5. Conclusion

Although the size effect may signal important differences between the economic structure of small and large corporations, how these differences are manifested in the stochastic properties of their equity returns cannot be reliably determined through data analysis alone. Much more convincing would be the empirical significance of size, or any other quantity, that is based on a model of economic equilibrium in which the characteristic is related to the behavior of asset returns endogenously. Our findings show that tests using securities grouped according to theoretically motivated correlations between X_i and $\hat{\alpha}_i$ can be powerful indeed—interestingly, tests of the APT with portfolios sorted by such characteristics (own-variance and dividend yield) no longer reject the null hypothesis [see Lehmann and Modest (1988)]. Sorting on size yields rejections whereas sorting on theoretically relevant characteristics such as own-variance and dividend yield does not. This suggests that data-instigated grouping procedures should be employed cautiously.

It is widely acknowledged that incorrect conclusions may be drawn from procedures violating the assumptions of classical statistical inference, but the nature of these violations is often as subtle as it is profound. In observing that economists (as well as those in the natural sciences) tend to seek out anomalies, Merton (1987, p. 104) writes: "All this fits well with what the cognitive psychologists tell us is our natural individual predilection to focus, often disproportionately so, on the unusual. . . . This focus, both individually and institutionally, together with little control over the number of tests performed, creates a fertile environment for both unintended selection bias and for attaching greater significance to otherwise unbiased estimates than is justified." The recognition of this possibility is a first step in guarding against it. The results of our paper provide a more concrete remedy for such biases in the particular case of portfolio formation via induced ordering on data-instigated characteristics. However, nonexperimental inference may never be completely free from data-snooping biases since the attention given to empirical anomalies, incongruities, and unusual correlations is also the modus operandi for genuine discovery and progress in the social sciences. Formal statistical analyses such as ours may serve as primitive guides to a better understanding of economic phenomena, but the ability to distinguish between the spurious and the substantive is likely to remain a cherished art.

References

Aldous, D., 1989, *Probability Approximations via the Poisson Clumping Heuristic,* Springer, New York.

Banz, R. W., 1978, "Limited Diversification and Market Equilibrium: An Empirical Analysis," Ph.D. dissertation, University of Chicago.

Banz, R. W., 1981, "The Relationship Between Return and Market Value of Common Stocks," *Journal of Financial Economics,* 9, 3–18.

Berger, J., and R. Wolpert, 1984, *The Likelihood Principle,* Lecture Notes—Monograph Series Volume 6, Institute of Mathematical Statistics, Hayward, Cal.

Bhattacharya, P. K., 1974, "Convergence of Sample Paths of Normalized Sums of Induced Order Statistics," *Annals of Statistics,* 2, 1034–1039.

Bhattacharya, P. K., 1984, "Induced Order Statistics: Theory and Applications," in P. R. Krishnaiah and P. K. Sen (eds.), *Handbook of Statistics 4: Nonparametric Methods,* North-Holland, Amsterdam.

Black, F., M. Jensen, and M. Scholes, 1972, "The Capital Asset Pricing Model: Some Empirical Tests," in M. Jensen (ed.), *Studies in the Theory of Capital Markets,* Praeger, New York.

Brown, P., A. Kleidon, and T. Marsh, 1983, "New Evidence on the Nature of Size Related Anomalies in Stock Prices," *Journal of Financial Economics,* 12, 33–56.

Campbell, J. Y., 1987, "Stock Returns and the Term Structure," *Journal of Financial Economics,* 18, 373–400.

Chamberlain, G., 1983, "Funds, Factors, and Diversification in Arbitrage Pricing Models," *Econometrica,* 51, 1305–1323.

Chan, K., N. Chen, and D. Hsieh, 1985, "An Exploratory Investigation of the Firm Size Effect," *Journal of Financial Economics,* 14, 451–471.

Chen, N., R. Roll, and S. Ross, 1986, "Economic Forces and the Stock Market," *Journal of Business,* 59, 383–403.

Connor, G., and R. Korajczyk, 1988, "Risk and Return in an Equilibrium APT: Application of a New Test Methodology," *Journal of Financial Economics,* 21, 255–290.

David, H. A., 1973, "Concomitants of Order Statistics," *Bulletin of the International Statistical Institute,* 45, 295–300.

David, H. A., 1981, *Order Statistics* (2nd ed.), Wiley, New York.

David, H. A., and J. Galambos, 1974, "The Asymptotic Theory of Concomitants of Order Statistics," *Journal of Applied Probability,* 11, 762–770.

Fama, E., and J. MacBeth, 1973, "Risk, Return, and Equilibrium: Empirical Tests," *Journal of Political Economy,* 71, 607–636.

Gibbons, M. R., 1982, "Multivariate Tests of Financial Models: A New Approach," *Journal of Financial Economics,* 10, 3–27.

Gibbons, M. R., and W. Ferson, 1985, "Testing Asset Pricing Models with Changing Expectations and an Unobservable Market Portfolio," *Journal of Financial Economics,* 14, 217–236.

Gibbons, M. R., S. A. Ross, and J. Shanken, 1989, "A Test of the Efficiency of a Given Portfolio," *Econometrica,* 57, 1121–1152.

Huberman, G., and S. Kandel, 1987, "Mean Variance Spanning," *Journal of Finance,* 42, 873–888.

Iyengar, S., and J. Greenhouse, 1988, "Selection Models and the File Drawer Problem," *Statistical Science,* 3, 109–135.

Lakonishok, J., and S. Smidt, 1988, "Are Seasonal Anomalies Real? A Ninety-Year Perspective," *Review of Financial Studies,* 1, 403–426.

Leamer, E., 1978, *Specification Searches,* Wiley, New York.

Lehmann, B. N., and D. Modest, 1988, "The Empirical Foundations of the Arbitrage Pricing Theory," *Journal of Financial Economics,* 21, 213–254.

MacKinlay, A. C., 1987, "On Multivariate Tests of the CAPM," *Journal of Financial Economics,* 18, 341–372.

Merton, R., 1987, "On the Current State of the Stock Market Rationality Hypothesis," in R. Dornbusch, S. Fischer, and J. Bossons (eds.), *Macroeconomics and Finance: Essays in Honor of Franco Modigliani,* M.I.T. Press, Cambridge, Mass.

Nagaraja, H. N., 1982a, "Some Asymptotic Results for the Induced Selection Differential," *Journal of Applied Probability,* 19, 233–239.

Nagaraja, H. N., 1982b, "Some Nondegenerate Limit Laws for the Selection Differential," *Annals of Statistics,* 10, 1306–1310.

Nagaraja, H. N., 1984, "Some Nondegenerate Limit Laws for Sample Selection Differential and Selection Differential," *Sankhyā,* 46, Series A, 355–369.

Ross, S., 1987, "Regression to the Max," Working Paper, Yale School of Organization and Management.

Sandström, A., 1987, "Asymptotic Normality of Linear Functions of Concomitants of Order Statistics," *Metrika,* 34, 129–142.

Sen, P. K., 1976, "A Note on Invariance Principles for Induced Order Statistics," *Annals of Probability,* 4, 474–479.

Sen, P. K., 1981, "Some Invariance Principles for Mixed Rank Statistics and Induced Order Statistics and Some Applications," *Communications in Statistics,* A10, 1691–1718.

Shanken, J., 1985, "Multivariate Tests of the Zero-Beta CAPM," *Journal of Financial Economics,* 14, 327–348.

Stambaugh, R. F., 1982, "On the Exclusion of Assets from Tests of the Two Parameter Model," *Journal of Financial Economics,* 10, 235–268.

Stuart, A., and J. Ord, 1987, *Kendall's Advanced Theory of Statistics,* Oxford U.P., New York.

Wang, T., 1988, *Essays on the Theory of Arbitrage Pricing,* unpublished doctoral dissertation, Wharton School, University of Pennsylvania.

Watterson, G. A., 1959, "Linear Estimation in Censored Samples from Multivariate Normal Populations," *Annals of Mathematical Statistics,* 30, 814–824.

Yang, S. S., 1977, "General Distribution Theory of the Concomitants of Order Statistics," *Annals of Statistics,* 5, 996–1002.

Yang, S. S., 1981a, "Linear Functions of Concomitants of Order Statistics with Application to Nonparametric Estimation of a Regression Function," *Journal of the American Statistical Association,* 76, 658–662.

Yang, S. S., 1981b, "Linear Combinations of Concomitants of Order Statistics with Application to Testing and Estimation," *Annals of the Institute of Statistical Mathematics,* 33 (Part A), 463–470.