

NBER WORKING PAPER SERIES

DATA-SNOOPING BIASES IN TESTS OF FINANCIAL
ASSET PRICING MODELS

Andrew W. Lo

A. Craig MacKinlay

Working Paper No. 3001

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
June 1989

We thank David Aldous, Herbert David, Mike Gibbons and seminar participants at Columbia University, M.I.T., Princeton University, and Stanford University for useful comments and suggestions. Research support from the Geewax-Terker Research Fund (MacKinlay), the John M. Olin Fellowship at the National Bureau of Economic Research (Lo), and the National Science Foundation is gratefully acknowledged. This paper is part of NBER's research program in Financial Markets and Monetary Economics. Any opinions expressed are those of the authors not those of the National Bureau of Economic Research.

NBER Working Paper #3001
June 1989

DATA-SNOOPING BIASES IN TESTS OF FINANCIAL ASSET PRICING MODELS

ABSTRACT

We investigate the extent to which tests of financial asset pricing models may be biased by using properties of the data to construct the test statistics. Specifically, we focus on tests using 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. We present both analytical calculations and Monte Carlo simulations that show the effects of this type of data-snooping to be substantial. Even when the sorting characteristic is only marginally correlated with individual security statistics, 5 percent tests based on sorted portfolio returns may reject with probability one under the null hypothesis. This bias is shown to worsen as the number of securities increases given a fixed number of portfolios, and as the number of portfolios decreases given a fixed number of securities. We provide an empirical example that illustrates the practical relevance of these biases.

Andrew W. Lo
Sloan School of Management
M.I.T.
50 Memorial Drive
Cambridge, MA 02138

A. Craig MacKinlay
Department of Finance
Wharton School
University of Pennsylvania
Philadelphia, PA 19104

1. Introduction.

The reliance of economic science upon non-experimental inference is, at once, one of the most challenging and most nettlesome aspects of the discipline. Because of the virtual impossibility of controlled experimentation the importance and influence of statistical data analysis in economics 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. For example, the classical statistical approach to selecting a method of estimation generally involves minimizing an expected loss function irrespective of the actual data at hand. 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 aspects of the data into our estimation and testing strategies 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 pre-test biases discussed in Leamer (1978). A meta-corollary of this proposition is that the degree of such biases is an increasing function of 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 seem to 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.¹ 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

¹ See Black, Jensen, and Scholes (1972), and Fama and MacBeth (1973).

one example. Constructing classical statistical tests on portfolios formed this way creates potentially significant biases in the test statistics. These are 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 thought of as a vector of zeros and ones that indicate which securities are to be included or excluded 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 empirically plausible assumptions our calculations show that this kind of data-snooping can yield test sizes of unity for commonly used nominal 5 percent tests.

Although the term “data-snooping” has 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, one purpose of the public dissemination of scientific discovery is to add to a store of common knowledge on which future research may build.

However, when scientific discovery is statistical in nature we must take care to evaluate 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.² 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. 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.” But 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 re-phrase this question in the following way: Are standard tests of significance valid

² See, for example, Siegmund (1985).

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 2 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 3 we report the results of Monte Carlo experiments designed to gauge the accuracy of the previous section's asymptotic approximations. Section 4 provides an empirical example that illustrates the potential importance of data-snooping biases in existing tests of asset pricing models, and Section 5 shows how these biases can arise naturally from our tendency to focus on the unusual. We conclude in Section 6.

2. 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."³ 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.⁴ 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,⁵ particularly when the empirical motivation is derived from the very data set on which the test is based. 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. Section 2.1 provides a brief summary of the asymptotic properties of induced order statistics. Section 2.2 presents results for tests based on individual securities, and Section 2.3 reports corresponding results for portfolios. In Section 2.4 we discuss directions in which the assumptions of the previous sections may be relaxed, and in Section 2.5 we provide a more positive interpretation of data-snooping biases as power against deviations from the null hypothesis.

2.1. 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:

³ The following is only a partial list of the more recent studies using size-sorted portfolios: Chan and Chen (1988), Chan, Chen, and Hsieh (1985), Chen, Roll, and Ross (1986), Connor and Korajczyk (1988), Kandel and Stambaugh (1987), Lehmann and Modest (1988), MacKinlay (1987), and Shanken (1985).

⁴ See Bans (1982), Brown, Kleidon, and Marsh (1983), and Chan, Chen, and Hsieh (1985) for example.

⁵ Unfortunately the use of "size" to mean both market value of equity and type I error is unavoidable. Readers beware.

$$R_{it} - R_{ft} = \alpha_i + (R_{mt} - R_{ft})\beta_i + \epsilon_{it} \quad (2.1)$$

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 to 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.

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 $\{\hat{\alpha}_i\}$. Otherwise, extreme realizations of estimation error will be confused with a violation of the null hypothesis. If instead of choosing $\hat{\alpha}_i$ based on 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 are used to form the test statistic and these n are chosen by sorting the X_i 's. That is, let us re-order the bivariate vectors $[X_i \quad \hat{\alpha}_i]'$ according to their first components, yielding the sequence:

$$\begin{pmatrix} X_{1:N} \\ \hat{\alpha}_{[1:N]} \end{pmatrix}, \begin{pmatrix} X_{2:N} \\ \hat{\alpha}_{[2:N]} \end{pmatrix}, \dots, \begin{pmatrix} X_{N:N} \\ \hat{\alpha}_{[N:N]} \end{pmatrix} \quad (2.2)$$

where $X_{1:N} < X_{2:N} < \dots < X_{N:N}$ and the notation $X_{i:N}$ follows that of the statistics literature in denoting the i -th order statistic from the sample of N observations $\{X_i\}$.⁶ The notation $\hat{\alpha}_{[i:N]}$ denotes the i -th *induced order statistic* corresponding to $X_{i:N}$ or the i -th *concomitant* of the order statistic $X_{i:N}$.⁷ 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 i -th ordered vector. The $\hat{\alpha}_{[i:N]}$ are not themselves ordered but correspond to the ordering of the $X_{i:N}$'s.⁸ 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 (2.2) the sampling theory cannot be the same as that of n securities selected independently of the data. Due to the following remarkably simple result by Yang (1977) an asymptotic sampling theory for test statistics based on induced order statistics may be derived analytically:⁹

⁶ It is implicitly assumed throughout that both $\hat{\alpha}_i$ and X_i have continuous joint and marginal c.d.f.'s, hence strict inequalities suffice.

⁷ The term *concomitant* of order statistics was introduced by David (1973), who was perhaps the first to systematically investigate its properties and applications. The term *induced order statistics* 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.

⁸ 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 $\hat{\alpha}_i$. In fact, it may readily be shown that for any fixed k :

$$\text{plim}_{N \rightarrow \infty} (2\sigma_\alpha^2 \log N)^{-1/2} X_{N-k:N} = 1, \quad \text{plim}_{N \rightarrow \infty} (2\sigma_\alpha^2 \log N)^{-1/2} \hat{\alpha}_{[N-k:N]} = \rho$$

thus the largest X_i is almost never paired with the largest $\hat{\alpha}_i$. See David (1981), David, O'Connell, and Yang (1977) and Galambos (1987) for further details.

⁹ See also David and Galambos (1974) and Watterson (1959). Yang (1977) also provides the exact finite-sample distribution of any finite collection of induced order statistics. However, even assuming bivariate normality does not yield a tractable form for this distribution. For example, the exact joint c.d.f. of the i -th and j -th induced order statistics is given by:

$$\Pr(\hat{\alpha}_{[i:N]} < a_i, \hat{\alpha}_{[j:N]} < a_j) = E \left[\Phi \left(\frac{a_i - \alpha - \rho \frac{\sigma_\alpha}{\sigma_x} (X_{i:N} - \mu_x)}{\sigma_\alpha^2 (1 - \rho^2)} \right) \cdot \Phi \left(\frac{a_j - \alpha - \rho \frac{\sigma_\alpha}{\sigma_x} (X_{j:N} - \mu_x)}{\sigma_\alpha^2 (1 - \rho^2)} \right) \right]$$

where $\Phi(\cdot)$ is the standard normal c.d.f. and the expectation is taken with respect to the joint distribution of the order statistics $(X_{i:N}, X_{j:N})$.

Theorem 2.1. Let the vectors $[X_i \ \hat{\alpha}_i]'$, $i = 1, \dots, N$ be independently and identically distributed and let $1 < i_1 < i_2 < \dots < i_n < N$ be sequences of integers such that, as $N \rightarrow \infty$, $i_k/N \rightarrow \xi_k \in (0, 1)$ ($k = 1, 2, \dots, n$). Then

$$\lim_{N \rightarrow \infty} Pr(\hat{\alpha}_{[i_1:N]} < a_1, \dots, \hat{\alpha}_{[i_n:N]} < a_n) = \prod_{j=1}^n Pr(\hat{\alpha}_j < a_j | F_x(X_j) = \xi_j) \quad (2.3)$$

where $F_x(\cdot)$ is the marginal c.d.f. 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 (2.3) it is evident that $\{\hat{\alpha}_{[i:N]}\}$ are mutually independent in large samples. If X_i were the market value of equity, or “size,” of the i -th security Theorem 2.1 shows that the $\hat{\alpha}_i$ of the security with size at the 27-th percentile is asymptotically independent of the $\hat{\alpha}_j$ corresponding to the security with size at the 45-th percentile.¹⁰ If the characteristics $\{X_i\}$ and $\{\hat{\alpha}_i\}$ are statistically independent, clearly the joint distribution of the latter 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 2.2 and 2.3.

2.2. Biases of Tests Based on Individual Securities.

We evaluate the bias of induced ordering under the following assumption:¹¹

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

The null hypothesis H is then:

¹⁰ This is a limiting result and, in particular, implies that the identities of the stocks with 27-th and 45-th percentile sizes may change as N increases.

¹¹ Some of our results are robust to this parametric specification but more general distributional assumptions will not admit explicit numerical evaluation.

$$H: \quad \alpha = 0.$$

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.¹² Since the sampling theory provided by Theorem 2.1 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:

$$\theta \equiv \frac{1}{\hat{\sigma}_\alpha^2} \sum_{i=1}^n \hat{\alpha}_i^2 \stackrel{a}{\sim} \chi_n^2 \quad (2.4)$$

where $\hat{\sigma}_\alpha^2$ is any consistent estimator of σ_α^2 .¹³ 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:

$$F_{\chi_n^2}(C_{.05}^n) = .95 \quad (2.5)$$

and $F_{\chi_n^2}(\cdot)$ is the c.d.f. of a χ_n^2 variate.

Suppose we construct θ from the induced order statistics $\hat{\alpha}_{[i_k:N]}$, $k = 1, \dots, n$, instead of the $\hat{\alpha}_i$'s. Specifically, define the following test statistic:

$$\tilde{\theta} \equiv \frac{1}{\hat{\sigma}_\alpha^2} \sum_{k=1}^n \hat{\alpha}_{[i_k:N]}^2. \quad (2.6)$$

Using Theorem 2.1, the following proposition is easily established:

¹² See Chamberlain (1983), Huberman and Kandel (1987), and Lehmann and Modest (1988) for the exact factor pricing model. Examples of tests which fit into the framework of H are those in Campbell (1987), Connor and Korajczyk (1988), Gibbons (1982), Gibbons and Person (1985), Gibbons, Ross, and Shanken (1987), Huberman and Kandel (1987), MacKinlay (1987), Lehmann and Modest (1988), Stambaugh (1982), and Shanken (1985).

¹³ 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 σ_α^2 may be estimated cross-sectionally hence the relation ' $\tilde{\theta}$ ' in (2.4) need only represent N -asymptotics.

Proposition 2.1. Under the null hypothesis H and assumption (A), as N increases without bound the induced order statistics $\hat{\alpha}_{[i_k:N]}$ ($k = 1, \dots, n$) converge in distribution to independent Gaussian random variables with mean μ_k and variance σ_k^2 where:

$$\mu_k \equiv \rho \frac{\sigma_\alpha}{\sigma_x} \left[F_x^{-1}(\xi_k) - \mu_x \right] = \rho \sigma_\alpha \Phi^{-1}(\xi_k) \quad (2.7)$$

$$\sigma_k^2 \equiv \sigma_\alpha^2 (1 - \rho^2) \quad (2.8)$$

which implies:

$$\tilde{\theta} \stackrel{a}{\sim} (1 - \rho^2) \cdot \chi_n^2(\lambda) \quad (2.9)$$

with non-centrality parameter:

$$\lambda = \sum_{k=1}^n \left(\frac{\mu_k}{\sigma_k} \right)^2 = \frac{\rho^2}{1 - \rho^2} \sum_{k=1}^n \left[\Phi^{-1}(\xi_k) \right]^2. \quad (2.10)$$

where $\Phi(\cdot)$ is the standard normal c.d.f.¹⁴

Proposition 2.1 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 95-th percentile is 0.164.

To evaluate the size of a 5 percent test based on the statistic $\tilde{\theta}$ we need only evaluate the c.d.f. of the non-central $\chi_n^2(\lambda)$ at the point $C_{.05}^n / (1 - \rho^2)$, where $C_{.05}^n$ is given in (2.5). Observe that the non-centrality 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 (2.4). This is consistent with common intuition since sorting on a characteristic that is statistically independent of the $\hat{\alpha}_i$'s cannot disturb its null distribution. As $\hat{\alpha}_i$ and X_i

¹⁴The second equality in (2.7) follows from the fact that $\Phi(\xi_k) = F_x(\xi_k \sigma_x + \mu_x)$.

become more highly correlated the non-central 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 .¹⁵

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 consider three sets of $\{\xi_k\}$, yielding three distinct test statistics $\tilde{\theta}_1$, $\tilde{\theta}_2$, and $\tilde{\theta}_3$:

$$\tilde{\theta}_1 \quad \Leftrightarrow \quad \xi_k = \frac{k}{n+1} \quad (2.11)$$

$$\tilde{\theta}_2 \quad \Leftrightarrow \quad \xi_k = \begin{cases} \frac{k}{(m+1)(n_o+1)} & \text{for } k = 1, 2, \dots, n_o. \\ \frac{k + m(n_o+1) - n_o}{(m+1)(n_o+1)} & \text{for } k = n_o + 1, \dots, 2n_o. \end{cases} \quad (2.12)$$

$$\tilde{\theta}_3 \quad \Leftrightarrow \quad \xi_k = \begin{cases} \frac{k + n_o + 1}{(m+1)(n_o+1)} & \text{for } k = 1, 2, \dots, n_o. \\ \frac{k + (m-1)(n_o+1) - n_o}{(m+1)(n_o+1)} & \text{for } k = n_o + 1, \dots, 2n_o. \end{cases} \quad (2.13)$$

where $n \equiv 2n_o$ and n_o is an arbitrary positive integer. The first method (2.11) simply sets the ξ_k 's so that they divide the unit interval into n equally-spaced increments. The second procedure (2.12) 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 $1/(m+1)(n_o+1)$.

¹⁵In 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.

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 increases 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 $\hat{\alpha}_{[i:N]}$'s in the second lowest and second highest $100/(m+1)$ -percentiles.

Table 1 reports 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 first panel of Table 1 show that the size of a 5 percent test using $\tilde{\theta}_1$ is indeed 5 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.

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, when securities with extreme characteristics are used the size can deviate substantially. 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 highest two 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 (2.7): the more extreme induced order statistics have means farther away from zero, hence a statistic based on evenly-distributed $\hat{\alpha}_{[i:N]}$'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 $\hat{\alpha}_{[i:N]}$'s may appear to be statistically incompatible with the null.

2.3. 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 securities are chosen independently of the data. Under the null hypothesis H we have the following:

$$\phi_k \equiv \frac{1}{n_o} \sum_{j=(k-1)n_o+1}^{kn_o} \hat{\alpha}_j \sim N\left(0, \frac{\sigma_\alpha^2}{n_o}\right) \quad k = 1, 2, \dots, q \quad (2.14)$$

$$\theta_p \equiv \frac{n_o}{\hat{\sigma}_\alpha^2} \sum_{k=1}^q \phi_k^2 \stackrel{a}{\sim} \chi_q^2. \quad (2.15)$$

To perform a 5 percent test of H using θ_p we compare it with the critical value $C_{.05}^q$ defined by:

$$F_{\chi_q^2}(C_{.05}^q) = .95. \quad (2.16)$$

Suppose we compute this test statistic using the induced order statistics $\{\hat{\alpha}_{[i;N]}\}$ instead of $\{\hat{\alpha}_i\}$. From Theorem 2.1 we have:

Proposition 2.2. *As N increases without bound, under the null hypothesis H the statistics $\tilde{\phi}_k$ ($k = 1, 2, \dots, q$) and $\tilde{\theta}_p$ converge in distribution to the following:*

$$\tilde{\phi}_k \equiv \frac{1}{n_o} \sum_{j=(k-1)n_o+1}^{kn_o} \hat{\alpha}_{[i_j;N]} \sim N\left(\sum_{j=(k-1)n_o+1}^{kn_o} \frac{\mu_j}{n_o}, \frac{\sigma_\alpha^2(1-\rho^2)}{n_o}\right) \quad (2.17)$$

$$\tilde{\theta}_p \equiv \frac{n_o}{\hat{\sigma}_\alpha^2} \sum_{k=1}^q \tilde{\phi}_k^2 \stackrel{a}{\sim} (1-\rho^2) \cdot \chi_q^2(\lambda) \quad (2.18)$$

with non-centrality parameter:

$$\lambda = \frac{n_o \rho^2}{1 - \rho^2} \sum_{k=1}^q \left(\frac{1}{n_o} \sum_{j=(k-1)n_o+1}^{kn_o} [\Phi^{-1}(\xi_j)] \right)^2. \quad (2.19)$$

The non-centrality parameter (2.19) is similar to that of the statistic based on individual securities; it is increasing in ρ^2 and reduces to the central χ_q^2 of (2.19) 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 j in the k -th portfolio. To see the importance of this consider the case where the relative ranks ξ_j are chosen to be evenly-spaced in $(0,1)$, i.e.,

$$\xi_j = \frac{j}{n_o q + 1}. \quad (2.20)$$

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 2a 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 an R^2 as low as 1 percent, implying a correlation of only ± 10 percent between $\hat{\alpha}_i$ and X_i , a 5 percent test based on 50 portfolios each with 50 securities 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 2a 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 in each portfolio is held fixed.

To understand why forming portfolios yields much higher rejection rates than using individual securities, recall from (2.7) and (2.8) that the mean of $\hat{\alpha}_{[i:N]}$ is a function of its relative rank i/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 $\hat{\alpha}_{[i:N]}$'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 2a. A more dramatic illustration is provided in Table 2b, which reports the appropriate 5 percent critical values for the tests in Table 2a: when $R^2 = .05$ the 5 percent critical 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 of 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 $\hat{\alpha}_{[i:N]}$. Using 100 securities, Table 1 reports a size of 4.3 percent with evenly-spaced $\hat{\alpha}_{[i:N]}$'s; Table 2a shows that placing those 100 securities into 5 portfolios each with 20 securities increases the size to 56.8 percent. Computing $\tilde{\theta}_p$ with extreme $\hat{\alpha}_{[i:N]}$ would presumably yield even higher rejection rates. The biases reported in Tables 2a,b are even more surprising in view of the limited use we have made of the data. The only data-related information impounded into the induced order statistics is the ordering of the characteristics $\{X_i\}$. Nowhere have we exploited the values of the X_i 's, which contains considerably more precise information about the $\hat{\alpha}_i$'s.

2.4. Relaxing Assumption (A).

The sharp results of Sections 2.2 and 2.3 are based on the assumption that the vectors $[X_i \ \hat{\alpha}_i]'$ are bivariate normal and independently and identically distributed. These requirements may be relaxed to some extent. In particular, X_i and $\hat{\alpha}_i$ need not be Gaussian, but need only satisfy a linear regression equation such as the following to yield the asymptotic sampling theory derived in Propositions 2.1 and 2.2:¹⁶

$$\hat{\alpha}_i = \mu_\alpha + \beta(X_i - \mu_x) + Z_i \quad (2.21)$$

where $E[Z_i] = 0$ and X_i and Z_i are independent. Of course, Yang's (1977) theorem applies

¹⁶ See David (1973), David and Galambos (1974), and Watterson (1969).

to any bivariate distribution for X_i and $\hat{\alpha}_i$, and although departures from linearity will change the limiting distributions of our test statistics $\tilde{\theta}$ and $\tilde{\theta}_p$, a sampling theory may still be obtained in principle. In such cases, the magnitude of data-snooping biases need not be directly related to the squared correlation between X_i and $\hat{\alpha}_i$, as it is under assumption (A). To see this, suppose the $\hat{\alpha}_i$'s were i.i.d. standard normal random variables, and consider the characteristic $X_i \equiv \hat{\alpha}_i \cdot Z$ where Z is also a standard normal variate independent of all the $\hat{\alpha}_i$'s. Clearly X_i and $\hat{\alpha}_i$ are uncorrelated and the vectors $[X_i \ \hat{\alpha}_i]'$ are independently and identically distributed. But the induced order statistics $\hat{\alpha}_{[i:N]}$ will be perfectly ordered! Not surprisingly, the correlation coefficient need not play an important role when X_i and $\hat{\alpha}_i$ are not linearly related.

Unfortunately, the assumptions of independence and identical distributions are not as easily relaxed. When X_i and $\hat{\alpha}_i$ satisfy (2.21), the induced order statistics have the following representation:¹⁷

$$\hat{\alpha}_{[i:N]} = \mu_\alpha + \beta(X_{i:N} - \mu_x) + Z_{[i]} \quad (2.22)$$

where $Z_{[i]}$ denotes the particular Z_i paired with the order statistic $X_{i:N}$. To develop the properties of induced order statistics for dependent and non-identically distributed X_i 's and $\hat{\alpha}_i$'s then requires an understanding of how $X_{i:N}$ and $Z_{[i]}$ behave under these weaker conditions. Although David (1981, chapters 2.8 and 5.6) does present some results for the non-identically distributed and the dependent cases separately, combining and applying them to (2.22) is a formidable task which we leave to the more industrious.

2.5. 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 re-interpreted as describing the power of tests based on induced ordering against certain alternative hypotheses. Recall that $\hat{\alpha}_i$ is defined to be a consistent estimator of some parameter α_i , hence:

¹⁷ See David (1981, chapter 5.7) and David and Galambos (1974).

$$\hat{\alpha}_i = \alpha_i + \zeta_i \quad (2.23)$$

where ζ_i denotes estimation or, more generally, “measurement” error. 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 non-zero and vary across i , then sorting by some characteristic X_i related to α_i and forming portfolios does yield a more powerful test. Grouping by X_i increases dispersion of the α_i ’s and forming portfolios reduces the estimation error through diversification (or the law of large numbers). Therefore what were called biases in Sections 2.1–2.3 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 2a show that grouping on a marginally correlated characteristic can increase the power substantially.¹⁸

Of course, the essence of the data-snooping problem lies in our inability to distinguish between this situation and that considered in the preceding sections. 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 measurement 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 . We shall return to this issue in our empirical example.

3. Monte Carlo Results.

Although the values in Tables 1 and 2a,b plainly show the magnitude of the biases associated with induced ordering their practical relevance may be limited in three respects. First, although the test statistics considered are similar in spirit to those used in empirical tests of asset pricing models, the assumption of cross-sectional independence has been used in their construction. In practice the covariances of N asset returns are estimated

¹⁸ However, implicit in Table 2a 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 would tend to be positively cross-sectionally correlated as a result of the omitted factor (assuming that the included factor is the return on the market). This positive correlation reduces the benefits of grouping in much the same way that positive correlation among stock returns reduces the benefits of diversification in a portfolio. Grouping by induced ordering does tend to cluster $\hat{\alpha}_i$ ’s with similar (non-zero) 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 3.3.

with a finite number T of time series observations which, under certain assumptions, yields an F-distributed test statistic. Both sampling error from the covariance matrix estimator and cross-sectional dependence will affect the null distribution in finite samples. Second, the sampling theory of Section 2 is based on asymptotic approximations, and few results on rates of convergence for Theorem 2.1 are available.¹⁹ 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.²⁰

In this section we report results from Monte Carlo experiments that show the asymptotic approximations of Section 2 to be quite accurate in practice despite these three reservations. Section 3.1 evaluates the quality of the asymptotic approximations for the $\tilde{\theta}_p$ test used in calculating Tables 2a,b. Section 3.2 considers 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. Section 3.3 considers the effects of relaxing the independence assumption.

3.1. Simulation Results for $\tilde{\theta}_p$.

The $\chi_q^2(\lambda)$ limiting distribution of $\tilde{\theta}_p$ obtains because any finite collection of induced order statistics, each with fixed unique 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. However, in practice all securities within a particular characteristic fractile are included in the sorted

¹⁹ However, see Bhattacharya (1984) and Sen (1981).

²⁰ However, 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 when n/N is large. 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).

portfolios, hence the theoretical sizes of Table 2a 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 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 2a. Except when both n_o and q are small, the empirical sizes of Table 3 match their asymptotic counterparts in Table 2a closely. Consider, for example, the $R^2 = .05$ panel; with 5 portfolios each with 5 securities, the difference between 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 10.3 percent when $R^2 = .30$. However, for the more relevant values of R^2 the empirical and theoretical sizes of the $\tilde{\theta}_p$ test are virtually identical.

3.2. Effects of Induced Ordering on F-Tests.

Although the results of Section 3.1 support the accuracy of our asymptotic approximation to the sampling distribution of $\tilde{\theta}_p$, the closely related F-statistic is more frequently used in practice. In this section we consider the finite-sample distribution of this test 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 $i = 1, 2, \dots, N$ and $t = 1, 2, \dots, T$. We assume that for all assets i and dates t the following obtains:

$$r_{it} = \alpha_i + \sum_{j=1}^k \beta_{ij} r_t^j + \epsilon_{it} \quad (3.1)$$

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 (3.1) may be viewed as a factor model where the factors correspond to particular portfolios of traded

assets, often called the “spanning portfolios” of an “exact factor pricing model.”²¹ In matrix-form, we have:

$$r_t = \alpha + B r_t^p + \epsilon_t \quad , \quad E[\epsilon_t] = 0 \quad , \quad E[r_t^p] = \mu_p \quad (3.2)$$

$$E[\epsilon_s \epsilon_t'] = \begin{cases} \Sigma & \text{for } s = t. \\ 0 & \text{otherwise.} \end{cases} \quad (3.3)$$

$$E[(r_s^p - \mu_p)(r_t^p - \mu_p)'] = \begin{cases} \Omega & \text{for } s = t. \\ 0 & \text{otherwise.} \end{cases} \quad (3.4)$$

where r_t is the $N \times 1$ -vector of asset returns at 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 means 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.²² Such restrictions often reduce to the following null hypothesis:

$$H: \quad g(\alpha, B) = 0$$

where the function g is model-dependent.²³ 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:

$$\psi = \kappa \cdot \frac{\hat{\alpha}' \hat{\Sigma}^{-1} \hat{\alpha}}{1 + \hat{r}^p \hat{\Omega}^{-1} \hat{r}^p} \sim F_{N, T-k-N} \quad \kappa \equiv \frac{T-k-N}{N} \quad (3.5)$$

²¹ See, for example, Lehmann and Modest (1988).

²² For example, under the Sharpe-Lintner CAPM the vector of disturbances ϵ_t is temporally independent, $k = 1$, and the single factor is the return on the value-weighted market portfolio. Alternatively, under the exact factor-pricing version of the APT the k factors are the returns of spanning portfolios, portfolios for which there exists a linear combination that is a tangency portfolio. See, for example, Chamberlain (1983), Huberman and Kandel (1987), and Lehmann and Modest (1988).

²³ Examples of tests which fit into this framework are those in Campbell (1987), Connor and Korajczyk (1988), Gibbons (1982), Gibbons and Ferson (1985), Gibbons, Shanken, and Ross (1987), Huberman and Kandel (1987), MacKinlay (1987), Lehmann and Modest (1988), Stambaugh (1982), and Shanken (1985).

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 N of available securities is greater than the number of time series observations T less $k + 1$ the estimator $\hat{\Sigma}$ is singular and the test statistic (3.5) 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 2 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 $\hat{\alpha}_i = \sum_{t=1}^T r_{it}/T$. For simplicity we also assume that the covariance matrix Σ of ϵ_t is equal to the identity matrix I ; this assumption is relaxed in Section 3.3. 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 values .01, .05, .10, .20, and .30) and then sorting the $\hat{\alpha}_i$'s by this characteristic. To do this we define:

$$X_i \equiv \hat{\alpha}_i + \eta_i \quad , \quad \eta_i \text{ i.i.d. } N(0, \sigma_\eta^2) \quad \sigma_\eta^2 = \frac{1 - \rho^2}{T\rho^2} \quad (3.6)$$

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

$$\phi_k = \frac{1}{n_o} \sum_{i=(k-1)n_o+1}^{kn_o} \hat{\alpha}_{[i:N]} \quad , \quad N \equiv n_o q \quad (3.7)$$

from which we form the F-statistic:

$$\psi = \kappa \cdot \hat{\phi}' \hat{\Sigma}^{-1} \hat{\phi} \sim F_{q, T-q}, \quad \kappa \equiv \frac{T-q}{q} \quad (3.8)$$

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 5,000 times and the mean and standard deviation of the resulting distribution for the statistic ψ are reported in Table 4a, 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 1,000 securities grouped into 10 portfolios the empirical rejection rate of 36.7 percent deviates substantially from 5 percent both economically and statistically. The size is somewhat lower – 26.8 percent – when the 1,000 securities are grouped into 20 portfolios, matching the pattern in Table 2a. 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 is more than quintupled for 1,000 securities. When the squared correlation between $\hat{\alpha}_i$ and X_i increases to 10 percent, the size of the F-test is essentially one 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 over-emphasized.

3.3. F-Tests Without Cross-Sectional Independence.

The intuition for the substantial bias that induced ordering imparts on the size of portfolio-based F-tests is that the induced order statistics $\{\hat{\alpha}_{[i:N]}\}$ generally have non-zero means,²⁴ hence the average of these statistics within sorted portfolios have non-zero means and 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, 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 (3.2) are cross-sectionally independent implies that the portfolio residual variance approaches zero rather quickly (at rate $\frac{1}{n_o}$). But in many applications (such as the CAPM) cross-sectional independence is too strong an assumption. Firm size and industry membership are but two characteristics that might induce cross-sectional

²⁴ Only those $\hat{\alpha}_{[i:N]}$ for which $\frac{i}{N} \rightarrow \frac{1}{2}$ will have zero expectation under the null hypothesis H.

correlation of security residuals. In particular, when the residuals are (positively) cross-sectionally correlated the bias is likely to be smaller since the variance-reduction due to portfolio-formation is less sharp than in the cross-sectionally independent case.

To provide a measure of the importance of relaxing the independence assumption, in this section we simulate a data-generating process in which disturbances are cross-sectionally correlated. The design is identical to that of Section 3.2 except that the residual covariance matrix Σ is not diagonal. Instead, we set:

$$\Sigma = \delta\delta' + I \quad (3.9)$$

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 :

$$\epsilon_t = \delta\Lambda_t + \nu_t \quad (3.10)$$

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 (3.6) with:

$$\sigma_\eta^2 = \frac{(1 - \rho^2)\sigma^2(\alpha)}{\rho^2} \quad (3.11)$$

$$\sigma^2(\alpha) \equiv \frac{1}{NT} \sum_{i=1}^N (\delta_i^2 + 1) \quad (3.12)$$

where ρ^2 is fixed at .05.

With this design, the results of the simulation experiments may be compared to the second panel of Table 4a and are reported in Table 4b.²⁵ Despite the presence of the cross-sectional dependence the impact of induced ordering on the size of the F-test is 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 4b demonstrate that the effects of data-snooping may still be substantial even in the presence of cross-correlation.

4. An Empirical Example.

To illustrate the potential empirical relevance of data-snooping biases associated with induced ordering we draw an example from the recent literature. Lehmann and Modest (1988) present a multivariate test of a 15-factor APT model which rejects the zero-intercept null hypothesis using five portfolios formed by grouping securities ordered by market value of equity.²⁶ We select this example because under its null hypothesis the disturbances ϵ_t of (3.2) are cross-sectionally independent, an assumption we adopted in Sections 2 and 3.²⁷ Moreover, since 15 factors are included in the cross-sectional regressions, a diagonal covariance matrix for ϵ_t is a plausible assumption.

It is a well-known fact 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.²⁸ 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.²⁹

²⁵ The correspondence between the two tables is not exact because the dependency introduced in (3.9) induces cross-sectional heteroscedasticity in the $\hat{\alpha}_i$'s hence $\rho^2 = .05$ yields an R^2 of .05 only approximately.

²⁶ See Lehmann and Modest (1988, Table 1, last row). Connor and Korajczyk (1988) report similar findings.

²⁷ Although Ross's (1976, 1977) original formulation assumes the cross-sectional uncorrelatedness of the disturbances (which is equivalent to independence under multivariate normality), generalizations by Chamberlain (1983), Chamberlain and Rothschild (1983), and Wang (1988) allow for cross-sectional dependence. We conjecture that the asymptotic results of Section 2 also obtain under the more general framework, however the finite sample properties deduced in Section 3 will be sensitive to cross-sectional dependence.

²⁸ See, for example, Bans (1982) and Brown, Kleidon, and Marsh (1983).

²⁹ 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.

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.³⁰ As an alternative we estimate the cross-sectional R^2 of the estimated CAPM alpha, $\hat{\alpha}_i$, with log-size of market value X_i and 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 5-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 5. One cross-sectional regression of $\hat{\alpha}_i$ on log-size X_i is run for each 5-year time period using monthly NYSE-AMEX data from CRSP. We run regressions only for those stocks having complete return histories within the relevant 5-year period.

Table 5 contains the test statistics for a 15-factor APT framework using five size-sorted portfolios. The first four rows contain results for each of the four sub-periods and the last row contains aggregate test statistics. To apply the results of Sections 2 and 3 we transform Lehmann and Modest's (1988) F-statistics into (asymptotic) χ^2 variates.³¹ The total number of available securities portfolios ranges from a minimum of 1001 for the first 5-year sub-period to a maximum of 1359 for the second sub-period. For each test statistic in Table 5 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 5 show that the 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 sub-periods 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 sub-periods range from 0.272 to 1.000, yielding an aggregate p-value of 1.000! Therefore, whether or not induced

³⁰ 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 (1986).

³¹ 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.

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 5. To see how this may affect our conclusions recall from (2.23) in Section 2.5 that the cross-sectional correlation between $\hat{\alpha}_i$ and log-size can arise from two sources: the 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 5, 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 5. 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 sub-periods, 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 5 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. If our maintained assumption (A) holds, this will yield a perfectly valid test of the null hypothesis H since the X_i 's are then independent of the $\hat{\alpha}_i$'s and induced ordering by an independent characteristic cannot disturb the null distribution of the test statistics.

5. How the Data Gets Snooped.

Whether the probabilities of rejection in Table 2a 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 2 was the assertion that this correlation is “spurious,” the values of Table 2a represent probabilities of falsely rejecting the null hypothesis. We suggested in Section 2.5 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. 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 created.

More formally, suppose for each of N securities we have a collection of K distinct and mutually independent characteristics Y_{ik} , $k = 1, 2, \dots, K$, where Y_{ik} is the k -th characteristic of the i -th security. Let the null hypothesis obtain so that $\alpha_i = 0$ for all i , and assume that all characteristics be 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:

$$\hat{\rho}_k = \frac{\sum_{i=1}^N (Y_{ik} - \bar{Y}_k)(\hat{\alpha}_i - \bar{\hat{\alpha}})}{\sqrt{\sum_{i=1}^N (Y_{ik} - \bar{Y}_k)^2} \cdot \sqrt{\sum_{i=1}^N (\hat{\alpha}_i - \bar{\hat{\alpha}})^2}}, \quad k = 1, 2, \dots, K \quad (5.1)$$

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 given by the relation:

$$\hat{\rho}_{k^*}^2 = \text{Max}_{1 \leq k \leq K} \hat{\rho}_k^2. \quad (5.2)$$

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.³² 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 , $E[\hat{\rho}_k^2] = \frac{1}{N-1}$ under our normality assumption. In fact, since the $\hat{\rho}_k^2$'s are well-known to be independently and identically distributed Beta $\left(\frac{1}{2}, \frac{1}{2}(N-2)\right)$ variates, the distribution and density functions of $\hat{\rho}_{k^*}^2$, denoted by $F_*(v)$ and $f_*(v)$ respectively, may be readily derived as:³³

$$F_*(v) = [F_\beta(v)]^K, \quad v \in (0, 1) \quad (5.3)$$

$$f_*(v) = K [F_\beta(v)]^{K-1} f_\beta(v), \quad v \in (0, 1) \quad (5.4)$$

where F_β and f_β are the c.d.f. and p.d.f. 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 $\hat{\rho}_{k^*}^2$ is then given by:

$$\gamma \equiv E[\hat{\rho}_{k^*}^2] - E[\hat{\rho}_k^2] = \int_0^1 v f_*(v) dv - \frac{1}{N-1}. \quad (5.5)$$

For 25 securities and 50 characteristics, γ is 20.5 percent!³⁴ With 100 securities, γ is still

³²In 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 K -th order statistic $\hat{\rho}_{K:K}^2 = \hat{\rho}_{k^*}^2$. As in the scalar case, induced ordering does change the distribution of the vector concomitants.

³³That 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.

³⁴Note that γ is only an approximation to the squared population correlation:

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

5.4 percent and only declines to 1.1 percent for $N = 500$. Although there is in fact no statistical relation between 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 2a that as the sample size grows the bias increases if the number of portfolios is held fixed, hence a larger N and thus a smaller γ does not necessarily 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.

It can be argued that 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.³⁵ 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 pre-test bias from alternative hypotheses is perhaps the most compelling criticism of "measurement without theory."

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 (5.5) 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.

³⁵ The "Anomalies" section of the *Journal of Economic Perspectives* is perhaps the most obvious example of our deliberate search for the unusual in economics.

6. 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 purely through data analysis. 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 shown to be related to the behavior of asset returns. 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.³⁶ Sorting on size yields rejections whereas sorting on theoretically relevant characteristics such as own-variance and dividend yield does not. This suggests that data-motivated 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 (1988, 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, non-experimental 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.

³⁶ See Lehmann and Modest (1988). Chen, Roll, and Ross (1986, footnote 8) report similar negative findings for the APT with portfolios formed by sorting on CAPM betas, the standard deviation of returns in a market-model regression, and the level of the stock price.

References

- Aldous, D., 1989, *Probability Approximations via the Poisson Clumping Heuristic*, New York: Springer-Verlag.
- Banz, R. W., 1981, "The Relationship Between Return and Market Value of Common Stocks," *Journal of Financial Economics* 9, 3-18.
- Bhattacharya, P. K., 1974, "Convergence of Sample Paths of Normalized Sums of Induced Order Statistics," *Annals of Statistics* 2, 1034-1039.
- , 1984, "Induced Order Statistics: Theory and Applications," in P. R. Krishnaiah and P. K. Sen (eds.) *Handbook of Statistics 4: Nonparametric Methods*, Amsterdam: North-Holland.
- Black, F., Jensen, M., and M. Scholes, 1972, "The Capital Asset Pricing Model: Some Empirical Tests," in M. Jensen (ed.) *Studies in the Theory of Capital Markets*, New York: Praeger.
- Brown, P., Kleidon, A. 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.
- , and M. Rothschild, 1983, "Arbitrage, Factor Structure, and Mean-Variance Analysis on Large Asset Markets," *Econometrica* 51, 1281-1304.
- Chan, K. and N. Chen, 1988, "An Unconditional Asset-Pricing Test and the Role of Firm Size as an Instrumental Variable for Risk," *Journal of Finance* 43, 309-325.
- , and D. Hsieh, 1985, "An Exploratory Investigation of the Firm Size Effect," *Journal of Financial Economics* 14, 451-471.
- Chen, N., Roll, R. 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.
- , 1981, *Order Statistics*, Second Edition, New York: John Wiley and Sons.
- , and J. Galambos, 1974, "The Asymptotic Theory of Concomitants of Order Statistics," *Journal of Applied Probability* 11, 762-770.
- , O'Connell, M. J. and S. S. Yang, 1977, "Distribution and Expected Value of the Rank of a Concomitant of an Order Statistic," *Annals of Statistics* 5, 216-223.

- Fama, E. and J. MacBeth, 1973, "Risk, Return, and Equilibrium: Empirical Tests," *Journal of Political Economy* 71, 607-636.
- Galambos, J., 1987, *The Asymptotic Theory of Extreme Order Statistics*, Second Edition, Malabar: Krieger Publishing Company.
- Gibbons, M. R., 1982, "Multivariate Tests of Financial Models: A New Approach," *Journal of Financial Economics* 10, 3-27.
- , and W. Ferson, 1985, "Testing Asset Pricing Models With Changing Expectations and an Unobservable Market Portfolio," *Journal of Financial Economics* 14, 217-236.
- , Ross, S. A. and J. Shanken, 1987, "A Test of the Efficiency of a Given Portfolio," to appear in *Econometrica*
- Huberman, G. and S. Kandel, 1987, "Mean Variance Spanning," *Journal of Finance* 42, 873-888.
- Kandel, S. and R. Stambaugh, 1987, "On Correlations and Inferences About Mean-Variance Efficiency," *Journal of Financial Economics* 18, 61-90.
- Leamer, E., 1978, *Specification Searches*, New York: John Wiley and Sons.
- 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*, Cambridge: M.I.T. Press.
- Nagaraja, H. N., 1982a, "Some Asymptotic Results for the Induced Selection Differential," *Journal of Applied Probability* 19, 233-239.
- , 1982b, "Some Nondegenerate Limit Laws for the Selection Differential," *Annals of Statistics* 10, 1306-1310.
- , 1984, "Some Nondegenerate Limit Laws for Sample Selection Differential And Selection Differential," *Sankhyā* 46, Series A, 355-369.
- 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.
- , 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.
- Siegmund, D., 1985, *Sequential Analysis*, New York: Springer-Verlag.

- 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*, New York: Oxford University Press.
- 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.
- , 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.
- , 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.

Table 1

Theoretical sizes of nominal 5 percent χ^2_n -tests of $H: \alpha_i = 0 (i = 1, \dots, n)$ using the test statistics $\bar{\theta}_j$, where $\bar{\theta}_j \equiv \sum_{k=1}^n \bar{\alpha}_{[i_k(j):N]}^2 / \sigma_{\alpha}^2$, $j = 1, 2, 3$ for various sample sizes n . The statistic $\bar{\theta}_1$ is based on induced order statistics with relative ranks evenly spaced in $(0,1)$; $\bar{\theta}_2$ is constructed from induced order statistics ranked in the lowest and highest $100/(m+1)$ -percent fractiles; and $\bar{\theta}_3$ 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 $\bar{\alpha}_i$ and the sorting characteristic.

n	$\bar{\theta}_1$ -Test	$\bar{\theta}_2$ -Test ($m = 4$)	$\bar{\theta}_3$ -Test ($m = 4$)	$\bar{\theta}_2$ -Test ($m = 9$)	$\bar{\theta}_3$ -Test ($m = 9$)	$\bar{\theta}_2$ -Test ($m = 19$)	$\bar{\theta}_3$ -Test ($m = 19$)
$R^2 = .01$							
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
$R^2 = .30$							
10	0.021	0.134	0.010	0.307	0.050	0.520	0.163
20	0.023	0.212	0.005	0.511	0.053	0.786	0.242
50	0.026	0.428	0.001	0.863	0.060	0.989	0.447
100	0.029	0.696	0.000	0.989	0.069	1.000	0.692

Table 2a

Theoretical sizes of nominal 5 percent χ^2_q -tests of $H: \alpha_i = 0 (i = 1, \dots, n)$ using the test statistic $\hat{\theta}_p$, where $\hat{\theta}_p \equiv q \sum_{k=1}^q \hat{\phi}_k^2 / \sigma_k^2$, and $\hat{\phi}_k \equiv \frac{1}{n_o} \sum_{j=(k-1)q+1}^{kq} \hat{\alpha}_{[i_j; N]}$ 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^2_q to $(1 - R^2) \cdot \chi^2_q(\lambda)$ where the non-centrality 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.

q	$n_o = 5$	$n_o = 10$	$n_o = 20$	$n_o = 25$	$n_o = 50$
$R^2 = .01$					
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	0.928	1.000	1.000	1.000	1.000
$R^2 = .30$					
5	0.285	0.747	0.993	0.999	1.000
10	0.547	0.970	1.000	1.000	1.000
20	0.848	1.000	1.000	1.000	1.000
25	0.918	1.000	1.000	1.000	1.000
50	0.997	1.000	1.000	1.000	1.000

Table 2b

Critical values $C_{.05}$ for 5 percent χ^2 -tests of $H: \alpha_i = 0$ ($i = 1, \dots, n$) using the test statistic $\hat{\theta}_p$, where $\hat{\theta}_p \equiv q \sum_{k=1}^q \hat{\phi}_k^2 / \sigma_{\alpha}^2$, and $\hat{\phi}_k \equiv \frac{1}{n_o} \sum_{j=(k-1)q+1}^{kq} \hat{\alpha}_{[i_j; N]}$ 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 non-centrality 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 $Pr(\hat{\theta}_p > C_{.05}) = 1 - F_{\chi_q^2(\lambda)}\left(\frac{C_{.05}}{1-R^2}\right) = .05$. For comparison, we also report the 5 percent critical value of for the central χ_q^2 distribution in the second column.

q	$C_{.05}-\chi_q^2$	$C_{.05}-\chi_q^2(\lambda)$ ($n_o = 5$)	$C_{.05}-\chi_q^2(\lambda)$ ($n_o = 10$)	$C_{.05}-\chi_q^2(\lambda)$ ($n_o = 20$)	$C_{.05}-\chi_q^2(\lambda)$ ($n_o = 25$)	$C_{.05}-\chi_q^2(\lambda)$ ($n_o = 50$)
$R^2 = .01$						
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
$R^2 = .30$						
5	11.07	17.38	26.77	44.17	52.52	92.59
10	18.31	31.84	50.11	84.66	101.41	182.73
20	31.41	58.63	93.72	160.99	193.88	354.84
25	37.65	71.55	114.85	198.18	239.01	439.25
50	67.50	134.10	217.65	380.06	460.06	854.55

Table 3

Empirical sizes of nominal 5 percent χ^2_q -tests of $H: \alpha_i = 0$ ($i = 1, \dots, n$) using the test statistic $\tilde{\delta}_p$, where $\tilde{\delta}_p \equiv q \sum_{k=1}^q \tilde{\phi}_k^2 / \sigma_{\alpha}^2$, and $\tilde{\phi}_k \equiv \frac{1}{n_o} \sum_{j=(k-1)q+1}^{kq} \hat{\alpha}_{[i_j; N]}$ 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 $\tilde{\delta}_p$ from χ^2_q to $(1 - R^2) \cdot \chi^2_q(\lambda)$ where the non-centrality 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 5,000 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.

q	$n_o = 5$	$n_o = 10$	$n_o = 20$	$n_o = 25$	$n_o = 50$
$R^2 = .01$					
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
$R^2 = .30$					
5	0.388	0.790	0.992	1.000	1.000
10	0.620	0.971	1.000	1.000	1.000
20	0.868	1.000	1.000	1.000	1.000
25	0.922	1.000	1.000	1.000	1.000
50	0.997	1.000	1.000	1.000	1.000

Table 4a

Empirical size of $F_{q,T-q}$ tests based on q portfolios sorted by a random characteristic whose squared correlation with δ_i is R^2 . n_0 is the number of securities in each portfolio and $n \equiv n_0 q$ 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 5,000 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.

q	n_0	n	Mean	Std. Dev.	Size-10%	Size-5%	Size-1%
$R^2 = .01$							
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	0.406	0.268	0.098
$R^2 = .05$							
10	20	200	1.980	0.883	0.549	0.342	0.189
20	10	200	1.506	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
$R^2 = .30$							
10	20	200	6.710	2.119	1.000	0.999	0.997
20	10	200	3.841	1.204	0.996	0.987	0.928
10	50	500	15.619	3.996	1.000	1.000	1.000
20	25	500	8.535	2.366	1.000	1.000	1.000
10	100	1000	30.671	7.156	1.000	1.000	1.000
20	50	1000	16.183	4.112	1.000	1.000	1.000

Table 4b

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. 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 5,000 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.

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	0.566	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

Table 5

Comparison of p-values for Lehmann and Modest's (1988) tests of the APT with and without correcting for the effects of induced ordering. 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 labelled ' $\chi^2(\lambda_i)$ p-value' ($i = 1, 2, 3$) [using the non-central χ^2 distribution], where λ_1 , λ_2 , and λ_3 are non-centrality parameters computed with \hat{R}^2 , $\frac{1}{2}\hat{R}^2$, and $\frac{1}{4}\hat{R}^2$ respectively. In all cases, 5 portfolios are formed from the total number of securities; this yields 5 degrees of freedom for the χ^2 statistics in the first four rows, and 20 degrees of freedom for the aggregate χ^2 statistics.

Sample	N	\hat{R}^2	$\hat{\sigma}_p$	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