THE JOURNAL OF FINANCE • VOL. XXXVI, NO. 4 • SEPTEMBER 1981

Performance Hypothesis Testing with the Sharpe and Treynor Measures

J. D. JOBSON and BOB M. KORKIE

ABSTRACT

Asymptotic distributions of the estimators of the Sharpe and Treynor performance measures are derived. Multivariate performance comparison measures, based on the traditional Sharpe and Treynor measures, are developed with their corresponding asymptotic distributions. The behavior of these new performance measures are evaluated in small samples. For single comparisons, a Sharpe z statistic is well behaved and for multiple comparisons a Sharpe chi-square statistic is reasonably well behaved. The powers of the tests are quite sensitive to the population coefficients of variation. Multivariate test statistics based upon the Treynor measure were not very satisfactory.

THIS STUDY DEVELOPS SIGNIFICANCE tests of portfolio performance utilizing the Sharpe and Treynor measures. The approximate bias and asymptotic distributions of the estimators of the traditional Sharpe and Treynor performance measures are derived. Multivariate performance comparison measures based on these traditional measures are proposed for comparing the performance of n portfolios. The approximate bias and asymptotic distributions of the multivariate comparison measures are then derived and test statistics proposed. The behavior and usefulness of the statistics for hypothesis testing are evaluated in small samples with simulation experiments.

We find that for single comparisons a z statistic based on Sharpe's measure is well behaved at small sample sizes although its power in detecting typical differences with monthly data is small. The z statistic, based on the Treynor measure, is not well behaved in small samples and also lacks power. For multiple comparisons a chi-square statistic, obtained from the Sharpe measures, is reasonably well behaved at small samples and its power increases as the number of portfolios increases and/or the coefficients of variation decrease. A chi-square statistic derived from the Treynor measure is not so well behaved.

Section I of the paper discusses the availability of test statistics for various classes of performance measures. It also discusses the use of the Sharpe and Treynor statistics in light of certain undesirable properties they possess and the section concludes with a brief literature survey. Section II derives the performance measure moments, asymptotic distributions, and proposed test statistics. Section III contains the results and conclusions of the univariate and multivariate simulations.

I. Perspective on Performance Measurement and Significance Tests

In this section the classes of two parameter performance measures are reviewed, the measures permitting significance tests of performance are identified, and some empirical problems with the performance measures are discussed.

We thank Steve Beveridge and Richard Roll for their helpful comments.

A. Classification of Performance Measures

Performance measures are used to compare a portfolio's performance in some time period relative to another period or to compare different portfolios in the same period. There are three general classes of two parameter performance measures dependent on their inherent usage and definition of risk. The first class includes performance measures based on total (standard deviation) risk of return. The second class is comprised of measures based on systematic (beta or covariance) risk of return. The third class does not require a risk pricing model. In the first category is the Sharpe [29] index and its variations. The second category of measures can be partitioned into measures which are prediction error based and those which are not. The common characteristic of the prediction error measures is the requirement of an ex ante expected return generating model that is usually estimated with data prior to the test period.[1] Prediction errors are then computed as the difference between the ex ante expected returns and the actual returns in the test period. These prediction errors are then aggregated into a performance measure. Some members of this class are the additive index of Fama, Fisher, Jensen, and Roll [8], the multiplicative index of Pettit [22], and the Ball and Brown [2] multiplicative-additive index. The members of the nonprediction error group are the Treynor [32] and Jensen [13, 14] measures including the extensions by Black [4] and Langetieg [16]. Finally, in the third category is the Cornell [7] procedure which computes the sample mean return prior to the test period and computes the sample mean's prediction errors in the test period. By assuming normality of returns, t tests may be used.

B. Significance Tests and Measurement Problems

Of these popular performance measures only the Cornell, Jensen, and additive prediction error indices permit statistical hypotheses tests of performance, although the latter two measures are sensitive to measurement error in the market index. The simple additive versions of the prediction error models have well defined sampling distributions because they are consecutive sums of independent, identically-distributed random variables. The predominant problem with the more complex prediction error models is their unknown sampling distributions.[2] In addition, all the prediction error indices require a substantial amount of data outside the performance measurement period and require the assumption inter alia, that this external period represents the norm. If the portfolio performs abnormally over both periods, the prediction error models will not reject the normal performance hypothesis. Thus, the use of prediction error indices may be

[1] The prediction errors may come from such return models as the capital asset pricing model, the market model, or the Fama-MacBeth [10] model including the Jaffe [12], Mandelker [17] approach. Brown and Warner [6] have a summary and description of some classes of prediction error models. The Cornell procedure is also a prediction error method, so our classifications are not mutually exclusive.

[2] For some performance measures, such as the Ball and Brown and the Pettit index, higher moments have been defined but not the statistical test. See Beveridge [3]. See also Roll [25] and Jobson and Korkie [15] for some suggestions for testing a portfolio's mean-variance efficiency.

limited. The Jensen index suffers from the "leverage bias" problem discussed, for example, in Modigliani and Pogue [20], which may limit its usefulness.

An additional problem with the Jensen, Treynor, and prediction error models is the requirement of a market portfolio proxy. The measurement error attendant in the proxy could cause major decision errors. Roll [26] concludes that the value of the Jensen and Treynor indices is not to test whether the dependent variable portfolio performs abnormally, but rather to test whether the independent variable market proxy is ex ante efficient. Thus, if the performance of a portfolio is to be measured, it should be treated as the market proxy and tested for mean-variance efficiency. On the other hand, Mayers and Rice [18] argue that the security market line can be used in performance evaluation, providing the chosen market portfolio reflects the uninformed segment of the market's assessment of efficiency. Roll [24] demonstrates that performance evaluation can also be accomplished with the market model prediction errors as well as the Sharpe index. Thus, the measurement problems associated with the market index may preclude the use of the security market line for performance evaluation, whether or not a statistical test exists.

Another problem occurs with the Sharpe and Treynor performance measures when the market risk premium is negative and the portfolio evaluated has a larger risk, a lower mean return, and a larger performance value than the market portfolio.[3] This is a problem that results from treating sample statistics as parameters. Ex ante, market premiums cannot be negative, but due to sampling variation (risk) one should occasionally expect this outcome. Significance tests will allow an otherwise impossible comparison of the portfolios for this perverse sample outcome.

In summary, all of the performance measures have shortcomings. The Sharpe measure appears to have a relatively small number of theoretical objections, but has no accompanying significance test. Some research has proceeded toward the goal of multivariate tests of the equality of portfolios' performance. Miller and Gehr [19] have identified the exact percentage bias in the Sharpe performance measure. Saniga, Gressis, and Hayya [27] have examined the effect of sample size and correlation on the probability of observing that one portfolio's mean-variance statistics dominate a second portfolio's statistics. However, no higher order moments or sampling distributions or significance tests are available for either the Sharpe or Treynor statistics. The following section derives some of the relevant moments and distributions.

II. Performance Measure Moments, Distributions, and Test Statistics

Consider the general situation in which the relative performance of a finite number of portfolios is to be evaluated. Let r_{it} represent the return premium (above the risk free rate) from the ith portfolio in period t, $i = 1, 2, \cdots, n$.[4] A

[3] See for example Tinic and West [31].

[4] Throughout the paper, returns and their moments are understood to be obtained from the difference between a portfolio's return and the risk free rate.

random sample of T return premiums on the n portfolios is denoted by

$$\mathbf{r}_t' = [r_{1t}, r_{2t}, \cdots, r_{nt}], \qquad t = 1, 2, \cdots, T$$

where $\mathbf{r}_t$ is assumed multivariate normal with mean vector $\boldsymbol{\mu} = \{\mu_i\}_{nx1}$, $i = 1, 2, \cdots, n$ and covariance matrix $\boldsymbol{\Sigma} = \{\sigma_{ij}\}_{nxn}$, $i, j = 1, 2, \cdots, n$. The unbiased estimators of the mean vector and covariance matrix are

$$\bar{\mathbf{r}} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{r}_t$$

and

$$\mathbf{S} = \{s_{ij}\}_{n\bar{x}n} \frac{1}{T-1}\sum_{t=1}^{T} (\mathbf{r}_t - \bar{\mathbf{r}})(\mathbf{r}_t - \bar{\mathbf{r}})'$$

respectively. These sample estimators are usually employed to form the estimators of the traditional Sharpe and Treynor performance measures.

A. *Traditional Sharpe and Treynor Measures*

The performance measures of Sharpe and Treynor for a portfolio i are defined by

$$Sh_i = \frac{\mu_i}{\sigma_i}$$

and

$$Tr_i = \frac{\mu_i \sigma_m^2}{\sigma_{im}}, \qquad i = 1, 2, \cdots, n$$

where the subscript m denotes the market proxy portfolio. Conventional sample estimators of these performance measures are

$$\widehat{Sh}_i = \frac{\bar{r}_i}{s_i}$$

and

$$\widehat{Tr}_i = \frac{\bar{r}_i s_m^2}{s_{im}}, \qquad i = 1, 2, \cdots, n$$

The expectations of these ratio estimators can be approximated by employing Taylor series expansions to order $\frac{1}{T^2}\left(\text{i.e. } O\left(\frac{1}{T^2}\right)\right)$ using the procedure outlined in the Appendix. That is, the expected values are

$$E(\widehat{Sh}_i) \approx \left(1 + \frac{3}{4T} + \frac{100}{128T^2}\right)\frac{\mu_i}{\sigma_i} \tag{1}$$

and

$$E(\widehat{Tr}_i) \approx \left[1 + \frac{1}{T}\left(\frac{1}{\rho_{im}^2} - 1\right) + \frac{1}{T^2}\left(1 - \frac{4}{\rho_{im}^2} + \frac{3}{\rho_{im}^4}\right)\right]\frac{\mu_i \sigma_m^2}{\sigma_{im}} \tag{2}$$

where ρ_{im} is the true correlation between the market proxy and portfolio premium returns.

Miller and Gehr have shown that the exact mean of the Sharpe estimator is

$$E(\widehat{Sh}_i) = \left[\frac{\Gamma\left(\frac{T-2}{2}\right)}{\Gamma\left(\frac{T-1}{2}\right)} \left(\frac{T-1}{2}\right)^{1/2} \right] \frac{\mu_i}{\sigma_i}$$

where Γ is the gamma function. A simpler approximation to the Sharpe mean is obtained by omitting the $O\left(\frac{1}{T^2}\right)$ term in (1). Comparison calculations show this simple approximation to be accurate to three decimals even at sample size 12. Thus, an approximately unbiased estimator of the Sharpe performance is given by

$$\widehat{Sh}_i^* = \frac{\bar{r}_i}{s_i} \frac{1}{\left(1 + \frac{.75}{T}\right)}$$

The advantages of this bias correction, in comparison to Miller and Gehr, are its simplicity as well as clearly revealing the effect of sample size on the estimator without a large sacrifice in accuracy.

The percentage bias in the Treynor measure is from (2) a function of both the sample size and the correlation which usually requires estimation. As the portfolio's correlation with the market tends to zero, the Treynor measure itself tends to infinity. For a finite sample size T, the Taylor series approximation of the expected value and bias diverges for small values of ρ_{im}. (See the Appendix.) The quality of this Treynor bias approximation, derived from (2), is determined in the simulation experiments of Section III.

Progress to this point has been the identification of the respective means and biases of the Sharpe and Treynor estimators. The asymptotic distributions of the Sharpe and Treynor performance measures are obtained by observing that they are estimators which are functions of the elements of $\bar{\mathbf{r}}$ and $\mathbf{S}$, where returns are assumed multivariate normal. (See the Appendix.) Under these conditions the asymptotic distributions may be derived as

$$\left(\frac{\bar{r}_i}{s_i}\right) \sim N\left(\frac{\mu_i}{\sigma_i}, \frac{1}{T}\left\{1 + \frac{\mu_i^2}{2\sigma_i^2}\right\}\right)$$

and

$$\left(\frac{s_m^2 \bar{r}_i}{s_{im}}\right) \sim N\left(\frac{\sigma_m^2 \mu_i}{\sigma_{im}}, \frac{1}{T}\left\{\sigma_i^2 + \mu_i^2\left(1 - \frac{1}{\rho_{im}^2}\right)\right\} \frac{\sigma_m^4}{\sigma_{im}^2}\right)$$

where the variances are equivalent to the $O\left(\frac{1}{T}\right)$ Taylor series approximations, to the true variances of the sampling distributions.[5]

[5] The conventional notation for an asymptotic distribution subtracts the mean from the statistic and multiplies by $\sqrt{T}$. For the sample mean example, the central limit theorem states that $\sqrt{T}(\bar{x} - \mu)$ converges in distribution to a normal distribution with mean 0 and variance σ^2. We choose to write the asymptotic distributions in the form $\bar{x} \sim N\left(\mu, \frac{\sigma^2}{T}\right)$ omitting the subtraction and multiplication convention.

Statistical procedures are developed next which compare two or more portfolios based on the equality of their Sharpe or Treynor measures. The procedure for two portfolios is discussed first.

B. Performance Comparison Measures for Two Portfolios

For two portfolios i and n we wish to test the hypotheses

$$H_{OS}: Sh_i - Sh_n = 0$$

and

$$H_{OT}: Tr_i - Tr_n = 0$$

The obvious choices for the test statistics are the sample differences $(\widehat{Sh}_i - \widehat{Sh}_n)$ and $(\widehat{Tr}_i - \widehat{Tr}_n)$. However, the statistical properties of the Treynor difference are poor. The presence of the estimated covariance term in the denominator causes large unpredictable fluctuations in the statistic. Substantial improvement in statistical properties was obtained by using the transformed difference,[6]

$$\widehat{Tr}_{in} = \frac{s_{nm}\bar{r}_i}{s_m^2} - \frac{s_{im}\bar{r}_n}{s_m^2} \tag{3}$$

The transformed difference for the Sharpe measure

$$\widehat{Sh}_{in} = s_n\bar{r}_i - s_i\bar{r}_n \tag{4}$$

provided a marginal improvement over the regular difference in small samples. For the remainder of the study, the transformed differences are employed in both cases.

For statistical testing, the moments and sampling distribution of the new comparison measures are required. Taylor series approximations, to $O\left(\frac{1}{T^2}\right)$, of the expectations of the transformed differences are

$$E(\widehat{Sh}_{in}) \approx (\sigma_n\mu_i - \sigma_i\mu_n)\left(1 - \frac{1}{4T} + \frac{1}{32T^2}\right) \tag{5}$$

and

$$E(\widehat{Tr}_{in}) = \left(\frac{\sigma_{nm}}{\sigma_m^2}\mu_i - \frac{\sigma_{im}}{\sigma_m^2}\mu_n\right) \tag{6}$$

Thus, the transformed Treynor difference is unbiased and the bias to $O\left(\frac{1}{T}\right)$ of the transformed Sharpe difference is $-\frac{\mathrm{Sh}_{in}}{4\mathrm{T}}$.

[6] Other transformations, such as $\bar{r}_1^2 s_2^2 - \bar{r}_2 s_1^2$, $\log\left(\frac{r_1/s_1}{\bar{r}_2/s_2}\right)$, and $\bar{r}_1 s_{2m} - \bar{r}_2 s_{1m}$ were tried in the study with less success than those presented here.

The asymptotic distributions of the transformed difference statistics are available from the procedure outlined in the Appendix. For the Sharpe statistic, the asymptotic distribution is normal with mean Sh_{in} and variance given by

$$\theta = \frac{1}{T}\left[2\sigma_i^2\sigma_n^2 - 2\sigma_i\sigma_n\sigma_{in} + \frac{1}{2}\mu_i^2\sigma_n^2 + \frac{1}{2}\mu_n^2\sigma_i^2 - \frac{\mu_i\mu_n}{2\sigma_i\sigma_n}(\sigma_{in}^2 + \sigma_i^2\sigma_n^2)\right] \tag{7}$$

For the Treynor measure, the asymptotic distribution is normal with mean Tr_{in} and variance

$$\psi = \frac{1}{T}[\sigma_i^2\sigma_{nm}^2 + \sigma_n^2\sigma_{im}^2 - 2\sigma_{im}\sigma_{nm}\sigma_{in} + \mu_i^2(\sigma_n^2\sigma_m^2 - \sigma_{nm}^2)$$

$$+ \mu_n^2(\sigma_i^2\sigma_m^2 - \sigma_{im}^2) - 2\mu_i\mu_n(\sigma_{in}\sigma_m^2 - \sigma_{im}\sigma_{nm})] \tag{8}$$

In practice, these variances are unknown because they are functions of the population means and covariances. Estimators denoted by $\hat{\Theta}$ and $\hat{\psi}$ are obtained by substituting sample estimators of the means and covariances in Equations (7) and (8). The behavior of these estimators is examined in Section III.

To test the single comparison hypotheses H_{OS}:$Sh_{in} = 0$ or H_{OT}:$Tr_{in} = 0$, the asymptotic normality of $\widehat{Sh}$ and $\widehat{Tr}$ suggest employing the test statistics

$$z_{\text{Sin}} = \frac{\widehat{Sh}_{in}}{\sqrt{\hat{\theta}}} \quad \text{or} \quad z_{Tin} = \frac{\widehat{Tr}_{in}}{\sqrt{\hat{\Psi}}} \tag{9}$$

The distributions of these approximate z statistics and the powers of the tests are also studied in the simulation experiment in Section III.

C. Performance Comparison Measures for n Portfolios

For n portfolios, the hypotheses of interest are

$$H_{OS}:Sh_1 = Sh_2 = \cdots = Sh_n$$

and

$$H_{OT}:Tr_1 = Tr_2 = \cdots = Tr_n$$

which are equivalent to

$$H_{OS}:\mathbf{Sh} = \mathbf{0}$$

and

$$H_{OT}:\mathbf{Tr} = \mathbf{0}$$

where **Sh** and **Tr** are the $(n-1)\times 1$ vectors containing the transformed differences Sh_{in} and Tr_{in}, $i = 1, 2, \cdots, n-1$ respectively.

The sampling distributions of the estimators $\widehat{\mathbf{Sh}}$ and $\widehat{\mathbf{Tr}}$ are again derived from the procedure outlined in the Appendix. For the Sharpe measure $\widehat{\mathbf{Sh}} \sim N(\mathbf{Sh}, \boldsymbol{\Theta})$, where an element of the mean vector is

$$Sh_{in} = (\sigma_n\mu_i - \sigma_i\mu_n), \qquad i = 1, 2, \cdots, n-1$$

and an element of the covariance matrix is

$$\theta_{ij} = \frac{1}{T}\left[\sigma_n^2\sigma_i\sigma_j - \sigma_{jn}\sigma_n\sigma_i - \sigma_{in}\sigma_n\sigma_j + \sigma_n^2\sigma_{ij} + \frac{1}{2}\mu_i\mu_j\sigma_n^2 - \frac{\mu_n\mu_j}{4\sigma_n\sigma_i}(\sigma_{in}^2 + \sigma_i^2\sigma_n^2)\right.$$
$$\left. - \frac{\mu_n\mu_i}{4\sigma_n\sigma_j}(\sigma_{jn}^2 + \sigma_j^2\sigma_n^2) + \frac{\mu_n^2}{4\sigma_i\sigma_j}(\sigma_{ij}^2 + \sigma_i^2\sigma_j^2)\right] \qquad i, j = 1, 2, \cdots, n-1 \quad (10)$$

For the Treynor measure $\widehat{\mathbf{Tr}} \sim N(\mathbf{Tr}, \boldsymbol{\Psi})$, where an element of the mean vector is

$$Tr_{in} = \left(\frac{\sigma_{nm}}{\sigma_m^2}\mu_i - \frac{\sigma_{im}}{\sigma_m^2}\mu_n\right), \qquad i = 1, 2, \cdots, n-1$$

and an element of the covariance matrix is

$$\psi_{ij} = \frac{1}{\sigma_m^4 T}[\sigma_{nm}^2\sigma_{ij} - \sigma_{im}\sigma_{nm}\sigma_{jn} - \sigma_{nm}\sigma_{in}\sigma_{jm} + \sigma_{im}\sigma_{jm}\sigma_n^2 + \mu_n^2\sigma_{ij}\sigma_m^2$$
$$- \mu_n\mu_i\sigma_{jn}\sigma_m^2 - \mu_n\mu_j\sigma_{in}\sigma_m^2 + \mu_i\mu_j\sigma_n^2\sigma_m^2 - \sigma_{nm}^2\mu_j\mu_i$$
$$- \sigma_{im}\sigma_{jm}\mu_n^2 + \sigma_{jm}\sigma_{nm}\mu_i\mu_n + \sigma_{nm}\sigma_{im}\mu_j\mu_n], \qquad i, j = 1, 2, \cdots, n-1 \quad (11)$$

In practice, these covariances are unknown because they are functions of the population means, variances, and covariances. Covariance estimators $\hat{\boldsymbol{\Theta}}$ or $\hat{\boldsymbol{\Psi}}$ are obtained by substituting the sample estimators $\bar{r}_i$, s_i^2 and s_{ij} into Equations (10) and (11). The behavior of these estimators is examined in the simulation experiment to follow.

To test the general multivariate hypotheses, H_{OS}:$\mathbf{Sh} = \mathbf{0}$ or H_{OT}:$\mathbf{Tr} = \mathbf{0}$, of the equality of performance of a set of portfolios, two test statistics seem potentially useful. First, since $\widehat{\mathbf{Sh}}$ and $\widehat{\mathbf{Tr}}$ are asymptotically multivariate normal, with means $\mathbf{Sh}$ and $\mathbf{Tr}$ and covariance matrices $\boldsymbol{\Theta}$ and $\boldsymbol{\Psi}$, the quadratic forms $\widehat{\mathbf{Sh}}'\boldsymbol{\Theta}^{-1}\widehat{\mathbf{Sh}}$ and $\widehat{\mathbf{Tr}}'\boldsymbol{\Psi}^{-1}\widehat{\mathbf{Tr}}$ are chi-square random variables with $(n-1)$ degrees of freedom under the null hypothesis. Because the elements of $\boldsymbol{\Theta}$ and $\boldsymbol{\Psi}$ are unknown, estimators of these matrices, $\hat{\boldsymbol{\Theta}}$ and $\hat{\boldsymbol{\Psi}}$ are employed. The behavior of these approximate chi-square statistics is determined in Section III.[7] Alternative test statistics are the Z sum statistics, $Z_S = \sum_{i=1}^{n-1} z_{\mathrm{Sin}}$ and $Z_T = \sum_{i=1}^{n-1} z_{Tin}$ which are asymptotically normal with zero means and variances $\mathbf{e}'\boldsymbol{\Theta}\mathbf{e}$ and $\mathbf{e}'\boldsymbol{\Psi}\mathbf{e}$ respectively, where $\mathbf{e}$ is the unit vector. The covariance matrices $\boldsymbol{\Theta}$ and $\boldsymbol{\Psi}$ once again require estimation. The test statistics are therefore

$$\frac{Z_S}{\sqrt{\mathbf{e}'\hat{\boldsymbol{\Theta}}\mathbf{e}}} \quad \text{and} \quad \frac{Z_T}{\sqrt{\mathbf{e}'\hat{\boldsymbol{\Psi}}\mathbf{e}}} \qquad (12)$$

The behavior of these approximately normal test statistics is studied in the following section. A choice between the two proposed test statistics is to be made on the basis of the statistics' conformity with their distributional assumptions and their statistical power in detecting performance differences.

In summary, the approximate bias and asymptotic distributions of the traditional Sharpe and Treynor performance measures have been derived. Perform-

[7] The resulting statistic is asymptotically chi-square and is called the Wald [34] statistic. See the Appendix for a brief discussion.

ance comparison measures have also been specified with a view to multivariate performance testing. The approximate bias, asymptotic distributions, and two alternative test statistics for these comparison measures have also been derived. These test statistics do not rely on the assumption of cross-sectional independence of the n portfolios which plagued other attempts to assess performance. To assess their behavior and usefulness for hypotheses testing with small samples, a simulation experiment was performed and its results are described next.

III. Simulation Experiments

The more important questions to be answered by the Monte Carlo study are: (i) whether the estimators of the performance comparison variances are well behaved; (ii) whether the test statistics for performance in small samples are comparable to their asymptotic distributions; (iii) whether sample size has much effect on the test's power; and (iv) how the test power behaves with the magnitude of unequal performance. This section is divided into two parts, the first dealing with the single comparison of two portfolios and the second studying the multiple comparison of n portfolios.

A. The Single Comparison Study

Two portfolios and a market proxy portfolio were specified by their population mean premium return vector and covariance matrix. The specification includes two cases where no difference and differences exist in performance. Within each of the two cases, subcases were constructed by changing the means, variances, correlations, and sample size T. A summary of the population parameter cases is shown in Table I.

For each case, the population values of the performance comparison measures were computed together with their population asymptotic variances from Equations (7) and (8). Then 500 samples of size T were generated from the population mean and covariance parameters.[8] For each of these 500 samples, unbiased estimates of the mean premium return vector and covariance matrix were obtained. The sample estimates of the performance comparisons were computed and the estimates of the performance comparison variances were computed by inserting sample estimates in Equations (7) and (8). The z test statistics in (9) were computed for each sample. Then K-S normal goodness of fit tests, over the sample z's, were performed and the powers of the test statistic under normality were computed for the 5% alpha level.

Tables II and III summarize the results of the simulation experiment by the case numbers defined in Table I. Comparison of the asymptotic population variance (Column 3) and the sampling mean of the estimated variance (Column 4) for the Sharpe comparison statistic shows they are approximately equal at all sample sizes, but particularly for size $T = 24$ and above. There seems to be a

[8] The multivariate normal random number generator from the IMSL Subrouting package was used for both the univariate and multivariate tests of performance. All computations were performed in double precision Fortran.

Table I

Summary of Population Parameters for Single Comparisons

Case	T	Means μ_1, μ_2, μ_m	Variances $\sigma_1^2, \sigma_2^2, \sigma_m^2$	Correlations $\rho_{1,2}, \rho_{1m}, \rho_{2m}$	Traditional Sharpe Sh_1	Traditional Sharpe Sh_2	Traditional Treynor Tr_1	Traditional Treynor Tr_2
No Difference in Performance								
1. a	12	1.0, 1.0, 1.0	16, 16, 16	.5, .5, .5	.25	.25	2.0	2.0
b	24	1.0, 1.0, 1.0	16, 16, 16	.5, .5, .5	.25	.25	2.0	2.0
c	60	1.0, 1.0, 1.0	16, 16, 16	.5, .5, .5	.25	.25	2.0	2.0
d	60	.5, .5, 1.0	16, 16, 16	.5, .5, .5	.125	.125	1.0	1.0
e	60	1.0, 1.0, 1.0	36, 36, 16	.5, .5, .5	.167	.167	1.33	1.33
f	60	1.0, 1.0, 1.0	16, 16, 16	.25, .25, .25	.25	.25	4.0	4.0
g	60	1.0, 1.0, 1.0	16, 16, 16	−.25, −.25, −.25	.25	.25	−4.0	−4.0
Different Performances								
2. a	12	1.2, .6, 1.0	64, 36, 16	.5, .5, .5	.15	.10	1.2	.8
b	24	1.2, .6, 1.0	64, 36, 16	.5, .5, .5	.15	.10	1.2	.8
c	60	1.2, .6, 1.0	64, 36, 16	.5, .5, .5	.15	.10	1.2	.8
d	120	1.2, .6, 1.0	64, 36, 16	.5, .5, .5	.15	.10	1.2	.8
e	480	1.2, .6, 1.0	64, 36, 16	.5, .5, .5	.15	.10	1.2	.8
f	60	2.4, 1.2, 1.0	64, 36, 16	.5, .5, .5	.30	.20	2.4	1.6
g	60	1.2, .6, 1.0	16, 9, 16	.5, .5, .5	.30	.20	2.4	1.6
h	60	1.2, .6, 1.0	64, 36, 16	.25, .25, .25	.15	.10	2.4	1.6
i	60	1.2, .6, 1.0	64, 36, 16	−.25, −.25, −.25	.15	.10	−2.4	−1.6
j	60	4.8, 2.4, 4.0	16, 9, 16	.5, .5, .5	1.2	.80	9.6	6.4
k	120	4.8, 2.4, 4.0	16, 9, 16	.5, .5, .5	1.2	.80	9.6	6.4

Table II

Results of Simulation Experiment: Behavior of Sharpe Performance Comparison Statistic

Case	T	Asymptotic Variance	Monte Carlo Variance	z Statistic Bias	z Statistic Std. Dev.	*K-S* Goodness of Fit	Power
1. a	12	24.36	25.52	.00	.95	.64	.06
b	24	11.65	11.68	.01	1.00	.69	.07
c	60	4.54	4.62	.01	1.00	.71	.05
d	60	4.39	4.47	.01	1.01	.68	.05
e	60	22.42	22.85	.01	1.01	.69	.05
f	60	6.76	6.90	.02	1.00	.84	.06
g	60	11.10	11.44	.02	1.02	.96	.06
2. a	12	212.07	222.36	−.02	0.96	.98	.06
b	24	101.43	100.80	−.04	0.96	.45	.05
c	60	39.54	40.29	.00	1.01	.61	.08
d	120	19.60	19.82	.04	1.03	.78	.09
e	480	4.87	4.87	−.02	0.99	.52	.19
f	60	41.00	41.75	.00	1.00	.72	.14
g	60	2.56	2.61	.00	1.00	.72	.14
h	60	59.17	60.47	.01	1.02	.60	.08
i	60	98.23	101.34	.01	1.03	.99	.07
j	60	4.39	4.43	−.02	1.02	.54	.64
k	120	2.18	2.19	.02	1.02	.92	.90

Table III

Results of Simulation Experiment: Behavior of Treynor Performance Comparison Statistic

Case	T	Asymptotic Variance	Monte Carlo Variance	z Statistic Bias	z Statistic Std. Dev.	*K-S* Goodness of Fit	Power
1. a	12	0.46	0.92	−0.01	0.79	0.04	.00
b	24	0.22	0.31	0.02	0.87	0.45	.01
c	60	0.09	0.10	0.01	0.95	0.62	.02
d	60	0.07	0.09	0.01	0.93	0.87	.02
e	60	0.38	0.45	0.01	0.94	0.75	.02
f	60	0.05	0.07	0.02	0.88	0.39	.02
g	60	0.09	0.10	−0.04	0.95	0.21	.03
2. a	12	3.51	7.33	−0.06	0.77	0.01	.00
b	24	1.68	2.35	−0.07	0.83	0.06	.00
c	60	0.65	0.77	−0.03	0.94	0.63	.03
d	120	0.32	0.35	0.02	1.00	0.86	.07
e	480	0.08	0.08	−0.02	0.98	0.53	.18
f	60	0.79	0.91	0.00	0.95	0.51	.06
g	60	0.05	0.06	0.00	0.95	0.51	.06
h	60	0.29	0.46	0.01	0.85	0.08	.01
i	60	0.48	0.63	0.01	0.91	0.88	.02
j	60	0.22	0.23	−.04	0.94	0.67	.21
k	120	0.11	0.11	−.02	0.98	0.51	.44

slight positive bias in the variance estimator. The Treynor variance is not so well behaved, since the mean of the variance estimate does not approximate the asymptotic variance until at least sample size $T = 60$. This also causes overestimation bias in the Treynor comparison statistic.[9]

Tables II and III also record the bias and standard deviation of the z statistic for the Sharpe comparison measure and the Treynor comparison measure (Columns 5 and 6). In general, the Sharpe z has zero bias and unit standard deviation for all sample sizes, whereas the Treynor z requires a sample size of at least $T = 60$ for equivalent performance due to the variance overestimation. K-S two-tailed rejection probabilities for testing if z is normal are shown in Column 7 of Tables II and III for each of the Sharpe and Treynor comparison statistics. For the Sharpe statistic, the null hypothesis of normality would not be rejected under any of the specifications unless the Type I error level was large (~.40). The Treynor z statistic, on the contrary, is not a standard normal variable until at least sample size $T = 24$.

The powers of the test statistics are shown in Tables II and III, Column 8, for a Type I error level of 5%. The Sharpe statistics do not show any powerful ability to distinguish the differences in performance parameters which are similar to monthly return parameters for stocks (Cases 2a through i). The probability of rejecting the equal performance hypothesis for sample size $T = 60$ and a regular Sharpe difference of .05 or .10, is about 10%. This conclusion seems relatively unaffected by the population mean, variance, or correlation structure. Extremely large sample sizes do not substantially increase the power. The Treynor measure is virtually powerless except at large sample sizes and the actual Type I error level (Cases 1a to 1g) is much smaller than the 5% level used to determine the critical values of z. This result is expected because the standard deviation of the z statistics were less than 1.0. If the performance parameters are similar to those encountered with portfolios or larger data intervals with smaller coefficients of variation (Cases 2j and k), the Sharpe statistic is quite powerful although moderately large samples are required. These results emphasize the dubious practice of evaluating differences in investment performance with Sharpe and

[9] The $O\left(\frac{1}{T}\right)$ Taylor series approximation to the variance is too small. Addition of the $O\left(\frac{1}{T^2}\right)$ term yields the approximation

$$\psi = \frac{1}{\sigma_m^4 T}[\sigma_i^2\sigma_{nm}^2 + \sigma_n^2\sigma_{im}^2 - 2\sigma_{in}\sigma_{im}\sigma_{nm} + \mu_i^2(\sigma_{nm}^2 + \sigma_n^2\sigma_m^2) + \mu_n^2(\sigma_{im}^2 + \sigma_i^2\sigma_m^2) - 2\mu_i\mu_n(\sigma_{im}\sigma_{nm} + \sigma_{nm} + \sigma_{in}\sigma_m^2)] + \frac{1}{T^2}\left[\frac{\sigma_i^2\sigma_n^2}{\sigma_m^2}\left(1 - \frac{\sigma_{in}}{\sigma_i\sigma_n}\right)\right]$$

For the Case (1b), the $O\left(\frac{1}{T^2}\right)$ approximation is $\psi = .217 + .042 = .259$ which compares more favorably with the simulation value of .307. The use of a variance estimator which includes the $O\left(\frac{1}{T^2}\right)$ term, however, provides an inferior estimator of the sample variance. The resulting Z and χ^2 statistics, which employ the estimates of the $O\left(\frac{1}{T^2}\right)$ term, also demonstrated inferior performance compared to the estimators based only on $O\left(\frac{1}{T}\right)$ terms.

Treynor statistics with small samples and large coefficients of variation. It is apparent that large sample sizes and/or large differences in performance are required for detection.

The asymptotic bias from (2), the estimated bias using a sample estimate in (2), and the actual sample bias were computed for the traditional Treynor measure. The results were not very successful and are not reported in detail. Generally, the population asymptotic bias never approached the actual sample bias mean until very large samples of $T > 480$. Consequently, estimators of the actual bias did not do well in small samples. In fact, samples of $T > 120$ were necessary before the estimated bias approached the asymptotic bias. Thus, higher order terms in Equation (2) are important and estimators of the Treynor bias based on (2) are extremely poor in small samples.

In summary, the Treynor test statistic is not well behaved for sample sizes less than $T = 60$. Thus, there is little good that can be said about this performance statistic from either a theoretical point of view or a statistical point of view. On the other hand, the Sharpe statistic has more desirable qualities as a test statistic.

To illustrate the use of the Sharpe statistic, consider the two mutual funds from Sharpe [28] with the largest and smallest sample Sharpe measures. They are Boston Fund with $\bar{r}_1 = 9.4\%$/year, $s_1 = 12.1\%$/year, and Incorporated Investors with $\bar{r}_2 = 14.0\%$ and $s_2 = 25.5$. These statistics were estimated with $T = 10$ observations. Assuming an estimate of the correlation of .5 between the funds' returns, the transformed performance and its estimated variance are

$$\widehat{Sh}_{in} = 9.4(25.5) - 14(12.1) = 70.3$$

and

$$\hat{\theta} = \frac{(146.41)(650.25)}{10}\left[2(.5) + \frac{1}{2}\left\{\left(\frac{9.4}{12.1}\right)^2 + \left(\frac{14.0}{25.5}\right)^2 - \left(\frac{9.4}{12.1}\right)\left(\frac{14.0}{25.5}\right)(1.5^2)\right\}\right] = 11290.1$$

Under the null hypothesis of no difference in performance, the value of the test statistic is

$$z_{\text{Sin}} = \frac{70.3}{\sqrt{11290.1}} = .66$$

Unless the Type I error level is set at $\alpha \geq .51$, the null hypothesis of equal performance cannot be rejected.

B. Multivariate Study

To test the behavior of the multivariate comparison statistics, χ^2 and Z sum outlined in Section II*C*, $n = 20$ stocks (henceforth called portfolios) were drawn from the CRSP monthly data file. The values of the mean premium return vector and covariance matrix were computed from the 313 monthly return vectors beginning January 1950 and ending January 1976. These are displayed in Table IV, together with the CRSP value weighted market return mean, variance, and the covariances with the CRSP market return index. This mean vector and covariance

Table IV

Portfolio and Market Parameters for the Multivariate Simulation

Means μ_i

0.50	0.90	1.10	1.74	1.82	1.11	0.91	1.18	1.35	1.07	1.16	1.23	0.81	1.18	0.88	1.20	0.72	1.16	0.92	1.25

Covariances σ_{ij}*:*

53.64																			
6.60	29.84																		
19.34	16.68	82.88																	
34.14	20.66	48.01	178.07																
6.32	6.48	18.66	27.54	118.09															
5.76	11.33	21.04	19.28	26.29	57.07														
16.92	8.43	22.16	32.35	23.88	20.24	52.05													
15.26	9.26	16.21	26.87	12.41	11.69	15.29	48.25												
9.57	10.80	16.26	18.29	14.16	15.23	12.10	9.69	29.80											
10.12	11.22	18.94	22.39	23.13	16.27	17.74	9.37	11.21	35.12										
10.33	9.59	21.59	21.75	31.03	13.72	17.95	8.59	13.05	22.58	47.64									
18.89	8.76	27.01	41.73	13.08	19.29	21.39	14.42	13.83	12.96	16.56	65.62								
8.45	13.50	8.80	17.34	5.40	7.78	9.58	9.86	7.27	7.92	5.97	7.92	23.51							
14.58	14.06	23.31	42.93	20.36	21.53	26.36	11.34	16.74	17.62	19.75	23.10	12.03	51.20						
14.64	16.47	17.40	26.27	9.88	11.33	16.24	13.33	11.42	10.70	9.26	11.55	14.27	16.42	28.72					
14.34	8.76	22.27	30.30	14.33	13.21	15.17	16.96	8.21	12.57	13.46	25.84	8.54	14.72	12.20	56.03				
27.85	14.93	36.71	65.96	17.07	13.45	25.64	32.08	15.70	16.21	20.46	35.76	15.22	26.17	19.87	32.30	109.46			
25.08	16.74	41.43	47.56	20.20	12.32	24.80	21.66	20.61	21.51	18.76	26.41	14.23	25.60	24.34	24.46	50.78	131.75		
11.77	22.82	21.44	20.67	13.46	16.84	15.50	14.34	14.80	14.23	13.32	16.96	15.80	20.42	22.36	13.14	18.58	26.97	44.66	
16.92	10.26	27.74	43.93	18.45	18.08	25.27	15.75	10.72	14.74	17.69	23.65	9.72	20.93	13.76	14.48	32.30	29.18	16.09	58.69

Market Mean μ_m: 0.94
Market Variance σ^2_m: 15.22
Covariances with the Market σ_{im}:

11.38	9.77	18.72	26.09	16.55	15.93	17.57	11.34	12.09	13.99	15.11	16.90	7.53	17.75	11.70	13.41	20.64	21.56	14.34	17.34

matrix were employed as the population values parametizing the simulation with reasonable values for monthly data. The behavior of the test statistics for the first 5 and 10 portfolios of the original 20 were also studied. For an estimated covariance matrix to have full rank, the sample size T must exceed the number of portfolios n. Thus, the sample sizes used for the simulations were $T = 12$, 24, 36, 60, and 120 for $n = 5$; $T = 24$, 36, 60, and 120 for $n = 10$, and $T = 36$, 60, 120, 240, and $T = 480$ for $n = 20$.

The population values of the Sharpe and Treynor transformed differences were computed using the nth portfolio as the base. Also, the population asymptotic covariance matrices $\boldsymbol{\Theta}$ and $\boldsymbol{\Psi}$ of the performance comparisons were computed.

The simulation generated 500 samples of size T from a given population $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. For each sample, unbiased estimates using $\bar{\mathbf{r}}$ and $\mathbf{S}$ were computed which were then used to compute the sample values of the transformed comparisons $\mathbf{Sh}$ and $\mathbf{Tr}$ from Equations (3) and (4). The estimated covariance matrices of the performance comparisons $\hat{\boldsymbol{\Theta}}$ and $\hat{\boldsymbol{\Psi}}$ and their inverses $\hat{\boldsymbol{\Theta}}^{-1}$ and $\hat{\boldsymbol{\Psi}}^{-1}$ were also computed using estimated values in Equations (10) and (11). The Z sum statistics divided by their standard deviations were computed from (12) and used for goodness of fit tests, to check if the test statistics were normal. In addition, the power of this test with alpha = .05 was computed. Finally, the χ^2 test statistics $\widehat{\mathbf{Sh}}'\hat{\boldsymbol{\Theta}}^{-1}\widehat{\mathbf{Sh}}$ and $\widehat{\mathbf{Tr}}'\hat{\boldsymbol{\Psi}}^{-1}\widehat{\mathbf{Tr}}$ were computed from the estimated performances and the estimated inverses. These 500 test statistics were used to compute the powers of the tests with an alpha level of .05 and to construct the observed frequency distributions of the statistics, $(\widehat{\mathbf{Sh}} - \mathbf{Sh})'\hat{\boldsymbol{\Theta}}^{-1}(\widehat{\mathbf{Sh}} - \mathbf{Sh})$ and $(\widehat{\mathbf{Tr}} - \mathbf{Tr})\hat{\boldsymbol{\Psi}}^{-1}(\widehat{\mathbf{Tr}} - \mathbf{Tr})$. These observed frequency distributions were then compared to the χ^2 distribution. The results of the multivariate simulation are summarized in Table V.

Although the primary purpose of the simulation was to examine the behavior of the χ^2 and Z sum statistics, the single comparison z_{Sin} statistics for each portfolio compared to the nth portfolio were also studied. As observed in the univariate simulation, the z_{Tin} statistic for the Treynor comparison tended to have variances less than one for $T \leq 36$. The Sharpe z_{Sin} variances were comparable to those observed in Table II. In general, the normal goodness of fit K-S statistics for both Sharpe and Treynor were insignificant for all portfolios and all sample sizes.

For the Treynor χ^2 statistic, the observed frequency distributions were less skewed and showed smaller means and variances than their theoretical counterparts. For this reason, the power of the 5% level tests were generally less than .05. As the size of the population increased from $n = 5$ to $n = 20$, a slight improvement in χ^2 was observed. A comparison of samples of size $T = 36$, 60, and 120 showed little difference in the observed χ^2 behavior. With $n = 20$ and $T = 240$ or 480, the theoretical and observed distributions were comparable and the power improved substantially to .22 and .80.

For the Sharpe χ^2 statistic, the observed frequency distribution in general showed more skewness, a higher mean, and a higher variance than the corresponding theoretical distribution. As the sample size T increased, however, the observed χ^2 distribution approached the theoretical distribution. For samples of size $T = 60$ and larger, the fit is reasonably good. As the number of portfolios n decreased, for a given T, the χ^2 fit improved. At $T = 36$ and $n = 5$, the observed

Table V

Results of Simulation Experiment: Behavior of Multivariate Test Statistics

T	Statistic	χ^2 Tail Areas 0.10	0.05	0.01	Mean	Variance	Power
Case: $N = 5$ Portfolios							
	Theoretical χ^2	7.78	9.49	13.28	4.0	8.0	–
12	Sharpe χ^2	9.51	11.55	15.87	4.9	11.6	.11
	Treynor χ^2	5.94	7.34	12.36	3.6	86.2	.02
24	Sharpe χ^2	9.37	11.70	15.48	4.7	12.0	.10
	Treynor χ^2	5.67	7.13	9.95	3.1	4.5	.01
36	Sharpe χ^2	8.31	10.17	14.11	4.4	9.0	.09
	Treynor χ^2	5.57	6.51	9.26	3.1	3.4	.01
60	Sharpe χ^2	7.84	9.96	14.70	4.1	9.0	.07
	Treynor χ^2	6.08	7.21	9.10	3.3	4.3	.02
120	Sharpe χ^2	7.36	8.91	12.10	3.9	7.0	.08
	Treynor χ^2	6.31	7.62	9.39	3.6	5.3	.03
Case: $N = 10$ Portfolios							
	Theoretical χ^2	14.68	16.92	21.67	9.0	18.0	–
24	Sharpe χ^2	17.19	20.42	24.80	10.8	23.8	.14
	Treynor χ^2	12.38	14.12	22.72	8.2	18.0	.02
36	Sharpe χ^2	16.33	19.45	23.71	10.2	23.6	.12
	Treynor χ^2	12.79	14.38	18.94	8.2	11.9	.03
60	Sharpe χ^2	15.06	17.94	21.94	9.7	18.6	.14
	Treynor χ^2	13.08	14.86	19.43	8.4	13.2	.04
120	Sharpe χ^2	14.78	17.60	22.48	9.3	19.3	.20
	Treynor χ^2	13.31	15.47	18.59	8.5	14.1	.08
Case: $N = 20$ Portfolios							
	Theoretical χ^2	27.20	30.14	36.19	19.0	38.0	–
36	Sharpe χ^2	36.01	40.32	46.02	25.8	68.4	.32
	Treynor χ^2	24.90	27.91	36.11	17.0	34.1	.02
60	Sharpe χ^2	30.09	33.76	42.45	21.4	50.2	.22
	Treynor χ^2	23.05	25.49	30.80	16.8	24.7	.02
120	Sharpe χ^2	32.45	36.26	44.17	22.4	54.4	.31
	Treynor χ^2	25.02	27.24	32.06	18.1	27.5	.03
240	Sharpe χ^2	27.05	36.00	37.36	19.4	38.4	.48
	Treynor χ^2	24.69	27.30	37.98	17.8	27.2	.22
480	Sharpe χ^2	28.03	30.83	35.60	19.5	40.6	.83
	Treynor χ^2	26.70	28.68	36.07	18.8	35.4	.80
Case: $N = 20$ Portfolios, Second Population							
	Theoretical χ^2	27.20	30.14	36.19	19.0	38.0	–
36	Sharpe χ^2	35.62	40.63	54.05	24.6	78.5	.84
	Treynor χ^2	43.26	49.65	76.85	28.9	149.6	.36
60	Sharpe χ^2	32.08	35.87	46.06	22.1	62.8	.96
	Treynor χ^2	34.98	39.88	51.80	24.1	81.5	.55
120	Sharpe χ^2	30.34	34.63	41.71	20.9	53.3	1.00
	Treynor χ^2	31.58	34.56	39.16	21.4	51.0	.98

χ^2 fit was good. The power of the χ^2 statistic for a .05 level test and $T = 60$ increased from .07 at $n = 5$ to .14 at $n = 10$ and .22 at $n = 20$. The sample size T did not seem to substantially influence the power until $T = 240$ and $T = 480$, in which cases the test powers were .48 and .83.

The Z sum statistic for both Sharpe and Treynor were well behaved for all 3 populations and at all sample sizes. The power for a .05 level test, however, was seldom above .05.

A second multivariate population was chosen by simply mutliplying the mean vector of Table IV by 4.0. This enables us to determine the effect of smaller coefficients of variation which accompany larger data intervals and larger portfolios. These results are presented as the fourth panel in Table V. The Sharpe χ^2 has consistently larger skewness than the theoretical distribution and the Treynor χ^2 has changed to also have more skewness than the theoretical. The powers of the tests have increased dramatically and are particularly large for the Sharpe comparison measure.

We conclude, therefore, that the optimum test procedure for simultaneously comparing the performance of n portfolios, is the Sharpe χ^2 statistic. The sample size employed should be at least $T = 36$ for evaluating portfolio performance and larger if portfolios with larger coefficients of variation are evaluated.[10] As expected, the more portfolios compared the greater the power.

APPENDIX

This appendix briefly summarizes the statistical procedures employed to generate asymptotic approximations to the distributions and the moments of the performance estimators.

A. Asymptotic Moments of the Estimators

Asymptotic approximations are obtained for the expectation and variance of an estimator by employing expectations of the Taylor series expansion of the estimator, about its true value. For the estimators $\hat{X}$ and $\hat{Y}$, the Taylor series expansion of the ratio $\frac{\hat{X}}{\hat{Y}}$ including terms of $O\left(\frac{1}{T^2}\right)$ is given by

$$\frac{\hat{X}}{\hat{Y}} = \frac{X}{Y}\left[\frac{1+\frac{\delta X}{X}}{1+\frac{\delta Y}{Y}}\right] = \frac{X}{Y}\left(1+\frac{\delta X}{X}\right)\left(1-\frac{\delta Y}{Y}+\frac{\delta Y^2}{Y^2}-\frac{\delta Y^3}{Y^3}+\frac{\delta Y^4}{Y^4}\pm\cdots\right)$$

where $\delta Y^r = (\hat{Y} - Y)^r$ and $\delta X = (\hat{X} - X)$.

[10] If continuously compounded nominal returns are used, then the coefficients of variation should not vary with the data interval. If effective returns are used, then returns which are normal for monthly data intervals will not be normal for longer data intervals. Hence, any advantageous reduction in the coefficients of variation comes at the expense of departures from normality. The robustness of our performance tests with nonnormal returns has not been investigated.

The expectation is

$$E\left(\frac{\hat{X}}{\hat{Y}}\right) = \frac{X}{Y}\left(1 + \frac{E(\delta X)}{X} - \frac{E(\delta Y)}{Y} - \frac{E(\delta X \delta Y)}{XY} + \frac{E(\delta Y^2)}{Y^2} + \frac{E(\delta X \delta Y^2)}{XY^2} - \frac{E(\delta Y^3)}{Y^3} - \frac{E(\delta X \delta Y^3)}{XY^3} + \frac{E(\delta Y^4)}{Y^4} \pm \cdots\right)$$

The variance is obtained by squaring the expression for $\frac{\hat{X}}{\hat{Y}}$ and taking expectations. The approximate variance is then

$$V\left(\frac{\hat{X}}{\hat{Y}}\right) = E\left(\frac{\hat{X}^2}{\hat{Y}^2}\right) - E\left(\frac{\hat{X}}{\hat{Y}}\right)^2$$

where

$$E\left(\frac{\hat{X}^2}{\hat{Y}^2}\right) = \frac{X^2}{Y^2}\left(1 + \frac{2E(\delta X)}{X} - \frac{2E(\delta Y)}{Y} + \frac{E(\delta X^2)}{X^2} + \frac{3E(\delta Y^2)}{Y^2} - \frac{4E(\delta X \delta Y)}{XY} + \frac{6E(\delta X \delta Y^2)}{XY^2} - \frac{2E(\delta X^2 \delta Y)}{X^2 Y} - \frac{4E(\delta Y^3)}{Y^3} - \frac{8E(\delta X \delta Y^3)}{XY^3} + \frac{3E(\delta X^2 \delta Y^2)}{X^2 Y^2} + \frac{5E(\delta Y^4)}{Y^4} \pm \cdots\right)$$

Since $\hat{X}$ and $\hat{Y}$, in our application, are elements of $\bar{\mathbf{r}}$ or $\mathbf{S}$, then $E(\delta X) = E(\delta Y) = 0$. The moments for products of the elements of $\bar{\mathbf{r}}$ up to the fourth power are available in, for example, Anderson [1]. The moments of products of the elements of $\mathbf{S}$ are available in Wishart [35]. Since $\bar{\mathbf{r}}$ and $\mathbf{S}$ are independent, moments of products of the elements of $\bar{\mathbf{r}}$ and $\mathbf{S}$ are obtained from the products of the moments. These moments are substituted into the relevant Taylor series approximation and simplified to obtain the results shown in the paper.

The resulting expressions for the expectation and variance do not necessarily converge. The convergence usually depends on the magnitude of the parameters relative to the sample size T. For the traditional Treynor measure, the series expression for the expectation is dominated by terms in powers of $\frac{1}{T\rho_{im}^2}$. Convergence, therefore, requires that $\frac{1}{T\rho_{im}^2} < 1$ or $T > \frac{1}{\rho_{im}^2}$. For small values of ρ_{im}, large values of T are required. In the case of the traditional Sharpe measure, the series expression for the expectation is a power series in $\frac{1}{T}$, which coverges for all $T > 1$.

B. Asymptotic Distributions and Hypotheses Tests

Let γ denote the $(p \times 1)$ vector of unique elements of μ and $\mathbf{\Sigma}$ $\left(p = n + \frac{n(n+1)}{2}\right)$ and let $\hat{\gamma}$ be the corresponding estimator of γ based on $\bar{\mathbf{r}}$ and $\mathbf{S}$. Let the vector $\hat{\mathbf{f}}' = [\hat{f}_1(\hat{\gamma}), \hat{f}_2(\hat{\gamma}), \cdots, f_k(\hat{\gamma})]$ denote a set of k estimators, which are

functions of the elements of $\hat{\gamma}$. The asymptotic distribution of $\hat{\mathbf{f}}$, for "nice" functions under normality, is multivariate normal with mean vector $\mathbf{f}' = [f_1(\gamma), f_2(\gamma), \cdots f_k(\gamma)]$ and covariance matrix $\mathbf{D}'\Gamma\mathbf{D}$, where Γ is the $(p \times p)$ covariance matrix of $\hat{\gamma}$ and $\mathbf{D}$ is the $(k \times p)$ matrix of partials $d_{ij} = \frac{\partial f_i(\hat{\gamma})}{\partial \gamma_j}$, $i = 1, 2, \cdots, k, j = 1, 2, \cdots, p$, evaluated at γ. (See Rao [23] and Anderson.) The elements of $\mathbf{D}'\Gamma\mathbf{D}$ are equivalent to $O\left(\frac{1}{T}\right)$ Taylor series approximations to the second moments.

For the purpose of testing hypotheses regarding the elements of $\mathbf{f}$, for instance $H_O: \mathbf{f} = \mathbf{f}_0$, we may employ the Wald [34] statistic $(\hat{\mathbf{f}}, -\mathbf{f}_0)'(\hat{\mathbf{D}}\hat{\Gamma}\hat{\mathbf{D}})^{-1}(\hat{\mathbf{f}} - \mathbf{f}_0)$, which converges to a χ^2 distribution with k degrees of freedom as $T \to \infty$. The matrices $\hat{\mathbf{D}}$ and $\hat{\Gamma}$ are estimators of $\mathbf{D}$ and Γ based on $\hat{\gamma}$. See Stroud [30].

REFERENCES

1. T. W. Anderson. *An Introduction To Multivariate Analysis.* New York: John Wiley and Sons Inc., 1958.
2. Ray Ball and Philip Brown. "An Empirical Evaluation of Accounting Numbers." *Journal of Accounting Research* (Autumn 1968).
3. Stephen Beveridge. "A Note on Random Walks and Indices of Investment Performance." Working Paper: Faculty of Business Administration and Commerce, University of Alberta, 1978.
4. Fisher Black. "Capital Market Equilibrium with Restricted Borrowing." *Journal of Business* 45 (1972), 444–54.
5. Marshall E. Blume. "The Relative Efficiency of Various Portfolios: Some Further Evidence." Working Paper No. 26-1979a: Rodney L. White Center for Financial Research, University of Pennsylvania.
6. Stephen J. Brown and Jerold B. Warner. "Measuring Security Price Performance." *Journal of Financial Economics* 8 (September 1980), 205–58.
7. Bradford Cornell. "Asymmetric Information and Portfolio Performance Measurement." *Journal of Financial Economics* 7 (December 1979), 381–90.
8. Eugene Fama, Lawrence Fisher, Michael Jensen, and Richard Roll. "The Adjustment of Stock Prices to New Information." *International Economic Review* (February 1967).
9. ———. "Components of Investment Performance." *Journal of Finance* 27 (June 1972), 551–67.
10. ———. and James MacBeth. "Risk Return and Equilibrium: Empirical Tests." *Journal of Political Economy* 38 (1973), 607–63.
11. W. Feller. *An Introduction to Probability Theory and its Applications, Vol. 1.* New York: John Wiley and Sons, 1957.
12. Jeffrey F. Jaffe. "Special Information and Insider Trading." *Journal of Business* 47 (July 1974), 410–28.
13. Michael C. Jensen. "The Performance of Mutual Funds in the Period 1945–1964." *Journal of Finance* 23 (May 1968), 389–416.
14. ———. "Risk, the Pricing of Capital Assets and Evaluation of Investment Portfolios." *Journal of Business* 42 (April 1967), 167–247.
15. J. D. Jobson and Bob Korkie. "Estimation for Markowitz Efficient Portfolios." *Journal of the American Statistical Association* 75 (September 1980), 544–54.
16. Terence C. Langetieg. "An Application of a Three Factor Performance Index to Measure Stockholder Gains from Merger." *Journal of Financial Economics* 6 (December 1978), 365–84.
17. Gershon Mandelker. "Risk and Return: The Case of Merging Firms." *Journal of Financial Economics* 1 (December 1974), 303–35.
18. David Mayers and Edward M. Rice. "Measuring Portfolio Performance and the Empirical Content of Asset Pricing Models." *Journal of Financial Economics* 7 (March 1979), 3–28.

19. Robert E. Miller and Adam K. Gehr. "Sample Size Bias and Sharpe's Performance Measure: A Note." *Journal of Financial and Quantitative Analysis* 12 (December 1978), 943–46.
20. Franco Modigliani and Gerald A. Pogue. "An Introduction to Risk and Return: Concepts and Evidence." *Financial Analysts Handbook* I, ed. Sumner E. Levine. Dow Jones-Irwin Inc. 1975, pp. 1296–1342.
21. Michael Parkinson. "The Extreme Value Method for Estimating the Variance of the Rate of Return." *Journal of Business* 53 (January 1980), 61–65.
22. Richardson R. Pettit. "Dividend Announcements, Security Performance and Capital Market Efficiency." *Journal of Finance* (December 1972).
23. C. R. Rao. *Linear Statistical Inference and Its Applications*, 2nd. ed. New York: John Wiley and Sons Inc., 1973.
24. R. Roll. "A Reply to Mayers and Rice (1979)." *Journal of Financial Economics* 7 (December 1979), 391–400.
25. ———. "A Critique of the Asset Pricing Theory's Tests." Working Paper: Graduate School of Management UCLA, Los Angeles, 1976.
26. ———. "Ambiguity when Performance is Measured by the Securities Market Line." *Journal of Finance* 33 (September 1978), 1051–69.
27. Erwin Saniga, Nicolas Gressis, and Jack Hayya. "The Effects of Sample Size and Correlation on the Accuracy of the EV Efficiency Criterion." *Journal of Financial and Quantitative Analysis* 14 (September 1979), 615–28.
28. William F. Sharpe. "Mutual Fund Performance." *Journal of Business, A Supplement*, No. 1, Part 2 (January 1966), pp. 119–38.
29. ———. *Portfolio Theory and Capital Markets.* McGraw-Hill, 1970.
30. T. W. F. Stroud. "On Obtaining Large Sample Tests from Asymptotically Normal Estimators." *Annals of Mathematical Statistics* 42 (1971), 1412–24.
31. Seha M. Tinic and Richard R. West. *Investing in Securities: An Efficient Markets Approach.* Addison Wesley, 1979.
32. Jack L. Treynor. "How To Rate Management of Investment Funds." *Harvard Business Review* 43 (January–February 1965), 63–75.
33. Robert E. Verrecchia. "The Mayers-Rice Conjecture: A Counterexample." *Journal of Financial Economics* 8 (March 1980), 87–100.
34. A. Wald. "Tests of Statistical Hypotheses Concerning Several Parameters When the Number of Observations is Large." *Transactions of the American Mathematical Society* 54 (1943), 426–82.
35. John Wishart. "The Generalised Product Moment Distribution in Samples From a Normal Multivariate Population." *Biometrika* 15 (1928), 32–52.