

**Computational Finance 1999**

edited by Yaser S. Abu-Mostafa, Blake LeBaron, Andrew W. Lo,  
and Andreas S. Weigend

The MIT Press  
Cambridge, Massachusetts  
London, England

# 3 Confidence Intervals and Hypothesis Testing for the Sharpe and Treynor Performance Measures: A Bootstrap Approach

H. D. Vinod and Matthew R. Morey

The Sharpe (1966) and Treynor (1965) portfolio performance measures are widely cited and used in the literature and pedagogy of finance. However, due of the presence of random denominators in the definitions of the performance measures and the difficulty in determining the sample size needed to achieve asymptotic normality, they do not easily allow for the construction of confidence intervals and hypothesis testing. This paper uses the various forms of the bootstrap methodology including the single, studentized, and double bootstrap to construct confidence intervals and conduct hypothesis testing on these performance measures. This paper improves upon the previous efforts of Jobson and Korkie (1981) who use the delta method to develop hypothesis tests for these performance measures. We illustrate our methodology with several actual mutual funds.

## 3.1 Introduction

The Sharpe (1966) and Treynor (1965) portfolio performance measures are widely used and cited in the literature and pedagogy of finance. However, due of the presence of random denominators in their definitions and the difficulty in determining the sample size needed to achieve asymptotic normality, they do not easily allow for the construction of confidence intervals and hypothesis testing. This paper uses various forms of the bootstrap methodology including the single, studentized, and double bootstrap to construct confidence intervals and conduct hypothesis testing on these performance measures. In this way, the paper improves upon the previous efforts of Jobson and Korkie (JK) who use the delta method to develop hypothesis tests for these performance measures.

The paper is organized as follows. Section I of the paper discusses the Sharpe and Treynor measures and provides some motivation for the use of confidence intervals and hypothesis testing on the performance measures. Section II explains the bootstrap and then applies this methodology to constructing confidence intervals. Section III describes JK (1981) hypothesis testing approach and explains how our methodology improves upon their efforts. Section IV concludes the paper.

## 3.2 The Sharpe and Treynor Performance Measures

Consider the following scenario in which the relative performance of  $n$  portfolios is to be evaluated.<sup>1</sup> In this scenario,  $r_{i,t}$  represents the excess return from the  $i$ -th

1. The notation in this section is the same as Jobson and Korkie (1981).

portfolio in period  $t$ , where  $i = 1, 2, \dots, n$ . A random sample of  $T$  excess returns on the  $n$  portfolios is then illustrated by  $r'_t = [r_{1t}, r_{2t}, \dots, r_{nt}]$ , where  $t = 1, 2, \dots, T$  and where  $r_t$  is assumed to be multivariate normal random variable, with mean  $\mu = \{\mu_i\}, i = 1, 2, \dots, n$  and a covariance matrix  $\Sigma = (\sigma_{ij})$  where  $i, j = 1, 2, \dots, n$ . It is well-known that the unbiased estimators of the mean vector and the covariance matrix are

$$\bar{r} = \frac{1}{T} \sum_{t=1}^T r_t, \text{ and } S = \{s_{ij}\} \frac{1}{T-1} \sum_{t=1}^T (r_t - \bar{r})(r_t - \bar{r})'. \quad (3.1)$$

These two estimators are then used to form the estimators of the traditional Sharpe and Treynor performance measures.

The population value of the Sharpe (1966) performance measure for portfolio  $i$  is defined as  $Sh_i = \frac{\mu_i}{\sigma_i}, i = 1, 2, \dots, n$ . It is simply the mean excess return over the standard deviation of the excess returns for the portfolio. The population value of the Treynor (1965) performance measure is  $Tr = \frac{\mu_i \sigma_m^2}{\sigma_{im}^2}, i = 1, 2, \dots, n$ , where  $m$  is the market proxy portfolio often denoted by the Standard and Poor's 500 index (S&P 500) and is the familiar Beta from the Capital Asset Pricing Model (CAPM). The conventional sample-based point estimates of the Sharpe and Treynor performance measures used in (1) are then

$$\hat{Sh}_i = \frac{\bar{r}_i}{s_i} \text{ and } \hat{Tr}_i = \frac{\bar{r}_i s_m^2}{s_{im}} \text{ for } i = 1, 2, \dots, n. \quad (3.2)$$

The Sharpe and Treynor measures do not permit small-sample confidence intervals, because of the presence of  $s_i, s_n$  in their definitions of (2). Moreover, the sample size needed achieve asymptotic normality is difficult to determine.

Despite this difficulty, the ability to construct confidence intervals on these performance measures is valuable. A confidence interval provides additional information for comparing portfolios. For example, consider an investor who uses the Sharpe measure to evaluate portfolios. Such an investor will identify two different portfolios as having very similar performance, because they have similar Sharpe point estimates. However, such an appraisal does not at all consider the range of uncertainty behind these point estimates; one portfolio may have a very narrow confidence interval on its Sharpe measure while the other has a wide confidence interval. Such knowledge, may sway an investor to prefer the portfolio with the shorter interval.

Furthermore, there is a possibly serious problem with the use of the Sharpe and Treynor measures in cases of sampling in a bear market.<sup>2</sup> For example, if we are comparing portfolios, the following scenario can arise. A portfolio which has a higher risk and a more negative return can have a higher Sharpe measure than a portfolio which has a lower risk and a higher (but still negative) mean return. Due to the possibility of such results occurring, confidence intervals and hypothesis testing have further merit, as they clarify the situation and allow for a rational comparison of portfolios despite perverse ranking by point estimates.

### 3.3 Creating Confidence Intervals using the Bootstrap

A confidence interval around a population parameter is a well-known method of statistical inference. Traditional small-sample parametric confidence intervals start with the assumption that the sampling distribution of the test statistic is normal, at least asymptotically. However, since the small-sample sampling distributions of the Sharpe and Treynor measures are non-normal in that they have significant skewness and kurtosis, the usual method based on ratio of the statistic to its standard errors is biased and unreliable. Various bootstrap resampling techniques can help solve the non-normality problem. These techniques have only recently become available due to declining costs of computing. For a survey of the bootstrap literature see Davison and Hinkley (1997) and Vinod (1993).

To understand the basic idea behind the bootstrap consider the following. Consider a statistic  $\hat{B}$ , based on a sample of size,  $T$ . In the bootstrap methodology, instead of assuming the shape of the sampling distribution of  $\hat{B}$ , one empirically approximates the entire sampling distribution of  $\hat{B}$  by investigating the variation of  $\hat{B}$  over a large number of pseudo samples obtained by resampling. For the resampling, a Monte-Carlo type procedure is used on the available sample values. This is conducted by randomly drawing, with replacement, a large number of resamples of size  $T$  from the original sample. Each resample has  $T$  elements, however any given resample could have some of the original data points represented more than once and some not at all. Note that each element of the original sample has the same probability,  $(1/T)$ , of being in a sample. The initial idea behind bootstrapping was that a relative frequency distribution of  $\hat{B}$ 's calculated from the resamples can be a good approximation to the sampling distribution of  $\hat{B}$ . This idea has since been

2. Jobson and Korkie (1981) p. 891, also make this point.

extended to achieve better approximations and to conditional models and one-step conditional moments.<sup>3</sup>

In this paper we use three variations of the bootstrap to arrive at confidence intervals for the Sharpe and Treynor performance measures. These are the single bootstrap (s-boot), the studentized bootstrap (boot-t) and the double bootstrap (d-boot). Efron's original bootstrap, described in numerous journal articles and econometrics textbooks, is called the s-boot here.<sup>4</sup>

In terms of the s-boot, the resampling for the Sharpe measure is done "with replacement" of the original excess returns themselves for  $j = 1, 2, \dots, J$  or 999 times. Thus we calculate 999 Sharpe measures from the original excess return series. The choice of the odd number 999 is convenient, since the rank-ordered 25-th and 975-th values of estimated Sharpe ratios arranged from the smallest to the largest, yield a useful s-boot 95% confidence interval.

For the Treynor measure we obtain  $J$  estimates of the numerator  $\mu$ , similar to the Sharpe measure described above. However, the denominator is the Beta estimated by the regression of  $R_i = \alpha_i + \beta_i R_m + e_i$ , where  $R_i$  is the portfolio return and  $R_m$  is the market return (we use the S&P500 index returns as the market proxy). Hence the s-boot for the Treynor measure needs resampling of the residuals of the above regression  $J (= 999)$  times. Replacing the original residuals by the resampled residuals, keeping the original ordinary least squares (OLS) estimates of intercept, slope coefficients, and using the same right hand side regressors, yields  $J$  versions of the dependent variable. The procedure then creates  $J$  separate regression problems for which the intercept and slopes can be again estimated, as explained in Judge et al (1988). These  $J$  estimates of Beta yield the  $J$  denominators of the Treynor measures. Again we rank-order these measures from the smallest to the largest and choose the 25-th and 975-th order statistics to yield the simple s-boot 95% confidence interval.

Unfortunately, a typical 95% s-boot confidence interval for non-standard sample statistics similar to (2) may not cover the true parameter with the coverage probability of 0.95. Consider a standard statistic similar to the sample mean or a regression coefficient  $b$ , which estimates  $\beta$  with the standard error  $S$ . Now the quantity  $q = (b - \beta)/S$  is pivotal in the sense that its distribution does not depend on  $\beta$ .

3. For example, let  $b^*$  denote resampled regression coefficients,  $b$  the original coefficients, and the unknown parameters. The extended bootstrap approximates the properties of  $(b^*)$  by the observable  $(b^* - b)$ . See Davison and Hinkley (1997) for recent references and Vinod (1993) for references to earlier attempts.

4. Judge et al (1988, p.416) or Greene (1997, p.184).

Under normality of regression errors,  $q$  has the well-known  $t$ -distribution leading to the usual confidence intervals. If we wish to relax the normality assumption, simple  $s$ -boot can provide good confidence intervals that are first-order correct, i.e., correct to order  $(1/\sqrt{T})$  when  $T$  is the sample size. For further improvements, especially for non-standard statistics similar to (2), bootstrap theory suggests that a 95%  $s$ -boot confidence interval may need further adjustments to make sure that it covers the unknown parameter with the correct coverage probability of 0.95

One such adjustment called studentized bootstrap (boot- $t$ ) has been developed.<sup>5</sup> It is implemented as follows. The first step is to convert the  $J = 999$   $s$ -boot values into a studentized scale by subtracting the mean and dividing by the standard deviation of the  $J$  resampled values. Next, we select the 25-th and 975-th ordered values from the studentized transformed scale. Finally, we undo the studentization by multiplying 25-th and 975-th values by the standard deviation and adding the mean. Davison and Hinkley (1997, p. 212) offer an elegant proof that the boot- $t$  is second-order correct, in the sense that the error in coverage by the estimated confidence interval of the true parameter is of order  $(1/T)$ . This second-order accuracy result depends on availability of a reliable estimate of the standard deviation used as the denominator of the studentization transformation.

Another method of improving the coverage probability of confidence intervals for nonstandard statistics similar to those in (2) is bootstrapping the bootstrap, or  $d$ -boot.<sup>6</sup> The  $d$ -boot involves  $K$  further replications of each  $j = 1, 2, \dots, J$ . We choose  $K = 174$  replications based on optimal choice suggested in the literature and described in McCullough and Vinod (1998) who give a step-by-step description of the  $d$ -boot. Clearly, the  $d$ -boot is computationally intensive, since it will require 174 times 999 computations of the underlying statistic; indeed a supercomputer may be needed to study the properties of  $d$ -boot by simulation.<sup>7</sup>

Roughly speaking, the method used in the  $d$ -boot is the following. If the original estimate is  $\hat{\theta}$  and the single bootstrap estimate is denoted by  $\theta_j$  we resample each  $\theta_j$   $K$  times and denote the resampled estimates as  $\theta_{jk}$ . One keeps track of the proportion of times  $\theta_{jk}$  is less than or equal to  $\hat{\theta}$ . The theory of  $d$ -boot proves that under ideal conditions, the probability distribution of the proportions in which  $\theta_{jk}$  is less than or equal to  $\hat{\theta}$  must be a uniform random variable. The appeal of the double

5. An excellent up-to-date description of the related theory with examples is given in Davison and Hinkley (1997).

6. Vinod (1995) is one of the first applications of  $d$ -boot in econometrics.

7. Letson and McCullough (1998) have implemented such a simulation and shown its superiority. Vinod (1998) also uses  $d$ -boot for refined inference.

bootstrap method is that when conditions are not ideal, the resulting non-uniform distribution remains easy to handle. All that happens is that the probabilities 0.025 and 0.975 occur at some nearby numbers (0.028 and 0.977, say). For adjusting the s-boot confidence interval the d-boot involves choosing the nearby (e.g., 28-th and 977-th) order statistics instead of the 25-th and 975-th values indicated by s-boot. The point in using the d-boot is that one does not need to try to know the form of the non-normal sampling distribution of the statistic. In fact the distribution need not have any known form. The statistics based on (2) may not have any known form when the normality assumption is relaxed.

We now turn to an application of our approach where we examine the confidence intervals on the Sharpe and Treynor performance measures for several well-known mutual funds. In order to illustrate the usefulness of the procedure we examine the confidence intervals of similarly rated funds. Hence, we use two sets of "growth" mutual funds for our data.<sup>8</sup> Our first set is made from two mutual funds, the Fidelity Magellan and The Neuberger/Berman Partners, which have similar Sharpe measures. Our second set also has two funds, Putnam Vista and Elfun Trust, which have similar Treynor measures. Excess monthly returns for the funds were calculated by subtracting the monthly three-month T-Bill rates of returns from the monthly mutual fund returns. The data spans from January 1978 to December 1997 (240 monthly observations) and is taken from the Morningstar Principia program.

### 3.3.1 The Sharpe Measure

Table 1 reports estimates of the mean monthly excess returns,  $\bar{r}_i$ ; standard deviation of the excess monthly returns,  $s_i$ ; the Sharpe measure point estimates,  $\hat{S}h_i$ ; and the confidence intervals for the three bootstrap methods for the of the first set of funds (Fidelity Magellan and Neuberger/Berman Partner). On the basis of the Sharpe measure, the Fidelity Magellan Fund performance would be ranked just slightly higher than the Neuberger/Berman Fund. In terms of confidence intervals, the two funds have similar 95 percent confidence interval widths for both the single bootstrap and studentized bootstrap approaches. Only with the double bootstrap is a considerable difference revealed: the confidence interval width of the Neuberger/Berman Fund is twenty percent larger than the Magellan fund. Hence, the Magellan fund not only has a higher performance measure, but also we can be slightly more confident in the accuracy of this point estimate.

8. We follow Morningstar's (1998) definition of a growth fund.

3  
F  
H  
r  
e  
T  
re  
fu  
th  
Su  
rel  
  
3.  
Th  
use  
—  
9.  
Mo

**Table 3.1**  
Confidence Intervals (C.I.) for the Sharpe Performance Measure

Sharpe Performance Measures			
Fund Name	$\bar{r}_i$	$s_i$	$\hat{S}h_i$
Fidelity Magellan	1.3868	5.4104	0.2563
Neuberger/Berman Partners	0.8727	3.7414	0.2333
Single Bootstrap Confidence Intervals			
Fund Name	$\hat{S}h_i$	C. I.	Interval Width
Fidelity Magellan	0.2563	(0.1212, 0.4135)	0.2923
Neuberger/Berman Partners	0.2333	(0.0995, 0.3824)	0.2829
Studentized Bootstrap Confidence Intervals			
Fund Name	$\hat{S}h_i$	C.I.	Interval Width
Fidelity Magellan	0.2563	(0.1045, 0.3968)	0.2923
Neuberger/Berman Partners	0.2333	(0.0883, 0.3711)	0.2829
Double Bootstrap Confidence Intervals			
Fund Name	$\hat{S}h_i$	C.I.	Interval Width
Fidelity Magellan	0.2563	(0.2421, 0.2654)	0.0233
Neuberger/Berman Partners	0.2333	(0.2174, 0.2454)	0.0280

### 3.3.2 The Treynor Measure

For the Treynor measure we run the following regression  $R_{i,t} - R_{f,t} = \alpha_i + \beta_i(R_{m,t} - R_{f,t}) + \varepsilon_t$ , where  $R_{i,t}$  are the monthly returns of the mutual fund,  $R_{f,t}$  are the 3-month T-Bill rates, and  $R_{m,t}$  is the S&P 500 monthly returns. Table 2 lists the mean excess returns, standard deviations, results of the regressions, Treynor measures,  $\hat{T}r_i$ , and confidence intervals for the Putnam Vista and Elfun Trust Funds. The results show that despite its slightly lower Treynor measure, the Elfun Trusts fund has a much narrower confidence interval width than the Putnam Vista Fund; the Putnam Vista interval width is 75 percent larger than the Elfun Trust fund. Such information, if available to the investor, can help them better determine the relative performances of various portfolios.<sup>9</sup>

### 3.4 Hypothesis Testing

The bootstrap approach on the Sharpe and Treynor measures can also, of course, be used for hypothesis testing. In this section of the paper we discuss JK's (1981) efforts

9. An analytic approach to generating confidence intervals has been developed by Morey and Morey (1999).

**Table 3.2**  
Confidence Intervals (C.I.) for the Treynor Performance Measure

Descriptive Statistics			
Fund Name	$\bar{r}_i$	$s_i$	
Putnam Vista	0.9095	4.9465	
Elfun Trust	0.8224	4.1833	
Regression Analysis			
Results from running $R_{i,t} - R_{f,t} = \alpha_i + \beta_i(R_{m,t} - R_{f,t}) + \varepsilon_t$ . The standard errors are in parenthesis; a * indicates the variable is significantly different from zero at the 5 percent level of significance.			
Fund Name	$\hat{\alpha}_i$	$\hat{\beta}_i$	$R^2$
Putnam Vista	0.1284 (0.1476)	1.0329* (0.0341)	0.7937
Elfun Trust	0.1071 (0.0723)	0.9460* (0.0167)	0.9307
Single Bootstrap Confidence Intervals			
Fund Name	$\hat{T}r_i$	C.I.	Interval Width
Putnam Vista	0.8805	(0.6076, 1.1567)	0.5491
Elfun Trust	0.8693	(0.7266, 1.0264)	0.2998
Studentized Bootstrap Confidence Intervals			
Fund Name	$\hat{T}r_i$	C.I.	Interval Width
Putnam Vista	0.8805	(0.5961, 1.1452)	0.5491
Elfun Trust	0.8693	(0.7193, 1.0191)	0.2998
Double Bootstrap Confidence Intervals			
Fund Name	$\hat{T}r_i$	C.I.	Interval Width
Putnam Vista	0.8805	(0.8488, 0.8987)	0.0499
Elfun Trust	0.8693	(0.8573, 0.8858)	0.0285

to conduct hypothesis testing and then use the bootstrap to develop a hypothesis testing approach which improves upon their efforts.

JK (1981) developed hypothesis tests for whether one portfolio's performance measure is significantly different from another portfolio.<sup>10</sup> Their hypothesis test is not for the actual differences in the performance measures of the two portfolios, but for a certain transformation of the differences. The transformation is needed

10. Jobson and Korkie (1981) also develop a test statistic that examines multi-comparisons of one portfolio to many others. For the multiple comparison measure, a chi-square statistic is developed for a transformation of both performance measures. They found that the Sharpe multi-comparison measure was well behaved and that its power to detect differences increased as the number of portfolios increased. The Treynor multi-comparison measure, on the other hand, was not well behaved and, in general, Jobson and Korkie did not recommend its use. Moreover, Cadsby (1986) states that these have low power. Since the focus of our research is only on the single comparison method, we do not address multi-comparisons, although an extension from a comparison of two funds is quite feasible.

be  
de  
tra  
sta  
sho  
the  
reg  
Th  
der  
seri  
F

$H_0$   
 $H_a$

Sinc  
form  
beh.

$F(\hat{S})$   
 $F(\hat{C})$

Note  
that  
How  
vari  
vari  
test  
inste  
depe  
serio  
prod  
be la  
the c  
unde  
are r

11. Fe

because of random denominators in the original performance measures. They then derive the asymptotic distribution and approximate bias of the estimators of the transformed difference using a Taylor series expansion. This method, known in the statistics literature the delta method, requires that the original random variable should be normally distributed.<sup>11</sup> It should be noted that to use the delta method on the transformed performance measures, JK (1981) implicitly assume Taylor series regularity conditions, which state that ignoring higher order terms is appropriate. These regularity conditions are not always satisfied. For example, when random denominators are close to zero, the ratios diverge, and the higher order Taylor series terms also diverge. We now summarize their approach.

For two portfolios,  $k$  and  $n$ , they examine the following hypotheses:

$$H_{OS} : Sh_k - Sh_n = 0, \quad (3.3)$$

$$H_{OT} : Tr_k - Tr_n = 0 \quad (3.4)$$

Since the sampling behavior of these differences is nonstandard, they use the transformed  $F(\cdot)$  differences for the sample differences in the hope of finding better behaved statistics. The transformations used by JK (1981) are:

$$F(\hat{Sh}_{kn}) = s_n \bar{r}_k - s_k \bar{r}_n, \text{ and} \quad (3.5)$$

$$F(\hat{Tr}_{kn}) = \frac{s_{nm} \bar{r}_k}{s_m^2} - \frac{s_{km} \bar{r}_n}{s_m^2} \quad (3.6)$$

Note that (4a) is obtained from (3a) by multiplying through by  $s_k s_n$ . Also note that, when the null hypothesis is true, this multiplication makes no difference. However, when the alternative hypothesis is true, the difference in (3a) is a random variable and can assume any value. When we accept the null, we permit the random variable to assume all statistically insignificant values near zero. The power of the test depends on its performance for such departures from the null (e.g., 0.0001 instead of 0). The equivalence of (3a) and (3b) with (4a) and (4b), respectively, depends on the numerical magnitudes of variances and covariances and may be seriously compromised, especially in small samples. For example, zero times a large product of standard deviations  $s_k s_n$  can be ignored, but 0.0001 times  $s_k s_n$  may well be large. A truly equivalent test should maintain both the size and the power of the original test. Since the multiplication by  $s_k s_n$  changes the value of the statistic under alternative hypothesis, we assert that JK's tests based on (4a) and (4b) are not truly equivalent to the tests for (3a) and (3b). Hence, it does change the

11. For more on the delta method see Greene (1997, p.278).

power. Clearly, one needs regularity conditions which do not permit the random denominators to become zero.

JK (1981) constructed the means and standard errors for their test statistics (4a) and (4b) by using Taylor series approximations to order  $\frac{1}{T}$ . The means (expectations) of the transformed differences in (4a) and (4b) are

$$E(F(\hat{S}h_{kn})) \equiv (\sigma_n \mu_k - \sigma_k \mu_n) \left( 1 - \frac{1}{4T} + \frac{1}{32T^2} \right) \text{ and} \quad (3.7)$$

$$E(F(\hat{T}r_{kn})) = \left( \frac{\sigma_{nm}}{\sigma_m^2} - \frac{\sigma_{km}}{\sigma_m^2} \mu_n \right) \quad (3.8)$$

The mean and variance of asymptotic distributions of the transformed difference statistics are then defined. For the Sharpe transformed difference, the asymptotic distribution is approximately normal with the mean from (5a) and variance given by

$$\theta = \frac{1}{T} \left[ 2\sigma_l^2 \sigma_n^2 - 2\sigma_k \sigma_n \sigma_{kn} + \frac{1}{2} \mu_k^2 \sigma_n^2 + \frac{1}{2} \mu_n^2 - \frac{\mu_k \mu_n}{2\sigma_k \sigma_n} (\sigma_{kn}^2 + \sigma_k^2 \sigma_n^2) \right] \quad (3.9)$$

Since the expectation (5a) does not equal the population value, the sample statistic (4a) is biased. For the transformed Treynor difference, the asymptotic distribution is approximately normal with the mean equal to  $F(Tr_{kn})$ , i.e. (5b), and variance given by

$$\psi = \frac{1}{T\sigma_m^2} \left[ \sigma_k^2 \sigma_{nm}^2 + \sigma_n^2 \sigma_{km}^2 - 2\sigma_{km} \sigma_{kn} + \mu_k^2 (\sigma_n^2 \sigma_m^2 - \sigma_{nm}^2) + \mu_n^2 (\sigma_l^2 \sigma_m^2 - \sigma_{km}^2) - 2\mu_k \mu_n (\sigma_m^2 - \sigma_{km} \sigma_{nm}) \right] \quad (3.10)$$

Following Cadsby (1986) we have corrected an error in the original version of (7). Note that equations (6) and (7) involve unknown population means, variances and covariances. Estimators denoted by  $\hat{\theta}$  and  $\hat{\psi}$  are obtained by substituting sample estimators of means, variances and covariances.

Assuming asymptotic normality of  $F(\hat{S}h)$  and  $F(\hat{T}r)$  the test statistics on the null hypotheses:  $H_{OS} = Sh_{kn} = 0$ , and  $H_{OT} = 0$  are then:

$$A_{Skn} = \frac{F(\hat{S}h_{kn})}{\sqrt{\hat{\theta}}} \quad (3.11)$$

or

$$Z_{Tkn} = \frac{F(\hat{T}r_{kn})}{\sqrt{\hat{\psi}}} \quad (3.12)$$

In examining the distributions of these Z statistics, JK (1981) found that the Sharpe Z statistic (8a) was well behaved at small sample sizes. However, the Treynor measure (8b), was not well behaved. Additionally, in both the Sharpe and Treynor cases, they found that the power in detecting differences between portfolios was small.

Our approach is as follows. Instead of using a Taylor series (delta method) approximation to compute the standard errors, we use the bootstrap to conduct hypothesis testing. The advantages of the bootstrap over the JK method are twofold. One, our methodology considers the sampling distribution of the actual difference between Sharpe and Treynor performance measures themselves. By contrast, JK provide the expected value and a standard error for a transformed quantity:  $F(\hat{S}_{h_{kn}}) = s_n \bar{r}_k - s_k \bar{r}_k$ . Again, the reason why these authors used the transformation was that sampling properties of the original difference are not tractable by the Taylor series methods and the results are unstable and unreliable. The second advantage of our methods is that the bootstraps are robust or distribution-free. In other words, the assumption of normality, which is clearly invalid in our small sample case with a random denominator, is avoided.

To illustrate the bootstrap approach for hypothesis testing we use the same two sets of funds used in the confidence interval approach plus a third set, the Fidelity Magellan Fund and the Rainbow Fund.<sup>12</sup> This additional set allows a clear illustration of two funds which have significantly different performance measures using the bootstrap yet are not considered to be significantly different under the JK (1981) approach.

Table 3 shows, for both fund sets 1 and 3, the Sharpe measures and the difference in the two fund's Sharpe measures. Also, the table presents the 95 percent confidence intervals of the Sharpe measure differences for the single, studentized and double bootstrap methods, and the transformed Sharpe difference and corresponding confidence intervals using the JK approach. If the confidence interval of the difference does not contain zero, it indicates that there is a significant difference in the performance measures of the two funds.

The two sets of funds in Table 3 show some interesting results. The difference between the Sharpe measures of the Fidelity Magellan/Neuberger Berman Partners is quite small and accordingly the JK (1981), single and studentized bootstrap show no significant difference. The double bootstrap on the other hand, with its increased precision (see Letson and McCullough (1998)), does show a significant difference

12. The data period for the Rainbow fund is the same as the other funds and the excess returns are calculated by subtracting the three-month T-Bill rates.

**Table 3.3**  
Hypothesis Testing for the Sharpe Performance Measure

Two sets of funds presented: Set 1: Fidelity Magellan versus Neuberger/Berman Partners. Set 2: Fidelity Magellan versus Rainbow.

Data Period: January 1978 to December 1997. Frequency: Monthly.

*Set 1: Fidelity Magellan against Neuberger/Berman Partners*

Fund Name	Sharpe Measure
Fidelity Magellan	0.2563
Neuberger/Berman Partners	0.2333
Difference	0.0230
Bootstrap Method	Confidence Interval of the Difference in the Sharpe Measures
Single Bootstrap	(-0.1878, 0.2352)
Studentized Bootstrap	(-0.1878, 0.2351)
Double Bootstrap	(0.0008, 0.0422)*

Jobson and Korkie (1981)

Transformed Difference of the Sharpe Measures of the two funds: -0.4671

Confidence Interval of the Transformed Difference: (-0.6334, 1.5675)

*Set 2: Fidelity Magellan against Rainbow*

Fund Name	Sharpe Measure
Fidelity Magellan	0.2563
Rainbow	0.0404
Difference	0.2159
Bootstrap Method	Confidence Interval of the Difference in the Sharpe Measures
Single Bootstrap	(0.0240, 0.4189)*
Studentized Bootstrap	(0.0107, 0.4056)*
Double Bootstrap	(0.1911, 0.2296)*

Jobson and Korkie (1981)

Transformed Difference of the Sharpe Measures of the two funds: 5.6879

Confidence Interval of the Transformed Difference: (3.5283, 7.8475)\*

\* indicates the two funds are significantly different from each other

between the two funds. For the Fidelity Magellan/Rainbow set, which is located in the lower part of Table 3, the difference between the Sharpe measures is quite large. Correspondingly, there is a significant difference found in all of the forms of the bootstrap and in the JK (1981) approach as none of the confidence intervals contain zero.

Table 4 presents the 95 percent confidence intervals of the Treynor measure differences for JK (1981), single bootstrap, studentized bootstrap and double bootstrap methods. The results for the first set of funds—the set of funds which have similar Treynor measures—show that all 4 methods show no significant difference in the

J.  
T  
C  
\*  
T  
a  
p  
f  
s  
p  
s  
i  
n

**Table 3.4**  
Hypothesis Testing for Treynor Performance Measure

Two sets of funds presented: Set 1: Putnam Vista versus Elfun Trust. Set 2: Fidelity Magellan versus Rainbow.

Data Period: January 1978 to December 1997. Frequency: Monthly.

*Set 1: Putnam Vista and Elfun Trust*

Fund Name	Treynor Measure
Putnam Vista	0.8805
Elfun Trust	0.8693
Difference	0.0112

Bootstrap Method	Confidence Interval of the Difference in the Treynor Measures
Single Bootstrap	(-0.3166, 0.3511)
Studentized Bootstrap	(-0.3440, 0.3238)
Double Bootstrap	(-0.0297, 0.0253)

*Jobson and Korkie (1981)*

Transformed Difference of the Treynor Measures of the two funds: -0.2631

Confidence Interval of the Transformed Difference: (-14.162, 13.637)

*Set 2: Fidelity Magellan and Rainbow*

Fund Name	Treynor Measure
Fidelity Magellan	1.2065
Rainbow	0.2212
Difference	0.9853

Method	Confidence Interval of the Difference in the Treynor Measures
Single Bootstrap	(0.4861, 1.4709) <sup>13</sup>
Studentized Bootstrap	(0.4929, 1.4779) <sup>14</sup>
Double Bootstrap	(0.9267, 1.0462) <sup>15</sup>

*Jobson and Korkie (1981)*

Transformed Difference of the Treynor Measures of the two funds: 5.6820

Confidence Interval of the Transformed Difference: (-17.967, 29.259)

\* indicates the two funds are significantly different from each other

Treynor values of the two funds. For the other set, Fidelity Magellan/Rainbow, all forms of the bootstrap method illustrate a significant difference. The JK approach, however, indicates that there is no significant difference between the two funds; an extremely unsatisfactory result given the Fidelity Magellan Treynor measure is about 5 times the level of the Rainbow fund. The fact that bootstrap approach works relatively well for the Treynor performance measure is not surprising, since Jobson and Korkie's method is admittedly not well behaved for the Treynor measure.

### 3.5 Concluding Remarks

This paper has used the bootstrap methodology to construct confidence intervals and conduct hypothesis testing on the well-known Sharpe and Treynor portfolio performance measures. Due to the presence of random denominators in the definitions of the performance measures, and the difficulty in determining the sample size needed to achieve asymptotic normality, the Sharpe and Treynor performance measures have not allowed for the construction of confidence intervals. With this paper, however, we show that the bootstrap methodology can build confidence intervals around the point estimates of these measures, allowing investors further information on the performance of their portfolios. We have illustrated the use of the single, studentized and double bootstrap by constructing confidence intervals on the Sharpe and Treynor measures of several actual growth mutual funds.

The bootstrap methodology also allows for simple hypothesis testing for comparing one portfolio's Sharpe or Treynor measure against another portfolio. Since the bootstrap is largely distribution-free and requires no transformation of the difference of the performance measures of the two funds, the paper extends and improves upon the efforts of Jobson and Korkie (1981).

### References

- Cadsby, C. B., 1986, "Performance Hypothesis Testing with the Sharpe and Treynor Measures: A Comment," *Journal of Finance*, 41, no. 5, 1175-1176.
- Davison, A. C. and D. V. Hinkley, 1997, *Bootstrap Methods and Their Applications*, New York: Cambridge University Press.
- Greene, W. H., 1997, *Econometric Analysis*, 3rd ed., Upper Saddle River, NJ: Prentice Hall.
- Jobson, J.D. and B.M. Korkie, 1981, "Performance Hypothesis Testing with the Sharpe and Treynor Measures," *Journal of Finance*, 36, no. 4, 889-908.
- Judge, G. G., R.C. Hill, W. E. Griffiths, H. Lutkepohl and T-C Lee (1988) *Introduction to the Theory and Practice of Econometrics*. New York: Wiley.
- Letson, David and B. D. McCullough, 1998, "Better Confidence Intervals: The Double Bootstrap with No Pivot," *American Journal of Agricultural Economics*, 80, 552-559.
- McCullough B. D. and H. D. Vinod, 1998, "Implementing the Double Bootstrap," *Computational Economics*, 12, 79-95.
- Morey, Matthew R. and Richard C. Morey, 1999, "An Analytical Confidence Interval for the Treynor Index: Formula, Conditions and Properties," *Journal of Business Finance and Accounting*, forthcoming.
- Morningstar Principia Manual, Chicago, IL, 1998.
- Roy S.N. and R.F. Potthoff, 1958, "Confidence Bounds on Vector Analogues of the 'Ratios of Means' and the 'Ratio of Variances' for Two Correlated Normal Variates and Some Associated Tests," *Annals of Mathematical Statistics*, vol. 29, 829-841.

Sharpe, W. F., 1966, "Mutual Fund Performance," *Journal of Business*, vol. 39, No. 1, Part 2, January, 119-138.

Treynor, J. L., 1965, "How to Rate Management of Investment Funds," *Harvard Business Review*, vol. 43, January-February, 63-75.

Vinod, H. D., 1993, "Bootstrap, jackknife resampling and simulation: applications in econometrics," in G.S. Maddala, C.R. Rao, and H.D. Vinod (eds.), *Handbook of Statistics: Econometrics*, 11, 629-661, New York: North Holland.

Vinod, H. D., 1995, "Double bootstrap for shrinkage estimators", *Journal of Econometrics*, 68, 287-302.

Vinod, H. D., 1998, "Foundations of statistical inference based on numerical roots of robust pivot functions (Fellow's Corner)", *Journal of Econometrics*, 86, 387-396.