An Alternative Route to Performance Hypothesis Testing
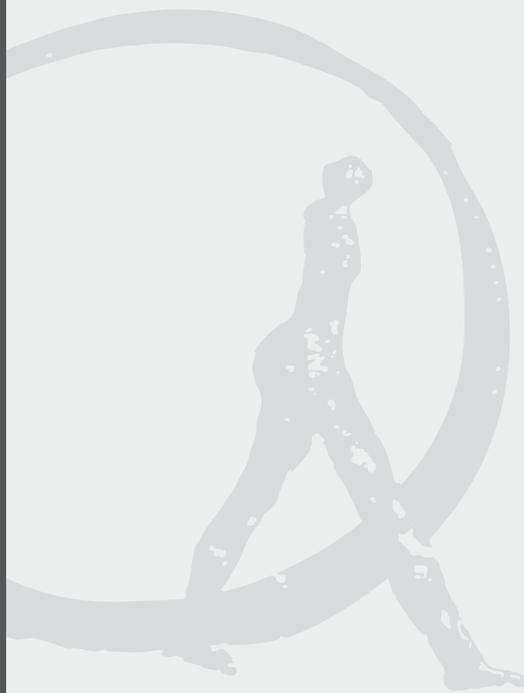
November 2009

Bernhard Scherer
Professor of Finance, EDHEC Business School

EDHEC
BUSINESS SCHOOL

Abstract

A wide variety of risk-return ratios is routinely reported in sales pitches as well as academic publications. Few attempts have been made, however, to look at the small sample distributions of these estimators in order to derive confidence bands. The reason for this has been the extreme difficulty of working out the required statistics for most risk-return ratios. Rather than following classic statistics, this paper relies on a general and robust method which not only provides confidence intervals for arbitrary risk-return ratios, sample sizes and distribution, but is also fairly easy to implement.

EDHEC is one of the top five business schools in France. Its reputation is built on the high quality of its faculty and the privileged relationship with professionals that the school has cultivated since its establishment in 1906. EDHEC Business School has decided to draw on its extensive knowledge of the professional environment and has therefore focused its research on themes that satisfy the needs of professionals.

EDHEC pursues an active research policy in the field of finance. The EDHEC-Risk Institute carries out numerous research programmes in the areas of asset allocation and risk management in both the traditional and alternative investment universes.

1. Introduction

Reported risk–return ratios relate average returns to alternative measures of risk, and hence involve the ratio of a random (owing to sampling error) nominator and denominator. As such, point estimates of these ratios are easy to calculate, but confidence intervals are much more difficult to arrive at. Confidence intervals are needed, however, for any kind of statistical inference and decision making. While Lo (2002) has shown that an asymptotic distribution exists for the Sharpe ratio, this result provides a special case rather than a general concept. Traditional methods only work well if the sampling distribution of the statistic in question is asymptotically normally distributed. There is little guidance on the small sample behaviour of risk-adjusted performance measures or how many data points one needs to rely on for asymptotic results. Useful exceptions are Pedersen and Satchell (2000), Miller and Gehr (1978) and Jobson and Korkie (1981). Moreover, these analytical solutions are either extremely difficult to work out or simply do not exist for modifications of the popular Sharpe ratio which focus more on downside risk. An example is the well-known Sortino ratio (it relates average return to standard deviation of downside returns). A general method is needed which provides confidence intervals for arbitrary risk–return ratios, sample sizes and distributions.

2. Bootstrapping theory as an alternative

Suppose a series of returns (total return minus risk-free rate) $r_1, r_2, \ldots, r_m$ is observed. *Ex post* risk–return ratios ($\hat{s}$) are calculated as the ratio of average return per unit of risk. This paper will focus on Sharpe and Sortino ratios (Sharpe, 1994; Sortino and Price, 1994). Both ratios differ with respect to the employed risk measure. The Sharpe ratio applies a symmetric risk concept, equally penalising (squaring) downside and upside deviations from the sample mean return, while the Sortino ratio only includes negative performance in their calculation of squared returns. The Sortino ratio is included for three reasons. First, it provides a better capture of risk if returns are non-normally distributed, as is the case for hedge fund returns. Secondly, it is known to suffer more from estimation error, as it uses less data (extreme returns are by definition rare). Thirdly, no large sample approximations exist. The sample calculations for both ratios are given below.

$$\text{Sharpe ratio} = \frac{\dfrac{1}{m}\sum_{i=1}^{m} r_i}{\sqrt{\dfrac{1}{m-1}\sum_{i=1}^{m}(r_i - \bar{r})^2}} \qquad (1)$$

$$\text{Sortino ratio} = \frac{\dfrac{1}{m}\sum_{i=1}^{m} r_i}{\sqrt{\dfrac{1}{m-1}\sum_{i=1}^{m} I(r_i < 0)(r_i)^2}} \qquad (2)$$

where $I(r_i < 0)$ denotes the indicator function. High ratios are preferable (everything else equal) as they indicate a better return per unit of risk taken. If the small sample distribution of $\hat{s}$ is far from normal, classical methods are biased and unreliable. In any case, the analytical formula for the large sampling distribution of (1) is extremely hard to come by, while it is unknown for (2). In order to overcome the above problem, bootstrapping techniques are relied on. Resampling treats the current sample as a good approximation of the true distribution (in the absence of further information, it is the best available). It then repeatedly draws from the empirical distribution to recalculate the statistic of interest many times to arrive at the bootstrap sampling distribution that can now be used for hypothesis testing. Suppose one is given monthly returns on the HFR fund of funds index ranging from January 1990 to April 2003 (160 observations). The JPM one-month

cash rate from DataStream is used to calculate a risk-free return. The bootstrapping procedure is as follows. Randomly draw 160 (original sample size) returns with replacement from the original sample. Calculate a new risk–return ratio $\hat{s}_b^\star$ based on the resampled returns. Repeat this procedure for $b = 1, \ldots, B$ times arriving at $\hat{s}_1^\star$, $\hat{s}_1^\star$, L, $\hat{s}_{b,L}^\star$ $\hat{s}_B^\star$ resampled ratios. The bootstrap sampling distribution of $\hat{s}_b^\star$ can now be taken to judge whether the sampling distribution $\hat{s}$ of in small samples is normal (and hence traditional approaches might not be so bad after all) or not. With $B = 1,000$, the following results are obtained (see Figure 1).[1] While the Sharpe ratio is well behaved (all resampled realisations plot on a straight line, equalising hypothetical and empirical percentiles), the same cannot be said about the Sortino ratio.

Deviations from normality are large at both ends. Normal (small sample) approximations look reasonable only for the traditional Sharpe ratio. They seem to be largely misleading, however, for the Sortino ratio. Taking the 2.5th and 97.5th percentile, one can now arrive at a symmetric 95% confidence band $CI(\hat{s}_{2.5\%}^\star, \hat{s}_{97.5\%}^\star)$. Note the apparent non-normality of the sample distribution for the Sortino ratio in Figure 2. As suspected, the Sortino ratio shows a much larger dispersion in resampled outcomes and hence estimation error. The corresponding values can be cut off from the histograms in Figure 2. For the Sortino (Sharpe) ratio, the 95% interval ranges from 0.11 to 0.92 (0.08–0.41). Both ratios are significantly different from zero.

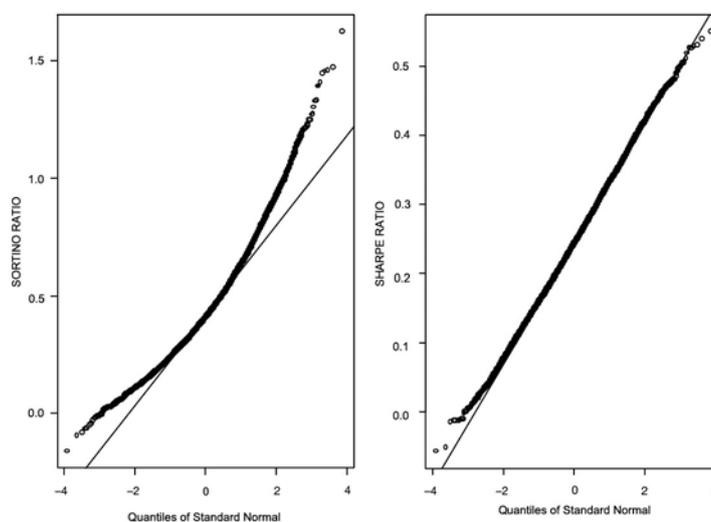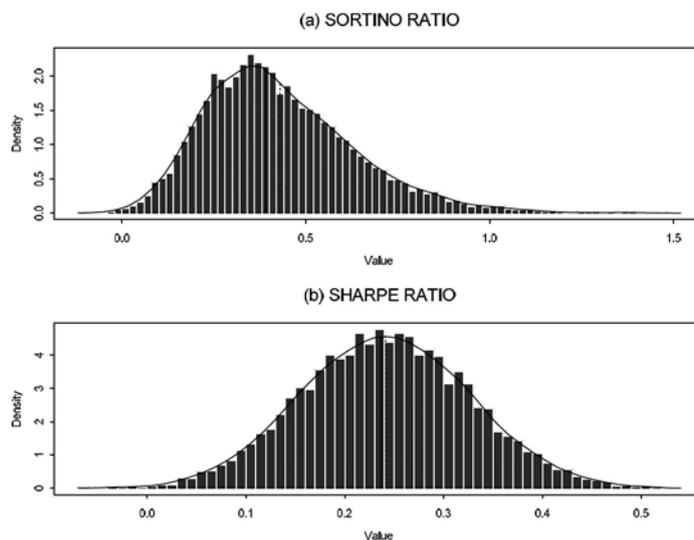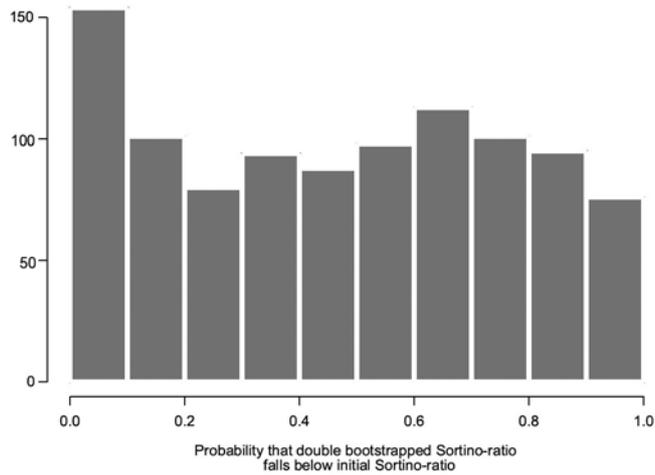Figure 1: QQ-plot for bootstrapped sampling distribution

Figure 2: Bootstrapped sampling distributions

Figure 3: Sample histogram of $u_b = \frac{1}{Z}\Sigma_{z=1}^Z I(\hat{s}_{bz}^{\star\star} < \hat{s})$

Probability that double bootstrapped Sortino-ratio
falls below initial Sortino-ratio

3. Increasing the coverage ratio with the double bootstrap

So far, reliance has been on the 95% interval from a simple bootstrap procedure. It is not guaranteed, however, that the 95% confidence band calculated above indeed covers the true ratio with 95% probability. The true ratio might only fall 85% of all times into the estimated confidence band. One method of increasing the coverage probability is the double bootstrap, which can be thought of as bootstrapping the bootstrap. The double bootstrap as described by Nankervis (2002) involves the following calculations. Perform the simple bootstrap as described above. Save all $b = 1, \ldots, B$ resampled data sets as well as resampled ratios $\hat{s}_b^{\star}$. This is called the first stage resampling. For each of the B resampled datasets, start a second round of $z = 1, \ldots, Z$ resamples, leading to a total of BZ resamples denoted as $\hat{s}_{bz}^{\star\star}$. For each $\hat{s}_b^{\star}$, there exists a new set of Z resampled $\hat{s}_{b1}^{\star\star}, \mathrm{L}, \hat{s}_{bZ}^{\star\star}$ ratios $\hat{s}_{b1}^{\star\star}, \mathrm{L}, \hat{s}_{bZ}^{\star\star}$. These are the second stage resamples. For each $\hat{s}_b^{\star}$, calculate the percentage of second stage resamples $\hat{s}_{b1}^{\star\star}, \mathrm{L}, \hat{s}_{bZ}^{\star\star}$ that fall below the original sample estimate of the risk–return ratio $\hat{s}$, ie calculate

$$u_b = \frac{1}{Z}\sum_{z=1}^Z I(\hat{s}_{bz}^{\star\star} < \hat{s}) \qquad (3)$$

Choose $B = 1,000$ and $Z = 200$. Under ideal conditions u_b follows a uniform distribution. Figure 3 shows that this assumption is clearly violated for the double bootstrapped Sortino ratios $\hat{s}_{b1}^{\star\star}, \mathrm{L}, \hat{s}_{bZ}^{\star\star}$. Formal tests such as the Kolmogorov and Smirnov test as well as the χ^2 adjustment test provide p values close to 0%. Hence, the null hypothesis that Figure 3 comes from a uniform distribution can be safely rejected.

Finally, calculate the 2.5th and 97.5th percentiles of u_b. Use these values to adjust the first stage resample confidence band to $CI(\hat{s}_{u2.5\%}^{\star}, \hat{s}_{u97.5\%}^{\star})$. With the double bootstrap, the confidence interval changes to a much higher bound of 0.18 (instead of 0.11) representing the 8% quantile (rather than the 2.5% quantile). In effect, one gets $CI(\hat{s}_{u7.9\%}^{\star} = 0.18, \hat{s}_{u96\%}^{\star} = 0.96)$.

## 4. What if one needs to deal with autocorrelated data?

So far, it has been implicitly assumed that return data $r_1, r_2, \ldots, r_m$ have been drawn independently, *i.e.*, neither returns nor volatilities are autocorrelated. Bootstrapping implicitly removes any time dependence by its very definition. All draws are unconditional on the result of the previous draw. After a strong positive draw, there is no mechanism that would favour another positive draw (in case of positive autocorrelation) or another large draw (in case of ARCH effects). Many financial time series (real estate, corporate bonds, high yield, hedge funds), however, show strong autocorrelation due to the illiquidity (infrequent trading) of the underlying instruments. Bootstrapping fails if one does not account for return dependence.

Figure 4: Autocorrelation for hedge fund series

(a) RAW DATA   (b) FILTERED DATA
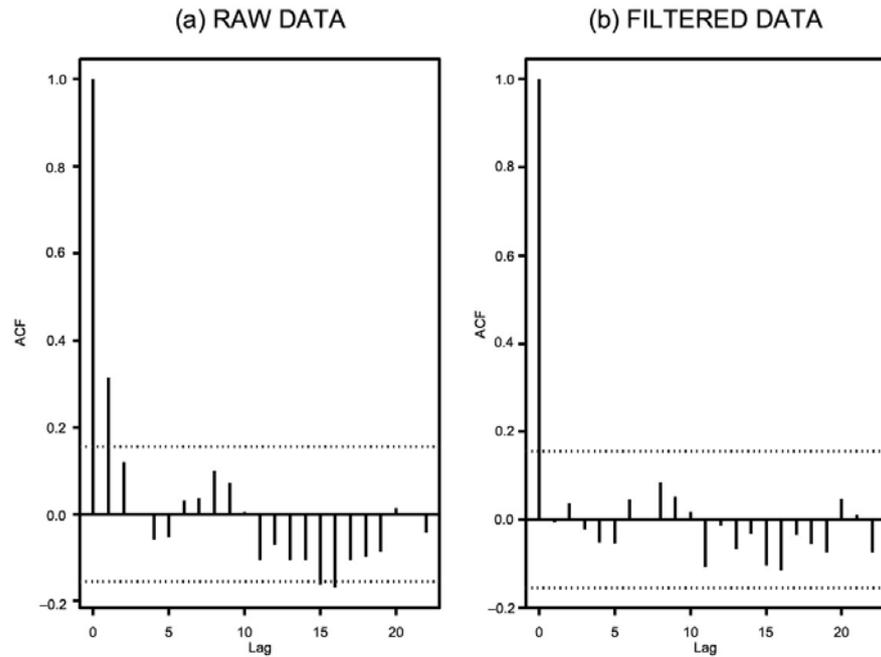
For the data analysed so far, the autocorrelation function can be plotted as in Figure 4(a). The first-order autocorrelation coefficient amounts to 0.31 and is well outside the confidence limits (given by dotted line), *i.e.*, it is statistically significant.

Figure 5: Lower confidence bound (2.5%) on Sharpe ratio for raw and adjusted data
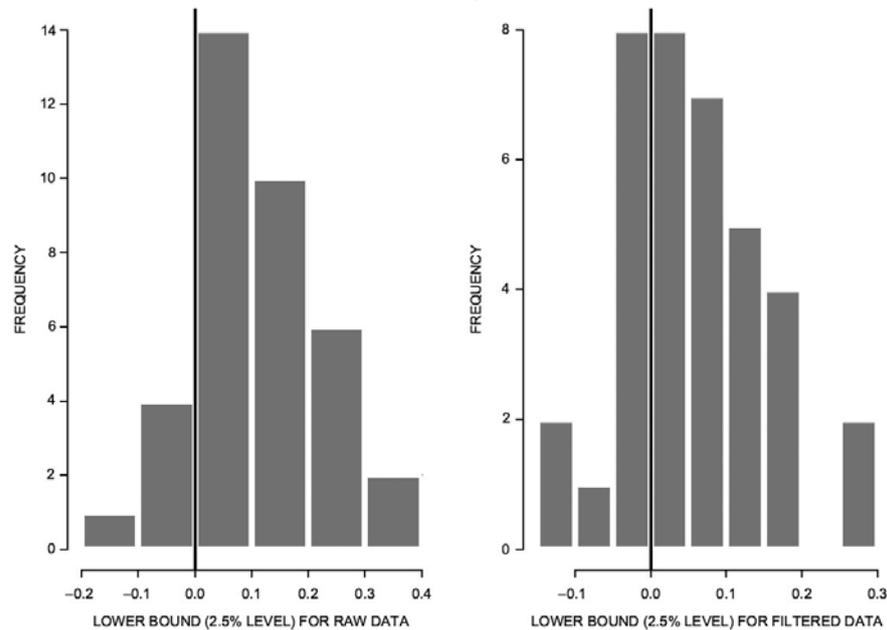
One way to deal with this issue is to remove the autocorrelation using a simple filter that adjusts for the inherent AR(1) process (lag one is statistically significant, while lag two is not) and reapply the methods established in the previous sections to the transformed series. It is known that the first-order autocorrelation can be removed. Estimate the AR(1) coefficient from a linear regression

$$r_t = \theta_0 + \theta_1 r_{t-1} + \varepsilon_t \quad (4)$$

Save the regression coefficient θ_1 and create a new filtered series $r_t^{\star}$ according to

$$r_t^{\star} = \frac{1}{1-\theta_1} r_t - \frac{\theta}{1-\theta_1} r_{t-1} \quad (5)$$

Obviously, this represents a break with the non-parametric approach used so far. As a compromise, one might want to estimate Equation (4) using robust methods. The above procedure will leave the average return of a series constant, but effectively increase (decrease) its variance for positive autocorrelation (negative correlation) as described in Scherer (2002). The effect of this procedure can be seen in the right-hand autocorrelation plot in Figure 4(b). Virtually all significant autocorrelation (confidence bounds are given by dotted lines) has been removed. After the removal of autocorrelation, one can proceed as in the previous section. Rather than focusing on a single series, all HFR series are used for the described frequency and data period (see the HFR website for a description of the underlying data series).

Figure 5 plots the distribution of lower confidence bounds (2.5% percentile) of Sharpe ratios for raw data as well as adjusted (filtered) data. The Sharpe ratios of 5 out of 37 series (expected: 2.5% x 37≈1) are not statistically different from zero. This number increases to 11 if filtered data (with higher volatility) are used. If this is combined with a suspicion of upward bias in hedge fund returns, many hedge fund styles fail to show statistically significant Sharpe ratios of even monthly returns. An almost identical picture can be found for the Sortino ratio in Figure 6. Again, the historical track record of the hedge fund industry as a whole leaves some doubt about the industry's ability to create added value.

Figure 6 Lower confidence bound (2.5%) on Sortino ratio for raw and adjusted data
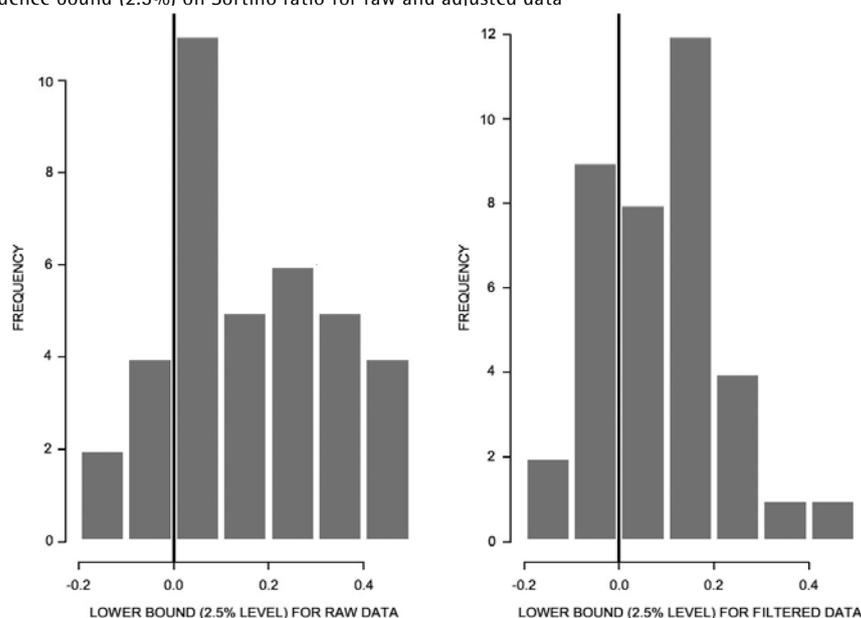
5. Summary
Renewed interest in the significance of risk–return ratios has been focusing on closed form solutions for the well-known Sharpe ratio. This paper provided a robust methodology for evaluating the properties of the Sharpe ratio's sampling distribution, as well as how to derive confidence intervals without having to rely on asymptotic approximations (as in reality, samples are small). It has also been shown that the sampling distribution of other risk–return measures for which asymptotic results do not exist might be highly non-normal. While the double bootstrap methodology leads to significantly refined confidence bands, a more realistic modelling of hedge fund returns leaves the case for hedge funds less optimistic than providers of these services might wish.

References
• Jobson, J. and Korkie, B. (1981) 'Performance Hypothesis Testing with the Sharpe and Treynor Measures', *Journal of Finance*, 36, 889–908.

• Lo, A. W. (2002) 'The Statistics of Sharpe Ratios', *Financial Analysts Journal*, July/August, 58(4).

• Miller, R. and Gehr, A. (1978) 'Sample Bias and Performance Measures: A Note', *Journal of Financial and Quantitative Analysis*, 13, 943–6.

• Nankervis, J. (2002) 'Stopping Rules for Double Bootstrap Confidence Intervals, University of Surrey, http://www.bus.qut.edu.au/esam02/program/papers-/Nankervis_John.pdf.

• Pedersen, C. and Satchell, S. (2000) 'Small Sample Measures of Performance Measures in the Asymmetric Response Model', *Journal of Financial and Quantitative Analysis*, 35(3), 425–50.

• Scherer, B. (2002) *Portfolio Construction and Risk Budgeting*, Riskwater, London.

• Scherer, B. and Martin, D. (2003) *Portfolio Optimization Using Nuopt for S-Plus*, Springer, New York.

• Sharpe, W. F. (1994) 'The Sharpe Ratio', *Journal of Portfolio Management*, Fall, 49–58.

• Sortino, F. A. and Price, L. N. (1994) 'Performance Measurement in a Downside Risk Framework', *Journal of Investing*, Fall.

—