How Consistent are the Judges of Portfolio Performance?

Matthew Brigida[*] & Chin Yang [†]

June 17, 2014

Abstract

This analysis tests whether the rank ordering of a set of portfolios, with varying number of assets, differs depending on the performance measure used. To test this hypothesis, we construct a Friedman test over 4,950 portfolios (50 portfolios for each portfolios sizes 2 to 100) of S&P 500 stocks. We find, that while measures may differ in their sensitivity to portfolio size, all measures consistently prefer larger portfolios. Therefore, the rank ordering of a set of portfolios of various sizes is consistent across performance measures.

1 Introduction

The goal of this analysis is to compare the rankings of four popular portfolio performance measures as portfolio size increases. These measures considered are the Sharpe ratio (Sharpe (1966)), information ratio (Goodwin (1988)), Treynor ratio (Treynor (1965)), and Jensen's alpha (Jensen (1969)). Note, these are four unique measures—they are not linear transformations of each other.

[*] Associate Professor of Finance, Clarion University of Pennsylvania, 840 Wood St., Clarion, PA 16214, Email: mbrigida@clarion.edu or mdbrigida@gmail.com

[†] Professor of Economics, Clarion University of Pennsylvania

This analysis is intended to assist practitioners (portfolio managers, and other investment professionals) when ranking portfolios of differing sizes. Specifically, it will give them a better understanding of how a varying number of assets in a set of potential portfolios will affect the rank ordering of the set provided by each performance measure. Thus, given a set of portfolios of differing sizes, this analysis will answer the question: Is the rank ordering of this set of portfolios dependent on the performance measure used?

Note our focus on ranks and not the value of the performance measure. First, values of one performance measure cannot be compared to another. Moreover, for practical purposes, it is the ranking of the portfolios which matters, and not the comparative values of the performance measures. So to compare the behavior of various measures, we compare the respective rankings they provide, and test whether these rankings are statistically different among performance measures.

Previous studies have shown portfolio managers may trade in such as way as to increase the value of performance measures, despite not having any private information. For example Ferson and Siegel (2001) and Lhabitabt (2000) use separate methods to increase the value of a portfolio's Sharpe ratio. We don't consider such trading strategies, since the portfolio manager is using the measures as a way to rank portfolios for potential investment.

2 Portfolio Construction

Our data set consists of daily log returns for every stock in the S&P 500 index, as well as returns for major indices themselves (S&P 500, Wilshire 5000, Dow Jones Industrial Average, etc.) over the three years beginning January 2010 and ending December 2012. The log returns were calculated using stock price data from Interactive Brokers.

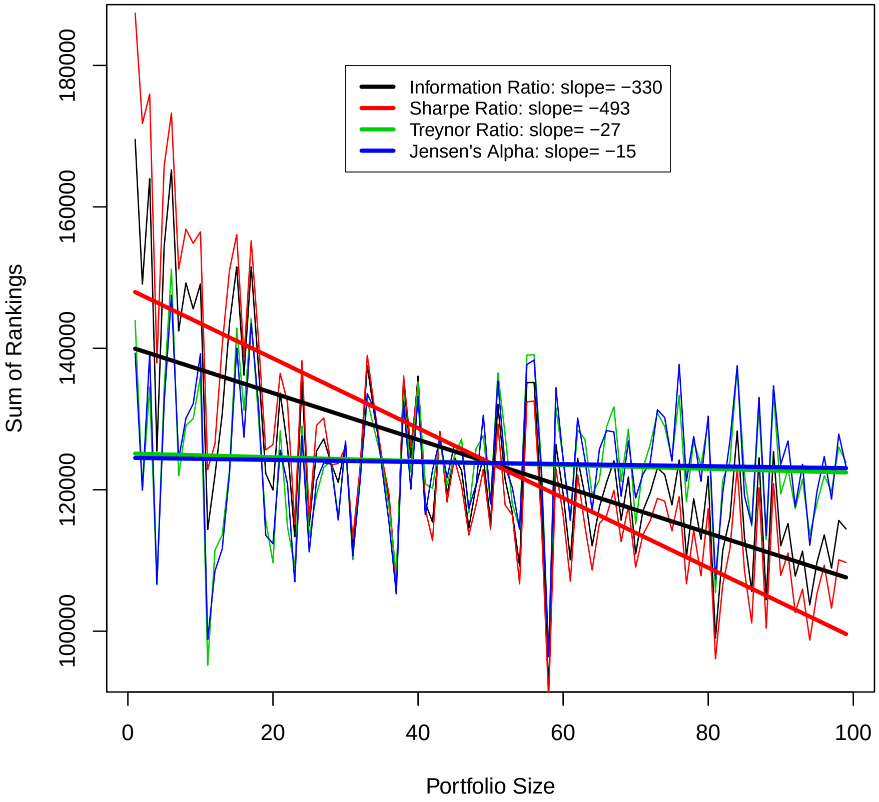
First we created 50 random portfolios of size two from the set of stocks in the S&P 500. We then calculate each performance measure for each of the 50 random portfolios. We repeat this procedure for portfolio sizes up to 100. This gives us the performance measure values for each of the 50 random portfolios for each portfolio size 2 to 100 (4,950 portfolios).

We then rank the 4,950 portfolios using each measure. So for each performance measure we have a ranking of the random portfolios from 1 to 4,950, where 1 denoted the portfolio with the highest ranking (most preferred). Letting $r_{n,i}$ denote the rank of the i^{th} portfolio of size n (where $2 \leq n \leq 50$). Then $r_{n,i}$ is an integer between 1 and 4,950 inclusive.

Lastly for each portfolio size we sum the ranks of the portfolios of that size (for each n denote this as $\varphi_n = \sum_{i=1}^{50} r_{n,i}$). So for all 99 portfolio sizes we have a sum of the rankings of the 50 random portfolios of that size (φ_n) for each portfolio measure.

This methodology allows us to compare the how each portfolio measure ranking changes as the portfolio size changes. A plot of the sum of each ranking (φ_n) for each measure, over the 99 portfolio sizes, is in figure 1 below. Figure 1 also contains a linear trendline showing how each portfolio measure is affected by increasing portfolio size.

Figure 1: Sum of portfolio rankings over portfolio sizes 1 to 100, with trendlines. The negative slope of each trendline shows that each portfolio measure tends to rank larger portfolios higher.



The figure shows that the Sharpe and Information ratios are more sensitive to portfolio size than the Treynor ratio and Jensen's alpha. We see that, on average, φ_{n+1} is 493 and 330 less than φ_n for the Sharpe and Information ratios respectively. This compares with 27 and 15 for the Treynor ratio and Jensen's alpha respectively.

Note that a decrease in φ_n means an increase in rank (because for each portfolio measure the highest value has rank 1 and lowest value rank 4,950). Thus, our results so far are evidence that Sharpe and Information ratios have a tendency to give a higher rank to larger portfolios. The Treynor ratio and Jensen's alpha have the same tendency, though to a much less degree.

3 The Friedman Test

In this section we will formally test whether the portfolio measures form consistent rankings of the portfolios. We may think of each portfolio measure as a judge, and each portfolio as a contestant. We wish to test whether each judge will rank the contestants consistently across portfolio sizes. Or rather, we are testing whether a particular judge has a bias toward large, or small, portfolios.

To test this we'll use the Friedman test (Friedman 1937, 1939, 1940) which is designed to test whether judges' ranking are consistent with one another. A classic example is to test whether n judges rank k wines consistently. In our case the n judges are our portfolio measures, and the k wines are our 4,950 portfolios. Note, the Friedman test is nonparametric, which is necessary given nonnormality of some performance measures.

To conduct the Friedman test we sum the sum of rankings over each portfolio measure, for each portfolio size. For example, consider the 3 stock portfolio size in table 1 below. We sum each column across the performance measures, and have a sum of $151,887 + 135,609 + 116,304 + 105,176 = 508,976$. For each portfolio size we denote this value as ϖ_n.

The null hypothesis is that the average ranking is equal across all portfolio sizes, that is $\varpi_2 = \varpi_3 = \cdots = \varpi_{100}$. Under this null hypothesis our test statistic $Q = \frac{12S}{nk\lambda^2(n\lambda+1)} \sim \chi^2_{n-1}$, where n is the number of portfolios (here 99), k is the number of performance measures (here 4), λ is the number of repetitions (here 50), and $S = \sum_{\varpi_i=2}^{100} \varpi_i^2 - \frac{k^n\lambda^2(n\lambda+1)^2}{4}$.

Table 1. Structure of Friedman Test

Measure/Portfolio Size	2	3	4	...	100
Sharpe ratio	171,463	151,887	142,251	...	104,580
Information ratio	147,382	135,609	131,635	...	108,492
Treynor ratio	118,937	116,304	129,489	...	116,066
Jensen's alpha	114,552	105,176	122,702	...	119,816
ϖ_n	552,334	508,976	526,077	...	448,954

If we reject the null of the Friedman test, this is evidence that the judges consistently rank the contestants. Conversely if we cannot reject the null this is evidence that the rankings are inconsistent.

3.1 Results

Using the Friedman test, from portfolio size 2 to 100, we reject the null at the less than 0.1% level of significance (the test statistic $Q = 462.67$). This is evidence that all judges are on average ranking the portfolios consistently, that is, all judges consistently prefer larger portfolios. Further, the null hypothesis is also rejected if the test is run over any set of 10 consecutive portfolio sizes.

4 Conclusion

Our four portfolio measures are consistent in their preference for larger portfolios, i.e. the measures are affected by portfolio size, but in a consistent manner. They only differ in their sensitivities to portfolio size. That is, the portfolio rankings provided by the Sharpe and information ratios afford a greater preference for larger portfolios than do the Treynor ratio and Jensen's alpha.

In sum, however, the answer to our question posed above is: On average, the rank ordering of a given set of portfolios of differing numbers of assets is *not* dependent on the performance measure used. That is, while these are four unique measures, each of the performance measures is in itself adequate.

References

Ferson, W., and A. Siegel (2001). “The efficient use of conditioning information in portfolios”. *Journal of Finance* 3, 967–982.

Friedman, Milton (December 1937). “The use of ranks to avoid the assumption of normality implicit in the analysis of variance”. *Journal of the American Statistical Association (American Statistical Association)* 32 (200): 675-701.

Friedman, Milton (March 1939). “A correction: The use of ranks to avoid the assumption of normality implicit in the analysis of variance”. *Journal of the American Statistical Association (American Statistical Association)* 34 (205): 109.

Friedman, Milton (March 1940). “A comparison of alternative tests of significance for the problem of m rankings”. *The Annals of Mathematical Statistics* 11 (1): 86-92.

Goodwin, T., (1988). “The information ratio”. *Financial Analysts Journal* July/August, 34–43.

Jensen, M. (1969). “Risk, the pricing of capital assets, and the evaluation of investment portfolios”. *Journal of Business* 42, 167–247.

Lhabitant, F., (2000). “Derivatives in portfolio management: Why beating the market is easy”. *Derivatives Quarterly* 6, 39–45.

Sharpe, W.F., (1966). “Mutual fund performance” *Journal of Business* 39, 119–138.

Treynor, J., (1965). “How to rate management of investment funds” *Harvard Business Review* 43, 63–75.