EJW
ECON JOURNAL WATCH
Scholarly Comments on Academic Economics

ECON JOURNAL WATCH 21(1)
March 2024: 113–155

Revisiting Hypothesis Testing with the Sharpe Ratio

Michael Christopher O'Connor

LINK TO ABSTRACT

If you manage a portfolio, or if you are invested in a portfolio that is managed by others, such as a fund, you would like it to show a high Sharpe ratio on an ongoing basis. The statistic is a popular device for assessing and rating portfolios and their managers. The Sharpe ratio is the mean of an investment's per-period returns in excess of the returns on cash, divided by the standard deviation of those excess returns, during some interval of time that has been divided into periods (e.g., months).[1]

Perhaps the ratio owes much of its popularity to the fact that it combines two familiar statistics that are each of great interest to investors, the mean and the variance (the square of the standard deviation), into one. Combining the two familiar statistics suggests a simple rating scheme whose use, in selecting investment alternatives, seemingly honors both high cash-beating returns and low variance.

But of course, the wisdom of basing investment decisions upon a single performance measure is doubtful. Here I focus on comparing the performance of one investment alternative with the performance of another—using the paired difference between the Sharpe ratios of the two portfolios as the performance measure.

When the Sharpe ratio of one portfolio substantially exceeds that of another there is a need to consider whether the exceedance is likely to be sustained. Did

1. If P_i is the value of the investment at the end of period i (which may consist of the market value of the asset and income that the asset has generated), and C_i is the value at the end of period i of cash held by anyone during period i, then the investment's excess returns of period i can be taken to be $ln(P_i / P_{i-1}) - ln\left(C_i / C_{i-1}\right)$, where $ln(\cdot)$ is the natural logarithm.

it merely happen by chance? The statistical properties of the ratio that delimit the proper use of the test of the happened-by-chance hypothesis are hardly as widely understood and appreciated as they should be. There are entire categories of studies of Sharpe ratio differences whose authors, having disregarded limitations on the power of the test, reached erroneous conclusions regarding statistical significance. I demonstrate that. But first I review oft-cited mathematical publications about hypothesis testing with Sharpe ratio differences, publications that contain peculiar errors and omissions that may account for the deficiencies in understanding. I present a legitimate means of post hoc analysis of hypothesis test findings, for sorting out misclassifications of statistical significance. And I show, categorically, using simulations with pertinent Sharpe ratio difference examples and other analysis, that when the power of the test is low then the very best estimators that can be put to the task of determining statistical significance can hardly perform better than random number generators.

I typically have in mind one of the portfolios being a suitable benchmark—e.g., if the investment alternative of interest involves long positions in big-cap stocks, a suitable benchmark would be one that would replicate the returns of the S&P500® index. The burning issue then becomes the question of whether past outperformance of the benchmark by the investment alternative of interest is likely to be sustained. The hypothesis that said outperformance was due to chance, and not likely to be sustained is the null hypothesis. And it is that hypothesis that is tested. The hope is to be able to reject it.

And to do so reliably the probability of rejecting the null hypothesis when it is false, which is called the power of the test, must be high. The statistical properties of Sharpe ratio differences differ substantially from those of Sharpe ratios. That's the rub. With differences, it's far easier to find yourself in circumstances in which the power of the test is too low for hypothesis testing to be conducted.

It is therefore necessary to be especially wary when wielding the Sharpe ratio difference hypothesis test. And that's really what this article is about.

The Sharpe ratio difference hypothesis test is the subject of the oft-cited Olivier Ledoit and Michael Wolf (2008) article "Robust Performance Hypothesis Testing with the Sharpe Ratio." Although the authors seem to have produced a "robust" way of performing hypothesis testing, within the technical meaning of that word, they: (1) utterly fail to discuss the crucial matter of the power of the test; (2) improperly demonstrate the use of their method in a circumstance in which the power of the test is obviously too low to offer reliable hypothesis testing; and (3) claim superiority over competing methods when the number of observations is low, despite the obvious fact that the power of the test is too low for hypothesis testing when the number of observations is low.

Ledoit and Wolf (2008, 851) describe themselves as statisticians and refer to

others who might be interested in their work as financial practitioners. The article has been influential, so that there are users who are seemingly not as mathematically capable as the authors who have made the gross error of following suit and applying the method when the power of the test is obviously too low for use. Herein I deal with that and how it came about.

The very next section below reviews the basics of hypothesis testing and can be skipped by readers who are well versed on the subject. Further on, the section "The power of the test in practice" contains a discourse on post hoc analyses of hypothesis testing, which has been a controversial topic. In that section the use of a secondary hypothesis test is proposed, to better cope with errors that occur with findings of statistical insignificance when the adequacy of the power of the test is open to question. I have included a simple method of incorporating investment horizons when using Sharpe ratio differences, which you will find in the section "The recourse to confidence intervals." Then further on, in the section "The last word," I thoroughly demonstrate, using the very code that was provided by Ledoit and Wolf (2008), that it is imprudent to base portfolio selection on hypothesis testing when the power is truly low. Those are highlights. Complete derivations are provided for all of the mathematics, either directly in this article in easy-to-follow steps or straightforwardly in cited articles.

The mathematics of hypothesis testing

The operative null hypothesis when the test statistic is the difference between the Sharpe ratios of two portfolios is simply the claim that it will not be positive in the long run: Its true value is zero or negative. If the null hypothesis is true, then any positive difference that was computed from a sample of historical data was due to chance—merely analogous to ten tosses of a fair coin happening to turn up heads six times instead of five. To assess the risk that the measured difference is due to chance, one computes, using sampled historical data, the highest probability that the difference would be equal to or greater than the measured value that could be computed *if the null hypothesis were true*. That probability is called the 'p-value.'[2] Low p-values, below some fixed cutoff value that the analyst is free to choose, indicate that it is likely that the null hypothesis isn't true and that it should be rejected. When it is thus rejected, 'statistical significance' is declared. Would that it were that simple.

2. I'm limiting the present discussion to my preferred form of the hypothesis test, which is referred to as a 'one-sided' or 'right-tailed' test, because when the stated null hypothesis is rejected, the conclusion is that the value of the Sharpe ratio difference is almost certainly positive, to the right of zero. With either a right-sided or two-sided form of the test, the highest probability that the test statistic would be equal to or greater than the measured value if the null hypothesis were true is computed by taking the true value to be zero.

The first complication involves the chosen cutoff value of the p-value, which is called the 'significance level.' Given the definition of the p-value, it is the probability of wrongly declaring statistical significance when the null hypothesis is true. Doing that is dubbed a 'type I' error or a 'false discovery,' and to avoid such errors the significance level is set rather low (e.g., 0.05)—but not so low as to fix it so that there would seldom be findings of statistical significance. So, already it's not so simple.

Then it takes some doing to compute the p-value. There is a head-first dive into that further on below. But it is equally imperative to heed another probability that is on the flip side of null hypothesis testing— the probability of erring by failing to reject the null hypothesis (due to the p-value being above the cutoff value) when the null hypothesis isn't true. And 1.0 minus that probability is the probability of rejecting the null hypothesis when it isn't true, the 'power of the test.'

Whereas the p-value can be estimated using the sample of historical data, the power of the test is based upon the true value, called the population value by statisticians (and, hereafter, in this article), of the test statistic. And that value can never be known. The goal is to find out as much about it as possible, such as bounds within which it is likely to be found. And mere hypothesis testing does just that, because a finding of statistical significance is evidence that the test statistic is probably not negative, which means that zero has been shown to be a lower bound of sorts.

The population value of the chosen hypothesis test statistic is the most important of the parameters on which the power of the test depends. But surprisingly, when the chosen test statistic is the Sharpe ratio difference, there is a very strong dependence of the power upon the correlation coefficient between the returns of the two portfolios. With the Sharpe ratio difference, factors that bring about inadequate power include the population value of the difference being too small, the correlation coefficient being low, and the time duration of the sample being small.

Certainly high power is wanted, so that the null hypothesis will usually be rejected when it isn't true. But exactly what sort of ill wind is it that blows if the power is low, not anywhere near 1.0? When there is a failure to reject the null hypothesis when it isn't true, a 'type II error' has been sustained. The immediate harm is that the investor is thereby led toward failing to invest in the portfolio with the seemingly better Sharpe ratio, even though it is destined to continue to outperform the portfolio with which it was compared. That other portfolio with the inferior Sharpe ratio could be a benchmark, such as a stock market index portfolio. Missing out on beating the market certainly involves some kind of opportunity cost. But there's more.

The low power can otherwise increase the likelihood of committing the error

of rejecting the null hypothesis when it is true, a 'type I error' or a 'false discovery.' How so? Consider an analyst who wants to go forward with an investment alternative that is found to be benchmark-beating, based upon Sharpe ratios compiled from sample returns of the alternative and of the benchmark. A single alternative, when considered alone with its benchmark as a Sharpe ratio difference, has a finite chance of generating a type I error. What happens if, owing to a type II error that is brought about by the low power of the test, the analyst is misled into having to sift further through several alternatives, looking for one whose benchmark beating was calculated to be statistically significant? The analyst encounters, with each alternative considered, the risk of a type I error. The overall probability of settling upon a false discovery is thereby enhanced because of the multiple tries.[3]

But even if the search doesn't finish with a type I error there is still a problem that is brought about by having to search for an alternative with statistically significant outperformance. Seeking statistical significance (that was denied due to the low power of the test causing type II errors) is, in the main, seeking a high sample value of the Sharpe ratio difference because the two go together. And selecting the alternative that performs best on sampled data is a formula for inducing selection bias: Any preference for the alternative to have scored high in performance on the sample is automatically also a preference for chance outliers, for alternatives whose performance on the sample is considerably better than what can be expected in the long term. The Sharpe ratio differences of such alternatives may have population values that are no better than those of alternatives that were passed over because of type II errors. Cross-validation methods and walk-forward methods can be used to counter selection bias, but they have their limitations.

Relatedly, very low power can even diminish the probability, when there is immediately a positive finding of statistical significance, of the true Sharpe ratio difference actually being positive. This is explained by Katherine Button, John Ioannidis, et al. (2013, 2), building on some simple mathematics from Ioannidis (2005, 696–697).[4]

---

3. Remediation is available for this problem. See for example Benjamini and Yekutieli (2001, 1168–1169), wherein the word "conservative" means that the method errs on the side of dismissing more findings of significance than would be necessary to control the false discovery rate. It is still advisable to do the remediation.

4. Ioannidis (2005) finds that the probability that a finding of statistical significance is really true (that the population value of the test statistic is positive), which has been dubbed the positive predictive value (PPV), depends upon the product of the power of the test and the odds ratio being high enough. Loosely put, the odds ratio is the expected frequency of good performance outcomes divided by the expected frequency of bad performance outcomes. Often financial analyses involve circumstances in which the odds ratio might as well be estimated as being $\cong 1$. For example, if an investor sets out to examine the past performance of a no-load mutual fund, one whose holdings seem to be of much the same character as those

Reviewing Sharpe ratio difference literature

Circa 2021, I was innocent of knowing anything of substance about the use of a Sharpe ratio difference as a test statistic. As for other test statistics involving financial time series data, I did at least know that the gold standard was that an analysis that properly establishes statistical significance is one that allows for investment portfolio returns not being independently and identically distributed (i.i.d.), having fat-tailed and skewed distributions, and being autocorrelated so that the current month's return is influenced by the returns of prior months. 'Heteroskedasticity' denotes the circumstance of the variance of a statistic not being constant but varying with time. So yes, there is a need to try to allow for that too—a varying variance. I searched the literature and I found Ledoit and Wolf's article (2008). It is all about computing the p-value when the test statistic is the difference between the Sharpe ratios of two portfolios. I had found the answer.

Or so I thought. Ledoit and Wolf (2008) have open-sourced their code (found **here**), in the computer languages Matlab and R, and their article has been cited something like 1,000 times. Thus their procedures and the code that they wrote to implement them have been well scrutinized and used. I opted for extending their checks on the validity of their code just a bit, in a relevant way. Their tests involved progressively more realistic simulated histories of portfolio returns but, unlike mine, their tests were limited to the case when the Sharpe ratio difference is actually zero (Ledoit and Wolf 2008, 857).

I tested only with simulated returns randomly drawn from the not very realistic bivariate-normal distribution, so that the sampled values are i.i.d. My simple tests are of interest, notwithstanding the limitation to idealized circumstances, because the imposed limitation doesn't leave us with just simulations to work with. The idealized circumstances make it possible to straightforwardly derive asymptotically valid expressions for the p-value and the power of the test. And I did not limit my tests to Sharpe ratio differences of zero, because only non-zero differences are of interest when hypothesis testing. Apart from the very small effects of the involved iterations not being infinite in number,

---

of the benchmark, then probably the odds ratio is about 1. Low power tends to drive the PPV below 1, but when the odds ratio is not much below 1 the power of the test must be *very* low if the PPV is to be driven much below 1.

But, if the investor is considering a mutual fund with a substantial load or is considering an investment advisor of the kind that are affiliated with brokers, who professes skill and consequently charges an annual fee of, say, 0.5 to 1.0 percent, then the odds ratio may be drastically lower than 1. That's with the mutual fund's load or the advisor's fee being fully accounted for in the performance measure that is the test statistic, thus lowering the odds ratio. In that circumstance the deleterious effect of low power on the PPV is greatly magnified by the odds ratio multiplier being low.

I found no disagreement between the p-values that I computed with Ledoit and Wolf's code and those that I computed with either my own simulations or with the asymptotic expressions that were derived using calculus.

In the introduction to this article I delineated the wholesale and consequential neglect of the power of the test in Ledoit and Wolf (2008). With the goal of understanding how it might have happened that the 2008 article won acceptance despite its fairly obvious deficiencies, I read prior articles that Ledoit and Wolf had cited, by J. D. Jobson and Bob M. Korkie (1981), Christoph Memmel (2003), and John D. Opdyke (2007).

And indeed, Jobson and Korkie (1981) were sowing confusion. Their paper contains three glaring errors: One is their claim that they had concocted an alternative to the Sharpe ratio difference statistic that had more power as a test statistic than the Sharpe ratio difference; the second was to state an incorrect formula for the asymptotic variance of their own statistic; and the third was to state that the power of the test is chronically low and insensitive to the value of the correlation coefficient between the returns of the two portfolios. In only three pages, Memmel (2003) corrected the first two errors; leveraging off of Memmel's work, Opdyke (2007) disclosed and fixed the third error, showing that when the correlation coefficient is high the power of the test is high (but also affirming that when it is not high the power is low).

But Ledoit and Wolf mention Opdyke's (2007) contribution without noting that he affirmed that a low correlation coefficient means that the power of the test is very low. The inappropriate use that Ledoit and Wolf made of their method was to compute the p-value for the paired Sharpe ratio difference between two hedge funds, using a decade of data. The authors provided the returns of the funds and I have calculated the correlation coefficient between the returns of the two hedge funds as 0.00—meaning, in conjunction with the use of only a decade of data, that the power was unusably low. That the power was too low is confirmed below, at the conclusion of the "Hedge funds are different" subsection of the section "The power of the test in practice."

Citations of these articles since the start of 2009 have been as follows: Ledoit and Wolf (2008) 981 times, Jobson and Korkie (1981) 794 times, Memmel (2003) 450 times, but Opdyke (2007) just 180 times.[5] Those who see in Ledoit and Wolf (2008) that Jobson and Korkie (1981) and Memmel (2003) are cited, who are diligent enough to obtain and read those articles but missed Opdyke (2007), would really have to be on their toes when figuring out what to think about Jobson and Korkie's power estimates because of all of their mistakes.

Furthermore, anyone who did read Opdyke (2007) would have been grossly

5. Google Scholar citations as of May 13, 2023.

misled by the author's discussion of the power of the Sharpe ratio difference test being confirmed, by his analysis of mutual fund data, to be high when the correlation coefficient is high. Opdyke well knew, from his derivations, that the power can be high when the coefficient is high if the population value of the Sharpe ratio difference is not too small and if the duration of the historical data is sufficient. But in circumstances where the portfolios being compared are the result of very similar portfolio formation schemes, the differences in performance are entirely unforeseeable and for numerous pairs the true Sharpe ratio difference is small. And not only were the funds that Opdyke studied of that character, but he had only three years of data. The power of the test must be presumed to be low in those circumstances, whatever the value of the correlation coefficient.[6] The relevant mathematics are reviewed in the next section.

These circumstances have led to numerous researchers using Ledoit and

6. Opdyke wielded the right-sided hypothesis test. He formed 190 ordered pairs out of the 20 mutual funds—every possible ordered pair, where 'ordered' simply means that the sample value of the Sharpe ratio of the first fund of each pair was bigger than that of the second. He seems to have maintained that the fact that the four or five out of the 20 funds that had the highest sample values of the Sharpe ratios ranked above about half of the other funds by statistically significant margins was indicative of the test generally having good power, because low power would mean that not many null hypotheses would have been rejected, whereas (of the pairs formed with the top four or five funds) about half were rejected. But on his Table IV we see that there was failure to reject the null hypotheses of 122 of the 190 pairs. Many of the other pairs, having small Sharpe ratio differences, simply didn't have adequate power.

How do I know that many pairs did not have adequate power? Well, it's obvious that the 20 mutual funds having very similar methods of portfolio formation within the same asset class rules out the true Sharpe ratio differences of the pairs being somehow fenced away from zero. If say, someone were to compare the performance of 20 hedge funds with that of a benchmark portfolio that tracked the S&P500®, then yes, it should be expected that the hedge funds would have higher Sharpe ratios than the benchmark—meaning that the distribution of the Sharpe ratio differences of the 20 hedge funds would be centered on some positive value of the Sharpe ratio difference.

But that's hardly the case with pairs of similar mutual funds. Ordered pairs could be formed based on the population values of the Sharpe ratios (if only we knew those values), instead of on the sample values. Every pair has a population value of the Sharpe ratio difference that is the main determinant of the power of the test in application to the pair. Suppose that we sort the list of funds in descending order by the population Sharpe ratio. Then the Sharpe ratio difference between the top fund and the fund immediately below will be small, because they are only one place apart on the list. The same will be true of the pair that consists of the funds that are placed second and third on the list, and so on. There will be 19 such pairs with small Sharpe ratio differences due to the funds being only one place apart on the list. What about pairs of funds that are two places apart? Yes, they will have bigger Sharpe ratio differences but there will be 18 of them, not 19. We can continue. When the funds are three, four, five places apart, etc., the Sharpe ratio differences will be getting larger still, but the number of pairs will be getting reduced (e.g., with the funds 10 places apart there will be 10 such pairs with large differences). The point is that the distribution of the population values of the Sharpe ratio differences would be peaked near zero. And for those fairly numerous pairs with differences near zero the power of the test will be too low.

That low Sharpe ratio differences always mean low power, however large the correlation coefficient, is demonstrated in the next section of this article. It's also shown there that Opdyke using weekly data instead of monthly data in his three-year backtest interval did not improve the power.

Wolf's (2008) article and code to calculate p-values for use with hypothesis testing, when the test statistic is the difference between two Sharpe ratios (or the absolute value of that), without understanding that in many commonly encountered circumstances the power of the test is simply way too low.

Mathematics of power—i.i.d. bivariate-normal returns

Here the mathematics of how to compute p-values and the power of the test whose statistic is either the difference between the Sharpe ratios of two portfolios or the absolute value of that difference is summarized, with some of the analysis being specific to the case in which the returns of the two portfolios are i.i.d. bivariate-normal. One of the portfolios may be a benchmark.

Dependencies of the power of the test

If Sh_a is the Sharpe ratio of the portfolio under study, and Sh_b is the Sharpe ratio of a suitable benchmark for it (or some other portfolio of particular comparative interest), then four test statistics will be considered: Sh_a, $|Sh_a|$, $Sh_a - Sh_b$, and $|Sh_a - Sh_b|$. The statistics that involve absolute values bring about two-sided tests; the other statistics establish one-sided, right-sided tests.

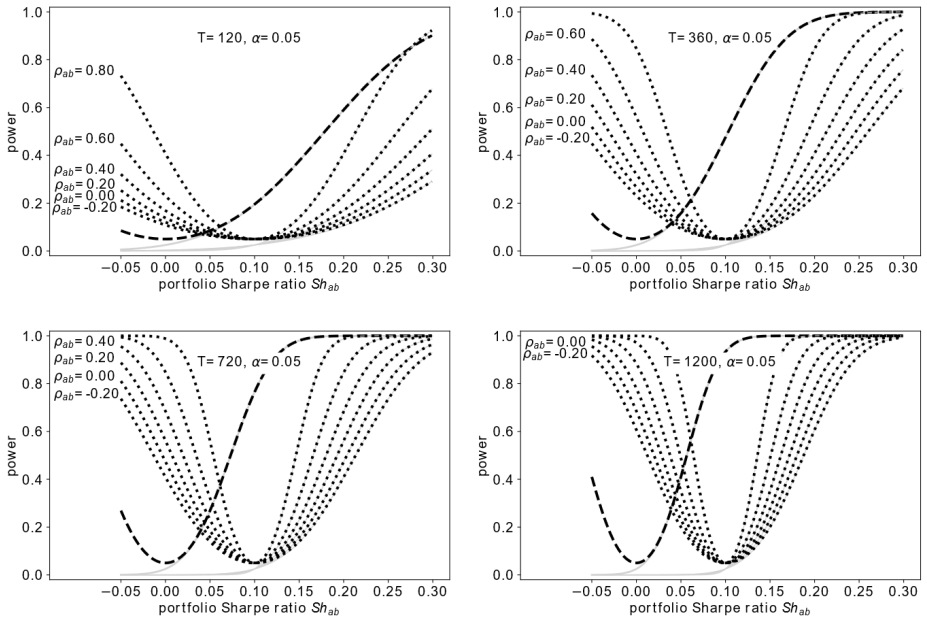
Figure 1 plots the power of the test versus the Sharpe ratio Sh_a of the portfolio under study, when the test statistic is any one of those four statistics—for various values of the number T of observations and of the correlation coefficient ρ_{ab} between the portfolio and the benchmark (when applicable). The dotted lines pertain to the two-sided test whose statistic is $|Sh_a - Sh_b|$; the dashed lines pertain to the two-sided test whose statistic is $|Sh_a|$. The thin gray lines that conform to the dotted and dashed lines at elevated values of the Sharpe ratio Sh_a of the portfolio under study plot the right-sided powers when the test statistic is Sh_a or $Sh_a - Sh_b$. And the figure pertains to the circumstance of $Sh_b = 0.10$ (when applicable).

Figure 1 is valid as-is, for any frequency of observations—e.g., daily, weekly, monthly, or yearly. If for example the raw data were monthly with the Sharpe ratios being computed monthly as well, then the annualized Sharpe ratios could be approximated as the monthly ratios multiplied by $\sqrt{12}$.[7] Thus, where the benchmark's Sharpe ratio Sh_b is taken to be 0.10 on the figure, that would correspond to an annualized Sharpe ratio of about $0.10\sqrt{12} \cong 0.35$—similar in

7. Lo (2002) cautions that if the returns are autocorrelated then the factor of $\sqrt{12}$ is not accurate.

magnitude to what we might expect of the long-term average Sharpe ratio of some capitalization-weighted portfolio of stocks belonging to some stock market index. The fixed value of the benchmark portfolio's monthly Sharpe ratio on the figure, of 0.10, was also the choice of Jobson and Korkie (1981) in almost all their examples. The shown number of months, especially T=120 and T=360 on the top half of the figure, would then respectively represent often-achievable and sometimes-achievable backtesting intervals of 10 years and 30 years.[8]

Figure 1. Dependencies of test power with Sharpe ratios and Sharpe ratio differences as test statistics



Notes: The dotted lines and the thin gray lines that conform to them as the portfolio Sharpe ratio Sh_a is increased respectively pertain to the hypothesis tests that are based on the statistics $|Sh_a - Sh_b|$ and $Sh_a - Sh_b$, where Sh_b has been set at 0.10. The dashed line and the thin gray line associated with it respectively pertain to the test statistics $|Sh_a|$ and Sh_a that involve just the portfolio Sharpe ratio.

8. But if instead the portfolio under study were a hedge fund, with the Sharpe ratios of it and a benchmark forming a difference, then if the data were weekly the benchmark's Sharpe ratio Sh_b of 0.10 on the figure would very roughly become $0.10\sqrt{52} \approx 0.72$. That would not be an unreasonable value for the annualized Sharpe ratio of a hedge fund. Then the shown number of weeks, T=720 and T=1200 on the bottom half of the figure, would respectively represent often-achievable and sometimes-achievable backtesting intervals of something like 14 years and 23 years.

I hasten to add that the Sharpe ratios in Figure 1 are not to be thought of as sample-derived values. The ratios are in fact the population values. It can be imagined instead that data have not yet been acquired but significance testing for some portfolio formation scheme that has been contrived is being planned. And one of the burning questions would be: How many months of data are needed? If the analyst were to have some idea of the population values of the ratios, even if it were just a rough idea, or even if the analyst were proceeding on a 'what if' basis, Figure 1 (or better yet the code that produces it) could be looked to for some helpful guidance concerning the very feasibility of hypothesis testing with the Sharpe ratio.

The returns of the two portfolios are assumed to be i.i.d. and drawn from a bivariate-normal distribution, as in the Jobson and Korkie (1981) article. The figure directly implements Memmel (2003)'s correct asymptotic expression for the 'standard error' of the Sharpe ratio difference. I present the math and the mathematical terms in the next subsection.

The horizontal axes pertain to the Sharpe ratio of the portfolio under study; the power is plotted vertically. The distinctions between the four kinds of power that are illustrated in the figure have been drawn above, where four test statistics were described. There was also imposed on the analysis that produced Figure 1 the value $\alpha = 0.05$, which is the 'significance level,' which was introduced above. Again, the analyst gets to choose it! If the test is two-sided, then if the two-sided p-value exceeds α there is failure to reject the null hypothesis. If such an α has been chosen for a two-sided test, then for purposes of side-by-side comparison with a one-sided test there is failure to reject the null of the one-sided test if the one-sided p-value exceeds $\alpha/2$. So α or $\alpha/2$, depending upon the sidedness of the test, is the probability of rejecting the null hypothesis if it is true (because the p-value is based on the null hypothesis being assumed to be true). Again, that would be a type I error, a false discovery.

That sidedness lingo is something that you will encounter in academic articles that present statistical tests based on Sharpe ratios. And it may be important for a reader to take note of which kind of test has been used, because whereas the reader might be quite understandably concerned about the sign of the Sharpe ratio or Sharpe ratio difference, rather than just its magnitude, there are circumstances in which academics prefer to implement a two-sided test which could lead to a finding of statistical significance with the Sharpe ratio or Sharpe ratio difference being of either sign. For example, there might be a factor that influences the makeup of one of the portfolios that the academician suspects of being incapable of producing a substantial effect. With a two-sided test there are two ways to reject the null hypothesis, so that the two-sided power is always greater, at least a tiny bit greater, than the one-sided power.

A most clearly discernible thing about Figure 1 is that the dotted lines show the power of the test, whose statistic is the absolute value of the difference between the Sharpe ratio of the portfolio under study and that of its benchmark, increasing markedly as the correlation coefficient ρ_{ab} is increased from −0.20 to 0.00 and on up to 0.80. That is what Opdyke (2007) noticed and called attention to.

Taking a closer look, suppose that the Sharpe ratio of the portfolio of interest is 0.15. That's 0.05 above the benchmark value of 0.10, a seemingly substantial improvement: If the frequency of observations is monthly, that would be annualized as the benchmark having a Sharpe ratio of 0.35 with the portfolio's ratio being 0.52. So, it might be supposed that the difference of 0.05 would be statistically significant just because of that. But looking, say, at the plot on the upper right-hand side that was prepared with the number of observations being 360—three decades of months, which is a fair amount of data—it is seen that the dotted line for the two-sided test when the correlation coefficient $\rho_{ab} = 0.80$ indicates that the power is well below 0.80 (which is often adopted as a minimally satisfactory power). The power is only 0.32, which means that the null hypothesis would be rejected only 32 percent of the time when it's false.

Now it's true that if ρ_{ab} the correlation coefficient is allowed to approach 1.0, then yes, the power will approach 1.0. But if the correlation coefficient is 0.80, which is not a low value for comparisons between equity portfolios, how many observations would be needed to achieve acceptable power? Well, if there are T = 1,200 observations, a century of months, the power is 0.78—close enough, but that's a prohibitively large span of time.

In comparison, starting again with T = 360, what sort of power is achieved when the test is based on the portfolio Sharpe ratio alone, with no consideration given to the benchmark at all? If the portfolio Sharpe ratio is again 0.15, the dashed line and the underlying light gray line tell the tale: The power is about 0.81—satisfactory. It's largely because the tests that only involve the portfolio Sharpe ratio effectively use zero as the point of reference that the power is so much higher than what it would be with the testing being done with the difference from the benchmark's Sharpe ratio. With tests involving that difference, the point of reference is moved up (all the way up to 0.10 on Figure 1). Of course, it's tougher to beat the higher reference point. And that has implications for the power of the test.

Therefore, correlation coefficients not being high, the duration of the time interval for backtesting not being long enough, and Sharpe ratio differences not being sizeable are controlling factors when Sharpe ratio differences are used as test statistics. Standing alone, any one of them can cause the power of the test to be unacceptably low.

The mathematical framework

Here again the 'portfolio Sharpe ratio,' as it is called above in Figure 1 and the discussion of that figure, is Sh_a. And the Sharpe ratio of the other portfolio, which was referred to as a benchmark above, is Sh_b. The other portfolio may not be a benchmark but could be any other portfolio that is of interest. Let $\Delta_- = Sh_a - Sh_b$ and $\Delta_+ = Sh_a + Sh_b$. These are not sample values of these quantities but can be thought of as typifying the long-term future performance of the portfolios. It is better to say that they are the population values. The values of these statistics will never be known, and it is only for limited purposes that they can be estimated as being approximately equal to the values that they take on in samples. Sample values are distinguished by $\hat{\Delta}_-$ and $\hat{\Delta}_+$ with a circumflex atop the symbols. In this notation the population value of the correlation coefficient is ρ_{ab}, but there will be occasion to refer to the sample value $\hat{\rho}_{ab}$ as well.

Equations (1a)–(1c) are straight from Memmel (2003, 23), which contains very clear derivations involving calculus. The term $SE(\Delta_-, \Delta_+, \rho_{ab}, T)$ in equations (1a) and (1b) is the standard error, the standard deviation of the distribution of the $\hat{\Delta}_-$ sample value about the population value Δ_-. And equation (1c) just announces that in the limit of $T \to \infty$ the distribution will take on a Gaussian form, a normal form.

$$T \cdot SE(\Delta_-, \Delta_+, \rho_{ab}\rho_{in}, T)^2 \equiv 2(1 - \rho_{ab}) + \frac{\Delta_-^2(1 + \rho_{ab}^2) + \Delta_+^2(1 - \rho_{ab}^2)}{4} \quad (1a)$$

$$Z \equiv \frac{\hat{\Delta}_- - \Delta_-}{SE(\Delta_-, \Delta_+, \rho_{ab}, T)} \quad (1b)$$

$$Z \sim \mathbb{N}(0, 1) \quad (1c)$$

With either a two-sided or a one-sided test involving a comparison between the Sharpe ratios of two portfolios, when computing the p-value the null hypothesis is implemented by taking Δ_- to be 0. Thus equations (1b) and (1c) become equations (2a) and (2b).[9] Only equation (1a) is dependent upon the returns of the

9. This assumes that when the null hypothesis is invoked the values of Δ_+ and ρ_{ab} are known. In this situation those parameters are called 'nuisance parameters.' That language refers to the fact that although they are dependencies, they are in fact not known. They are the true values of these parameters, not sample values. It is clear that if Δ_+ is not large, say under 1.0, then there is only a weak dependence of $SE(\Delta_-, \Delta_+, \rho_{ab}, T)$ upon Δ_+. And that suggests that it might be feasible to simply approximate Δ_+ by $\hat{\Delta}_+$.

two portfolios being i.i.d. and bivariate-normal. The other expressions are instead asymptotically valid if only the returns of the two portfolios are stationary, which roughly means that the bivariate return distribution does not change over time; see for example Andrew W. Lo (2002, 39) in that regard.

$$Z_0 \equiv \frac{\hat{\Delta}_-}{SE\left(0,\, \Delta_+,\, \rho_{ab},\, T\right)} \tag{2a}$$

$$Z_0 \sim \mathbb{N}(0,\, 1) \tag{2b}$$

In equations (3a)–(3c) Φ is the cumulative distribution function (cdf) whose derivative is the standard normal probability density function. And p_2 is the two-sided p-value, with equation (3a) following directly from the definition of the p-value and the meaning of equations (2a) and (2b) regarding the distribution of $\hat{\Delta}_-$ under the null hypothesis. More simply, $p_2 = 2\Phi\left(-\left|Z_0\right|\right)$.

$$p_2 = 1 - \Phi\left(\left|Z_0\right|\right) + \Phi\left(-\left|Z_0\right|\right) \tag{3a}$$

$$\pi_2(\alpha) = 1 - \Phi\left(Z_{\hat{\Delta}_- = \Delta_{-,r}(\alpha)}\right) + \Phi\left(Z_{\hat{\Delta}_- = -\Delta_{-,r}(\alpha)}\right) \tag{3b}$$

$$\Delta_{-,r}(\alpha) = SE\left(0,\, \Delta_+,\, \rho_{ab},\, T\right) \bullet \Phi^{-1}(1 - \alpha/2) \tag{3c}$$

Equation (3c) calculates the critical value of the test statistic. If $\left|\hat{\Delta}_-\right| \geq \left|\Delta_{-,\,r}(\alpha)\right|$ then with the two-sided test the outcome is declared to be statistically significant. It is just a re-expression of the definition of the right-sided p-value p_r, which is $p_r = 1 - \Phi\left(Z_0\right)$, evaluated with a particular p-value: Set $p_r = \alpha/2$, the critical value of the p-value, solve for $\Phi\left(Z_0\right)$, and then apply the inverse of Φ. And $\alpha/2$ is the cutoff that is mentioned above that the analyst gets to choose. Often the significance level $\alpha = 0.05$ is chosen, in which case $\Phi^{-1}(1 - \alpha/2)$ has the familiar value of ≈ 1.96. Once again, α and $\alpha/2$ are respectively the probabilities, with a two-sided test or a one-sided test, of rejecting the null hypothesis when it is true.

But there is a strong dependence upon ρ_{ab}. In Pearson and Filon (1898, 242) the standard error of ρ_{ab} is calculated as being $\left(1 - \rho_{ab}^2\right)/\sqrt{T}$. That might suggest that if with a large enough sample of the returns of the two portfolios under consideration they are found to be highly correlated, then it may be assumed that $\hat{\rho}_{ab} \cong \rho_{ab}$. In the next subsection it is explained that indeed, that is acceptable.

So, the meaning of equation (3c) is that it sets a critical value for $\hat{\Delta}_-$ that corresponds to setting the critical value $p_r = \alpha/2$. With that understanding $\pi_r(\alpha) = 1 - \Phi\left(Z_{\hat{\Delta}_- = \Delta_-,\, r(\alpha)}\right)$, with Z a function of $\hat{\Delta}_-$ as in equation (1b), would be the correct expression for the right-sided power because it is the probability, given the distribution of $\hat{\Delta}_-$ about Δ_-, that $\hat{\Delta}_-$ is bigger than its critical value $\hat{\Delta}_{-,\, r}(\alpha)$ —causing us to reject the null hypothesis. The $\Phi\left(Z_{\hat{\Delta}_- = -\Delta_-,\, r(\alpha)}\right)$ term of equation (3b) just adds the left-sided component of the power. And for it Z is evaluated at $\hat{\Delta}_- = -\Delta_{-,\, r}(\alpha)$ because $\Delta_{-,\, l}(\alpha) = -\Delta_{-,\, r}(\alpha)$ due to the symmetric form of the distribution.

Suppose that the data are monthly. If the power, computed with equation (3b) or the right-sided version of it, is too low, then the question arises whether it would help to instead make use of weekly, daily or even intraday data, as that would increase the number of observations. But the increased number of observations would be brought about by an increase in the frequency of observations within the same span of time. So, do the additional observations count the same as getting additional months of data to process? The short answer is no.[10]

---

10. The long answer: What exactly changes when the frequency of observations is increased but the duration of the backtesting interval is not? We'll start with the assumption that the returns are i.i.d. If for example the frequency of observations is increased to daily rather than monthly, the value of the correlation coefficient ρ_{ab} from daily returns would be the same as the monthly value. This follows if the returns of the portfolio are calculated, as they should be, as logarithms of the ratio of the value of the portfolio at the end of the given period to the value at the end of the prior period, so that compounding of returns is accomplished by summing. But with the same increase in the frequency of observations the daily Sharpe ratio difference would (assuming 21 trading days per month) become $\Delta_-/\sqrt{21}$ where Δ_- is the value from monthly returns, and likewise with Δ_+. This is derived by Lo (2002, 40). And of course, the new number of observations would be $21T$ in place of T where T is the number of months.

For securities that can be modeled as if the returns were i.i.d. bivariate-normal, begin by considering whether, with the starting monthly frequency of observations, the second term on the right-hand side of equation (1a) is sizable in comparison to the first term. The correlation coefficient ρ_{ab} being near 1 and the Sharpe ratio difference Δ_- being large are circumstances within which the second term could be sizable when compared with the first. If the second term isn't sizable relative to the first then further reducing its magnitude by resorting to a daily frequency of observations and thereby turning Δ_- into $\Delta_-/\sqrt{21}$ and Δ_+ into $\Delta_+/\sqrt{21}$ would not significantly change the magnitude of the right-hand side of equation (1a). And then the left-hand side should be $21T \bullet SE_{21}^2$ where the 21 subscript refers to daily observations, with the right-hand side being dominated by $2\left(1 - \rho_{ab}\right)$. Thus $SE_{21} \cong \sqrt{SE^2/21}$ which scales with the number of days in the month in the same way as $\Delta_-/\sqrt{21}$. This identical scaling factor has the consequence that the monthly value of the second term of $Z_{\hat{\Delta}_- = \Delta_-,\, r(\alpha)}$, which is also the second term of $Z_{\hat{\Delta}_- = -\Delta_-,\, r(\alpha)}$, which is $-\Delta_-/SE$, is not changed with the switch to daily values. Thus, the

Estimating the nuisance parameters from sample returns

By its very definition, the power of the test invokes a null hypothesis that has somehow been implemented. In equation (3c) above, which computes a quantity that is used in equation (3b) that computes the power, it's assumed that to implement the null hypothesis it suffices to set $\Delta_- = 0\ldots$ to compute the standard error using that value and the population values for Δ_+ and ρ_{ab}. And we also need to set $\Delta_- = 0$ to compute the p-value as in equations (2a) and (3a).

It's reasonable to worry a bit about whether that can properly be done. For one, can Δ_- be changed to 0 without also changing ρ_{ab}? After all, both parameters pertain to the same portfolio returns. The answer is yes. With $Sh_a = \frac{\mu_a}{\sigma_a}$ and $Sh_b = \frac{\mu_b}{\sigma_b}$, subtract the constant $(\sigma_b\mu_a - \sigma_a\mu_b)/(2\sigma_b)$ from the returns of portfolio *a* and subtract the constant $(\sigma_a\mu_b - \sigma_b\mu_a)/(2\sigma_a)$ from the returns of portfolio *b*. Because constants are being subtracted, ρ_{ab} the correlation coefficient of the returns is unchanged. And a bit of algebra shows that the new value of Δ_- is 0, but that Δ_+ is unchanged.

In all, that's a way with which the Sharpe ratio difference parameter Δ_- could be dealt with to implement the null hypothesis when computing the p-value and the power—with null-restricted data. But what, say, should be used for the value of Δ_+, the sum of the population values of the Sharpe ratios? And what for the correlation coefficient ρ_{ab} of the returns of the two portfolios? Note that these two parameters only affect the standard error.

I have been referring to asymptotic approximations that are valid only when the returns are i.i.d. bivariate-normal. Happily, the truth is that to compute the standard error of the sampling distribution of Δ_-, the Δ_+ and ρ_{ab} parameters can simply be approximated by their sample values $\hat{\Delta}_+$ and $\hat{\rho}_{ab}$. And furthermore, that is generally applicable—not just to idealized distributions of portfolio returns.

It is a matter of the estimators for $\hat{\Delta}_+$ and $\hat{\rho}_{ab}$ being 'consistent,' which

---

power is not changed.

Now what happens if the second term on the right-hand side of equation (1a) is dominant over the first term? That could happen if ρ_{ab} gets very close to 1.0. If it is—see again equation (1a)—then $T \bullet SE_{21}^2 \approx \Delta_-^2/2$ so that $SE \approx \Delta_-/\sqrt{2T}$, which yields $\Delta_-/SE \approx \sqrt{2T}$. In practice T is often 100 or more, so that the equation-(3b) power is already nearly 1.0 before we even consider upping the frequency of observations. Further improvement is scarcely possible. Things are a bit different with hedge funds, with portfolios whose returns are non-i.i.d. and both highly autocorrelated and heteroskedastic. Of course, increasing the frequency of observation would not improve the power of the test for such portfolios either. But such portfolios present the complication of the scaling factor that is the square root of the frequency of observation not being valid. That is shown by Lo (2002, 40).

roughly means that their values converge to those of Δ_+ and ρ_{ab} as T the number of samples is increased without limit. For example, again with i.i.d. bivariate-normal returns, and beginning with $\hat{\rho}_{ab}$, if ρ_{ab} is the true population value then $\hat{\rho}_{ab} = \rho_{ab} + \gamma\left(1 - \rho_{ab}^2\right)/\sqrt{T}$ can be written, where the added term consists of the adjustable factor γ multiplying Pearson and Filon's (1898) calculation of the standard error of the correlation coefficient. Obviously every possible value of $\hat{\rho}_{ab}$ can be written that way, with a suitable choice of the value of γ. But that equation can be solved for ρ_{ab} via the quadratic formula: $\rho_{ab} = \frac{\sqrt{T}}{2\gamma}\left[1 - \sqrt{1 - 4\gamma\left(\frac{\hat{\rho}_{ab}}{\sqrt{T}} - \frac{\gamma}{T}\right)}\right]$. And for large T that becomes $\rho_{ab} \approx \hat{\rho}_{ab} - \frac{\gamma}{\sqrt{T}}$ as $T \to \infty$. Of course γ is to be expected to be of the order of magnitude of 1. If for example γ is +2, then $\hat{\rho}_{ab}$ is two standard deviations above ρ_{ab}. It would be unlikely for $\hat{\rho}_{ab}$ to be higher still, so that to be very sure not to overestimate ρ_{ab} while using $\hat{\rho}_{ab}$ in its place, $\hat{\rho}_{ab} - \frac{2}{\sqrt{T}}$ could be substituted instead.

But the real point to make is that the standard error of the sampling distribution of the Sharpe ratio difference, as shown in equation (1a), is proportional to $\frac{1}{\sqrt{T}}$ because it is the first term of an asymptotic expansion in the variable $\frac{1}{\sqrt{T}}$. That's with it as a function of the population values of Δ_+ and ρ_{ab}. If $\hat{\rho}_{ab} - \frac{\gamma}{\sqrt{T}}$ is substituted for ρ_{ab} and the asymptotic expansion is formed again, the leading term will become the standard error as defined by equation (1a) but as a function of Δ_+ and $\hat{\rho}_{ab}$. And the same can be done with Δ_+ so that Δ_+ can be replaced by $\hat{\Delta}_+$.[11,12]

This is actually standard operating procedure in the standard error business;

---

11. I used the procedure of Memmel (2003), and the quadratic formula, and have assumed $\Delta_- = 0$ as for the null hypothesis case, to derive $\Delta_+ = \hat{\Delta}_+ + \gamma\sqrt{2\left(1 + \hat{\rho}_{ab}\right) + \hat{\Delta}_+^2\left(1 + \hat{\rho}_{ab}^2\right)/4}\ /\ \sqrt{T}$. Again, this permits us to substitute $\hat{\Delta}_+$ *for* Δ_- *in* the expression for the standard error of the sampling distribution of the Sharpe ratio difference because the last term, in $1/\sqrt{T}$, has a numerator that is of order of magnitude 1 and the term can't contribute to the first-order term of the asymptotic expansion of the standard error.

12. Even with i.i.d. bivariate-normal returns, the mean of the sample values of $\hat{\rho}_{ab}$ is a biased estimator of ρ_{ab}. Lehman and Casella (1998, 96) report that the expected value is approximately $\rho_{ab}\left[1 - \frac{\left(1 - \rho_{ab}^2\right)}{2T}\right]$. But note that the correction term can be quite small if, say, $T \geq 100$. More to the point, no such correction term that varies in inverse proportion to T could contribute to the first-order term in $1/\sqrt{T}$ for the standard error. Similarly, Pav (2021, 5) finds that the expected value of the Sharpe ratio is $Sh\left[1 + \frac{3}{4T}\right]$, which is similarly inconsequential.

see for example equation (10) of Lo (2002, 38). But some scholars have concerned themselves with whether such a simple resort is the best that can be done with nuisance parameters; see for example Peter Reinhard Hansen (2005). So, I have dwelled on the matter.

To summarize, if the histories of the returns of the two portfolios of the Sharpe ratio difference test statistic are long enough, then those data alone can be used to estimate its standard error. No knowledge of the population values of the nuisance parameters is needed because the departures of the sample values from the population values are asymptotically insignificant. Suitable estimators of the standard error might be theoretically derived, based on assumptions about the statistical properties of the returns, as Memmel did. Or the analyst might resort to using resampling methods, bootstrapping, as in Ledoit and Wolf (2008). Therefore, with any backtesting procedure it can be assumed that the standard error $\hat{SE}$ is known. And of course, $\hat{\Delta}_{-}$ is known, leaving unknown only the population value Δ_{-} of the Sharpe ratio difference.

Confidence intervals and the standard error

Given a sample value $\hat{\Delta}_{-}$ of the test statistic, the population value can be written as $\Delta_{-} = \hat{\Delta}_{-} + \gamma \bullet \hat{SE}$, with $\hat{SE}$ being the standard error as inferred from the sample—by choosing γ to make that true. If there are grounds for supposing that $\gamma \sim \mathbb{N}(0, 1)$ then confidence intervals for Δ_{-} can be defined, with a 95-percent confidence interval's endpoints being $\hat{\Delta}_{-} \pm 1.96 \bullet \hat{SE}$ where $1.96 \cong \Phi^{-1}(0.975)$.

A better approach might be to derive the interval using quantiles of the bootstrap distribution of $t^{*} = \frac{\left(\hat{\Delta}_{-}^{*} - \hat{\Delta}_{-}\right)}{\hat{SE}^{*}}$, where the asterisk refers to bootstrap samples derived from the original sample. For a two-sided test Ledoit and Wolf (2008) formed the bootstrap distribution of $|t^{*}|$ via a complex procedure that involved doing bootstrapping twice, and in effect they used it to form a symmetric confidence interval. The confidence intervals of Table 1 of the "Hedge funds are different" subsection of the next section of this article are thus derived. Whereas the thus-found limits of the confidence intervals for the mutual fund pair and the hedge fund pair are respectively (−0.020, 0.213) and (−0.390, 1.283), the $\hat{\Delta}_{-} \pm 1.96 \bullet \hat{SE}$ endpoints are (−0.016, 0.210) and (−0.388, 1.282)—not so very different.

That $\hat{SE}$ can be substituted for SE must be considered to be a generally valid approach, and the best available method for deriving $\hat{SE}$ should be used. But if

normalcy of the sampling distribution of Δ_- can be assumed then $\hat{SE}$ can be derived from a published p-value. Start by replacing SE in equation (2a) with $\hat{SE}$, so that $Z_0 = \hat{\Delta}_-/\hat{SE}$. Then equation (3a) rewritten as $p_2 = 2 - 2\Phi(|Z_0|)$ can be inverted to become $|Z_0| = \Phi^{-1}(1 - p_2/2)$. And so, it is found that the formula for deriving $\hat{SE}$ is $\hat{SE} = |\hat{\Delta}_-| \,/\, \Phi^{-1}(1 - p_2/2)$. There is a minor complication here in that the sample $\hat{\Delta}_-$ value may differ only very slightly from zero and so may be reported, with rounding off, as zero, with p_2 practically 1.0 and possibly reported as such. That would cause both the numerator and denominator of the right-hand side of the formula for $\hat{SE}$ to be zero. But there is actually no singular behavior at those limits. Taking into consideration the dependence of p_2 on $\hat{\Delta}_-$, the limit of the formula for getting $\hat{SE}$ from $\hat{\Delta}_-$ and p_2 as $\hat{\Delta}_- \rightarrow 0$ is indeed $\hat{SE}$.

And then, if only the analyst had some idea of the population value of the test statistic, that value of $\hat{SE}$ could be used to compute the power of the test. It would just be a matter of substituting $\hat{SE}$ for SE in equations (3c) and (1b). Once again, the Δ_- that remains in the numerators of the Z arguments of equation (3b) must represent the population value, not something derived from the sample.

The power of the test in practice

It might be supposed that the power of the test may be straightforwardly confronted, a priori. An analyst would like answers to questions such as: If the true Sharpe ratio difference Δ_- has approximately this value, then, with the available history of data, do I have satisfactory power? The problem that immediately arises is of course that of the analyst having hardly any proper basis on which to proceed to estimate the value of Δ_-.

But there are other ways to proceed. The analyst may want to indulge in a form of 'a posteriori' or 'post hoc' analysis regarding the power of the test, that is pursued *subsequent* to and on top of having assessed the statistical significance of a measured effect via the use of the p-value. The goal of the analyst might then be to further interpret any finding of insignificance by giving some sort of consideration to the power of the test. Is the negative finding trustworthy, or could the true effect be positive after all?

Those negative significance findings when the true effects are positive are of course type II errors. Whereas type I errors occur with the low frequency of the chosen significance level, analysts can easily stumble into circumstances in which the power of the test is so low that the frequency of type II errors exceeds 50

percent. And so, the general idea would be to give some sort of due consideration to the adequacy of the power, in order to control the type II errors—in a way that is analogous to the p-value having been used to control type I errors by setting the significance level α to a small value. Said control might just amount to labeling some findings of insignificance as suspect due to inadequate power.

Thus if I have been disappointed by seeing a portfolio formation scheme of mine produce a Sharpe ratio difference that was ruled to be statistically insignificant, and some form of post hoc power analysis suggests that the problem might have been that I simply didn't have enough power, then that would mean that I should persevere and try to find uses for my scheme in circumstances in which more adequate data are available. That at least would be the hope.

Some researchers, statisticians being among them, have condemned post hoc power analysis as deeply flawed, even delusional. John Hoenig and Dennis Heisey (2001) has been cited about two thousand times and has been particularly influential in that regard.[13] Others, such as Len Thomas (1997) and Magdalena M. Mair et al. (2020), have been less sanguinary. Hoenig and Heisey did offer an alternative approach. All of that will be sorted out next.

In search of a pro forma approach

Strikingly simple methods have been adopted by researchers, and even by some government agencies, that nonetheless look as though they should surely be helpful. For example, there is the idea of calculating a minimum detectable effect (MDE). That is the effect size—the Sharpe ratio difference—that would, with the sample value $\hat{SE}$ of the standard error, produce a test power of, say, 0.90 (which would be an adequate power). So, the natural thought is that a low MDE is wanted: The lower it turns out to be, the more evidence we have in support of the null hypothesis when there is a finding of statistical insignificance, because the power of the test will be quite high enough even when the population value of the test statistic is close to fulfilling the null hypothesis. That's the thought, such as it is. And that is spelled out in Hoenig and Heisey (2001) and Mair et al. (2020). Or, more simply, the lower the MDE the better because a low MDE value means that the statistical significance of even a small effect can be accurately assessed with adequate power. Is that going to work out? That's the question.

I'll use the notation for the Sharpe ratio difference here, even though the findings will be generally applicable. And I'll be dealing with the right-sided test, so the power function, cf. equation (3b), is

13. Google Scholar was accessed on 9/24/2023.

$$\pi_r(\alpha, \Delta_-) = 1 - \Phi\left(\Phi^{-1}(1 - \alpha/2) - \Delta_-/\hat{SE}\right).$$

Inverting that equation,

$$\Delta_-/\hat{SE} = \Phi^{-1}(1 - \alpha/2) - \Phi^{-1}(1 - \pi_r).$$

Writing $\Delta_{-,0.90}$ for the MDE at 90-percent power, we see that it is

$$\Delta_{-,0.90} = \hat{SE} \bullet \ [\Phi^{-1}(1 - \alpha/2) - \Phi^{-1}(1 - 0.90)].$$

It is simply proportional to $\hat{SE}$.

The use that is made of the MDE is to set some threshold value for it, as a matter of policy, and to, in effect, affix *trust* labels to findings of insignificance if the calculated MDE does not exceed the threshold. Tests with MDE values greater than the threshold are labeled *mistrust*. If the label is *trust*, then the analyst trusts that the finding of insignificance is not a type II error. What could go wrong with that?

Well, because $\Delta_{-,0.90}$ is simply proportional to $\hat{SE}$, and with the constant of proportionality being just that, a fixed number, filtering MDE values with a threshold is the same as filtering $\hat{SE}$ values with a threshold. There is utterly no control of Δ_-, the population value of the test statistic, which means that there seems to be no actual control of the power of the test.

There is a related approach that consists of computing the power of the test using the minimum value of the test statistic that would be of practical interest, in lieu of the unknown population value of the test statistic. If the thus-computed power is at or above, again, say 0.90, then that warrants the *trust* label. But that would mean that the MDE would be at or below the minimum value of practical interest, so that we are again talking about imposing an MDE threshold. Thus, this related approach is really the same approach.

There is also a quantity that is called the minimum detectable difference (MDD), which is the critical value of the test statistic, $\hat{SE} \bullet \ \Phi^{-1}(1 - \alpha/2)$, which is used in exactly the same way as MDE. But again, it is just proportional to the standard error and so, just as with the MDE and the related minimum value of practical interest, it does not provide direct control of the power of the test due to the lack of involvement of the population value of the test statistic (Δ_- in this article).

The Hoenig and Heisey (2001) article refers to these efforts to determine the reliability of findings of insignificance as the power approach—though only one determinant of the power of the test is dealt with, the standard error. The authors then mention an alternative approach based on confidence intervals. And

in that discussion they remark that if neither limit of the confidence interval is far from the null value of the test statistic, which is zero in this article, then one has confidence that the null hypothesis should not be refuted. The reasoning behind the confidence interval approach outwardly seemed to be irrefutably correct, and it appears to have inspired the authors of Mair et al. (2020) to make use of the upper limit of the confidence interval in the very same way that MDD and MDE have been used: to affix the *trust* label to findings of insignificance when the upper limit of the CI is below some predetermined threshold.

Exploring that, let $CI_r = \hat{\Delta}_- + \hat{SE} \bullet \Phi^{-1}(1 - \alpha/2)$ be the upper limit of the confidence interval. The finding of insignificance will be labeled *trust* if $CI_r < \Delta_{-,max}$ where $\Delta_{-,max}$ is the predetermined threshold at or above which the label is to be *mistrust*. By inspection of the expression for the right-sided power π_r, we then have that $CI_r < \Delta_{-,max}$ means that $\pi_r(\alpha, CI_r) < \pi_r(\alpha, \Delta_{-,max})$, because $\pi_r(\alpha, \Delta_-)$ increases in a strictly monotone way with Δ_-. Thus, if we progressively decrease $\Delta_{-,max}$ so as to confine the upper limit of the confidence interval to the immediate neighborhood of zero, which should provide confidence-interval-approach assurances to the effect that the null hypothesis should indeed not be refuted, the power to determine the reliability of that conclusion is diminished. So is the confidence interval approach at odds with the the so-called power approach? This is indeed perplexing. Hoenig and Heisey brought this to light, but their examples and discussion differ from what I'm presenting here. Read on. This will be resolved.

Trying again, I then again explored the use of the upper limit of the CI, again with regard to the right-sided test, but with the studentized version of the upper limit: The studentized version of the upper limit is $\hat{\Delta}_-/\hat{SE} + \Phi^{-1}(1 - \alpha/2)$, which is $\Phi^{-1}(1 - p_r) + \Phi^{-1}(1 - \alpha/2)$. In that form we see that trusting a finding of insignificance because the studentized upper limit of the CI is below some predetermined threshold is the same as trusting such findings when $\Phi^{-1}(1 - p_r)$ is is below some predetermined threshold, which is the same as trusting the findings when p_r is above some predetermined threshold.

Thus, the hope of being able to use the studentized upper limit of the CI to better determine the reliability of findings of insignificance went up in smoke. It boiled down to classifying the reliability based solely on the p-value, with higher p-values providing greater evidence of reliability. That is of course entirely consistent with prior expectations concerning the p-value, so that nothing new was discovered. And the computation and use of the p-value does not involve the population value of the test statistic (Δ_- in this article). And so again there is no direct relevance to the power of the test because the power is so very dependent upon the unknown population value of the test statistic.

In short, the appearances seem to be that there is no post hoc way to bring about an improved understanding of the reliability of findings of insignificance. It's all because the population value of the test statistic remains unknown and unknowable. According to Hoenig and Heisey (2001), one must leave what they call the power approach at that. But, I took a second look at the mutually-related MDE and minimum value of practical interest schemes, both of which amount to classifying findings of insignificance as trusted or mistrusted, respectively according to whether $\hat{SE} < SE_{\max}$. What should $SE_{\max}$ be? If $\Delta_{-,PRAC}$ is the minimum value of the Sharpe ratio difference that is of practical interest, suppose that $SE_{\max}$ is chosen to be the value of $\hat{SE}$ that would, with the value of the test statistic being $\Delta_{-,PRAC}$, produce a right-sided power of 0.90. Thus $SE_{\max} \equiv \Delta_{-,PRAC} / [\Phi^{-1}(1 - \alpha/2) - \Phi^{-1}(1 - 0.90)]$.

$\hat{SE}$ can be derived from the sample by inverting the definition of the right-sided p-value p_r, getting $\hat{SE} = \left|\hat{\Delta}_-\right| / |\Phi^{-1}(1 - p_r)|$. Or, rather than making use of a p-value estimator, an estimator for the standard error may be directly used. Therefore, a test based on $\hat{SE} < SE_{\max}$ could be implemented whenever the original hypothesis test has been concluded with a finding of statistical insignificance ($p_r \geq \alpha/2$)—with the goal being to characterize the finding of statistical insignificance with respect to its reliability, given that the power of the test may be suspect.

The starting point of this idea was the vague notion that an $\hat{SE}$ value being known to be low, via $\hat{SE} < SE_{\max}$, seems to imply that the only way that the power could be low would be if Δ_-, the population value of the Sharpe ratio difference, were to also be low—low but not necessarily zero or negative. If Δ_- were not low then the low $\hat{SE}$ would mean that power would be high. But a high power with Δ_- not being low should mean that the finding of insignificance would not have happened in the first place. Thus, perhaps low $\hat{SE}$ can be taken to truly mean low Δ_-, so that one might try to turn $\hat{SE} < SE_{\max}$ into a provision that more or less guarantees that if the finding of statistical insignificance doesn't mean that the population value of the test statistic is zero or negative then it at least means that the value is almost certainly less than $\Delta_{-,PRAC}$.

But of course, that is just a sketchy idea. To further investigate, I also implemented a *secondary* left-sided hypothesis test whose null hypothesis is H_0: $\Delta_- \geq \Delta_{-,PRAC}$. Hoenig and Heisey (2001, 4) propose the use of a two-tailed version of this non-traditional casting of a null hypothesis, and present it as amounting to abandonment of the flawed power approach in favor of the confidence interval approach. A finding of statistical significance with respect to the secondary hypothesis test is then reached with rejection of the secondary null

hypothesis if $\hat{\Delta}_- < \Delta_{-,CRIT}$ where $\Delta_{-,CRIT}$ is de rigueur, defined by choosing it so as to make the significance level of the left-sided secondary hypothesis test be $\alpha/2$, that hypothesis test's type I error rate, as in $\frac{\alpha}{2} = \Phi\left(\frac{\Delta_{-,CRIT} - \Delta_{-,PRAC}}{\hat{SE}}\right)$. Thus $\Delta_{-,CRIT} = \Delta_{-,PRAC} + \hat{SE} \bullet \Phi^{-1}(\alpha/2)$.

Does this look familiar? If we take the test that was introduced above, which was adapted from the formulation of Mair et al. (2020), which was $CI_r < \Delta_{-,max}$ with $CI_r = \hat{\Delta}_- + \hat{SE} \bullet \Phi^{-1}(1 - \alpha/2)$ being the upper limit of the confidence interval, the fact that $\Phi^{-1}\left(1 - \frac{\alpha}{2}\right) = -\Phi^{-1}\left(\frac{\alpha}{2}\right)$ means that if we set $\Delta_{-,max} \equiv \Delta_{-,PRAC}$ then indeed the two tests are one and the same. But now it's understood to be a bona fide hypothesis test of significance level $\alpha/2$ with a well-defined attendant null hypothesis.

The simultaneous use of a secondary hypothesis test makes all the difference. For clarity in the moment, I shall affix *accept* and *don't accept* labels, respectively according to whether $\hat{\Delta}_- < \hat{\Delta}_{-,CRIT}$, in lieu of *trust* and *mistrust*. The difference is that by *accept* I mean accept the finding of statistical insignificance with the use of the original null hypothesis, but with the understanding that the attendant rejection of the secondary hypothesis test's null hypothesis means that in reality Δ_- might in fact be as large as $\Delta_{-,PRAC}$ (and not restricted to being as large as zero as in the primary null hypothesis). But the *don't accept* label means much the same as the *mistrust* label—it is only to be taken to be an indication that it's best to assume that the available data are not sufficient for reliable hypothesis testing. Any Sharpe ratio difference that earns the *don't accept* label is to be regarded as not reliably replicable with regard to sign or magnitude; the *accept* label means that the population value of the difference very likely does not exceed $\Delta_{-,PRAC}$.

Consider what that means. The goal was to somehow dodge the adverse effect of the original hypothesis test having type II errors. With those type II errors there is a finding of statistical insignificance, even though the null hypothesis H_0: $\Delta_- \leq 0$ is false, meaning that $\Delta_- > 0$. With type I errors of the secondary hypothesis test there is a finding of statistical significance, appending the *accept* label to the original hypothesis test's finding of insignificance, even though the secondary null hypothesis H_0: $\Delta_- \geq \Delta_{-,PRAC}$ is true, meaning that $\Delta_- \geq \Delta_{-,PRAC}$. So the type II error of the original hypothesis test occurs with $\Delta_- > 0$ and the type I error of the secondary hypothesis test occurs with $\Delta_- \geq \Delta_{-,PRAC}$. But these are similar inequalities. There is just a shift in the right-hand side and the analyst controls the magnitude of the shift.

And now here's the big difference. With the original hypothesis test we have no way of knowing what the probability of occurrence of type II errors is, due to the power of the test not being known. But, notwithstanding the fact that the

secondary hypothesis test's type I error is a doppelgänger of the original test's type II error, we *do* know the probability of *its* occurrence: It's $\alpha/2$. And as such, *it's controlled.*

And, about the type II errors of the secondary hypothesis test, they consist of affixing the *don't accept* label to findings of statistical insignificance under the original null hypothesis of H_0: $\Delta_- \leq 0$, after having failed to reject the null hypothesis H_0: $\Delta_- \geq \Delta_{-,PRAC}$ of the secondary hypothesis test when in fact $\Delta_- < \Delta_{-,PRAC}$. Note especially that the secondary hypothesis test cannot be used backhandedly, in stand-alone fashion, to establish statistically significant outperformance. A finding of statistical insignificance under it only means that the *possibility* of $\Delta_- \geq \Delta_{-,PRAC}$ can't be dismissed. And the power of the test being unknown, uncontrolled, means that any such finding would be confounded by the potentially high type II error rate. The left-sided secondary hypothesis test is only suitable for its ability to lend confirmation, of a sort, to some of the findings of insignificance under the original right-sided hypothesis test.

Now it might come to mind that some analyst might elect to just proceed with the original hypothesis test alone, and to take a finding of statistical insignificance under it to just mean that further study with more adequate data was needed—declining to make a spot judgment to the effect that the investment alternative of interest was not likely to become an outperformer. Yes, but the improvement that is brought about with the use of the secondary hypothesis test is the immediate resolution of the statistical insignificance findings in some cases, permitting the investment alternative to immediately be categorized as unlikely to be destined to offer outperformance of practical significance. That is not a huge victory, but it is beneficial.

I did apply both the $\hat{SE} < SE_{\max}$ test and the secondary hypothesis test to real data—to 47 findings of insignificance that I found in three published articles. Details about the findings of the three studies are to be found in the "More literature" section below. Of those 47 findings of insignificance, the standard error test rated nine as *accept* and the secondary hypothesis test rated 13 as *accept*. In no case was the standard error *accept* result overruled by the secondary hypothesis test; the latter awarded the *accept* label more liberally.

Of the three articles, one was based on a backtesting interval that was longer by far than that of the other two. And the frequency with which the *accept* label was awarded using the secondary hypothesis test was highest by far with that article—suggesting a positive correlation between the secondary hypothesis test's frequency of acceptance of the original hypothesis test's findings of insignificance and the duration of the backtesting interval. Such a correlation can come about in practice through the role played by the standard error $\hat{SE}$ in the formula $\Delta_{-,CRIT} =$

$\Delta_{-,PRAC} + \hat{SE} \bullet \Phi^{-1}(\alpha/2)$ for the critical value of the secondary hypothesis test. Bearing in mind that $\Phi^{-1}(\alpha/2)$ is negative, $\Delta_{-,CRIT}$ is increased if $\hat{SE}$ is reduced, meaning that *accept* labeling should be more frequent. $\hat{SE}$ is reduced if the duration of the backtesting interval is increased. Hence the positive correlation.

Because the type I error rate of the secondary hypothesis test is known and controlled, I am inclined to use it rather than the standard error test. But on occasion power-approach-like consideration of the size of the standard error nonetheless suffices. For example, suppose that the analyst deliberately conjures up, out of whole cloth, a value for the population value of the test statistic that is implausibly high, and with it and the sample-derived standard error computes, using equation (3b) or its right-sided counterpart, the power that the test would have with that implausibly high value in place of the population value of the test statistic. If the thus-computed power were to be unacceptably low, then it would be reasonable to conclude that the true power of the test was unacceptably low. This would be much the same as finding the value of the MDE to be implausibly high for consideration as a population value of the test statistic.

The section "The last word" near the end of this article definitively demonstrates the folly of conducting hypothesis tests when the power of the test is really too low. It's not easy to come up with an example of a portfolio pair that is such that the analyst can be confident in advance that the true Sharpe ratio difference test statistic, be it Δ_- or $|\Delta_-|$, is quite substantially greater than zero, with investors seriously interested in seeing the difference tested for statistical significance. For the most part, we are left with understanding only that the duration of the backtesting interval being too short or the correlation coefficient between the returns of the paired portfolios not being high enough (meaning that the standard error is not small enough), and the population value of the test statistic not being high enough, singly or together in concert, are factors that can readily cause the test to have inadequate power. But the secondary hypothesis test that I have described, which is an adaptation of Hoenig and Heisey's (2001) confidence interval scheme as nearly emulated by Mair et al. (2020), can be helpful at resolving some outcomes as definitely being unpromising, even when it's not known in advance that the power of the test is adequate for that purpose.

Hedge funds are different

Lo (2002, 44) provides data on actual funds that show that the returns of hedge funds are autocorrelated to a much greater degree than those of mutual funds. Benjamin R. Auer and Frank Schuhmacher (2013, 201) confirm strong autocorrelations in hedge fund returns. It's hardly obvious as to why that might be

the case, but Lo and colleagues did present a feasible explanation: Mila Getmansky, Andrew W. Lo, and Igor Makarov (2004) investigated several mechanisms and found that only the effect of the illiquidity of assets held by hedge funds was enough to account for the observed degree of serial correlation. The fact that market prices are not frequently reestablished for illiquid investments means that there is effectively a sort of smoothing process affecting the returns, which acts like a trailing moving average. Since hedge funds hold securities both long and short, market risk is greatly diminished and that causes the funds to sport high Sharpe ratios. But the smoothing effect of the illiquidity further reduces the volatility and further boosts Sharpe ratios—by 73 percent in one studied example that involved smoothing over just two reporting periods.

Relatedly, Ledoit and Wolf (2008, 857) shows results pertaining to simulated portfolio returns. One pair of portfolios was given simulated i.i.d. returns drawn from a bivariate-normal distribution, as in the Jobson and Korkie (1981) work; other pairs were simulated so as to exhibit autocorrelations and other non-i.i.d., and non-bivariate-normal characteristics, such as might be exhibited by hedge funds. The Memmel (2003) corrected Jobson and Korkie model was configured to compute p-values, of iterates. And from the p-values the probability of rejecting the null hypothesis was computed, for the two-sided test whose statistic is the absolute value of the Sharpe ratio difference. The simulations emulated the null hypothesis being true, by forcing the two Sharpe ratios to be equal.

It is explained above that this rejection probability must be the significance level α. Ledoit and Wolf's findings, in their Table 1, are that the Memmel-corrected model brought about overestimates of the rejection probability when applied to simulated returns that had characteristics in common with those of hedge fund returns: For the five given flavors of departure from i.i.d. bivariate-normal returns the computed powers of the test when the assumed value of α was 0.05 were not 0.05 but were 10.7, 7.2, 7.4, 9.5, and 14.5. Understand that to get these particular estimates Ledoit and Wolf did not use their method for conducting hypothesis testing; they are the result of applying the Memmel-corrected model to simulated hedge-fund-like returns.

Auer and Schuhmacher (2013, 203) provide some confirmation of that result using real hedge fund data. Auer and Schuhmacher show that when the same Memmel-corrected model is applied to calculating the p-values of hedge funds, findings of statistical significance are reached much more often than with the method and code of Ledoit and Wolf which accounts for and compensates for non-i.i.d. and non-bivariate-normal characteristics. Assuming that the Ledoit and Wolf method and codebase doesn't overestimate p-values of hedge funds, this can be taken to mean that the Memmel-corrected model underestimates p-values of hedge funds. These findings and Ledoit and Wolf's rejection probability tests both

indicate that the Memmel-corrected model produces a standard error (a standard deviation of the Sharpe ratio difference) that is much too small for use with hedge funds. And that leads to p-values that are too low and power estimates that are too high.

And although it cannot always be counted on, by their very nature hedge funds tend to have returns that are weakly correlated with the returns of almost any other investment, including other hedge funds. As has been shown above, this has negative implications regarding the power of the test whose statistic is a Sharpe ratio difference.

In their own article Ledoit and Wolf (2008) show p-values for two pairs of funds. If there is a reason for supposing in advance that the true Δ of either of the two pairs of is well removed from zero, I don't know what it could be. So, it can't just be assumed that power of the hypothesis test was sufficient. I therefore revised Tables 2 and 3 of Ledoit and Wolf (2008, 857–858), adding confidence intervals. To do so I made minor changes to Ledoit and Wolf's R code for computing the p-value. My findings are shown in Table 1 below. The p-values and confidence intervals do not depend upon the significance level α. That's a hypothesis testing thing. If you don't test the hypothesis with the computed p-value, then you don't need and therefore don't have a significance level.

TABLE 1. P-values converted into confidence intervals

Fund pair	Sharpe ratio	P-value	Sharpe ratio difference
Fidelity Fidelity Aggressive Growth	0.108 0.011	0.092	0.097 [95.0% CI: −0.020, 0.213]
Coast Enhance Income JMG Capital Partners	1.461 1.014	0.294	0.447 [95.0% CI: −0.390, 1.283]

I did investigate the power of the test for the Coast Enhance Income–JMG Capital Partners hedge fund example—using the given sample values of the Sharpe ratios, the p-value as computed by Ledoit and Wolf, and the means described at the conclusion of the "Confidence intervals and the standard error" subsection of the "Mathematics of power—i.i.d. bivariate-normal returns" section above. I assumed various values of the population value of the Sharpe ratio difference, 0.40, 0.60, 0.80, and 1.0. These are conjectured monthly values, not annualized, and they are deliberately chosen to vary from being large to being very-very large (even for hedge funds). Surely the likes of one of those should produce adequate power, one might suppose. But no. The corresponding powers of the test are 0.16, 0.29, 0.47, and 0.65. That supports the claim that I made in the introduction of this article to the effect that it was certainly improper to have presented this particular example without disclosure of the likelihood of the power being too low. But also,

the secondary hypothesis test produced a *don't accept* label for this pair, which is consistent with the power being too low.

More literature

I was curious to see what others had done with the Ledoit and Wolf (2008) procedure. Whereas the article has been cited 981 times since 2008, it was cited 136 times in 2022.[14] It is not going out of style. I discovered that there is an entire category of studies, some quite recent, that use the Ledoit and Wolf (2008) method and codebase to determine if portfolio and benchmark Sharpe ratios differ by statistically significant margins, when at the outset there is no compelling reason to suppose that the ratios should substantially differ.

Why investigate when there's no compelling reason to suppose that there would be substantial differences? Perhaps the authors of such studies privately stand in disbelief of the claims of *others* about some innovation or supposedly superior way of investing. So, their intention could be to dispassionately demonstrate that the true believers are wrong by showing that the observed Sharpe ratio differences lack statistical significance.

But if that is indeed their plan, then what should their thoughts be about the power of the test? They should be concerned that the power of the test would often be too low because there would be a systematic tendency for the magnitudes of the population values of the Sharpe ratio differences, the Δ_- values, to not be well removed from zero (cf. Figure 1). I have explained, and make especially clear in "The last word" section below, that it is foolhardy to attempt significance testing if there is every reason to suspect the power of the test to be quite low.

The aforementioned article by Auer and Schuhmacher (2013) is a study of hedge fund performance that involves 4,322 funds divided into 19 categories with a different benchmark for each category. For each of the funds the authors calculated the statistical significance of the difference between the fund's Sharpe ratio and the benchmark's. In their abstract the authors state that "Only a small fraction of hedge funds in our large dataset can significantly outperform passive investments in corresponding hedge fund indices." And by "significantly" they did mean to refer to statistical significance. With that broad brush they condemn the entire hedge fund industry.

Remarkably, this study involves benchmarks that were each presumably designed to do a good job of representing hedge fund performance within each category. That should have caused Auer and Schuhmacher to have no reason to

14. Google Scholar accessed May 13, 2023.

suppose that the magnitudes of many of the Δ_- values were well removed from zero. Furthermore, the authors only eliminated funds having 24 or fewer months of historical returns data from their study, so that in some cases the number of observations was dismally low. Finally, unlike, say, long-only mutual funds, it can't be assumed that hedge funds will generally be well correlated with their benchmarks. These are all foreboding indications that suggest that the power of the test was too low in many of the cases considered.

Since the low power of the test means that false null hypotheses will rather often not be rejected, it is understandable why few of the hedge funds were found to have outperformed their benchmarks in a statistically significant way—type II errors. The authors would have been better off speculating that the dearth of findings of significance was due to the designers of the benchmarks having done their jobs well, leading to generally low magnitudes of the Δ_- values and low power, rather than to lackluster hedge fund performance.[15]

I began to write something to the effect that in recent years dozens upon dozens of analysts have undertaken the task of estimating the efficacy of investing in stocks of corporations that are governed in such a way as to qualify as 'ESG' according to certain guidelines. I was a bit off! Google Scholar finds 484,000 hits on 'ESG' and 48 of them are among the articles citing Ledoit and Wolf (2008).[16] Generally, in one way or another these investigations involve substituting an ESG-qualified security for one of an enterprise that is otherwise similar in ESG-unrelated ways. The studies are based on the returns on the stocks of the companies, not upon environmental or social benefits brought about by ESG operation.

Of course, the authors of these studies don't have good a priori reasons to suppose that ESG investing would be either beneficial or harmful to investors. Expecting a professor of finance to venture into the jungle to investigate whether growing coffee in the shade is going to be lucrative or not is asking too much. Then too, if ESG operations are good for investors and the stock market is efficient, an announcement of conversion to ESG operation by a company should immediately be greeted by a sudden large increase in the price of the company's stock—with the gains to investors being realized at once and not necessarily occurring within the

15. Curiously, the authors state on page 198 that "… the statistical power of the test is low, especially for small sample sizes. Thus, a significant test result can be seen as strong evidence of a difference in risk-adjusted performance." But they say this only in reference to the Jobson and Korkie findings for i.i.d. bivariate-normal returns. And it is odd that they only mention an upside to the problem of low power. They may have concluded that Ledoit and Wolf's method somehow improves the power of the test. And Ledoit and Wolf did claim that their method worked especially well with small sample sizes. That would explain why they are not concerned about low power and small sample sizes affecting their own Ledoit and Wolf test findings.

16. Accessed on May 6, 2023.

study periods of these various ESG articles.

So, again, were the authors of such ESG studies to have considered the power of the test they would have had to assume that the population values of the Δ_- Sharpe ratio differences would be systematically of low magnitude—meaning that the power of the test would be low.

Consider Ricardo de Souza Tavares and Joao Frois Caldeira (2023, 61), who ask "Is replacing standard investments with ESG substitutes a good choice?" There was only about a decade of monthly data. Table 6 of that article shows that the procedure of Ledoit and Wolf (2008) assigned Sharpe ratio difference p-values under $\alpha = 0.05$ to only three of the 12 market indexes that were modified by ESG substitutions. And although the authors do state that this means that generally there is a lack of statistical significance, they don't mention the power of the test. One is left wondering if the findings of insignificance can be trusted, due to the possibility of the power of the test possibly being systematically low.

Costanza Torricelli and Beatrice Bertelli (2022, 18) address "The trade-off between ESG screening and portfolio diversification in the short and in the long run." In their Table 5, the significance of Sharpe ratio differences is assessed despite the data being monthly and there being only 15–18 months of data. The method of Ledoit and Wolf (2008) was used. If somehow in any of the studied cases the population value of the Sharpe ratio difference was sizeable, the power of the test would still be quite low due to the very low number of observations. With $\alpha = 0.05$ the significance level, the authors find significance with only two out of 24 endpoints. Some such outcomes would be expected given the low power of the test.[17]

With all such considerations of ESG investment there are apt to be several alternatives that have been defined and tested—different guidelines for acceptability, substitutions made in different market indices, involvement with different alternative mean-variance optimization models, etc. Thus, the problems discussed above in "The mathematics of hypothesis testing" section of this article that are brought about by needing to choose among investment alternatives when the hypothesis tests are of low power come to the fore.

I had written all of the above, about the Opdyke (2007) article with his

17. A note at the bottom of their Table A1 states that the tests of *that* table, which do not include the Sharpe ratio difference test, were not applied to the short-term subperiods because the power would be inadequate given the small number of observations. Why then was significance testing conducted with the Sharpe ratio difference over the short-term subperiods on Table 5? Quite possibly it is because Ledoit and Wolf (2008) fails to mention the power of the test but contains statements that suggest that the Ledoit and Wolf method is especially good about working well with small sample sizes. Indeed, in the text Torricelli and Bertelli (2022) state that the method "is robust to non-normality, correlation and errors due to small samples," which would seem to mean that it was assumed that the method could cure the problem of the low power due to the small sample size.

analysis of just three years of weekly data on certain mutual funds and about the research by Torricelli and Bertelli (2022) and by Tavares and Caldeira (2023), before I had settled on the secondary hypothesis test that is described in the "In search of a pro forma approach" subsection of "The power of the test in practice" section above in this article. Does the secondary hypothesis test help with the interpretation of the findings of these authors?

For the monthly Torricelli and Bertelli (2022) and Tavares and Caldeira (2023) data I set $\Delta_{-,PRAC} = 0.04$ which when annualized would amount to a Sharpe ratio improvement of approximately $0.04\sqrt{12} \cong 0.14$ which would be a small but meaningful improvement. And for the weekly Opdyke (2007) data I chose $\Delta_{-,PRAC} = 0.02$ which would be nearly equivalent on an annualized basis. Of all of the 15 Opdyke mutual fund matchups with his two best-performing funds (in the first two columns of his p. 20 Table IV) that produced p-values indicating statistical insignificance (with $\alpha = 0.05$), exactly none earned the *accept* label. That is entirely consistent with the fact that he had only three years of data, which would have led to low power of the original hypothesis test, hence the denial of *accept* labeling.

Similarly, the study of Torricelli and Bertelli (2022) that involved only 15–18 months of data would surely have been affected by low power, and of the 22 findings of statistical insignificance only six earned the *accept* label. But Tavares and Caldeira (2023) had a decade of monthly data, which is an amount that could in some circumstances be adequate. And seven of the nine findings of statistical insignificance (with $\alpha = 0.05$) earned the *accept* label. That seems to demonstrate a certain efficacy of the secondary hypothesis test at reducing the uncertainty in findings of insignificance when the power of the test is suspect. It would have made it possible for Tavares and Caldeira to have written, with a known measure of certainty, that with those seven outcomes there would be no improvement in the annualized Sharpe ratio difference of more than about 0.14. They were instead only able to write that "we cannot say that there is any difference between the Sharpe ratios," which is decidedly less informative.

The recourse to confidence intervals

Larry Wasserman's *All of Statistics* (2004) is a textbook by an academician, a statistician. And in it Wasserman wrote this: "There is a tendency to use hypothesis testing methods even when they are not appropriate. Often, estimation and confidence intervals are better tools." But that was before the ball got rolling. Since then, others have chimed in, with statisticians that work in medical research leading the way. Valentin Amrhein et al. (2019) offered a letter that was signed by 800 researchers that appeared as a comment in Nature with the heading "Retire

statistical significance." John Ioannidis (2019) asked "What Have We (Not) Learnt from Millions of Scientific Papers with P Values?" And Stephen Ziliak and Deirdre McCloskey wrote a book with the title *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives* (2008).

To be sure, the objections are not all focused on the untoward aspects of hypothesis testing that appear when the adequacy of the power of the test is in question, because it is reasonable to simply complain about the black and white aspect of significance testing. That would be a reasonable complaint in and of itself. But John Ioannidis is a coauthor of Button et al. (2013), which is indeed particularly about the perils of testing when the power is too low. So yes, the complaints are in good measure about the damage done when the power is too low.

Means of computing confidence intervals have been presented above in the "Confidence intervals and the standard error" subsection of the "Mathematics of power—i.i.d. bivariate-normal returns" section. And it was shown that standard errors, and therefore confidence intervals, can be derived from published p-values. That means that whenever an article is encountered that lists p-values of Sharpe ratio differences and tells us whether the differences were found to be statistically significant or not, there is no need to worry about whether the power of the test was high enough. Approximate confidence intervals can be calculated and analysts and investors alike can be content with them.

I have been asked to show how 'investment horizons' might be worked into the topic of this article. There is nothing definite or exact about the bare notion of an investment horizon. That is, there are many entirely different ways of responding to the needs of investors who want to be assured that their money will be there when they need it at some point in the future, such as at retirement. But one way to implement an investment horizon is to modify the confidence intervals of the performance measures. The method that I show here is applicable to any performance measure that could serve as a hypothesis test statistic, not just to Sharpe ratio differences. But I'll use the notation that was used for Sharpe ratio differences.

The formula is simple. Above, confidence interval endpoints for a 95-percent confidence interval were found to be $\hat{\Delta}_- \pm 1.96\,\hat{SE}$ with $\hat{\Delta}_-$ being the sample value of the Sharpe ratio difference and $\hat{SE}$ being the sample-derived standard error of that statistic. That would be for normally distributed values of $\hat{\Delta}_-$. With, say, a bootstrap distribution of $t^* = \frac{\left(\hat{\Delta}_-^* - \hat{\Delta}_-\right)}{\hat{SE}^*}$, the 95-percent confidence interval endpoints would be $\left(\hat{\Delta}_- - z_{97.5}\hat{SE},\ \hat{\Delta}_- - z_{2.5}\hat{SE}\right)$, as Tim C. Hesterberg (2015, 381) demonstrates. Here $z_{2.5}$ and $z_{97.5}$ are quantiles of the distribution of t^*

so that $z_{2.5}$ is negative. To implement an investment horizon of T_{hor} periods into the future as a confidence-interval modification, simply multiply $\hat{SE}$ in the formula for the endpoints by $\sqrt{1 + T/T_{hor}}$ where T is, as before, the number of periods in the sample.

How does this come about? It's easier done than said. The random variable $\hat{\Delta}_- - \Delta_-$ is distributed with standard error SE, which pertains to T observations of historical data, whose estimator yields the sample value $\hat{SE}$. The distribution's 2.5 and 97.5 percentiles give us the 95-percent confidence interval endpoints. Similarly, the random variable $\Delta_- - \hat{\Delta}_{-,hor}$ is distributed with standard error SE_{hor}, which pertains to T_{hor} observations of future data—for which an estimator is needed that somehow yields a $\hat{SE}_{hor}$ value that is derived from the only data that is available, the data of the historical sample.

To get horizon-adjusted confidence intervals, simply get the 2.5 and 97.5 percentiles of the distribution of $\hat{\Delta}_- - \hat{\Delta}_{-,hor}$, which is $\left(\hat{\Delta}_- - \Delta_-\right) + \left(\Delta_- - \hat{\Delta}_{-,hor}\right)$ —after finding this distribution's standard error. And what is it? Well, the squares of the standard errors of the distributions are all just variances. Since these are sampling distributions, pertaining to entirely different time intervals at that, the values of the Sharpe ratio differences would not be correlated. So, the square of the standard error of the distribution of $\hat{\Delta}_- - \hat{\Delta}_{-,hor}$ is the sum of the squares of the standard errors of the distributions of $\hat{\Delta}_- - \Delta_-$ and of $\Delta_- - \hat{\Delta}_{-,hor}$. That is, $SE^2_{adj} = SE^2 + SE^2_{hor}$. And since the square of the estimator value of the standard error scales with the reciprocal of the number of observations, $\hat{SE}^2_{hor}$ may be taken to be $(T/T_{hor})\,\hat{SE}^2$. Thus $\hat{SE}^2_{adj} = \left(1 + \frac{T}{T_{hor}}\right)\hat{SE}^2$. Hence the investment-horizon confidence-interval adjusting factor of $\sqrt{1 + T/T_{hor}}$.

If T_{hor} is very large then the adjustment factor is about 1, which is consonant with remarks made above to the effect that Δ_-, the population value of the Sharpe ratio difference, can be regarded as the long-term future value. Or, if T is much smaller than T_{hor}, which would not be a particularly unusual circumstance to encounter in practice, then the adjustment factor would be not much more than 1. That would be because the initial uncertainty that is brought about by T being small would be dominant over that brought about by the larger T_{hor}, so that the initial uncertainty would not need much adjustment.

In practice, the horizon-conscious investor who needs to choose a safe-enough investment from a list of investments could rank the investments by, say, the lower bounds of the horizon-adjusted confidence intervals of the Sharpe ratio differences rather than by the Sharpe ratio differences. In general, the two rankings would be quite different. It's interesting to note that whereas the Sharpe ratio is a

risk-adjusted measure of returns, the lower bound of the Sharpe ratio difference confidence interval is a risk-adjusted measure of the Sharpe ratio difference (because the term in $\hat{SE}$ is subtracted, amounting to a penalty for volatility).

Unfortunately, selection bias must still be dealt with when there are multiple investment alternatives under consideration and confidence intervals are in use. Cross-validation and walk-forward methods can mitigate selection bias. But also, a problem is encountered that is analogous to the problem of the false error rate accumulating with multiple hypotheses that is discussed in "The mathematics of hypothesis testing" section of this article. There is a similar kind of accumulation of errors, of the failures of the population values of the performance statistic to lie within the confidence intervals. But again, there are means of dealing with that. The confidence intervals can be adjusted (see Benjamini and Yekutieli 2005; Benjamini, Hechtlinger, and Stark 2019). These remedial measures for dealing with the multiple hypothesis problem have nothing in particular to do with Sharpe ratio differences but are quite generally applicable.

The last word

The power of any given properly stated hypothesis test must always be regarded as not being up for grabs. That is, the power of the test is intrinsic and immutable once the test statistic and the null hypothesis are both defined and the significance level α has been chosen. As Sir Ronald Aylmer Fisher (1935) himself put it: "It is evident that the null hypothesis must be exact, that is free from vagueness and ambiguity, because it must supply the basis of the 'problem of distribution,' of which the test of significance is the solution." If the null hypothesis is indeed exact, then the power follows unambiguously for it is, by definition, a probability that can be straightforwardly thought of as a frequency of occurrence.

What remains that might be confusedly regarded as amounting to mutability of the power of the test is the struggle to invent improved estimators of the p-value, that are not grossly biased. Equation (1a) of this article is exact (in the limit of large T), provided that the returns of the two portfolios are i.i.d. bivariate-normal. That is, the estimator of equation (3a) is consistent if it incorporates (1a) via (2a). But if the returns are not distributed in that idealized way, then (3a) becomes a biased estimator. Auer and Schuhmacher (2013) confirm that it produces, with hedge funds whose returns are far from i.i.d. bivariate-normal, p-values that are very substantially lower than those computed using the method of Ledoit and Wolf (2008). Because of the way that it was derived it must be assumed that the Ledoit and Wolf estimator to be the least biased of the two estimators—able to cope with autocorrelations and heteroskedasticity.

But the p-value estimator of Ledoit and Wolf must in essence also be, and is, about as exact in application to i.i.d. bivariate-normal returns as equation (3a). That is confirmed by Figures 2 and 3 which appear further on below. Thus, for all its sophistication, if the power of the test with such idealized returns is low due to factors such as the duration of the backtesting interval not being long enough, the true Sharpe ratio difference not being high enough, or the correlation coefficient being too low, then the Ledoit and Wolf estimator can't and doesn't overcome that. The low power is innate.

One path to better understanding is to pay attention to the effect of the power of the test on the sampling distribution of the p-value. Figure 2 shows the sampling distribution of the two-sided p-value, computed via simulation of the histories of the returns of the two portfolios. Note that on the figure there are twenty histogram bars, so that the boundary between the first and second bars falls, by design, at a p-value of 0.05 which is the assumed significance level α. With each of the 1,000 trials the p-value was calculated using both the method and code of Ledoit and Wolf and the asymptotic method of equation (3a). Equation (1a) provided the estimate of the standard error for the asymptotic approach.[18]

Note especially, in the legend, that the population value of the Sharpe ratio difference is non-zero, positive. Therefore, the fraction of all the p-values that exceed $\alpha = 0.05$ is the type II error rate β. And from the legend, β is about 0.07 so that the power as computed from the sampling distribution of the p-value is about $1 - 0.07 = 0.93$, which is very good. And that agrees nicely with what is independently computed with the asymptotic expression of (3b), which is 0.929.

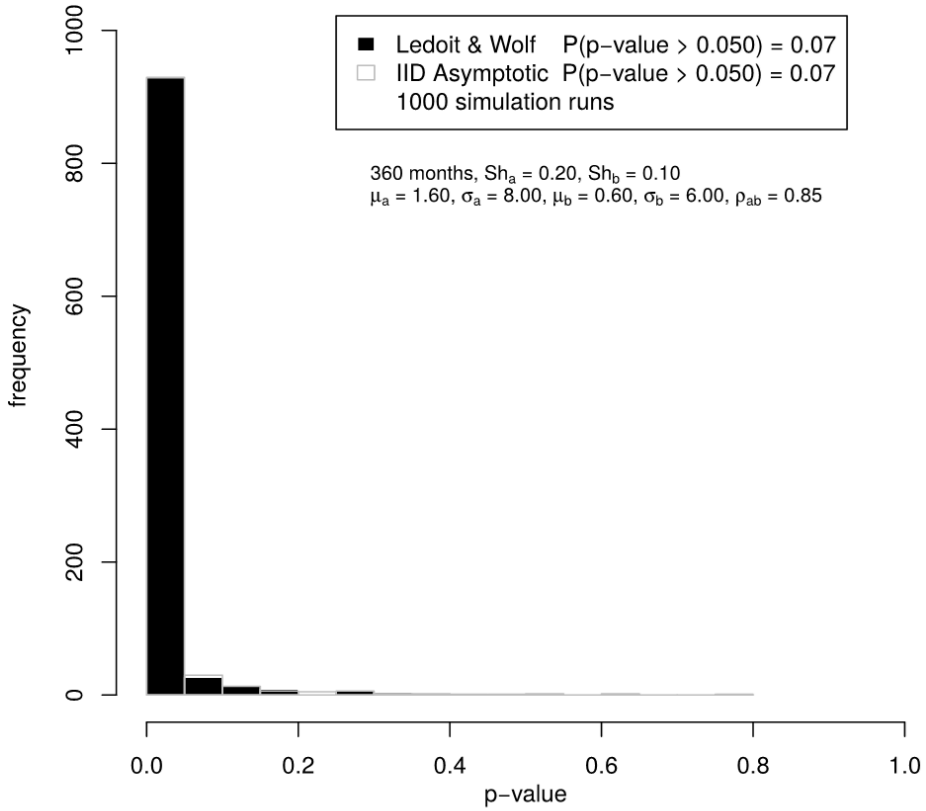
The following is found in Wasserman's text (2004, 158): "In other words, if H_0 is true, the p-value is like a random draw from a Uniform (0,1) distribution. If H_1 is true, the distribution of the p-value will tend to concentrate closer to 0." His H_0 can be taken to be the two-sided null hypothesis $\Delta_- = 0$, and then his H_1 would be $|\Delta_-| > 0$.

On Figure 2 the distribution of the p-value is well-concentrated close to 0. But it is not a really a matter of one or the other—of the p-values being either uniformly distributed or concentrated closer to 0. Rather, the higher $|\Delta_-|$, the greater the concentration near 0. Thus "if H_0 is true" can be taken to mean in the limit of $|\Delta_-| \to 0$, in which case there is no concentration near 0 but just an utterly

18. To simulate the return histories of two portfolios a sample a is prepared using random selection from $N(0, 1)$ and then a second one b is prepared in the same way but is then altered by replacing it by $\sqrt{1 - \rho_{ab}^2}$ times itself plus ρ_{ab} times the first, with $\rho_{ab} \subset [0, 1]$—thus establishing a substantial amount of cross correlation between the first sample and the altered second sample. The samples are then parameterized: The returns of the first sample are multiplied by σ_a and then μ_a is added; the altered second sample's returns are multiplied by σ_b and then μ_b is added.

flat histogram.

Figure 2. An example with 360 observations



Notes: Here $\Delta_{-} = Sh_a - Sh_b$ is the a priori population value of the Sharpe ratio difference. And P(p-value > 0.05) is the probability of a type II error, whose complement is the power of the test.

Wasserman states that it's clear that the histogram must be flat when H_0 is true because that is consistent with the probability of committing a type I error being α, the significance level (which is 0.05 in the figure), when the null hypothesis is rejected because the p-value is less than α. That is, the area under the first bar of the histogram must be one-twentieth of the total area under the 20-bar histogram, because that would be 0.05, the value of α. That is consistent with flatness. I have indeed rerun the code that produced the figure, forcing $\Delta_{-} = Sh_a - Sh_b = 0$ which fulfills H_0, and the histogram is utterly flat.

Professor Wasserman might *also* have gone on to relate the fact that if H_1

is true, even if $|\Delta_-|$ is quite substantially greater than zero, then the p-value distribution must also be flat if the power of the test is near α, which is as low as it can get.[19] With $\alpha = 0.05$ and the histogram having 20 bars it's a matter of the area under the last 19 bars, which is the probability of a type II error, needing to sum to $1 - \alpha = 0.95$. Flatness brings that about.

Figure 3 illustrates that tendency. By inspection of Figure 1, the power is decreased as the number of observations is reduced. To produce Figure 3, I lowered the number of observations to just 36, corresponding to just three years of monthly data, to deliberately lower the power. But I have retained the same $|\Delta_-|$ value that is substantially above zero, and the same ρ_{ab} value of 0.85 with is rather high. Now the power has not been decreased all the way to α, which is 0.05; from the legend, it is instead approximately $1 - 0.85$ or $1 - 0.80$, which is 0.15 or 0.20. That's why the histogram is not completely flat. And the discrepancy between the Ledoit and Wolf histogram and that of the asymptotic approximation, such as it is, may have arisen due to the asymptotic formula for the p-value not working as well with only 36 observations.

With 1,000 trials and 20 histogram bars an utterly flat histogram would show 50 p-values in each bar. I reran the code that produced Figure 3, again with 36 observations but with Δ_- reduced to 0.05 and ρ_{ab} set to 0. That still means that H_1 is true. The histogram became truly flat—with the number of p-values in each bar generally varying between 40 and 60 with no discernible downtrend in the direction of the higher p-values. And the power was about 0.055—very close to $\alpha = 0.05$.

If this is not yet clear, if in the circumstances of Figure 2 the Ledoit and Wolf (2008) code was run and a p-value under 0.05 was found, there was little chance that a value higher than 0.05 might instead have been produced because the few bars above 0.05 that are not empty don't contain many of the trials. But in the low-power circumstances of Figure 3 if the Ledoit and Wolf code had been run and some p-value had been found, there was every chance that a very different value might instead have been produced. So, in the circumstances of Figure 3 whatever is produced for the p-value is utterly unreliable.

Ledoit and Wolf (2008, 858) state the following: "We have discussed alternative inference methods which are robust. HAC inference uses kernel estimators to come up with consistent standard errors. The resulting inference works well with large samples but is often liberal for small to moderate sample sizes. In such applications, it is preferable to use a studentized time series bootstrap."[20,21] By

19. Wasserman (2004, 157) does state the following: "Warning! A large p-value is not strong evidence in favor of H_0. A large p-value can occur for two reasons: (i) H_0 is true or (ii) H_0 is false but the test has low power."

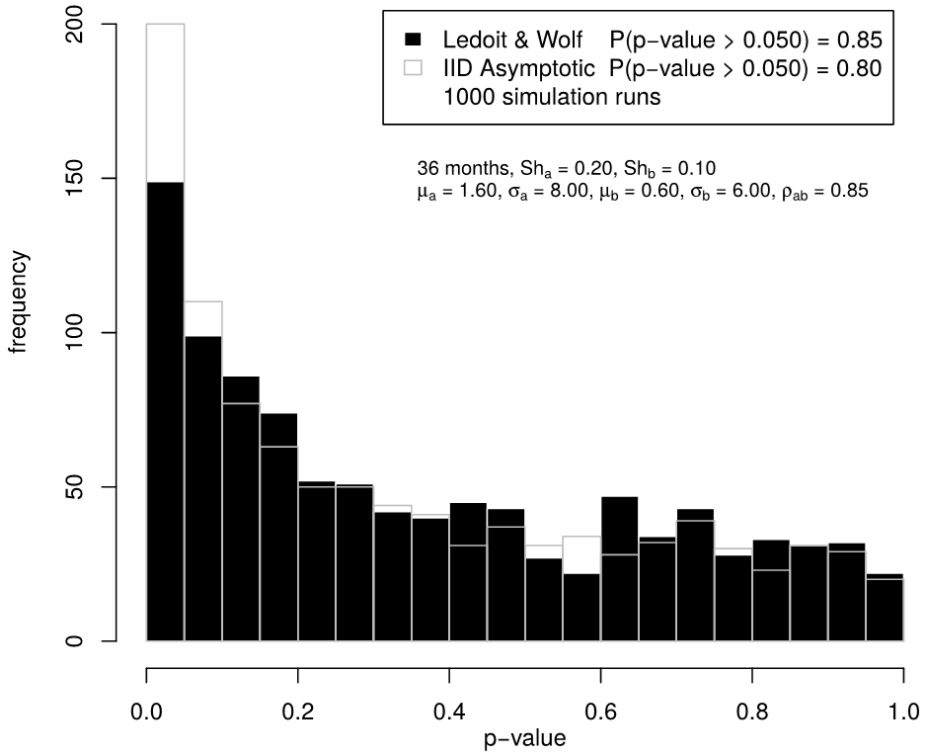
20. "HAC" means heteroskedasticity- and autocorrelation-consistent.

21. An identical statement appears in the subsequent article Ledoit and Wolf 2011. And in that follow-up

"studentized time series bootstrap" the authors are referring to the method of their article, which they do describe as being "quite complex."

But it is clear, when Sharpe ratio differences are the test statistics, that small and moderate sample sizes automatically have low or moderate power—making p-values randomized and hypothesis testing infeasible, as shown by Figure 3. Thus, it's hard to imagine circumstances in which it would be advantageous to use the quite complex Ledoit and Wolf method in lieu of one of the simpler HAC inference schemes.

Figure 3. A low-power example with just 36 observations



Notes: Here $\Delta_{-} = Sh_a - Sh_b$ is the a priori population value of the Sharpe ratio difference. And P(p-value > 0.05) is the probability of a type II error, whose complement is the power of the test.

In the low power limit—due to some combination of low $|\Delta_{-}|$, low ρ_{ab}, or the duration of the backtesting interval not being long enough—the p-value estimator of Ledoit and Wolf and equation (3a) with the standard error derived

article, which is about the variance rather than about the Sharpe ratio, the word *power* does not appear.

from equation (1a) both absurdly become random number generators.

The characteristic form of Figures 2 and 3 is not actually limited to i.i.d. bivariate-normal returns. Whatever the autocorrelations, heteroskedasticity, etc., in the returns of the paired portfolios, in the asymptotic limit the entire histogram is only dependent upon the power of the test and the significance level α.

To derive pdf_{p_2}, the sampling distribution of the two-sided p-value, we can start by presenting it as an alternative means of calculating the power of the test: $\pi_2(\alpha) = \int_0^{\alpha} pdf_{p_2}(p_2')\,dp_2'$. But we don't have to be limited to evaluating the power function π_2 at just α. Since α can be given any value, give it the value p_2. Then $\frac{d\pi_2}{dp_2} = pdf_{p_2}$. Writing $Z_{p_2} \equiv \Phi^{-1}(1 - p_2/2)$ and $Z_{true} \equiv \Delta_{-}/\hat{SE}$, then

$$\pi_2(p_2) = 1 - \Phi\left(Z_{p_2} - Z_{true}\right) + \Phi\left(-Z_{p_2} - Z_{true}\right).$$

And $\frac{dZ_{p_2}}{dp_2} = -\frac{1}{2}\,\Phi'\left(Z_{p_2}\right)^{-1}$.

Putting it all together,

$$pdf_{p_2}(p_2) = \tfrac{1}{2}\,\left[\Phi'\left(Z_{p_2} - Z_{true}\right) + \Phi'\left(-Z_{p_2} - Z_{true}\right)\right] / \Phi'\left(Z_{p_2}\right)$$

from the chain rule of differential calculus. For every value of the power π_2 there is a unique value of $|Z_{true}|$. If the $Z_{true} > 0$ choice is made, as in Figures 2 and 3, then the equation for the two-sided power $\pi_2(\alpha)$ can be inverted to find Z_{true}, and that value can be substituted for Z_{true} in pdf_{p_2}. Thus pdf_{p_2} is only dependent upon the power π_2 and the significance level α. Upon coding the expression for pdf_{p_2} I find that, with the $\pi_2(\alpha)$ values that can be inferred from Figures 2 and 3, it closely agrees with those histograms.

Conclusion

I have shown that there are formidable limitations on the use of hypothesis testing with the Sharpe ratio difference between a pair of portfolios being the test statistic, that are innate to that statistic. Investors should accordingly be wary of claims by portfolio managers that their Sharpe ratio exceeds the ratios of other managers. It is advisable to avoid hypothesis testing when there is good reason to believe that the power of the test is too low—such as the sample covering a small interval of time, the correlation coefficient between the returns of the two portfolios not being high, or the reasonably expectable size of the true Sharpe ratio

difference being small. If the power of the test is too low there is really no good fix for that, unless somehow pertinent additional data become available that span additional months. But confidence intervals do in any case provide a reasonable recourse. A secondary hypothesis test can be of some help in sorting out type II errors, but it's no panacea.

Code

The source code and data that were used to produce the figures is available from the journal website (**link**).

References

Amrhein, Valentin, Sander Greenland, Blake McShane, et al. 2019. Retire Statistical Significance. *Nature* 567: 305–307.

Auer, Benjamin R., and Frank Schuhmacher. 2013. Performance Hypothesis Testing with the Sharpe Ratio: The Case of Hedge Funds. *Finance Research Letters* 10(4): 196–208. **Link**

Benjamini, Yoav, Yotam Hechtlinger, and Philip B. Stark. 2019. Confidence Intervals for Selected Parameters. Working paper, June 2. **Link**

Benjamini, Yoav, and Daniel Yekutieli. 2001. The Control of the False Discovery Rate in Multiple Testing Under Dependency. *Annals of Statistics* 29(4): 1165–1188.

Benjamini, Yoav, and Daniel Yekutieli. 2005. False Discovery Rate Adjusted Multiple Confidence Intervals for Selected Parameters. *Journal of the American Statistical Association* 100(469): 71–81. **Link**

Button, Katherine S., John P. A. Ioannidis, Claire Mokrysz, Brian A. Nosek, Jonathan Flint, Emma S. J. Robinson, and Marcus R. Munafò. 2013. Power Failure: Why Small Sample Size Undermines the Reliability of Neuroscience. *Nature Reviews Neuroscience* 14: 365–376. **Link**

Fisher, Ronald Aylmer. 1971 [1935]. *The Design of Experiments*, 8th ed. New York: Hafner.

Getmansky, Mila, Andrew W. Lo, and Igor Makarov. 2004. An Econometric Model of Serial Correlation and Illiquidity In Hedge Fund Returns. *Journal of Financial Economics* 74(3): 529–609.

Hansen, Peter Reinhard. 2005. A Test for Superior Predictive Ability. *Journal of Business & Economic Statistics* 23(4): 365–380. **Link**

Hesterberg, Tim C. 2015. What Teachers Should Know About the Bootstrap: Resampling in the Undergraduate Curriculum. *American Statistician* 69(4): 371–386.

Hoenig, John M., and Dennis M. Heisey. 2001. The Abuse of Power: The Pervasive Fallacy of Power Calculations for Data Analysis. *American Statistician* 55(1): 19–24. **Link**

Ioannidis, John P. A. 2005. Why Most Published Research Findings Are False. *PLoS*

Medicine 2(8): 696–701. **Link**

Ioannidis, John P. A. 2019. What Have We (Not) Learnt from Millions of Scientific Papers with *P* Values? *American Statistician* 73(51): 20–25. **Link**

Jobson, J. D., and Bob M. Korkie. 1981. Performance Hypothesis Testing with the Sharpe and Treynor Measures. *Journal of Finance* 36(4): 889–908. **Link**

Ledoit, Olivier, and Michael Wolf. 2008. Robust Performance Hypothesis Testing with the Sharpe Ratio. *Journal of Empirical Finance* 15(5): 850–859. **Link**

Ledoit, Olivier, and Michael Wolf. 2011. Robust Performance Hypothesis Testing With the Variance. *Wilmott Magazine* 2011(55): 86–89. **Link**

Lehman, Eric Leo, and George Casella. 1998. *Theory of Point Estimation*, 2nd ed. New York: Springer.

Lo, Andrew W. 2002. The Statistics of Sharpe Ratios. *Financial Analysts Journal* 58(4): 36–52. **Link**

Mair, Magdalena M., Mira Kattwinkel, Oliver Jakoby, and Florian Hartig. 2020. The Minimum Detectable Difference (MDD) Concept for Establishing Trust in Nonsignificant Results: A Critical Review. *Environmental Toxicology and Chemistry* 38(11): 2109–2123.

Memmel, Christoph. 2003. Performance Hypothesis Testing with the Sharpe Ratio. *Finance Letters* (Global EcoFinance, Edinburgh) 1: 21–23.

Opdyke, John D. 2007. Comparing Sharpe Ratios: So Where Are the P-Values? *Journal of Asset Management* 8(5): 308–336.

Pav, Steven E. 2021. Notes on the Sharpe Ratio. Working paper, August 17. **Link**

Pearson, Karl, and Lewis Napoleon George Filon. 1898. On the Probable Errors of Frequency Constants and On the Influence of Random Selection on Variation and Correlation. *Philosophical Transactions of the Royal Society A* 191: 229–311. **Link**

Tavares, Ricardo de Souza, and Joao Frois Caldeira. 2023. Is Replacing Standard Investments with ESG Substitutes a Good Choice? *Brazilian Review of Finance* (Brazilian Society of Finance) 21(1): 49–75. **Link**

Thomas, Len. 1997. Retrospective Power Analysis. *Conservation Biology* 11(1): 276–280. **Link**

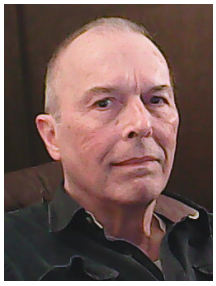
Torricelli, Costanza, and Beatrice Bertelli. 2022. The Trade-Off Between ESG Screening and Portfolio Diversification in the Short and in the Long Run. Working paper. **Link**

Wacholder, Sholom, Stephen Chanock, Montserrat Garcia-Closas, Laure El ghormli, and Nathaniel Rothman. 2004. Assessing the Probability That a Positive Report Is False: An Approach for Molecular Epidemiology Studies. *Journal of the National Cancer Institute* 96(6): 434–442. **Link**

Wasserman, Larry. 2004. *All of Statistics: A Concise Course in Statistical Inference.* New York Springer.

Ziliak, Stephen T., and Deirdre N. McCloskey. 2008. *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives.* Ann Arbor, Mich.: University of Michigan Press.

About the Author



With a B.S. in physics and mathematics from Tulane University, **Michael O'Connor** worked for the United States Navy as a physicist and then went on to graduate studies. His Ph.D. is in physics from Stanford University. For 25 years he owned and operated a Silicon Valley business that conducted noise and air quality studies, often of major transporation projects. Today, O'Connor runs a startup consultancy, MO'C Portfolio Analytics in Washington State. His email address is mike@mocpa.com.

Go to archive of Economics in Practice section
Go to March 2024 issue