

A Martingale Test for 'Alpha'

Dean P. Foster*, Robert Stine*, H. Peyton Young**

This version: February 15, 2010

*Department of Statistics, Wharton School, University of Pennsylvania

** Department of Economics, University of Oxford

Abstract

This paper describes a new way of testing whether the returns from a financial asset are systematically higher than a perfectly efficient market would allow. Such an asset is said to have *positive alpha*. The standard way of estimating alpha is to regress the asset's returns against the market returns over an extended period of time and to apply the *t*-test. The difficulty is that the residuals often fail to satisfy independence and normality. In fact, portfolio managers may have an incentive to employ strategies whose residuals depart *by design* from independence and normality. To address these problems we propose a robust test for alpha based on the martingale maximal inequality. Unlike the *t*-test, our test places no restrictions on the distribution of returns while retaining substantial statistical power. The method is illustrated for four assets: a stock, a mutual fund, a hedge fund, and a fabricated fund that is deliberately designed to fool standard tests of significance.

1. Estimating alpha

An asset that consistently delivers higher returns than a broad-based market portfolio is said to have *positive alpha*. This is the excess return that results from the asset manager's superior skill in exploiting arbitrage opportunities and judging the risks and rewards associated with various investments. But how can investors (and statisticians) tell from historical data whether a given portfolio actually is generating positive alpha relative to the market? To answer this question one must address four issues: i) multiplicity; ii) trends; iii) cross-sectional correlation; iv) robustness. We begin by reviewing standard adjustments for the first three; this will set the stage for our approach to the robustness issue, which involves a novel application of the martingale maximal inequality (Doob, 1953).

The first step in evaluating the historical performance of a financial asset requires adjusting for multiplicity. Assets are seldom considered in isolation: investors can choose among hundreds or thousands of stocks, bonds, mutual funds, hedge funds, and other financial products. Without adjusting for multiplicity, statistical tests of significance can be seriously misleading. To take a trivial example, if we were to test individually whether each of 100 mutual funds "beats the market" at level $\alpha = 0.05$, we would expect to find five statistically significant p -values when in fact *none* of them beats the market.

Statisticians are well-versed in the dangers of searching for the most statistically significant hypotheses in data and have developed a wide variety of procedures

to correct for multiple comparisons. The simplest and most easily used of these is the Bonferroni rule. When testing m hypotheses simultaneously, one compares the observed p -values to an appropriately reduced threshold. For example, instead of comparing each p -value p_i to a threshold such as $\alpha = 0.05$, one would compare them to the reduced threshold α/m .

Modern alternatives to Bonferroni have extended it in two directions. The first, called alpha spending, allows the splitting of the alpha into uneven pieces; see for example Pocock (1977) and O'Brien and Fleming (1979). The second group of extensions provides more power when several different hypotheses are being tested. These can be motivated from many perspectives: false discovery rate (Benjamini and Hochberg 1995), Bayesian (George and Foster, 2000), information theory (Stine, 2004) and frequentist risk (Abramovich, Benjamini, Donoho, and Johnstone, 2006). One can even apply several of these approaches simultaneously (Foster and Stine 2007). In the case of financial markets, however, the natural null hypothesis is that *no* asset can beat the market for an extended period of time because this would create exploitable arbitrage opportunities. Thus, in this setting, the key issue is whether *anything* beats the market, let alone whether multiple assets can beat the market.

A second key issue in evaluating the historical performance of different assets is the need to de-trend the data. This is particularly important for financial assets, which generally exhibit a strong upward trend due to compounding. Consider, for example, the price series shown in Figure 1 for four quite different types of assets. The top two panels represent the value over time of a long-running

general mutual fund (Fidelity Puritan), and the stock of a diversified company (Berkshire Hathaway) that is famous for its superior performance over a long period of time. The two assets in the lower panels represent managed funds that follow dynamic investment strategies that we shall describe in more detail later on. The Team Fund is based on a dynamic rebalancing algorithm that seeks to reduce volatility while maintaining high returns (Gerth, 1999; see also Agnew, 2002). The Piggyback Fund is based on an options trading strategy that is designed to fool investors into believing that the fund is generating excess returns when in expectation it is not (Foster and Young, 2009; for a related construction see Lo, 2001).

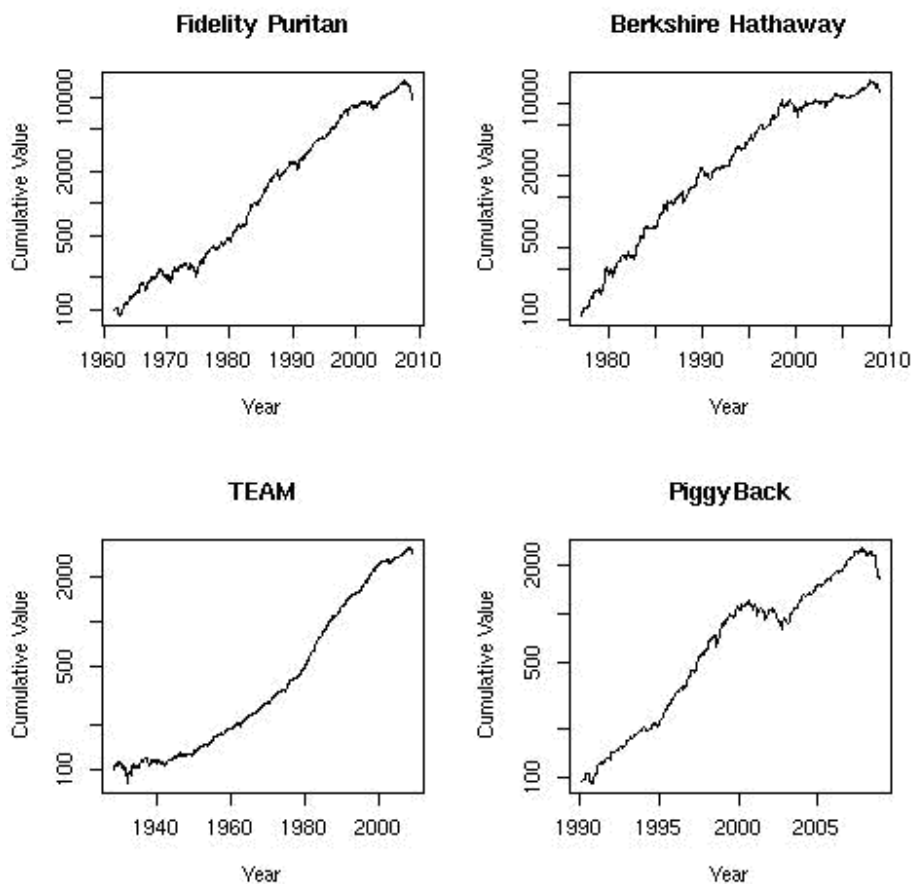


Figure 1. Value of four assets over time

The simplest way to de-trend such data is to study the period-by-period returns rather than the value of the asset itself. That is, instead of focusing on the value, V_t , one studies the sequence of returns $(V_t - V_{t-1})/V_{t-1}$ over successive time periods t . The hope is that the returns are being generated by a process that is sufficiently stationary for standard statistical tests to be applied. As we shall see later on, this hope may not be well-founded given that financial portfolios are often managed in a way that produces highly nonstationary behavior by design.

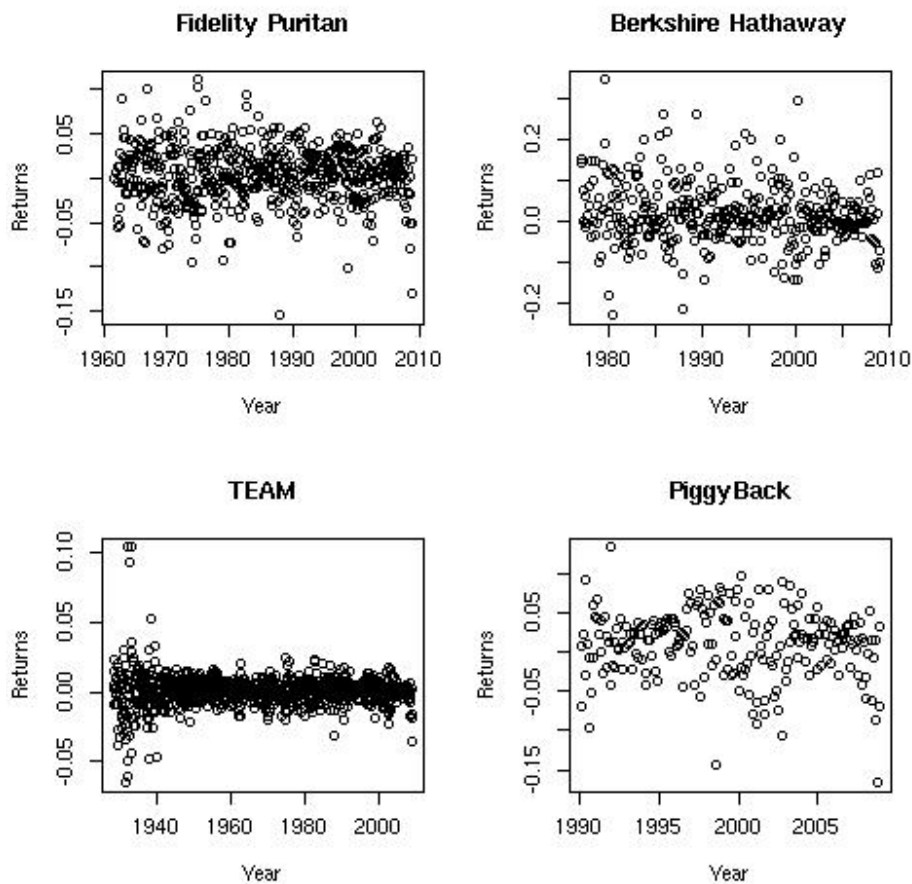


Figure 2. Monthly returns series for the four assets

The third issue, cross-sectional correlation, arises because returns on financial assets often exhibit a high degree of positive correlation. The standard way to deal with this problem is the Capital Asset Pricing Model (CAPM), which partitions the variation in asset returns into two orthogonal components: market risk, which is non-diversifiable and hence unavoidable, and idiosyncratic risk. By construction, idiosyncratic risk is orthogonal to market risk and measures the rewards and risks associated with a specific asset. It is the mean return on this idiosyncratic risk, known as *alpha*, that draws investors to specific stocks, mutual funds, and alternative investment vehicles such as hedge funds.

The standard way to estimate the alpha of a particular asset or portfolio of assets is to regress its returns against the returns from a broad-based market index such as the *S&P 500* after subtracting out the risk-free return. Specifically, let m_t be the return generated by the market portfolio in period t , and let r_t be the risk-free rate of return during the period, that is, the return available on a safe asset such as US Treasury bills. The *excess return* of the market during the t^{th} period is $m_t - r_t$. Let y_t be the return in period t from a particular asset (or portfolio of assets) that we wish to compare with the market. The portfolio's *excess return* in period t is defined as $y_t - r_t$. CAPM isolates the idiosyncratic return of the portfolio from the overall market return by estimating the regression equation

$$y_t - r_t = \alpha + \beta(m_t - r_t) + \varepsilon_t. \quad (1)$$

The *market-adjusted return* is defined to be the intercept plus the residual, that is, $\alpha + \varepsilon_t$.

The crucial question from the investor's standpoint is whether the intercept is positive ($\alpha > 0$). If it is, then the investor could increase his overall return by investing a portion of his wealth in this portfolio instead of in the market. Moreover, this increased return could be achieved without exposing himself to much additional volatility, provided he puts a sufficiently small *proportion* of his wealth in the portfolio. *Thus the relevant question is whether there exists any asset, or portfolio of assets, that systematically beats the market in the sense that it exhibits positive alpha at a high level of significance.*

As we have already noted, the standard way to answer this question is to apply a *t*-test to the estimate of the intercept. But this is only justified if the residuals satisfy quite restrictive assumptions, including independence and normality. When we graph the residuals from our four candidate assets over time, however, it becomes apparent that two of them (Team and Piggyback) exhibit highly erratic, nonstationary behavior (see Figures 3-4).

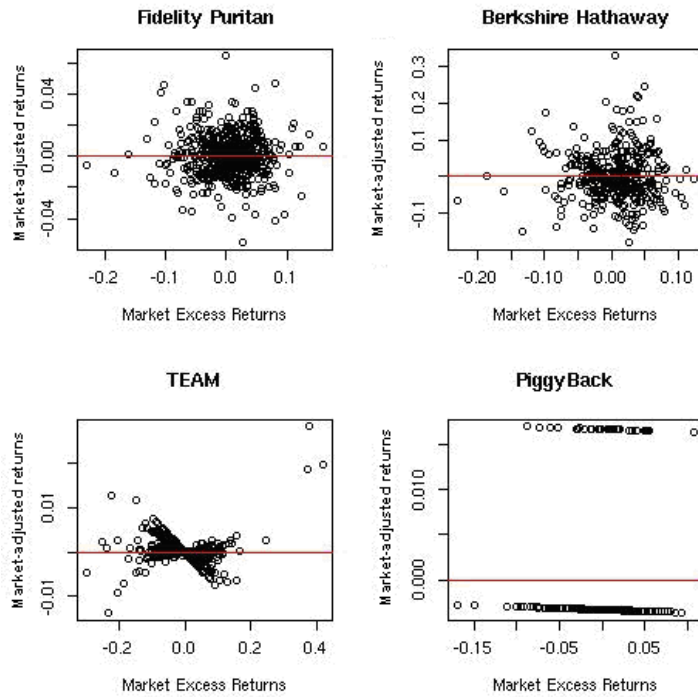


Figure 3. Market-adjusted returns versus market excess returns.

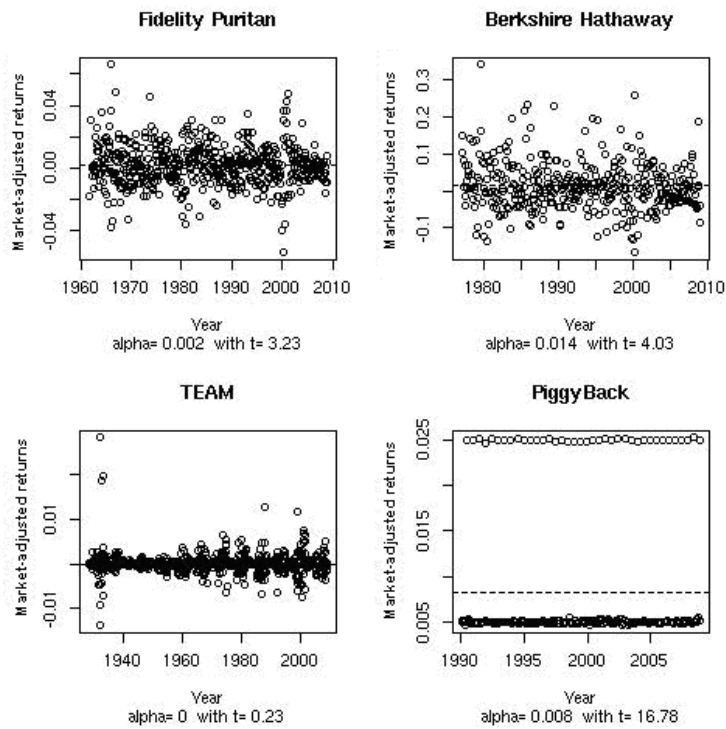


Figure 4. Time series of market-adjusted returns for the four assets.

In the case of Team, the market-adjusted returns are negatively correlated with the market returns (see Figure 3), and they exhibit a cyclic pattern with bursts of high volatility followed by periods of low volatility immediately thereafter (see Figure 4). This is a direct consequence of Team's dynamic rebalancing strategy. At any given time the fund is invested in a mixture of cash (earning the risk-free rate of return) and the *S&P 500*. At the end of each year funds are moved from cash to stock if the risk-free rate in that period exceeded the market rate of return, and from stock to cash if the reverse was true. The target proportions to maintain between stock and cash are adjusted at the end of each five-year interval i . (The five-year intervals and target proportions -- 70% stock, 30% cash - are chosen purely for purposes of illustration.)

Gerth (1999) proves that if stocks have returns that are independent and identically distributed, and if the risk-free rate of return is constant, then there exists a choice of targets such that this strategy yields higher returns than the market without increasing the level of volatility. Our purpose is not to critique this approach, but to take it as a concrete example showing why a natural portfolio management strategy could easily produce returns that *by design* differ substantially from the assumptions needed to apply the t -test.

The returns from the Piggyback Fund exhibit even more erratic behavior. The reasons for this will be discussed in section 3. Suffice it to say here that these returns are produced by a strategy that is designed to earn the portfolio manager a lot of money rather than to deliver superior returns to the investors. However, since the natural aim of portfolio managers *is* to earn large amounts of money, statistical tests of significance must accommodate this sort of behavior.

The purpose of this paper is to introduce a robust test for alpha that is immune to this type of manipulation. Our test assumes nothing about the distribution of market-adjusted returns except that they form a martingale difference: neither independence, constant variance, nor normality are required. The test is simple to compute and does not depend on the frequency of observations. A particularly important feature of the test is that it corrects for the possibility that the manager's strategy conceals a small risk of a large loss in the tail. This is not a hypothetical problem: empirical studies using multifactor risk analysis have shown that many hedge funds have negatively skewed returns and that as a result the t -test significantly underestimates the left-tail risk (Agarwal and Naik, 2004). Moreover, negatively skewed returns are to be *expected*, because standard compensation arrangements give managers an incentive to follow just such strategies (Foster and Young, 2009).

The plan of the paper is as follows. In section 2 we derive our test in its most basic form. The basic idea is first to correct for market correlation by computing the market-adjusted returns as in (1), then compound these returns and test whether they form a martingale. Section 3 shows why it is vital to have a test that, unlike the t -test, is robust to distributional assumptions. In particular, we show that the t -test leads to a false degree of confidence in the Piggyback Fund, which is constructed so that it looks like it produces positive alpha when this is not actually the case. Section 4 extends the approach to test whether one or more assets out of a given population of assets is producing positive alpha at a high level of significance.

In section 5 we show how to ramp up the statistical power of this approach by leveraging the asset. We show, in particular, that if the returns of the asset are lognormally distributed with known variance, then we can choose a level of leverage such that the loss in power is quite modest relative to the optimal test (which in this case is the t-test). In fact, for a p -value of .01 the loss in power is less than 30%, and for a p -value of .001 the loss in power is only about 20%.

In section 6 we consider the more general situation in which we know nothing *ex ante* about the asset's true distribution. In spite of this we can use leverage to advantage by applying different levels of leverage to the asset in question. This results in a population of leveraged assets to which we apply *PERT*. The essential point is that the best level of leverage (which we do not know in advance) results in exponentially higher rates of growth than other levels. Hence the compound excess return test applied to the population (*PERT*) can come quite close to the compound excess return test applied to the best level of leverage. (This result is closely related to Cover's work on universal portfolios (Cover, 1991.) We call this the *exponential population excess returns test (EXPERT)*.

In the final section we apply this test to our four candidate assets. It turns out that, after correcting for multiplicity, market correlation, and unrealized downside risk, the only asset that plausibly delivers positive alpha is Berkshire Hathaway, even though two of the others (Fidelity, and Piggyback) look quite impressive at first sight.

2. The compound excess returns test (CERT)

Consider a financial asset, such as a stock, a mutual fund, or a hedge fund whose performance we wish to compare with that of the market. The data consist of returns generated by the asset over a series of reporting periods $t=1,2,\dots,T$. Denote the market return in period t by the random variable M_t and the asset's return by the random variable Y_t , where by definition $M_t, Y_t \geq 0$. In applications, M_t would be the return on a broad-based portfolio of stocks such as the *S&P 500*.

The first step in our analysis is to subtract off the risk-free rate of return in each period, that is, the rate available on a safe asset such as Treasury bills. Letting r_t denote the risk-free rate in period t , define the random variables

$$\tilde{M}_t = M_t - r_t, \quad \tilde{Y}_t = Y_t - r_t. \quad (2)$$

The second step is to correct for correlation with the market. Let

$$\beta = \frac{\text{Cov}(\tilde{Y}, \tilde{M})}{\text{Var}(\tilde{M})}. \quad (3)$$

In practice, β can either be estimated directly from historical data or by analyzing the composition of the asset in question. An important point to notice is that β can be estimated accurately using short-term data (e.g., daily returns) because it measures the extent to which the returns are correlated with the market, not their magnitude. (Thus with sufficiently high frequency data one

could estimate the correlation coefficient period by period, i.e., monthly or quarterly.) The market-adjusted return or *alpha* in period t is

$$A_t = \tilde{Y}_t - \beta \tilde{M}_t. \quad (4)$$

The *null hypothesis* is that the conditional expectation of A_t is zero in every period t , that is,

$$\forall t \leq T, \quad E[A_t | A_1, \dots, A_{t-1}] = 0. \quad (5)$$

This is equivalent to saying that the compound growth of the market-adjusted returns forms a nonnegative martingale

$$\text{Null Hypothesis: } C_t = \prod_{1 \leq s \leq t} (1 + A_s) \text{ is a nonnegative martingale.} \quad (6)$$

Compound Excess Return Test (CERT). For each t , $1 \leq t \leq T$, let $C_t \geq 0$ be the compound market-adjusted return of a candidate asset through period t . The CERT p -value for testing the null hypothesis is

$$p = \min_{1 \leq t \leq T} (1/C_t). \quad (7)$$

The proof is a straightforward application of the martingale maximal inequality, which for convenience we shall derive here (Doob, 1953). Under our assumptions, C_t is a nonnegative martingale with conditional expectation 1 in every period. Given a real number $\gamma > 0$, define the random time $T(\gamma)$ to be the

first time $t \leq T$ such that $C_t \geq \gamma$ if such a time exists, otherwise let $T(\gamma) = T$. By the optional stopping theorem, $E[C_{T(\gamma)}] = 1$. (See Doob, 1953, Theorem 2.1.) Clearly, if $\max_{1 \leq s \leq t} C_s \geq \gamma$ then $C_{T(\gamma)} \geq \gamma$, hence $P(\max_{1 \leq s \leq t} C_s \geq \gamma) \leq P(C_{T(\gamma)} \geq \gamma)$. Since $C_{T(\gamma)}$ is nonnegative, $P(C_{T(\gamma)} \geq \gamma) \leq E[C_{T(\gamma)}] / \gamma = 1 / \gamma$. It follows that

$$P(\max_{1 \leq s \leq t} C_s \geq \gamma) \leq 1 / \gamma. \quad (8)$$

It follows that the sequence C_1, C_2, \dots, C_T of compound excess returns was generated with probability at most $\min_{1 \leq t \leq T} (1 / C_t)$. \square

Notice that the *length* of the series is immaterial: what matters is the *maximum compound return* that was achieved at some point during the series. Moreover, the compound returns must be very large for the null hypothesis to be rejected. For example, to reject the null at the 5% level of significance requires that the asset grow *twenty-fold* after subtracting off the risk-free rate and correcting for correlation with the market. Two of the four funds meet this test (Fidelity and Berkshire) as may be seen from Figure 5, which shows the compound value of each asset's market-adjusted returns.

Table 1 compares the p -values from *CERT* with those from the t -test. According to the latter, Berkshire, Fidelity, and Piggyback have positive alpha at a high level of significance and Piggyback is particularly impressive. By contrast, *CERT* says that we should be confident in Berkshire, which has a *CERT* p -value equal to .011, but not in any of the others.

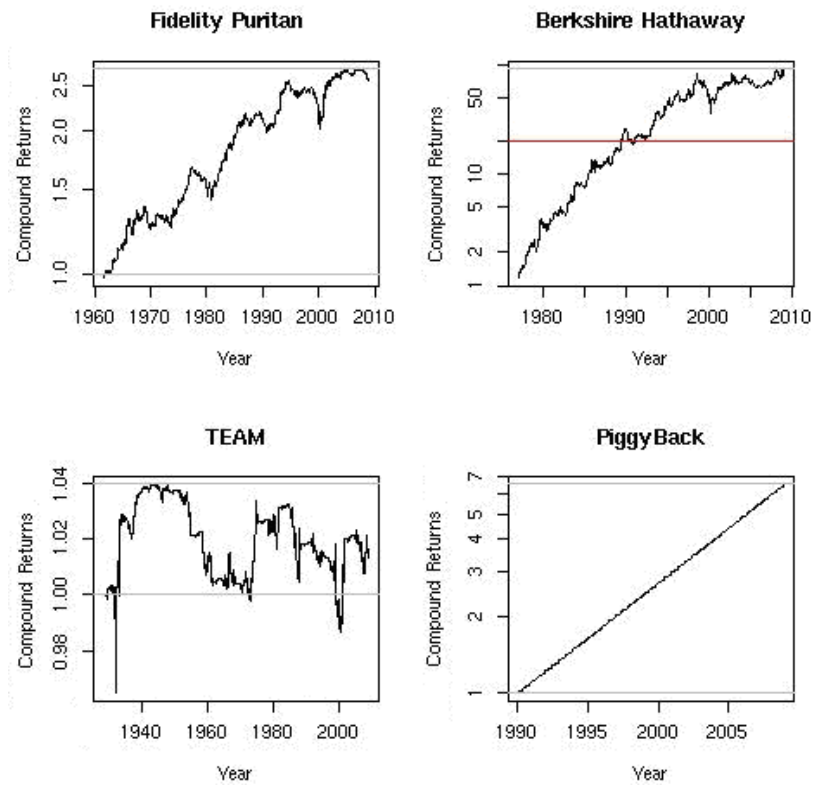


Figure 5. Compound value of market-adjusted returns for the four assets

<i>Asset</i>	<i>Regression t</i>	<i>Regression p-value</i>	<i>CERT p-value</i>	<i>CERT "t"</i>
Fidelity	3.23	0.0013	0.37	0.33
Berkshire	4.03	6.7×10^{-5}	0.011	2.30
Team	0.23	0.82	0.96	-1.77
Piggyback	16.79	1.3×10^{-41}	0.15	1.03

Table 1. Regression p-values versus CERT p-values for the four assets.

[CERT "t" is the value of the *t*-statistic that corresponds to the given *p*-value.]

3. Gaming the t-test.

At first it might seem counterintuitive that one would need to see a fund outperform the market by at least *twenty-fold* in order to be reasonably confident (at the 5% level of significance) that the outperformance is “for real.” The reason is that an apparently stellar performance can be driven by strategies that lead to a total loss with positive probability but a very long time can elapse before the loss materializes. Indeed, it is easy to construct strategies of this nature for which the *CERT* bound is tight. Choose a number $\gamma > 1$, and consider the following nonnegative martingale with conditional expectation 1

$$C_0 = 1, \quad P(C_t = \gamma^{1/T} C_{t-1}) = \gamma^{-1/T}, \quad P(C_t = 0) = 1 - \gamma^{-1/T} \text{ for } 1 \leq t \leq T. \quad (9)$$

In each period the fund compounds by the factor $\gamma^{1/T}$ with probability $\gamma^{-1/T}$ and crashes with probability $1 - \gamma^{-1/T}$. Thus the probability that the fund’s compound excess return C_t exceeds γ is precisely $1/\gamma$ over any number of periods T .¹

The Piggyback Fund is constructed along just these lines. Namely, the fund is invested in the *S&P 500*, and the returns are reported every month. However, once every six months the returns are artificially boosted by the factor 1.02. This can be done by taking an options position in the *S&P 500* that bankrupts the fund if the options are exercised. The strike price is chosen so that the probability of this event is $1/1.02 = .9804$, so the fair value of the option is zero.² This explains

¹ Returns series with this property can be constructed using standard options contracts (Foster and Young, 2009).

² With probability $1/1.02$ the fund grows by the factor 1.02 and with probability $.02/1.02$ it loses everything, which is a lottery with expectation zero.

the bizarre pattern of the market-adjusted residuals in Figure 3: one-sixth of the time they are + 2%, and five-sixths of the time they are - 2% . With less than 25 years of data there is a sizable probability that the downside risk will never be realized, and investors will be lulled into thinking that the fund is generating positive alpha. This is why *CERT* attaches a very modest p -value to the returns generated by the Piggyback Fund.

Of course, even a casual inspection of the residuals in Figure 3 suggests that the t -test should not be used in this case. However this is not the essence of the problem, because it is easy to construct “piggyback strategies” whose returns look i.i.d. normal. Namely, suppose that every month the manager boosts the fund’s returns by the factor 1.0033λ where λ is a lognormally distributed error with mean 1 and small variance. (The purpose of the error is to lend a plausible amount of variability to the realized returns.) As before, the boost comes at the cost of going bankrupt with probability $1/(1.0033\lambda) \approx .9967/\lambda$ each month. Assuming that the variance of λ is small, this scheme will run for about 300 months (25 years) before the fund goes bankrupt, and the residuals will look very convincing. Thus, in this case the t -test would seem to be appropriate, and the estimate of alpha will be about 4% per year at a very high level of significance. This is misleading, however, because in reality the distribution of returns is not approximately normal -- there is a large potential loss hidden in the tail. One of the main virtues of *CERT* is that it corrects for this “hidden volatility”: under *CERT* the p -value of this scheme after 25 years will only be about $p = (1.0033)^{-300} \approx .37$.

4. Multiplicity, Bonferroni, and the portfolio test.

The test described above is for returns generated by a single fund. In practice, investors would like to identify those funds *from a given population* that generate positive alpha with high probability. This requires a more demanding test of significance. To illustrate, suppose that we can observe the returns for each of n funds over the same time frame $t=1,2,\dots,T$. Let C_t^i be the compound excess return for fund i through period t . Suppose that we observe a particular fund, say i^* , whose returns are sufficiently high that we would reject the null at the 5% level using *CERT*. This does not mean we have 95% confidence that fund i^* is generating positive alpha. Suppose, in fact, that *none* of the n funds is able to generate positive alpha and that the excess returns are stochastically independent. It is straightforward to construct n independent nonnegative martingales, each with expectation 1, such that on average there will be $.05n$ funds that exceed the critical threshold $c_{.05}$. For example, out of 1000 funds there will, on average, be 50 funds that pass the single-fund threshold even though none of the funds is actually generating positive alpha.

More generally, suppose that we observe n nonnegative martingales $C_t^i, 1 \leq i \leq n$, over the periods $1 \leq t \leq T$. Let the null hypothesis H_0 be that $E[C_t^i | c_1^i, \dots, c_{t-1}^i] = 1$ for all i and for all t , where the C_t^i are assumed to be independent across i for each t . By the martingale maximal inequality we know that

$$\forall c > 0, \quad P(\max_{1 \leq t \leq T} C_t^i \geq c) \leq c. \quad (10)$$

Hence

$$\forall c > 0, \quad P(\max_{1 \leq i \leq n} \max_{1 \leq t \leq T} C_t^i \geq c) \leq n/c. \quad (11)$$

It follows that, to reject H_0 at level $p > 0$, there must exist one or more funds i such that

$$\max_{1 \leq t \leq T} C_t^i > n/p. \quad (12)$$

We call this test *Bonferroni-CERT*. (The idea is quite general and applies to any situation in which multiple hypothesis tests are being conducted; see Miller, 1981.) In the present case the test suggests that we should not be confident that *any* of the four assets exhibits positive alpha. The reason is that the only one that passes muster at an individual level is Berkshire, which we cherry-picked from the *S&P 500*. To be confident that the best of 500 stocks generated positive alpha at the 5% level of significance, would require that the stock's market-adjusted returns compound by a factor of at least $500 \times 20 = 10,000$. While this might seem like an impossibly high hurdle, we shall show in the next section that there are more powerful versions of the test that make such values achievable.

Before considering this issue, however, let us observe that there is a simple and powerful alternative to *Bonferroni-CERT* that works when one wants to know whether one or more assets in a given population has positive alpha without identifying *which* particular asset (or assets) that might be. This test, called the *Portfolio Excess Returns Test (PERT)*, relies on the fact that any weighted combination of assets that have zero alpha must also have zero alpha.

Portfolio Excess Returns Test (PERT). Consider a family of n funds, where each fund i generates a series of compound excess returns $C_1^i, C_2^i, \dots, C_T^i$ over T periods. Create a portfolio consisting of equal amounts invested initially in each of the funds. The portfolio's compound return series is given by $\bar{C}_t = (1/n) \sum_{1 \leq i \leq n} C_t^i$, $1 \leq t \leq T$. The null hypothesis that none of the funds has positive alpha has p -value $\min_{1 \leq t \leq T} (1/\bar{C}_t)$.

The proof is more or less immediate. Each of the n funds generates a nonnegative martingale of excess returns C_t^i that may or may not be independent across funds. The investor's portfolio is the nonnegative martingale $\bar{C}_t = (1/n) \sum_{1 \leq i \leq n} C_t^i$. The assumption that none of the funds exhibits positive alpha implies that the conditional expectation of \bar{C}_t equals 1 in every period. The conclusion follows at once from the martingale maximal inequality.

Note that this test is at least as powerful as the Bonferroni test: the latter rejects only if $C_t^i > n/p$ for some i , but in this event $\bar{C}_t > 1/p$, which implies that the portfolio test rejects also.

5. Power and leverage

While *CERT* and *PERT* are robust, they are also very conservative. In the next two sections we shall show that much more powerful variants of these tests can be devised by leveraging the asset under scrutiny. Let us begin by considering the special case in which the returns are known to be i.i.d. lognormal, in which case the optimal test of significance is the t -test applied to the logged returns. We shall show that by leveraging the asset at an appropriate level we obtain an

exponentiated form of *CERT* (*EXCERT*) that involves only a modest loss of power compared to the optimal test.

To be concrete, consider a manager whose fund is generating compound returns $C_t \geq 0$ relative to the risk-free rate and suppose for simplicity that there is no correlation with the market ($\beta = 0$). We shall assume that C_t is lognormally distributed:

$$\log C_t \sim N((\mu - \sigma^2 / 2)t, \sigma^2 t).^3 \quad (13)$$

When the asset is leveraged by the factor $\lambda > 0$, the compound returns at time t , $C_t(\lambda)$, are described by the process

$$\log C_t(\lambda) \sim N((\lambda\mu - \lambda^2\sigma^2 / 2)t, \lambda^2\sigma^2 t).^4 \quad (14)$$

Suppose that σ^2 is known and μ is not. The *null hypothesis* is that $\mu = 0$ and the *alternative hypothesis* is that $\mu > 0$. Choose a p-value $p > 0$ and a time t at which a test of significance is to be conducted. *CERT* rejects the null at level p if and only if

$$\log C_t(\lambda) > \log(1/p). \quad (15)$$

³ This is consistent with the traditional representation of asset returns as a geometric Brownian motion in continuous time $dC_t = \mu C_t dt + \sigma C_t dW_t$ (Berndt, 1996; Campbell, Lo, and MacKinlay, 1997).

⁴ To leverage an asset by the factor λ one borrows $\lambda - 1$ dollars at the risk-free rate and invests λ dollars in the asset. If $\lambda < 1$ this means that $1 - \lambda$ is invested in the risk-free asset and the remainder in the risky asset. Notice that to keep a constant level of leverage one will typically need to rebalance the absolute amounts invested in each asset over time. In continuous time this yields the process $dC_t = \lambda\mu C_t dt + \lambda\sigma C_t dW_t$.

Under the null hypothesis,

$$Z_t = \frac{\log C_t(\lambda) + (\lambda^2 \sigma^2 / 2)t}{\lambda \sigma \sqrt{t}} \text{ is } N(0,1). \quad (16)$$

Hence *CERT* rejects the null if and only if

$$Z_t > \frac{\log(1/p) + (\lambda^2 \sigma^2 / 2)t}{\lambda \sigma \sqrt{t}}. \quad (17)$$

To maximize the power of the test we choose the leverage so that the probability of rejection is maximized. This occurs when the right-hand side of (17) is minimized, that is, when

$$\lambda^* = \frac{\sqrt{2 \log(1/p)}}{\sigma \sqrt{t}}. \quad (18)$$

Definition. *EXCERT* (*exponential CERT*) is *CERT* applied to the asset leveraged by the amount λ^* .

Notice that λ^* depends on the variance of the process, the time at which the test is conducted, and the level of significance p . However, the corresponding z -value depends only on p , that is, *EXCERT* rejects if and only if

$$Z_t > \sqrt{2 \log(1/p)} \equiv c_p, \quad (19)$$

that is, c_p is the *critical value* for *EXCERT* at significance level p . Let us compare this with the critical value of the *t*-test, which rejects at level p if null at level p and only if

$$Z_t > \Phi^{-1}(1-p) \equiv z_p. \quad (20)$$

The *power loss* at significance level p , $L(p)$, is the maximum probability that for some combination of μ, σ, t , the *t*-test rejects the null at level p when *EXCERT* accepts.

Proposition 1. $L(p) < 2\Phi(.5(c_p - z_p)) - 1$ and $\lim_{p \rightarrow 0^+} L(p) = 0$.

The proof is given in the Appendix. The first expression in the proposition can be used to numerically compute an upper bound on $L(p)$; the results are illustrated in Figure 6. The loss in power is around 20-25% for p in the range 10^{-3} to 10^{-5} , which is the relevant range when we test for the best of n assets and n is on the order of several hundred or several thousand.

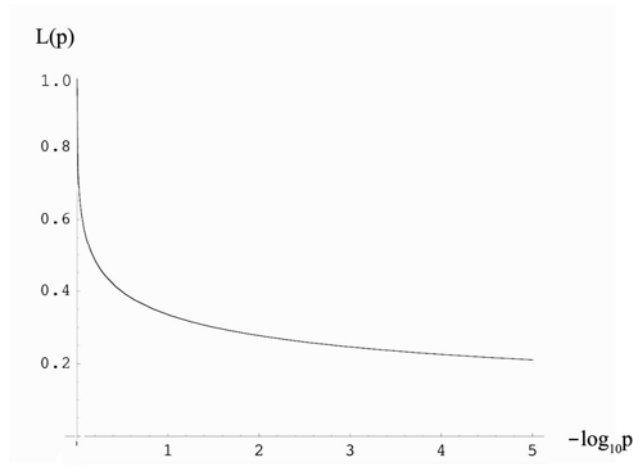


Figure 6. Power loss function $L(p)$ for *EXCERT* compared to the *t*-test.

6. Leveraging CERT when the variance is unknown

The situation considered in the preceding section is quite special in that the distribution of returns was assumed to be i.i.d. lognormal with known variance. Here we introduce an alternative version of the test that is more robust and is asymptotically just as powerful. To apply this test, all we need to know is the maximum leverage that can be applied to the asset under consideration without causing negative realizations; in other words it suffices to know the maximum downside loss that the asset can suffer in any given period.

Consider an asset whose market-adjusted returns A_t are bounded below by $-\phi > -1$. Then the random variable $M_t^\phi \equiv (1 + A_t / \phi)$ is nonnegative, that is, the maximum permissible level of leverage is $1 / \phi$. The null hypothesis is that $E[A_t] = 0$, which implies that $E[A_t / \phi] = 0$. Indeed the null hypothesis holds for every level of leverage in the range $0 \leq \lambda \leq 1 / \phi$:

Null hypothesis: $\forall \lambda, 0 < \lambda \leq 1 / \phi, C_t^\lambda = \prod_{1 \leq s \leq t} (1 + \lambda A_s)$ is a nonnegative martingale.

Let us now construct a *family* of funds that are based on the given asset $\{A_t\}$, where each fund operates under a different level of leverage $0 \leq \lambda \leq 1 / \phi$. Suppose, for example, that we weight all feasible levels of leverage equally. We then obtain a population of funds whose total value at time is given by

$$C_t = \phi \int_0^{1/\phi} \prod_{1 \leq s \leq t} (1 + \lambda A_s) d\lambda. \quad (21)$$

More generally, consider any density $f(\lambda)$ that is bounded away from zero on an interval $[a,b] \subseteq [0,1/\phi]$ where $\int_a^b f(\lambda)d\lambda = 1$. We can construct a family of funds with compound returns

$$C_t = \int_a^b f(\lambda) \prod_{1 \leq s \leq t} (1 + \lambda A_s) d\lambda. \quad (22)$$

Under the null hypothesis, any such fund has compound returns that form a nonnegative martingale. Hence we can reject the null hypothesis at level p if

$$C_t = \int_a^b f(\lambda) \prod_{1 \leq s \leq t} (1 + \lambda A_s) d\lambda > 1/p. \quad (23)$$

Any test of this form will be called an *exponential population excess returns test* (*EXPERT*). This type of test is very general and assumes nothing about the distribution of returns of the underlying asset $\{A_t\}$ except that they form a martingale difference and are bounded away from -1 .

Given its generality one might expect that the power of such a test is very low. In fact, however, it has the same power asymptotically as does *EXCERT* where the variance is assumed to be known. The only requirement is that the interval $[a,b]$ contains the actual value λ^* that optimizes *EXCERT*. By choosing the interval to be as wide as possible, namely $[0,1/\phi]$, this requirement will automatically be satisfied.

Proposition 2. Consider an asset whose returns are lognormally distributed $\log C_t \sim N((\mu - \sigma^2/2)t, \sigma^2 t)$. Suppose that EXPERT is applied at time t to a population of funds leveraged according to a density that is bounded away from zero on an interval $[a, b] \subseteq [0, 1/\phi]$ that contains the optimal leverage level. Then for all sufficiently large times t the loss in power relative to the t -test is well-approximated by $L(p)$.

This result follows from a more general theorem on “universal portfolios” due to Cover (1991). Let C_t be the size of the EXPERT portfolio at time t . Let C_t^* be the value at t of the portfolio that is leveraged at the optimal level λ^* using EXCER. (This depends, of course, on t , p , and σ). Cover’s theorem compares these values with the value C_t^{**} of a third portfolio in which the asset is leveraged at a level λ^{**} that is chosen ex post to maximize the value of the fund at time t given that the realized values of the returns from the underlying asset up through t are known. Clearly $C_t^{**} \geq C_t^*$ because C_t^{**} is optimized ex post whereas C_t^* is optimized ex ante. Cover’s theorem implies that, for every small $\varepsilon > 0$, there is a time T_ε such that $P(C_t / C_t^{**} \geq 1 - \varepsilon) > 1 - \varepsilon$ for all $t \geq T_\varepsilon$. It follows that $P(C_t / C_t^* \geq 1 - \varepsilon) > 1 - \varepsilon$ for all $t \geq T_\varepsilon$. By construction, EXCER rejects at level p if $C_t^* > 1/p$, whereas EXPERT rejects at level p if $C_t > 1/p$. It follows that EXPERT has approximately the same power loss as does EXCER, namely $L(p)$, at all sufficiently large times t .⁵

⁵ Cover considers portfolios that are convex combinations of a finite set of assets. This condition is satisfied in our set-up because any leveraged portfolio $1 + \lambda A_t$ can be represented as a convex combination of the maximally leveraged portfolio $1 + bA_t$ and the minimally leveraged portfolio $1 + aA_t$. Cover also assumes that the returns from each asset in any given period are nonnegative and bounded above. To meet this condition we can truncate the lognormal distribution of returns from the maximally leveraged fund at a level that is several orders of magnitude smaller than the level p we are testing.

7. Empirical analysis of the four assets

In this concluding section we shall apply the preceding framework to the four assets shown in Figures 1-4. As we have already seen, *CERT* without leveraging and without correcting for multiplicity leads to the p -values shown in Table 1. These suggest that Berkshire has positive alpha when viewed in isolation ($p = 0.011$), but not when culled from five hundred stocks (the *S&P 500*). However, we have additional information about the composition of Berkshire (as well as Fidelity and Team) that allows us to ramp up the leverage. Namely, in each of these cases they were primarily invested in common stocks and cash and (according to their annual reports) did not leverage their holdings to any appreciable extent.⁶ This allows us to estimate the amount of leverage that can be applied without the fund going bankrupt. We can then apply *EXPERT*.

More generally we propose the following four-step procedure for testing whether a given asset has positive alpha:

1. Estimate the correlation coefficient β between the asset's returns Y_t and the market's returns M_t .⁷
2. For each t compute the market-adjusted returns $A_t = (Y_t - r_t) - \beta(M_t - r_t)$.

⁶ Berkshire Hathaway is a diversified holding company that invests mainly in the common and preferred stock of other companies. The Fidelity Puritan Fund invests in a mixture of stocks and bonds, and rebalances the proportions periodically. By construction, Team holds a combination of cash and the S&P 500 at each point in time and is not leveraged.

⁷ Alternatively, one can estimate a rolling correlation coefficient β_t using a trailing subsample of data.

3. Estimate the maximum amount of leverage λ^+ that can be applied to the market-adjusted returns without going bankrupt. For an asset consisting of liquid common stocks, this can be estimated either from the cost of a put option, or from the size of the buffer needed to implement a stop-loss order. Here we shall assume that a stop-loss order on stocks can be executed within 3% of the limit price, so that one could leverage up to about 33 times without going negative. (It is not crucial to estimate the upper bound precisely; what matters is the optimal level of leverage, which will usually be lower than the maximum possible leverage.)

4. Compute the maximum compound value of the leveraged asset for a selection of leverage levels between 0 and λ^+ , and let \bar{C} be the average. The estimated p -value of the asset is $1/\bar{C}$. Although this estimate will depend on the assumed distribution of leverage levels, there will typically be a narrow band of leverage levels that yield vastly higher values of C than do the others. The average \bar{C} will depend largely on these critical levels and not on the particular form of the distribution.

For purposes of illustration we shall evaluate each of the four funds at seven leverage levels: 1/2, 1, 2, 4, 8, 16, 32.⁸ As noted above, the choice of these particular values is not important, what matters is that they cover the range of possible values reasonably well (in this case 0 - 33) and the same distribution of values is applied to all the assets being evaluated. The results for Berkshire, Fidelity, and Team are shown in Table 3. (Note that we cannot apply this method to the Piggyback Fund, because the maximum level of leverage is *not* 33; indeed,

⁸ This approximates the distribution in which log leverage is uniformly distributed.

the fund is already leveraged to the hilt because it is constructed from options that will bankrupt the fund with positive probability.)

Leverage	Max C-value		
	Fidelity	Berkshire	Team
0.5	1.7	12	1.0
1	2.7	92	1.0
2	6.7	2,400	1.1
4	31	110,000	1.2
8	290	230	1.3
16	1,200	3,600	1.5
32	44	200	1.5
Avg \bar{C}	225	16,648	1.22
<i>EXPERT</i> $p = 1/\bar{C}$.0044	.00006	0.82
<i>EXPERT</i> "t"	3.29	4.41	-0.89
<i>Regression t</i>	3.23	4.03	0.23
<i>Regression p-value</i>	.0017	.00007	0.41

Table 3. *EXPERT* applied to three assets at leverage levels (1/2, 1, 2, 4, 8, 16, 32) under the assumption that none of the leveraged assets can go negative at these levels.

Notice that the p -values estimated by our test and by the t -test are in quite close agreement even though our test makes none of the regularity assumptions required for the t -test.

While these p -values would be highly significant for a fund that is viewed in isolation, however, we need to correct for multiplicity using Bonferroni. In particular, Berkshire was cherry-picked from the *S&P 500*, so its *EXPERT* p -value, adjusted for multiplicity, is $.00006 \times 500 = .012$. This is still significant but not impressively so. (And if we consider that Berkshire is just one of thousands of listed stocks, then the adjusted p -value would not be significant.) Fidelity was selected from a pool of hundreds of stock mutual funds, so when adjusted for multiplicity its p -value is not even close to being significant under our test or the t -test. We conclude that Berkshire has some claim to delivering positive alpha after correcting for multiplicity, but even it must be viewed as a borderline case. The other three assets do not even come close.

Appendix: Proof of Proposition 1

We need to show that $L(p) < 2\Phi(.5(c_p - z_p)) - 1$ and $\lim_{p \rightarrow 0^+} L(p) = 0$. Under the null hypothesis ($\mu = 0$),

$$Z_t = \frac{\ln C_t + .5(\lambda * \sigma)^2 t}{\lambda * \sigma \sqrt{t}} = \frac{\ln C_t + .5c_p^2}{c_p} \text{ is } N(0,1). \quad (\text{A1})$$

Hence the t -test rejects the null at level p if and only if

$$z_p < \frac{\ln C_t}{c_p} + .5c_p. \quad (\text{A2})$$

By contrast, *EXCERT* accepts the null if and only if $\ln C_t \leq \ln(1/p) = .5c_p^2$, that is,

$$\frac{\ln C_t}{c_p} \leq .5c_p. \quad (\text{A3})$$

Power loss occurs in the region where both (22) and (23) hold. Assume now that $\frac{\ln C_t}{c_p} + .5c_p$ is distributed $N(\frac{\mu\sqrt{t}}{\sigma}, 1)$ for some $\mu > 0$. For this μ, σ, t the loss in power is given by the probability of the event

$$z_p - \frac{\mu\sqrt{t}}{\sigma} < \frac{\ln C_t}{c_p} + .5c_p - \frac{\mu\sqrt{t}}{\sigma} \leq c_p - \frac{\mu\sqrt{t}}{\sigma}. \quad (\text{A4})$$

The middle term is distributed $N(0, 1)$, so the probability of this event is

$$\Phi(c_p - \frac{\mu\sqrt{t}}{\sigma}) - \Phi(z_p - \frac{\mu\sqrt{t}}{\sigma}). \quad (\text{A5})$$

This probability is maximized when $z_p - \frac{\mu\sqrt{t}}{\sigma}$ and $c_p - \frac{\mu\sqrt{t}}{\sigma}$ are symmetrically situated about zero. It follows that the power loss function is

$$L(p) \leq 2\Phi(.5(c_p - z_p)) - 1. \quad (\text{A6})$$

In fact the inequality in (A6) holds strictly, because we arrived at this estimate by maximizing the rejection probability *at* time t , whereas CERT rejects if the maximum compound value is high enough at any time up to and including t .

To complete the proof we need to show that $c_p - z_p \rightarrow 0^+$ as $p \rightarrow 0^+$. Recall that when z is large the right tail of the normal distribution has the following approximation [Feller, 1957, p.193]:

$$P(Z \geq z) \approx \frac{e^{-z^2/2}}{z\sqrt{2\pi}}. \quad (\text{A7})$$

Therefore when p is reasonably small, say $p \leq .01$, we have the approximation

$$z_p \approx \sqrt{2 \log z_p + 2 \log(2\pi / p)}. \quad (\text{A8})$$

Combining (A6) and (A8) we find, after some manipulation, that

$$c_p - z_p \approx \frac{\log z_p + .5 \log(2\pi)}{2\sqrt{\pi} \sqrt{2 \log(1/p)}} \leq \frac{\log c_p + .5 \log(2\pi)}{2c_p \sqrt{\pi}}. \quad (\text{A9})$$

Hence $c_p - z_p \rightarrow 0^+$ as $p \rightarrow 0^+$, from which we conclude that $L(p) \rightarrow 0$ as $p \rightarrow 0^+$.

References

Abramovich, F., Benjamini, Y., Donoho, D. L., and Johnstone, I. M. (2006), "Adapting to unknown sparsity by controlling the false discovery rate," *Annals of Statistics*, 34, 584-653.

Agarwal, V., and Naik, N. Y., (2004), "Risks and portfolio decisions involving hedge funds," *Review of Financial Studies*, 17, 63-98.

Agnew, R. A. (2002), "On the TEAM approach to investing," *American Mathematical Monthly*, 109, 188-192.

Benjamini, Y. and Hochberg, Y. (1995), "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statist. Soc., Ser. B*, 57, 289–300.

Berndt, E. R. (1996), *The Practice of Econometrics: Classic and Contemporary*, New York, Addison-Wesley.

Campbell, J. C., Lo, A. W., and MacKinlay, A. C. (1997), *The Econometrics of Financial Markets*. Princeton NJ, Princeton University Press.

Cover, Thomas M. (1991), "Universal portfolios," *Mathematical Finance*, 1, 1-29.

Doob, J. L. (1953), *Stochastic Processes*, New York, John Wiley.

Feller, W. (1971), *An Introduction to Probability Theory and Its Applications*, Princeton University Press.

Foster, D. P. and Stine, R. A. (2008), "Alpha-investing: sequential control of expected false discoveries," *Journal of the Royal Statistical Society Series B*, 70, .

Foster, D. P., and Young, H. P. (2009), "Gaming Performance Fees by Portfolio Managers," Working Paper 08-041, Financial Institutions Center, Wharton School

of Business, University of Pennsylvania. Available at fic.wharton.upenn.edu/fic/papers/09/0909.pdf

George, E. and Foster, D. P. (2000), "Empirical Bayes Variable Selection," *Biometrika*, 87, 731 - 747.

Gerth, F. (1999), "The TEAM approach to investing," *American Mathematical Monthly*, 106, 553-558.

Lo, Andrew W., 2001, "Risk management for hedge funds: introduction and overview," *Financial Analysts' Journal*, Nov/Dec Issue, 16-33.

Miller, R. G. (1981), *Simultaneous Statistical Inference*, 2nd ed., New York: Springer-Verlag.

O'Brien, P.C., and Fleming T.R. (1979), "A multiple testing procedure for clinical trials," *Biometrics*, 35, 549-556.

Pocock, S. J. (1977), "Group sequential methods in the design and analysis of clinical trials," *Biometrika*, 64, 191-199.

Stine, R.A. (2004), "Model selection using information theory and the MDL principle." *Sociological Methods & Research*, 33, 230-260.