2015

Extreme Value Theory and Backtest Overfitting in Finance

Daniel C. Byrnes
danny31mibe@hotmail.com

Extreme Value Theory and Backtest Overfitting in Finance

An Honors Paper Presented for the Department of Mathematics

By Daniel Byrnes

Bowdoin College, 2015

Acknowledgements

I would like to thank professor Thomas Pietraho for his help in the creation of this thesis. The revisions and suggestions made by several members of the math faculty were also greatly appreciated. I would also like to thank the entire department for their support throughout my time at Bowdoin.

Contents

1 Abstract

In order to identify potentially profitable investment strategies, hedge funds and asset managers can use historical market data to simulate a strategy's performance, a process known as backtesting. While the abundance of historical stock price data and powerful computing technologies has made it feasible to run millions of simulations in a short period of time, this process may produce statistically insignificant results in the form of false positives. As the number of configurations of a strategy increases, it becomes more likely that some of the configurations will perform well by chance alone. The phenomenon of backtest overfitting occurs when a model interprets market idiosyncrasies as signal rather than noise, and is often not taken into account in the strategy selection process. As a result, the finance industry and academic literature are rife with skill-less strategies that have no capability of beating the market. This paper explores the development of a minimum criterion that managers and investors can use during the backtesting process in order to increase confidence that a strategy's performance is not the result of pure chance. To do this we will use extreme value theory to determine the probability of observing a specific result, or something more extreme than this result, given that multiple configurations of a strategy were tested.

2 Introduction

Spurious algorithmic investment strategies said to be mathematically sound and empirically tested are rampant within the finance industry and mathematical finance literature. Mathematical jargon such as "stochastic oscillators", "Fibonacci ratios", "Elliot waves", and "parabolic SAR" is strategically advertised to awe potential investors and create the illusion of a scientifically rigorous treatment of investment strategy development [1]. Trading strategies are often backtested, which is the process of using historical stock market data to test the performance of a strategy before it is backed with real money and deployed into the market.

A distinction is made between how the historical stock market data is used during the strategy development and selection process. The in-sample data, also known as the training set in the machine learning literature, is used to design a strategy. The out-of-sample data is the testing set that is unused during the creation of a strategy. A backtest is realistic if in-sample performance is consistent with out-of-sample performance. While backtesting is a valuable tool that can help investors identify profitable investment strategies, many firms and academic studies will only report the results of the in-sample performance.

Common methodologies for testing a collection of strategies can inflate hypothetical performance results and mislead unsuspecting investors into placing capital behind the best performing strategies in-sample. Due to computational advances and the increase in supercomputing technologies, researchers have the capability to search through thousands, or even millions of potentially profitable trading strategies. The improper use of scientific techniques and high-performance computing in the finance industry can yield unintended consequences. In particular, a common problem is backtest overfitting, the phenomenon of using too many variations of a strategy in a backtest relative to the amount of historical market data available.

Bailey et al. remarks in [1],

> We strongly suspect that such backtest overfitting is a large part of the reason why so many algorithmic or systematic hedge funds do not live up to the elevated expectations generated by their managers.

Current practices make it almost impossible for a researcher to not find a strategy that focuses on idiosyncrasies of market data, and thus produces attractive results. Many firms abuse sophisticated mathematical concepts and fail to uphold rigorous methodologies to test investment strategies, producing misleading and statistically insignificant results. Researchers need to control for multiple testing and selection bias when employing advanced computing techniques to find profitable trading strategies.

Stock market data is considered to be composed of signal, or values that are representative of an underlying market movement, and noise, random fluctuations in stock price and volume that are not indicative of any trends in the market. Investors and researchers are often unaware that increasing the number of parameters in a strategy makes it more likely that one of the configurations will happen to perform well in simulation as a result of noise rather than signal. The phenomenon of obtaining desired experimental results from increasing the number of parameters in a model is aptly characterized by the following statement made by mathematician John von Neumann [1]:

> With four parameters I can fit an elephant, and with five I can make him wiggle his trunk.

Commonly used model validation methods include the Akaike Information Criterion (AIC) and cross-validation. The AIC statistic measures the quality of a model by quantifying the trade off between model complexity and fit. However, as the number of strategies tested increases it becomes more likely to find a configuration of parameters that yield satisfying AIC results. The AIC test does not take the number of strategy configurations into account, and as a result is not a sufficient criterion to prevent backtest overfitting. Cross-validation is a technique that involves partitioning the historical dataset into in-sample and out-of-sample subsets as mentioned previously. The issue is that many firms and academic studies only

report in-sample performance results without the number of model configurations tested, which makes it difficult for investors to gauge the legitimacy of backtest performance results.

The issue of backtest overfitting is not specific to finance; other scientific disciplines involve experiments with a very large number of trials, making it difficult to not find results that support a flawed model. For example, a study in the medical research industry might selectively publish experimental results of the best outcomes of a drug trial involving thousands of patients, yet fail to mention the number of other trials that produced inconclusive or undesirable results [1]. For this reason, stringent criteria is necessary to estimate the statistical significance of a specific result given that multiple tests were conducted. For example, Campbell notes in [9] that physicists built the Large Hadron Collider in order to test various theories of particle physics and high-energy physics, such as the existence of the Higgs boson. By the nature of this experiment over a quadrillion tests needed to be conducted and many collisions from known processes produce the same decay-signature as the Higgs boson, indicating many potentially false discoveries of this subatomic particle [9]. Physicists thus required very rigorous standards in order to declare the discovery of the particle. Similar examples are prevalent in other fields that require numerous tests and large amounts of data, such as genetics and medicine. The bottom line is that higher standards for statistical significance are required when a large number of strategy configurations are tested. Unless the number of trials of an experiment are reported, one should be skeptical of backtests that indicate profitable results.

When researchers backtest on a collection of investment strategies they need to effectively minimize the chance that a finding is actually a fluke. The goal of this paper is to develop a criterion that investment managers can use to effectively identify trading strategies that produce promising backtest results yet have no capability of beating the stock market in the future. We will explore the development of a threshold statistic that can give an investment manager confidence that a trading strategy's positive backtest performance is not the result of overfitting.

3.1 The Sharpe Ratio

Performance metrics are used in finance to quantitatively analyze an investment strategy's performance. Given a collection of (potentially viable) trading strategies, an investment manager can use historical market data to backtest these algorithms and use a performance metric, such as the Sharpe ratio, to determine the most successful algorithm. The Sharpe ratio is a convenient summary of the risk and volatility a strategy assumes. We will say a backtest is overfit if a non-zero Sharpe ratio is produced in-sample while the strategy has an insignificant (less than or equal to zero) Sharpe ratio out-of-sample.

There are a variety of other performance measures that are also commonly used to evaluate an investment strategy's performance. Although this paper only considers the Sharpe ratio, similar analyses can be done using other statistics. We will use the Sharpe ratio since it will be assumed that the returns of an investment strategy are independently and identically distributed (iid). Furthermore, we will suppose that a strategy's excess returns are normally distributed. If instead we have non-normal returns then we would not be justified in using the Sharpe ratio. Other reasons why the Sharpe ratio was chosen for this analysis include its prevalence in the finance literature and the relatively few data requirements needed to compute it (in comparison to measures with higher moments). In the next section we discuss other measures used to quantify the success of investment strategies.

Definition. *The risk premium r_t of an investment is the return of the investment, R_a, in excess of the return that would be earned on a risk free investment, R_b. So $r_t = R_a - R_b$, where the risk free investment is often a U.S. treasury bill or an investment that involves no risk.*

Definition. *The Sharpe ratio is the ratio of the average risk premium and the standard deviation of the same risk premium[1]:*

$$SR = \frac{\mathbb{E}[R_a - R_b]}{\sqrt{\text{Var}[R_a - R_b]}},$$

where R_a is the sequence of returns on an asset and R_b is the sequence of returns on a benchmark. Simply put, the Sharpe ratio is the reward per unit of risk.

Let the random variables $X_1, X_2, ...X_n$ be iid returns of an investment strategy.
Then the sample mean is defined as $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and the sample variance is defined as $\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \hat{\mu})^2$. These tools allow us to make the following definition.

Definition. *The estimator of the Sharpe ratio, $\widehat{SR}$ is defined as*

$$\widehat{SR} = \frac{\hat{\mu} - R_b}{\hat{\sigma}},$$

where R_b is the return on the risk free asset.

The Sharpe ratio may be a misleading measure when the returns of an investment follows an asymmetric or fat-tailed distribution or when returns are not independently and identically distributed. To begin we will assume the risk premiums, r_t of any given investment strategy are iid and follow the normal distribution:

$$r_t \sim \mathcal{N}(\mu, \sigma^2)$$

Example 3.1. *The annualized average value of the daily returns of Apple (APPL) stock from December 2013 through December 2014 was -0.53, and the annualized standard deviation of daily returns was 0.88. Using a five year United States Treasury bond with an annual rate of 1.72% as the benchmark, the Share ratio of this investment is*

$$\widehat{SR} = \frac{(-0.53 - 0.0172)}{-0.88} = -0.62$$

Definition. *The annualized Sharpe ratio is defined as*

$$\widehat{SR} = \frac{\hat{\mu} - R_b}{\hat{\sigma}}\sqrt{q},$$

where q is the number of returns per year.

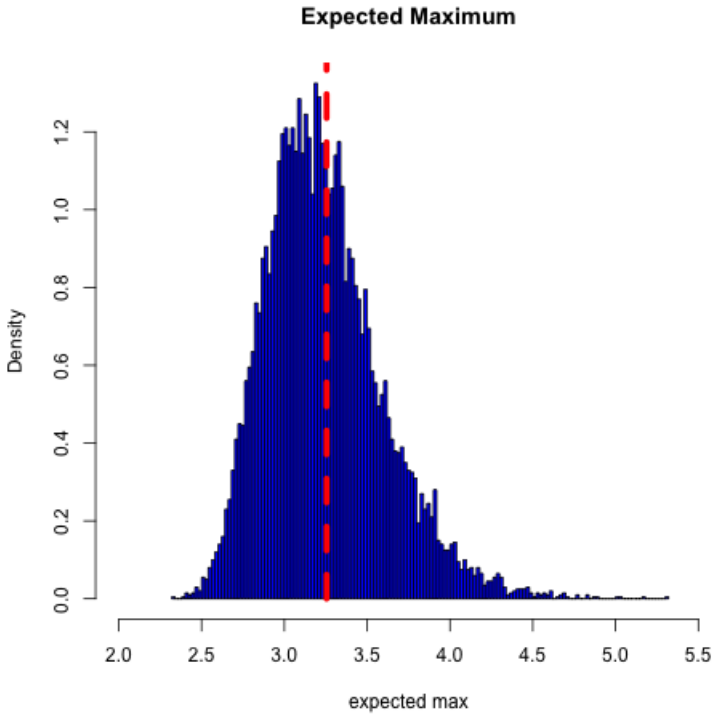
A big concern when testing many different configurations of a strategy with a given number of parameters is the increased likelihood that one of the strategies will produce a high Sharpe ratio purely by chance, and will thus be falsely regarded as profitable. A strategy that exhibits false positive performance is said to be *overfit* since it does not have any fundamental predictive power yet performed well on the training set.

Consider an investment strategy with n parameters and the collection of all configurations of this strategy. Suppose this is a skill-less strategy that does not identify any meaningful pattern in the market and thus should have a Sharpe ratio of zero. An important question asked in Bailey et al.[1] is how high is the expected maximum Sharpe ratio in-sample among a collection of strategies where the true Sharpe ratio is zero?

Suppose M_n is the observed maximum value of a sample of Sharpe ratios. Producing a Sharpe ratio greater than the expected maximum, $\mathbb{E}[M_n]$, is a minimum criterion for ensuring that the backtest is not overfit. Assuming that the distribution of Sharpe ratio is roughly normal, the expected maximum gives us an approximately 50% confidence that a Sharpe ratio greater than this result is not the result of an overfit backtest (Figure 1).

A method is presented in [1] for computing $\mathbb{E}[M_n]$ and the minimum backtest length, minBTL, or the minimum number of years worth of data needed to avoid selecting a trading strategy with this Sharpe ratio in-sample. This paper develops a stronger criterion that gives greater confidence that a backtest is not overfit than that allowed by $\mathbb{E}[M_n]$. This amounts to calculating the quantile $Q(p)$, which gives p confidence that the best Sharpe ratio of a series of backtests was not produced by a skill-less strategy.

Figure 1: Histogram of maxima of a collection of normally distributed random variables. The red line corresponds to $\mathbb{E}(M_N)$, which has a p-value of approximately 0.50.



3.2 Other Performance Measures

As mentioned above, the Sharpe ratio is used in the analysis presented in this paper, but other metrics could have been chosen. Below we describe similar performance metrics that also could have been used.

Definition. *The Sortino ratio is defined as*

$$\frac{\mathbb{E}(R_a) - R_b}{\sigma_N},$$

where $\mathbb{E}(R_a)$ is the expected return on the asset, R_b is the risk free rate of return, and σ_N is the standard deviation of the negative asset returns.

The Sortino ratio is a modified version of the Sharpe ratio that uses the downside deviation, or the standard deviation of negative asset returns. The Sortino differs from the Sharpe ratio in that it only penalizes for downside volatility, whereas the Sharpe takes general volatility into account.

Definition. *The Information ratio is defined as*

$$IR = \frac{R_a - R_i}{\sigma_i}$$

where R_a is the return of the portfolio, R_i is the return of an index (or benchmark), and σ_i is the standard deviation of the difference between the returns on portfolio and the returns on the index.

Similar to the Sharpe, the information ratio also measures excess returns per unit of risk. But rather than considering returns in excess of a risk-free investment, the information ratio measures the rate of return of an investment portfolio against a benchmark equity index. A commonly used benchmark is the S&P 500 index.

Performance metrics allow researchers to determine if an investment strategy is profitable by quantifying its performance results. As mentioned in the introduction, an investment strategy is backtested in order to gauge its ability to beat the market. As the number of configurations of a strategy increases it becomes more likely that a backtest is overfit, and as a result the Sharpe ratio will be inflated. We wish to determine the probability that a positive in-sample Sharpe ratio is zero out-of-sample. In order to do this we will study the distribution of Sharpe ratios and rely on convergence properties of certain random variables and their distributions to develop an in-depth analysis of this problem. Note that although we will only consider the Sharpe ratio these results can be reproduced using any other performance metric.

3.3 Example of an Overfit Strategy

Using a tool developed by D.H. Bailey et al. of Lawrence Berkeley National Laboratory we show how simple it is to optimize a trading strategy with a given number of parameters to make it fit a data set generated by a sequence of random numbers [1]. The parameters used for this strategy include the maximum holding period, maximum stop loss, entry day, and side. The maximum holding period is the number of days that stock can be held before it is sold, and the maximum stop loss is the percent of the initial price of an asset that

Figure 2: The in-sample performance of trading strategy. The green line represents the stock prices (given from a pseudorandom number generator) and the blue line represents the profit or loss of the trading strategy. Notice that over time the strategy becomes more profitable and is thus optimized.

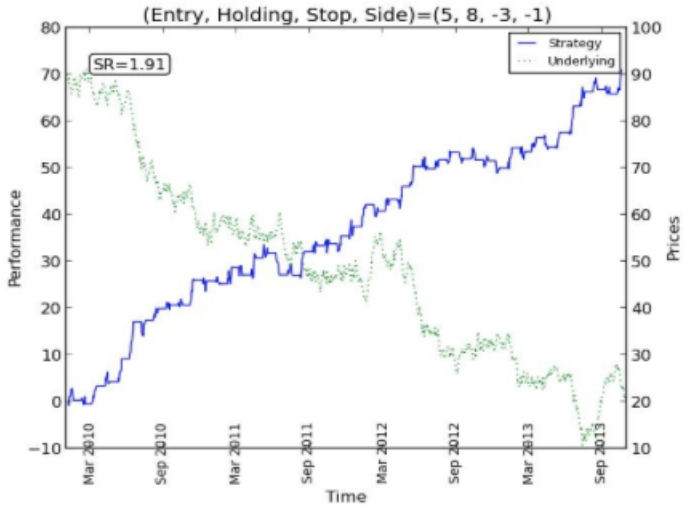
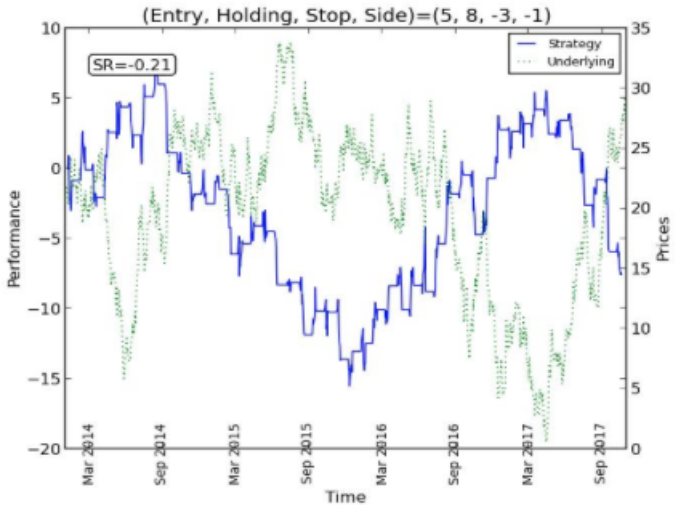


Figure 3: The out-of-sample performance of the strategy. Notice that the out-of-sample Sharpe ratio (-0.21) is lower than the in-sample Sharpe ratio (1.91)



can be lost before the stock is sold. The entry day is the day that the stock market is entered each trading month, and the side of a trading strategy refers to either going long or short. The long position in an investment strategy consists of buying a security with the expectation that the asset will increase in value. Similarly, the short position consists of selling a security that is expected to decrease in value.

This tool tries different configurations of the parameters until one yields a higher Sharpe ratio than all previous combinations, at which point the strategy parameters are updated. This optimal strategy was simulated over the course of 1000 trading days with a maximum holding period of 10 days and maximum stop loss of 20 percent. Combinations of all 22 entry days in a trading month were attempted along with both the long and short side options (Figures 2 and 3).

4 Modes of Convergence for Random Variables

Our goal is to analyze the distribution of the maximum Sharpe ratio of a collection of trading strategies. In order to do this we rely on the convergence properties of probability distributions to a special class of functions known as extreme value distributions. In this section we define convergence in distribution and other types of convergence for a sequence of random variables.

4.1 Random Variables and Distributions

Definition. *A probability space is an ordered triple* $(\Omega, \mathcal{E}, \mathbb{P})$ *where* Ω *is a sample space of all possible outcomes,* $\mathcal{E}$ *is a subset of the sample space which consists of all possible events, and* $\mathbb{P}$ *is a function that maps an event to a real number in the interval* $[0, 1]$.

A random variable is a mapping $X : \Omega \to \mathbb{R}$ from a probability space $(\Omega, \mathcal{E}, \mathbb{P})$ to the real numbers $\mathbb{R}$. So the function X assigns a real number to each outcome in the sample space.

Definition. *A random variable* X *has a continuos distribution if there exists a non-negative function* f *defined on the real line such that for any interval* A,

$$\mathbb{P}(X \in A) = \int_A f(x)dx.$$

The function f in the definition above is called the *probability density function* of X if it satisfies two requirements:

$$f(x) \geq 0 \text{ for all } x \in X$$

and

$$\int_{-\infty}^{\infty} f(x) = 1.$$

Definition. *The cumulative distribution function F of a random variable X is a function defined for each real number x as*

$$F(x) = \Pr(X \leq x) \text{ for } -\infty < x < \infty.$$

Similarly, $F(x) = \int_{-\infty}^{x} f(t)dt$ if X is a continuous random variable.

Example 4.1. *We say that a random variable X is normally distributed with parameters μ and σ, denoted by $X \sim \mathcal{N}(\mu, \sigma)$, if for all $x \in \mathbb{R}$,*

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

and

$$\mathbb{P}(X \leq x) = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} dx.$$

4.2 Convergence in Distribution

Definition. *Distributions F and G of random variables X and Y, respectively, are said to be of the same type if there exists $a > 0$, and $b \in \mathbb{R}$ such that $aX + b$ has the same distribution as Y. So then $F(aX + b) = G(x)$.*

If random variables X and Y are equivalent in distribution then we will write $X \stackrel{d}{=} Y$.

Definition. *Suppose that $(X_1, X_2, ...)$ and X are real-valued random variables with distribution functions $(F_1, F_2, ...)$ and F, respectively. We say that the distribution of X_n converges to the distribution of X as $n \to \infty$ if*

$$F_n(x) \to F(x) \text{ as } n \to \infty$$

for all x at which F is continuous. We write $F_n(x) \xrightarrow{d} F(x)$.

Example 4.2. *Consider a collection of iid random variables $\{X_i\}_{i=1}^n$ that are uniformly distributed in the interval $(0, 1)$ and define the random variable $M_n = max\{X_i\}_{i=1}^n$. Then the distribution of M_n, $F_n(x)$, converges to the distribution of an exponential random variable, $\lambda e^{-\lambda x}$ where λ is a constant.*

Proof. For any $\epsilon > 0$ we can see that,

$$\begin{aligned}\mathbb{P}(|M_n - 1| \geq \epsilon) &= \mathbb{P}(M_n \leq 1 - \epsilon) \\ &= \mathbb{P}(X_1, X_2, ..., X_n \leq 1 - \epsilon) \\ &= (1 - \epsilon)^n\end{aligned}$$

So then

$$\lim_{n\to\infty} \mathbb{P}(|M_n - 1| \geq \epsilon) = 0.$$

But if we let $\epsilon = \frac{t}{n}$ then we have,

$$\begin{aligned}\mathbb{P}(M_n \leq 1 - \epsilon) &= \mathbb{P}(M_n \leq 1 - \frac{t}{n}) \\ &= (1 - \frac{t}{n})^n\end{aligned}$$

and

$$\lim_{n\to\infty} (1 - \frac{t}{n})^n = e^{-t}.$$

So then we can see that $\mathbb{P}(n(1 - M_n) \leq t) = 1 - e^{-t}$ as $n \to \infty$. Thus, the distribution $F_n(x)$ converges to the exponential distribution as $n \to \infty$. □

4.3 Other Types of Convergence

A sequence of random variables A_n *converges in probability* to the random variable A, $A_n \xrightarrow{P} A$, if for all $\epsilon > 0$ the relation

$$\mathbb{P}(|A_n - A| > \epsilon) \to 0, \text{ as } n \to \infty$$

holds.

Convergence in probability is metrizable in that we can define the metric

$$\rho(X, Y) = \mathbb{E}\Big(\frac{|X - Y|}{1 + |X - Y|}\Big),$$

which implies convergence in probability if the distance between X and Y is zero.

We say that a sequence of random variables A_n converges *almost surely* (a.s.) to the random variable A if for outcome ω in the sample space Ω,

$$\mathbb{P}(A_n \to A) = \mathbb{P}(\{\omega | A_n(\omega) \to A(\omega)\}) = 1$$

Almost sure convergence implies convergence in probability, which implies convergence in distribution. These relations imply that convergence in distribution is the weakest mode of convergence considered in this paper.

Definition. *A distribution function F is a limit law if it is the limiting distribution for $Y \sim \max\{X_1, ..., X_n\}$ where X_i are iid with some distribution function G.*

The following theorem tells us that the limit law of a sequence of random variables is uniquely determined up to affine transformation (that is, up to changes of location and scale).

Theorem 4.3 (Convergence to types)**.** *If F_n, G, and H are distribution functions with G and H being non-degenerate, and there exist a_n, $a_n^{'} > 0$ and d_n, $d_n^{'} \in \mathbb{R}$ such that $F_n(a_n^{'}x+d_n^{'}) \xrightarrow{d} G(x)$ and $F_n(a_nx+d_n) \xrightarrow{d} H(x)$ at all points of continuity of G, respectively H, then*

$$\frac{a_n}{a_n^{'}} \to a > 0 \text{ and } \frac{(d_n - d_n^{'})}{a_n^{'}} \to b \in \mathbb{R},$$

and $G(ax + d) = H(x)$ for all $x \in \mathbb{R}$.

5 Limit Behavior of Maxima

5.1 Maximum Sharpe Ratio

Lo[11] uses asymptotic statistical theory to derive the limiting distribution of estimated annualized Sharpe ratios as $y \to \infty$:

$$\widehat{SR} \xrightarrow{d} \mathcal{N}\left[SR, \frac{1 + \frac{SR^2}{2q}}{y}\right]$$

where y is the number of years used to compute $\widehat{SR}$ and SR is the actual Sharpe ratio. So then for a sufficiently high number of years worth of data we should expect the values of $\widehat{SR}$ to be normally distributed with mean SR. Suppose we have a collection of strategy configurations that have true Sharpe ratio equal to zero and one year worth of data to calculate $\widehat{SR}$. Then $\mu = 0$, $y = 1$, and $\widehat{SR} \sim \mathcal{N}(0, 1)$. Thus, in order to determine the expected maximum of a collection of Sharpe ratios one can use the distribution of the maximum of a collection of iid random variables that are normally distributed with $\mu = 0$ and $\sigma^2 = 1$.

Suppose $X_1, X_2, ...$ is a sequence of iid non-degenerate random variables with common distribution function F. The maximum of this sample is denoted by $M_n = \max\{X_1, ..., X_n\}$. Unless we rescale the maximum, M_n, the resulting distribution will be trivial as we take the limit, meaning that $\mathbb{P}(M_n \leq x) \in \{0, 1\}$ for all values of x as $n \to \infty$. For example, suppose $x \in \mathbb{R}$, $n \in \mathbb{N}$ and let $x_F = \sup\{x \in \mathbb{R} | F(x) < 1\}$.

Then we have two cases:

If $x < x_F$

$$\lim_{n\to\infty} \mathbb{P}(M_n \leq x) = \lim_{n\to\infty} F^n(x) = 0$$

Or if $x \geq x_F$

$$\lim_{n\to\infty} \mathbb{P}(M_n \leq x) = \lim_{n\to\infty} F^n(x) = 1$$

One can use the limiting behavior of the scaled maxima to compute the expected maximum, or the probability that an observation is greater than the maximum of a distribution. These results are applicable to investment strategy backtesting if an appropriately large number of strategy configurations are considered. Our first goal is to determine the asymptotic behavior of $\mathbb{P}(M_n \leq x)$ as $n \to \infty$.

5.2 Connection to the Central Limit Theorem

The question above is similar to determining the distribution of the sum of a collection of iid random variables $\{X_i\}_{i=1}^n$ with well-defined mean and variance as $n \to \infty$.

Theorem 5.1 (Central Limit Theorem)**.** *Recall that the sample mean of a collection of random variables is defined as* $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ *with* $\mathbb{E}(X_i) = \mu$ *and* $\mathbb{V}(X_i) = \sigma$ $\forall i$*. Then*

$$\lim_{x\to\infty} \mathbb{P}\left[\frac{\overline{X}_n - \mu}{\sigma} \leq x\right] = \Phi(x)$$

where $\Phi(x)$ *is the CDF of the standard normal distribution.*

Proof. If we let $X_1, X_2, ..., X_n$ denote iid random variables with distribution D and with $\mathbb{E}(X_i) = \mu < \infty$, $\mathbb{V}(X_i) = \sigma^2 < \infty$ for all i, then it suffices to show that

$$\lim_{N\to\infty} \sqrt{n}\frac{(\overline{X}_n - \mu)}{\sigma} \overset{d}{=} Z,$$

where $Z \sim \mathcal{N}(0, 1)$.

Let $Z_n = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} Y_i$ where $Y_i = \frac{X_i - \mu}{\sigma}$. Note that $\mathbb{E}(Y_i) = 0$ and $\mathbb{V}(Y_i) = 1$. The characteristic

function of a random variable X is defined as,

$$\varphi_X(t) = \mathbb{E}(e^{itx}).$$

We will use the characteristic function of Z_n to show that Z_n is normally distributed.

$$\begin{aligned}
\varphi_{Z_n}(t) &= \varphi_{\frac{1}{\sqrt{n}}\sum_{i=1}^{n} Y_i}(t) \\
&= \prod_{i=1}^{n} \varphi_{\frac{Y_i}{\sqrt{n}}}(t) \\
&= \prod_{i=1}^{n} \varphi_{Y_i}\left(\frac{t}{\sqrt{n}}\right) \\
&= \varphi_{Y_i}\left(\frac{t}{\sqrt{n}}\right)^n
\end{aligned}$$

Note that,

$$\begin{aligned}
\varphi_{Y_i}\left(\frac{t}{\sqrt{n}}\right)^n &= \prod_{i=1}^{n}(1 - \frac{(t/\sqrt{n})^2}{2})) \\
&= \left(1 - \frac{t^2}{2n} + O(t^2)\right)^n
\end{aligned}$$

The Taylor series expansion of e^{itx} gives us:

$$e^{itx} = 1 + (it)x + \frac{(it)^2x^2}{2!} + ...$$

So then we can use the taylor expansion of e^{itx} as such:

$$\varphi_{Z_n}(t) = \left(1 - \frac{t^2}{2n} + O(t^2)\right)^n.$$

Recall that $\lim_{n\to\infty}\left(1 + \frac{x}{n}\right)^n = e^x$. Letting $x = -\frac{t^2}{2}$ we can see that

$$\varphi_{Z_n}(t) = e^x = e^{\frac{-t^2}{2}} \stackrel{d}{=} \mathcal{N}(0,1)$$

Thus,

$$\lim_{n\to\infty} \sqrt{n}\left(\frac{\overline{X_n}-\mu}{\sigma}\right) \sim \mathcal{N}(0,1)$$

□

Thus, the limiting distribution of the sum of a collection of iid random variables is the normal distribution. In the following section on extreme value theory we attempt to derive an analogous result for the limiting distribution of a collection of maxima rather than sums. Similar to the case for partial sums, we will need to scale M_n to account for the fact that M_n may tend to infinity.

5.3 Extreme Value Theory and Max-Stable Distributions

The distribution function of the maximum of a sequence of iid random variables $\{X_i\}_{i=1}^n$ is written as:

$$\mathbb{P}(M_n \leq x) = \mathbb{P}(X_1 \leq x, ..., X_n \leq x) = F^n(x),\ x \in \mathbb{R},\ n \in \mathbb{N}.$$

Consider probabilities of the form:

$$\mathbb{P}(c_n^{-1}(M_n - d_n) \leq x)$$

If we let $u_n = u_n(x) = c_n x + d_n$ the we can rewrite the above probability as

$$\mathbb{P}(M_n \leq u_n)$$

It is important consider the conditions on F needed to ensure that the limit of $\mathbb{P}(M_n \leq u_n)$ as $n \to \infty$ exists for appropriate constants u_n. For instance, there needs to be certain continuity conditions on F at its right endpoint, $x_F = \sup\{x \in \mathbb{R} | F(x) < 1\}$.

These specific conditions illustrate the crucial difference between the convergence of sums and maxima. As shown above, the central limit theorem states that the normal distribution will be the limit of the sum of iid random variables with finite mean and variance, which is implied by the very general moment condition $\mathbb{E}(X^2) < \infty$. In contrast to sums, specific

conditions on the right endpoint are needed to ensure that $\mathbb{P}(M_n \leq u_n)$ converges to a non-trivial limit. Specifically, a distribution function with a jump at its finite right endpoint cannot have a non-degenerate limit distribution for M_n, regardless of the normalization. For example, there is no non-trivial convergence of maxima for the Poisson or geometric distribution.

Definition. *The distribution of a non-degenerate random variable X is called max-stable if it satisfies*

$$\max\{X_1, ..., X_n\} \stackrel{d}{=} c_n X + d_n$$

for iid $X, X_1, ..., X_n$, appropriate $c_n > 0$ and $d_n \in \mathbb{R}$, and for every $n \geq 2$.

Theorem 5.2 (limit property of max-stable laws)**.** *The class of max-stable distributions coincides with the class of all possible (non-degenerate) limit laws for (properly normalized) maxima of iid random variables.*

A consequence of Theorem 5.2 is that if we want to find the limit law of a collection of iid random variables, then we simply need to look at the set of max-stable distributions.

Proof. If a distribution function of a random variable X is max-stable then it must be the limit-law of a collection of iid random variables by definition. It remains to prove that the limit law of affinely transformed maxima is max-stable. Assume that for appropriate norming constants,

$$\lim_{n\to\infty} F^n(c_n x + d_n) = H(x),$$

$x \in \mathbb{R}$. for some non-degenerate distribution function H. then for every $k \in \mathbb{N}$,

$$\lim_{n\to\infty} F^{nk}(c_n x + d_n) = \Big(\lim_{n\to\infty} F^n(c_n x + d_n)\Big)^k = H^k(x), x \in \mathbb{R}$$

and,

$$\lim_{n\to\infty} F^{nk}(c_{nk} x + d_{nk}) = H(x), x \in \mathbb{R}.$$

By the convergence to types theorem (theorem 4.3) there exist constants $\widetilde{c_k} > 0$ and $\widetilde{d_k} \in \mathbb{R}$ such that

$$\lim_{n\to\infty} \frac{c_{nk}}{c_n} = \widetilde{c_k} \text{ and } \lim_{n\to\infty} \frac{d_{nk} - d_n}{c_n} = \widetilde{d_k}$$

so for iid random variables $Y_1, ..., Y_k$ with distribution function H,

$$\max\{Y_1, ..., Y_k\} \stackrel{d}{=} \widetilde{c_k} Y_1 + \widetilde{d_k}.$$

Thus, the limit law of affinely transformed maxima is max-stable. □

The following theorem tells us that the possible limit distributions of normalized maxima consists of a small class of distribution functions. We will call these functions the extreme value distributions.

Theorem 5.3 (Fisher-Tippett-Gnedenko Theorem)**.** *Let (X_n) be a sequence of iid random variables. If there exist norming constants $c_n > 0$, $d_n \in \mathbb{R}$ and some non-degenerate cumulative distribution function H such that $c_n^{-1}(M_n - d_n) \xrightarrow{d} H$, then H belongs to type of one of the following three distribution functions:*

Fréchet:

$$\phi_\alpha(x) = \begin{cases} 0 & x \leq 0 \\ e^{-x^{-\alpha}} & x > 0 \end{cases}$$

Weibull:

$$\psi_{\alpha>0}(x) = \begin{cases} e^{-(-x)^{-\alpha}} & x \leq 0 \\ 1 & x > 0 \end{cases}$$

Gumbel:

$$\Lambda(x) = \begin{cases} e^{-e^{-x}} & x \in \mathbb{R} \end{cases}$$

Proof. Suppose that H is a non-degenerate, max-stable distribution function. We will show that under certain conditions H must be the Gumbel distribution. Similar arguments can be used to derive the Fréchet and Weibull distribution functions. Since H is max-stable we know that for all $s > 0$ and $n \in \mathbb{N}$,

$$H^{ns}(a_{ns}x + d_{ns}) = H(x)$$

for all $x \in \mathbb{R}$, and

$$H^{ns}(a_n x + d_n) = (H^n(a_n x + d_n))^{ns/n} = H^{ns/n}(x) \to H^s(x)$$

for all $x \in \mathbb{R}$ as $n \to \infty$. Since $H(x)$ and $H^s(x)$ are both non-degenerate distribution functions, we can apply Theorem 4.3:

$$\lim_{n\to\infty} \frac{a_n}{a_{ns}} = \alpha_s > 0, \lim_{n\to\infty} \frac{d_n - d_{ns}}{a_{ns}} = \delta_s \in \mathbb{R}$$

and so $H^s(\alpha_s x + \delta_s) = H(x)$.

So for any $s, t > 0$,

$$\begin{aligned} H^{ns}(x) &= H(\alpha_{ns} x + \delta_{ns}) \\ &\quad and \\ H^{ns}(x) &= (H^n(x))^s \\ &= (H(\alpha_n(x) + \delta_n))^s \\ &= H(\alpha_s[\alpha_n x + \delta_n] + \delta_s) \\ &= H(\alpha_s \alpha_n x + \alpha_s \delta_n + \delta_s) \end{aligned}$$

Since H is non-degenerate the arguments must be equal, so

$$\alpha_{ns} = \alpha_s \alpha_t, \ \delta_{ns} = \alpha_n \delta_s + \delta_n.$$

Rewriting these functional equations,

$$\alpha(ns) = \alpha(n)\alpha(s) \tag{1}$$

$$\delta(ns) = \alpha(n)\delta(s) + \delta(n). \tag{2}$$

Solving equations (1) and (2) leads to the three distribution types, $\Lambda(x), \psi_\alpha(x), \phi_\alpha(x)$.

The function $f(x) = \log \alpha(e^x)$ is continuous and satisfies the Cauchy Functional Equation

$$f(x+y) = f(x) + f(y)$$

which has solution $f(x) = \theta x$ for some $\theta \in \mathbb{R}$. So then

$$\log \alpha(e^x) = \theta x \implies \alpha(e^x) = e^{\theta x}$$

and so $\alpha(t) = t^\theta$, where $t = e^x$ and for some constant $\theta \in \mathbb{R}$. There are three cases that give rise to the three types of extreme value distributions, $\theta < 0$, $\theta = 0$, or $\theta > 0$. The case where $\theta = 0$ gives rise to the Gumbel distribution.

Suppose $\theta = 0$, then $\alpha_t = 1$ and so $\delta_{sn} = \delta_s + \delta_n$. If we let $g(x) = \delta(e^x)$ then $g(x)$ is again a solution of the Cauchy functional equation, and so $g(x) = cx$ and if we let $x = \log t$ then $d(t) = c \log t$ for some $c \in \mathbb{R}$.

So now

$$H^s(x + c \log s) = H(x)$$

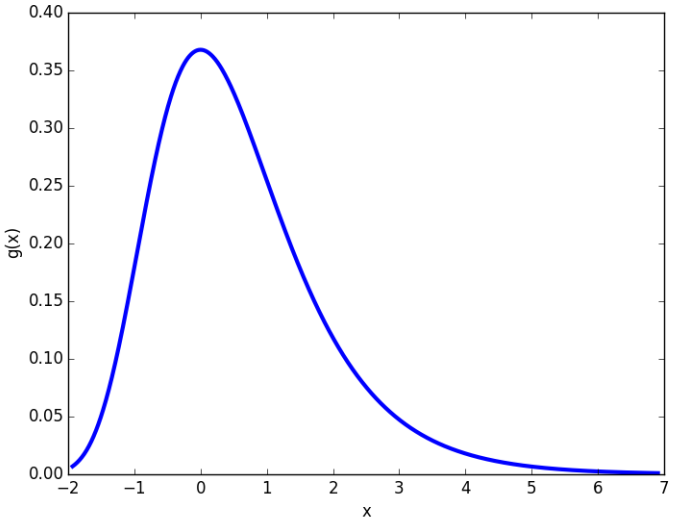
for all $s > 0, x \in \mathbb{R}$. Suppose $c = 0$, then $H(x) \in \{0, 1\}$ for all $x \in \mathbb{R}$, which implies that H is degenerate and contradicts our initial assumption, so $c \neq 0$. If H is non-degenerate then $H(x_0) \in (0, 1)$ for some $x_0 \in \mathbb{R}$. Assume without loss of generality that $x_0 = 0$. Also let $y = c \log s$. Then,

$$\begin{aligned} H(y) &= H(0)^{-\frac{y}{c}} \\ &= e^{-k^{-\frac{y}{c}}} \\ &= e^{-e^{-\frac{y}{c \log k}}} \\ &= e^{-e^x} \end{aligned}$$

Where $x = -\dfrac{y}{c \log k}$. Thus, $\theta = 0 \implies \Lambda(x) = e^{-e^{-x}}$, $x \in \mathbb{R}$.

□

Figure 4: The PDF of the standard Gumbel distribution



6 The Normal Distribution is in MDA(Λ)

Since we have assumed that the returns of our investment strategies are normally distributed we would like to know which of the three extreme distributions the maximum random variable, M_n, will converge to as $n \to \infty$. If the distribution of extreme normal random variables converges to one of the distributions of extreme value type then we can use the extreme distribution to model investment returns and the in-sample Sharpe ratios.

Definition (Maximum domain of attraction)**.** *We say that the random variable X belongs to the maximum domain of attraction (MDA) of the extreme value distribution H if there exist constants $c_n > 0$, $d_n \in \mathbb{R}$ such that $c_n^{-1}(M_n - d_n) \xrightarrow{d} H$.*

The following definition is of a class of functions known as von Mises functions. These functions will help us classify the maximum domain of attraction of the Gumbel distribution, and eventually prove that the normal distribution must belong to this domain of attraction.

Definition (Von Mises Function)**.** *Let F be a distribution function with right endpoint $x_F \leq \infty$, where $x_F = \sup\{x \in \mathbb{R} | F(x) < 1\}$. Suppose there exists some $z < x_F$ such that F can be written as*

$$1 - F(x) = c \cdot exp\left\{ - \int_z^x \frac{1}{a(t)} dt \right\}$$

where $z < x < x_F$, $c \in \mathbb{R}^+$ and $a(t)$ is a positive and absolutely continuous function with density a' and $\lim_{x \to x_F} a'(x) = 0$. Then F is called a von Mises function, and $a(t)$ is the auxiliary function of F.

The following lemma allows us to to characterize functions that are von Mises, and thus in the maximum domain of attraction of the Gumbel distribution.

Lemma 6.1 (Differentiability at the right endpoint)**.** *Let $F(x)$ be a distribution function with right endpoint $x_F \leq \infty$ and density function $f(x)$. Then F is a von Mises function if and only if*

$$\lim_{x \to x_F} \left(\frac{d}{dx} \frac{(1 - F(x))}{f(x)} \right) = 0$$

It is proven in [3] that if a distribution function F satisfies the lemma above then $F \in \text{MDA}(\Lambda)$.

Lemma 6.2 (Mill's ratio)**.** *Let $F(x)$ be the distribution function and $f(x)$ the density function of a random variable. Then*

$$\frac{(1 - F(x))}{f(x)}$$

is called Mill's ratio. If the random variable X is normally distributed, $X \sim \Phi$ (where Φ is the standard normal distribution) then,

$$\lim_{x \to \infty} \frac{(1 - \Phi(x))}{\phi(x)} = \frac{1}{x}$$

Using these lemmas it can be shown that the normal distribution is a von Mises function. From [5] we know that all von Mises functions are in the maximum domain of attraction of the Gumbel distribution, so we can now state the limiting distribution of properly

normalized maxima of normally distributed random variables:

Proposition 6.3. *The normal distribution belongs to the maximum domain of attraction of the Gumbel,* $\Phi(x) \in MDA(\Lambda)$.

Proof. Using lemma 6.1, we have that:

$$\begin{aligned}\frac{d}{dx}\frac{(1-\Phi(x))}{\phi(x)} &= \frac{-\phi(x)^2 - (1-\Phi(x))\phi^{'}(x)}{\phi(x)^2} \\ &= \frac{\phi^{'}(x)}{\phi(x)}\frac{(1-\Phi(x))}{\phi(x)} - 1\end{aligned}$$

Note that $\frac{\phi^{'}(x)}{\phi(x)} = x$ and $x_F = F^{-1}(1) = \infty$. So we must evaluate

$$\lim_{x\to\infty}\left(x\frac{(1-\Phi(x))}{\phi(x)} - 1\right)$$

Note that $\frac{(1-\Phi(x))}{\phi(x)}$ is Mill's ratio and this converges to $\frac{1}{x}$ as $x \to \infty$. So then,

$$\lim_{x\to\infty}\left(x\frac{(1-\Phi(x))}{\phi(x)} - 1\right) = x\frac{1}{x} - 1 = 0$$

And thus, the Φ is a Von Mises function and $\Phi(x) \in MDA(\Lambda)$. □

Proposition 6.4 ([4])**.** *Suppose a distribution function* F *is a von Mises function, so* $F \in MDA(\Lambda)$. *The norming constants for this function are*

$$d_n = \sqrt{n\ln n} - \frac{\ln\ln n + \ln(4\pi)}{2\sqrt{2\ln n}}$$

and

$$c_n = (2\ln n)^{-1/2}.$$

Thus for all $x \in \mathbb{R}$,

$$\lim_{n\to\infty}\mathbb{P}(c_n^{-1}(M_n - d_n) \le x) = e^{-e^{-x}}$$

In [1] Bailey et al. calculated the expected maximum for a large collection of calculated Sharpe ratios that follow the standard normal distribution:

$$\mathbb{E}(M_n) = (1-\gamma)Z^{-1}[1-\frac{1}{N}] + \gamma Z^{-1}[1-\frac{1}{N}e^{-1}],$$

where Z^{-1} is the inverse of the standard normal distribution function and $\gamma = 0.5772156649...$ is the Euler-Mascheroni constant. Suppose our null hypothesis is that the actual Sharpe ratio of an investment strategy is zero, and the strategy is thus skill-less. The probability that an in-sample Sharpe ratio greater than this criterion has a real Sharpe ratio equal to zero is approximately $\frac{1}{2}$. While the threshold $\mathbb{E}(M_n)$ serves as a minimum criterion for rejecting strategies with lower Sharpe ratios in-sample, producing an in-sample Sharpe ratio greater than this threshold does not guarantee that the positive performance is not a result of pure chance.

We wish to produce a statistic that will give an investor more confidence than $\mathbb{E}(M_n)$ that a strategy's great performance is not the result of backtest overfitting. In what follows below we generalize the concept of the expected maximum to obtain a result that depends on the desired p-value. We use the result from Theorem 6.3 that for very large sample sizes the limiting distribution of maximum normally distributed random variables is the Gumbel distribution. Later on we will consider the convergence of extremes of distributions with fat-tails, since end-tail behavior is prominent in financial returns.

7.1 Developing a More Stringent Criterion

Definition. *Let X be a random variable with distribution function F. The quantile function Q (or inverse CDF) is defined by,*

$$Q(p) = F^{-1}(p) = \inf\{x \in \mathbb{R} | F(x) \geq p\}$$

for probability $p \in [0,1]$.

Building off the work of Bailey et al. [1] we develop the following proposition to calculate the p-quantile:

Proposition 7.1. *Let $M_n = max\{X_1, ..., X_n\}$ where $X_i \sim \mathcal{N}(0, 1)$. The p-quantile of that sample, $Q(p)$, can be approximated for large n as*

$$Q(p) \approx d_n - c_n(\ln(-\ln(p))$$

for appropriate norming constants c_n, d_n.

Here we are using the Gumbel distribution with norming constants d_n and c_n. Note that in the case above $\mu = 0$ and $y = 1$. If instead we consider the case where $\mu = 0$ but $y \neq 1$ then the value above needs to be rescaled by the standard deviation of the annualized Sharpe ratio, $y^{-\frac{1}{2}}$. Thus, we can modify the p-quantile as such:

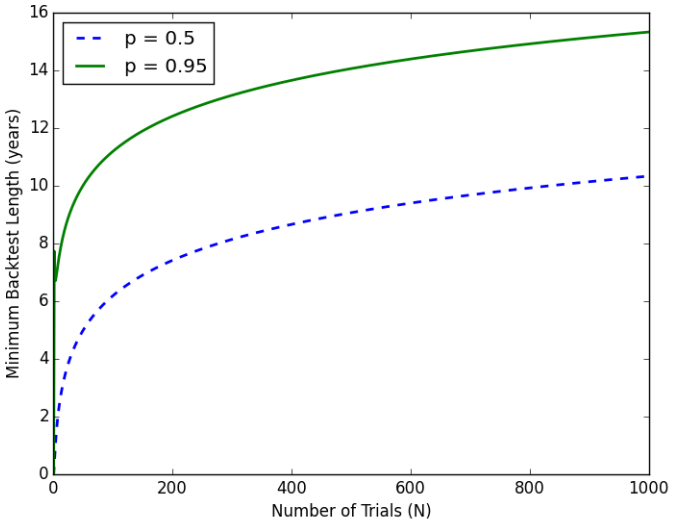
$$Q(p) \approx y^{-\frac{1}{2}}(\alpha - \beta(\ln(-\ln(p)))$$

Example 7.2. *Figure 5 plots various values of N, the number of strategy configurations tested, against the $Q(0.50)$ and $Q(0.95)$ Sharpe ratios of the optimal strategy in-sample for one year worth of market data. For $N = 10$ alternative strategy configurations $Q(0.50) = 1.53$ and $Q(0.95) = 3.82$. So if a researcher obtains a Sharpe ratio higher than 3.82 for one year worth of data and $N = 10$ than he should be 95% confidant that this strategy is not skill-less.*

Let $\overline{Q(p)}$ be a fixed Sharpe ratio. Using the same approach demonstrated in [1] we compute the Minimum Backtest Length (MinBTL), which is the number of years worth of market data needed to avoid selecting a strategy with an in-sample Sharpe ratio of $\overline{Q(p)}$ among N independent strategies with an expected OOS Sharpe ratio of zero:

$$MinBTL_p \approx \left(\frac{\alpha - \beta(\ln(-\ln(p))}{\overline{Q(p)}}\right)^2,$$

Figure 5: minBTL needed to prevent the generation of a skill-less strategy with Sharpe ratio equal to 1 with 0.5 and 0.95 confidence.



8 Missed Discoveries and Error of approximation

The issue with this Sharpe ratio rejection threshold is that while it is designed to prevent false discoveries, it also increases the chance of a missed discovery. This is due to the slow convergence of normed maxima of normally distributed random variables to the Gumbel distribution. The poor approximation towards the right tail of the distribution for small (realistic) sample sizes limits the usefulness of this criterion since most promising in-sample Sharpe ratios will be regarded as skill-less.

The rate of convergence of affinely transformed normally distributed maxima is $1/\log n$ [7], which is very slow convergence. Figure 6 shows the Gumbel distribution function and approximating functions for various values of n. Notice that even for $n = 250$ the Gumbel distribution serves as a poor approximation towards the tails of the distributions. All distributions, however, converge to the Gumbel near the point $x = \frac{1}{2}$, which validates the use of $\mathbb{E}(M_n)$ in [1]. Figure 2 shows the error of approximation and the slow rate of convergence near the tails of the distribution. The slow convergence of the Gumbel to extremes of the normal distribution needs to be taken into account if $Q(p)$ is going to be used to approximate the minimum backtest length for $p \neq 0.50$.

Figure 6: Convergence in distribution of maxima of n normally distributed random variables, M_n, to the Gumbel distribution. M_n converges to the Gumbel near to point $\Phi(x) = 1/2$ for all values of n simulated. This plot exhibits the slow convergence of M_n to the Gumbel (compare $n = 5$ to $n = 250$).

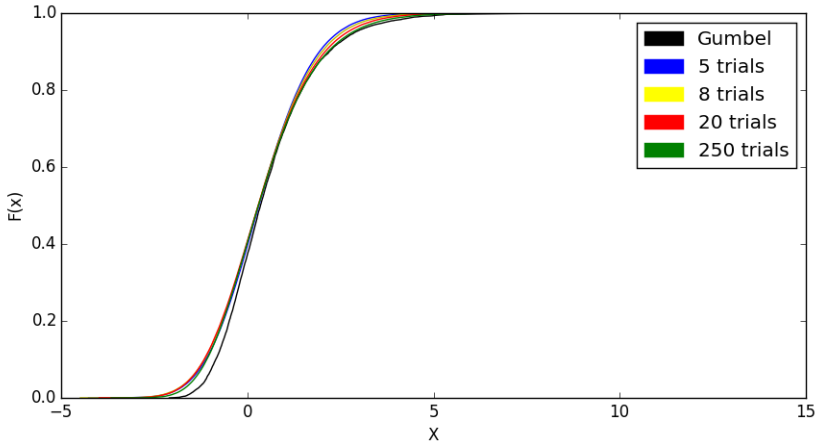
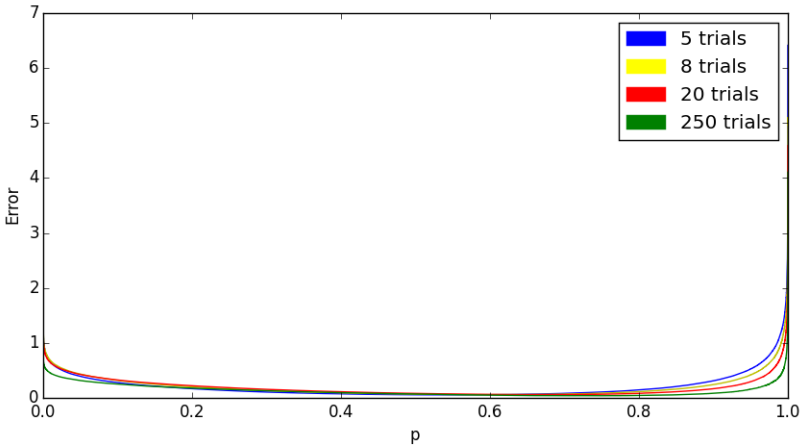


Figure 7: Error of approximation of the Gumbel to the maximum of normally distributed random variables. This figure shows the gap between $M_n(x)$ and $\Lambda(x)$ for each $p \in [0, 1]$. The error is defined as $\Lambda^{-1}(p) - M_n^{-1}(p)$ for all p.



8.1 Fat Tails and the Power-Law Distribution

Up until now we have assumed that the Sharpe ratio is normally distributed, an unrealistic assumption in finance. Mandelbrot addresses this issue in his 1963 seminal paper [13] on stock market prices, and notes that, "the empirical distributions of price changes are usually too "peaked" to be relative to samples from Gaussian populations." A tail, or rare, event is one that occurs with very small probability and is thus near the tail ends of a probability density function. For example, in [14] Nordhaus calculates the monthly returns of the U.S. stock market from the year 1871 to 2010 and notes that the extreme values are much larger than should be expected given normally distributed returns. He notes, for example, that if stock price changes follow a normal distribution then one should expect the see a 5% change in prices once every 14,000 years. For this reason, caution should be taken when using the normal distribution to model financial variables. Empirical data suggests that large deviations from the mean occur more frequently in finance than predicted by the normal distribution [12]. Furthermore, the tail ends of the distribution of market returns exhibits a ratio of returns that is constant up to a scaling factor. That is, the likelihood of a rise in the stock market exceeding 15% can be predicted from the likelihood of a rise exceeding 7.5%, and the same ratio applies to a 10% versus 5% increase. Mathematically speaking,

$$\frac{\mathbb{P}(X < x)}{\mathbb{P}(X < 2x)} = c$$

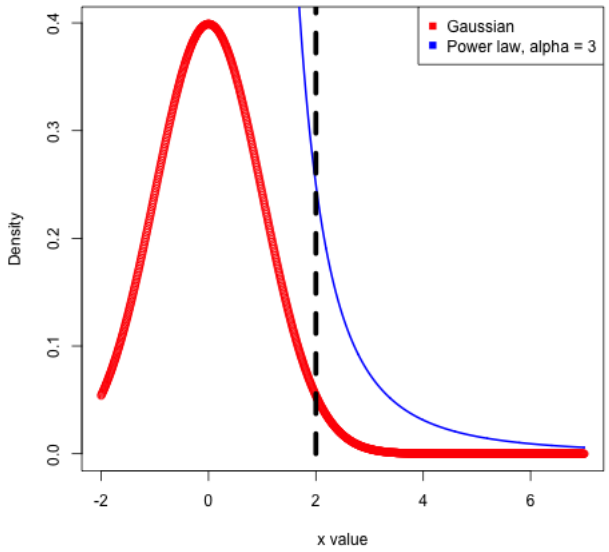
for large enough x and where c is independent of x. Let us consider functions that behave in this manner. Suppose we take a function F such that $\frac{F(2x)}{F(x)} = c$ and let $c' = c/F(2)$. So then if we set $c' = 1$ we can see that,

$$F(2x) = F(2)F(x)$$

which is satisfied by the power law equation $F(x) = \frac{\alpha}{x^\beta}$ where α and β are constants. A power law is a function that is proportional to the power of a variable.

In the context of finance, distributions with power law (fat) tails are more realistic than the Gaussian, or normal distribution, since they have fat tails, and thus the probability of

Figure 8: The probability density function of the Gaussian and power law with $\alpha = 3$.



an extreme event has a non-negligible chance of occurring. The returns of an investment strategy are expected to follow a fat tailed distribution, so the Sharpe ratios should be expected to follow a similar distribution.

Since the calculation of the p-quantile Sharpe ratio relies on the convergence of the limit behavior of extremes of normally distributed random variables, it would be valuable to reproduce this calculation using the limit behavior of a distribution with tails that follow a power law probability distribution. For example, the Pareto distribution is a heavy-tailed distribution that is commonly used to model financial variables.

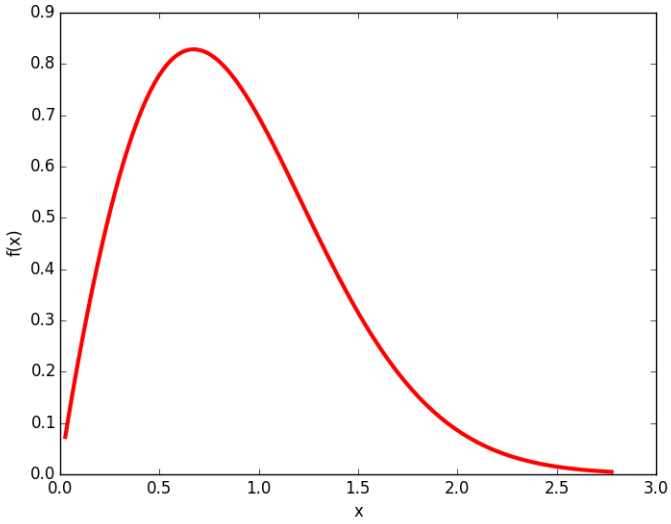
Definition. *The Pareto Distribution can be defined for* $\alpha > 0$ *as*

$$F(x) = \begin{cases} 1 - Kx^{-\alpha} & \text{if } x \geq K^{\frac{1}{\alpha}} \\ 0 & \text{otherwise} \end{cases}$$

for $1 \leq x < \infty$

What follows is a necessary and sufficient condition for a distribution function to be in the domain of attraction of the Fréchet distribution.

Figure 9: The PDF of the standard Fréchet distribution with $\alpha = 1.89281716035$.



Proposition 8.1. *Let* $x_F = \sup\{x|F(x) < 1\}$. *A distribution function* F *belongs to the domain of attraction of the Fréchet distribution if for* $x_F = \infty$,

$$\lim_{t\to\infty} \frac{1 - F(tx)}{1 - F(t)} = x^{-\alpha}$$

For all $x > 0$, *and* $\alpha > 0$.

Proof: Let F be the distribution function of the Fréchet and note that

$$\frac{1 - F(tx)}{1 - F(t)} = \frac{x^{-\alpha}t^{-\alpha}}{t^{\alpha}} = x^{-\alpha}$$

for $x > 0$. Thus, the Pareto distribution belongs to $MDA(\phi_\alpha)$, the maximum domain of attraction of the Fréchet distribution.

So then the quantile function of the Fréchet distribution can be written as

$$Q(p) = F^{-1}(p) = (-\ln p)^{-1/\alpha}.$$

The distribution of scaled maxima of the Pareto converges in distribution to the Fréchet with scaling constant c_n:

$$c_n^{-1} M_n \xrightarrow{d} \phi_\alpha(x),$$

where $c_n = (Kn)^{1/\alpha}$ for $K, \alpha > 0$. Thus, the convergence of distribution of the scaled maxima of the Pareto distribution can be written as,

$$\mathbb{P}((nK)^{-1/\alpha} M_n \leq x) \xrightarrow{d} e^{-x^{-\alpha}}$$

9 Conclusion

Backtesting is an essential tool for identifying profitable trading strategies, or at least avoiding capital loss due to skill-less strategies. The issue is that it becomes very easy to overfit a backtest as the number of configurations of strategies tested becomes unreasonable given the amount of historical data available. Hedge funds and academic journals often report the inflated in-sample backtest results without the number of strategies backtested, making it impossible to infer how a strategy will perform if released into the market. Building off the work of Bailey et al. [1], this paper used results from extreme value theory to develop a minimum backtest length, or the years worth of data necessary to increase an investor's confidence that the performance results of a backtest are not the result of backtest overfitting. We realize the limitations of this criterion given that it relies on the extreme value convergence of normally distributed random variables to the Gumbel distribution, which is very slow. The analysis presented in this paper also made the assumption that the returns of the investment in consideration are normally distributed, an assumption known to be incorrect in practice [13]. This paper concluded on the idea of performing a similar analysis with the assumption that the distribution of the maximum of a collection of Sharpe ratios has fat tails. Since we are particularly concerned with the analysis of the right tail of the distribution of maxima, a power law distribution, such as the Pareto, should be used to model the returns of an investment. The Fréchet distribution should be used to model performance metrics since power law-tailed distributions belong to $MDA(\phi_\alpha)$.

```
import numpy as np
import matplotlib.pyplot as plt
from scipy import stats
from math import exp, log, sqrt, pi, e

def computeScale(n):
    alpha = (2*log(n))**(-1.0/2)
    return alpha
def computeShift(n):
    beta = (sqrt(2*log(n)) - (log(4*pi) + log(log(n)))/(2*sqrt(2*log(n))))
    return beta

def inverseGumbel(prob, alpha, beta):
    return alpha - beta*(log(-log(prob)))

def minBTL(expSharpe, estSharpe):
    return (estSharpe/expSharpe)**2

if __name__ == '__main__':
    EM_constant = 0.5772156649
    prob_1 = 0.5
    prob_2 = 0.995
    expSharpe = 1

    t = np.arange(0, 1000)
    s_1 = []
    s_2 = []

    for i in range (0, 1000):
        if i < 8:
            s_1.append(0)
            s_2.append(0)
            continue

        a = computeShift(i)
        b = computeScale(i)

        s_1.append(minBTL(expSharpe, inverseGumbel(prob_1, a, b)))
        s_2.append(minBTL(expSharpe, inverseGumbel(prob_2, a, b)))

    plt.plot(t, s_1, lw = 2, label = 'p = 0.5', ls = '--')
    plt.plot(t, s_2, lw = 2, label = 'p = 0.95')
    plt.xlabel('Number of Trials (N)')
    plt.ylabel('Minimum Backtest Length (years)')
    plt.legend(loc=2)
    plt.show()
```

References

[1] D.H. Bailey, J.M. Borwein, M.P. Lopez, Q.J. Zhu, "Pseudo-Mathematics and Financial Charlatanism: The Effects of Backtest Overfitting on Out-of-Sample Performance", (April 1, 2014). Notices of the American Mathematical Society, 61(5), May 2014, pp.458-471; available from http://dx.doi.org/10.2139/ssrn.2308659.

[2] D.H. Bailey, M.P. Lopez, "The Deflated Sharpe Ratio: Correcting for Selection Bias, Backtest Overfitting and Non-Normality", (July 31, 2014). Journal of Portfolio Management, 40 (5), pp. 94-107. 2014 (40th Anniversary Special Issue). Available at SSRN: http://ssrn.com/abstract=246055.

[3] A.A. Balkema, L. De. Haan, "On R. Von Mises' Condition for the Domain of Attraction of exp(ex)1", Ann. Math. Statist. 43 (1972), no. 4, 1352–1354. doi:10.1214/aoms/1177692489. http://projecteuclid.org/euclid.aoms/1177692489.

[4] A. Bovier, "Extreme Values of Random Processes, Lecture Notes", Institut fur Angewandte Mathematik.

[5] P. Embrechts, C. Klueppelberg, T. Mikosch, Modeling Extremal Events - for Insurance and Finance, Springer-Verlag, New York, 2003.

[6] M. Haas, C Pigorsch, "Financial Economics, Fat-Tailed Distributions" Springer-Verlag, New York, (September 21, 2007).

[7] P. Hall, "On the Rate of Convergence of Normal Extremes" Journal of Applied Probability Vol. 16, No. 2 (Jun., 1979) , pp. 433-439 Published by: Applied Probability Trust Stable URL: http://www.jstor.org/stable/3212912.

[8] C.R. Harvey, Y. Liu, "Backtesting", (October 4, 2014). Available at SSRN: http://ssrn.com/abstract=2345489 or http://dx.doi.org/10.2139/ssrn.2345489.

[9] C.R. Harvey, Y. Liu, "Evaluating Trading Strategies", (August 25, 2014); available from http://ssrn.com/abstract=2474755orhttp://dx.doi.org/10.2139/ssrn.2474755.

[10] R. Jacobs, "The Dangerous Mathematical Con of Hedge Funds and Financial Advisors", Pacific Standard Magazine, April 28, 2014.

[11] A.W. Lo, "The Statistics of Sharpe Ratios", (July/August 2002). Financial Analysts Journal, Vol. 58, No. 4; available from http://ssrn.com/abstract=377260.

[12] B. Mandelbrot, N.N. Taleb, "How the Finance Gurus get risk all Wrong", Fortune Magazine, July 11, 2005.

[13] B. Mandelbrot, "The Variation of Certain Speculative Prices", The Journal of Business, Vol. 36, No. 4 (Oct., 1963), pp. 394-419, Stable URL: http://www.jstor.org/stable/2350970.

[14] W.D. Nordhaus, "The Economics of Tail Events with an Application to Climate Change", Review of Environmental Economics and Policy, volume 5, issue 2, summer 2011, pp. 240257 doi:10.1093/reep/rer004.

[15] S.I. Resnick, Extreme Values, Regular Variation, and Point Processes, Springer-Verlag, New York 1987.

[16] L. Wasserman, All of Statistics, A Concise Course in Statistical Inference, Springer 2014.

[17] J. Zweig, "Huge Returns at Low Risk? Not So Fast", The Wall Street Journal, June 27, 2014.