Inference-Based Performance Evaluation Using the Sortino Ratio

Olivier Ledoit [1] and Michael Wolf[1,2]

[1]University of Zurich

[2]ADIA Lab

Outline

1. Motivation
2. The (Generalized) Sortino Ratio
3. Difference of Two Sortino Ratios
4. Single Sortino Ratio
5. Multiple Testing
6. Empirical Applications
   - Difference of Two Ratios
   - Multiple Testing
7. Conclusions

Outline

1. Motivation
2. The (Generalized) Sortino Ratio
3. Difference of Two Sortino Ratios
4. Single Sortino Ratio
5. Multiple Testing
6. Empirical Applications
   - Difference of Two Ratios
   - Multiple Testing
7. Conclusions

Motivation

The Sharpe ratio remains the dominant measure of risk-adjusted performance of investment strategies.

However, in settings with skewed returns, alternative measures can be more appropriate.

Such measures still generally lack a well-developed and widely accepted inferential framework.

We aim to address this gap for the Sortino ratio:

"Give the people what they want!"

Outline

1. Motivation
2. The (Generalized) Sortino Ratio
3. Difference of Two Sortino Ratios
4. Single Sortino Ratio
5. Multiple Testing
6. Empirical Applications
   - Difference of Two Ratios
   - Multiple Testing
7. Conclusions

Measures of Downside Risk

The variance of a return r is a measure of total risk:

$$\mathrm{Var}(r) := \mathbb{E}\Big[(r-\mu)^2\Big]$$

whose root is called the standard deviation.

The earliest measure of downside risk is the semi-variance:

$$\mathrm{SemiVar}(r) := \mathbb{E}\Big[(\max(0,\mu-r))^2\Big]$$

whose root is called the semi-deviation.

More generally, lower partial moment of order p:

$$\mathrm{LPM}(\tau,p) := \mathbb{E}\Big[(\max(0,\tau-r))^p\Big]$$

for a threshold τ and a power $p \geq 1$.

Long-forgotten Interest

Markowitz (1959) stated:

> *In Chapter XIII, the standard deviation is compared with other measures of risk and variability. [. . .] One of the measures considered, the semi-deviation, produces efficient portfolios somewhat preferable to those of the standard deviation. Those produced by the standard deviation are satisfactory, however, and the standard deviation itself is easier to use, more familiar to many, and perhaps easier to interpret than the semi-deviation.*

Generalized Sortino Ratio

Generalized Sortino Ratio:

$$GenSo_p(\tau) := \frac{\mathbb{E}(r-\tau)}{\mathrm{LPM}(\tau,p)^{1/p}}$$

Special cases:

- Sortino ratio: $So(\tau) = GenSo_2(\tau)$
- Gains-loss ratio: $GLR(\tau) = GenSo_1(\tau) - 1$

To see the latter, with $x := r - \tau$:

$$GLR(\tau) := \frac{\mathbb{E}(x^+)}{\mathbb{E}(x^-)} = \frac{\mathbb{E}(\max(0,x))}{\mathbb{E}(\max(0,-x))} = \frac{\mathbb{E}(x)}{\mathbb{E}(\max(0,-x))} - 1$$

Implications

For symmetric return distributions, Sharpe and Sortino ratio agree:

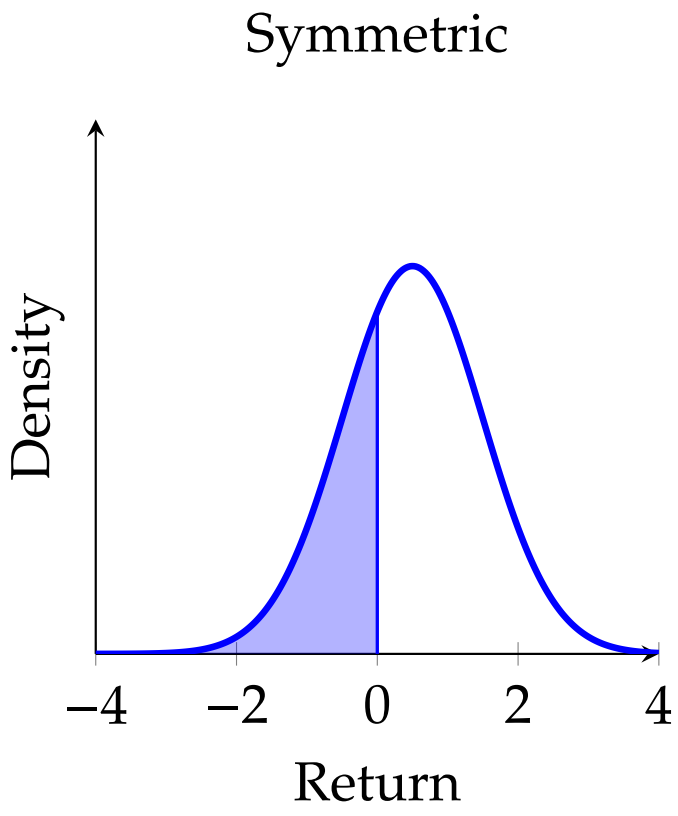
- If two distributions have the same mean,
  the ratio selects the one with the smaller variance.

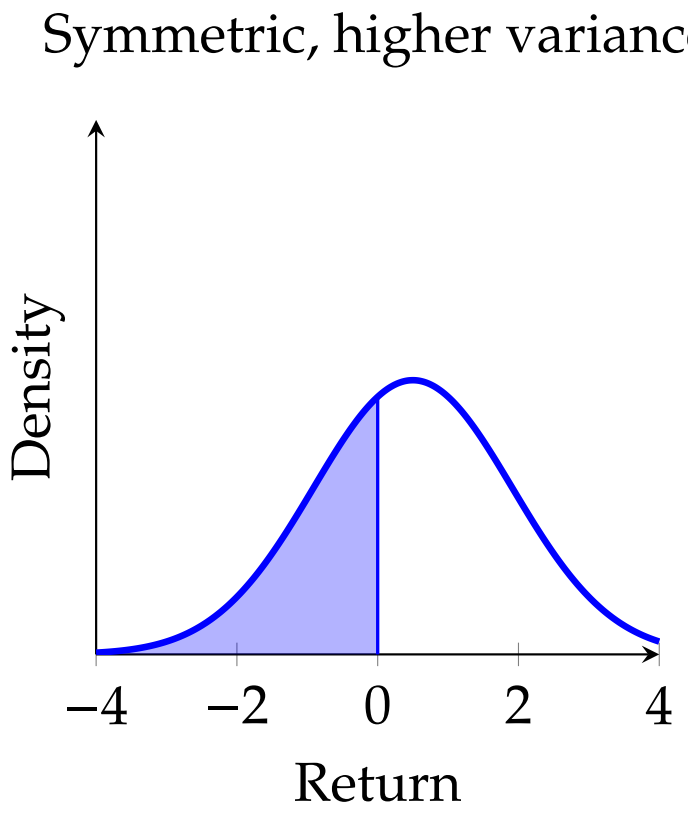
For skewed return distributions, this is no longer the case,
since asymmetry affects downside risk differently from total risk.

Implications:

- Right-skewed distributions favor the Sortino ratio
- Left-skewed distributions favor the Sharpe ratio

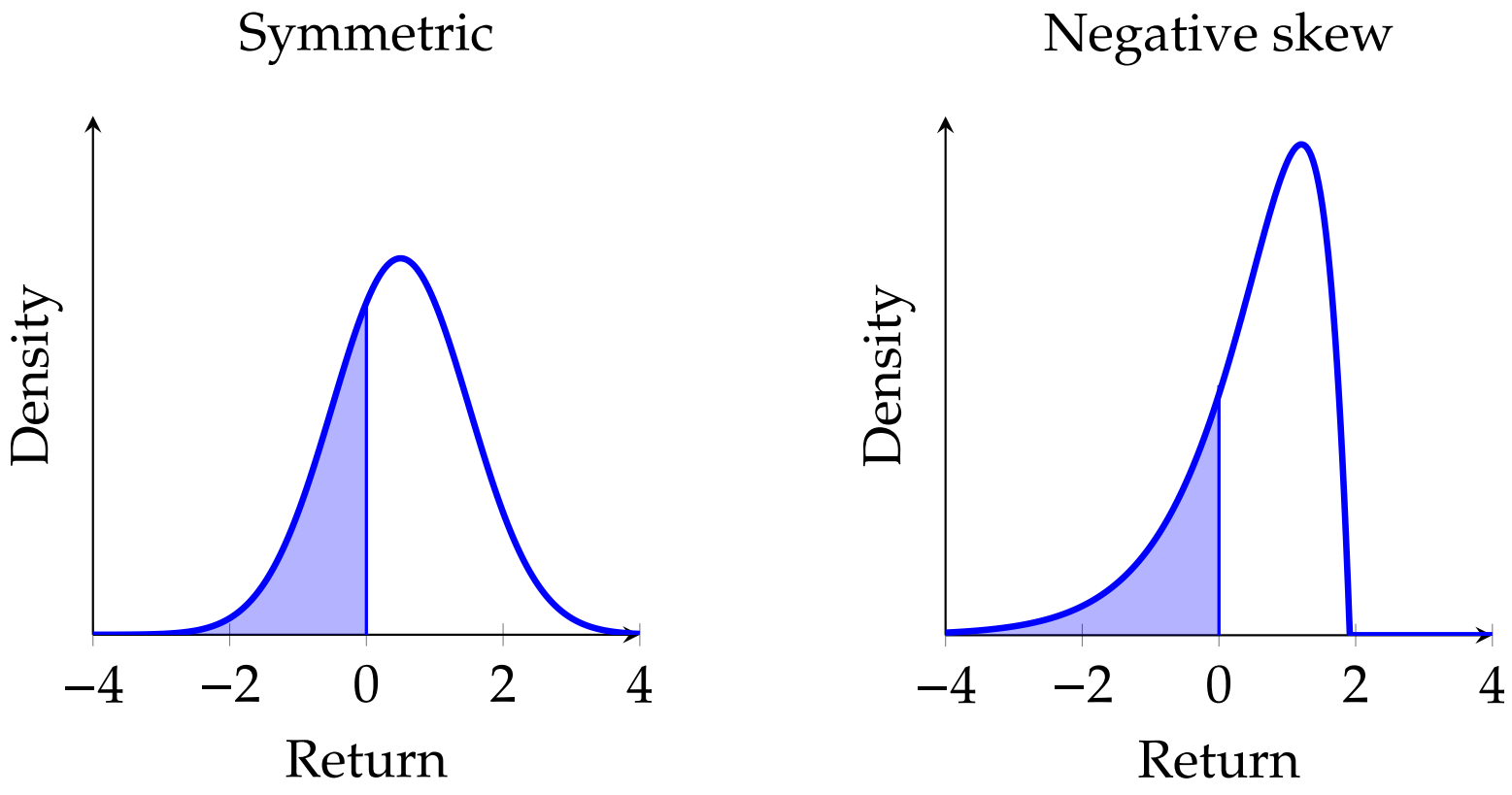
Different Sharpe and Sortino Ratios





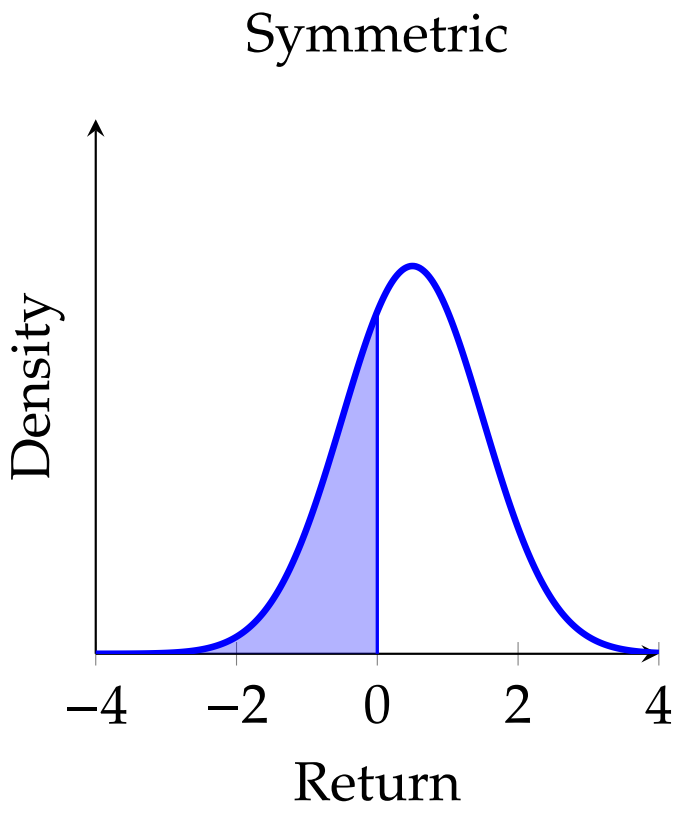
- Sharpe ratios: 0.5 vs. 0.35
- Sortino ratios: 1.09 vs. 0.67

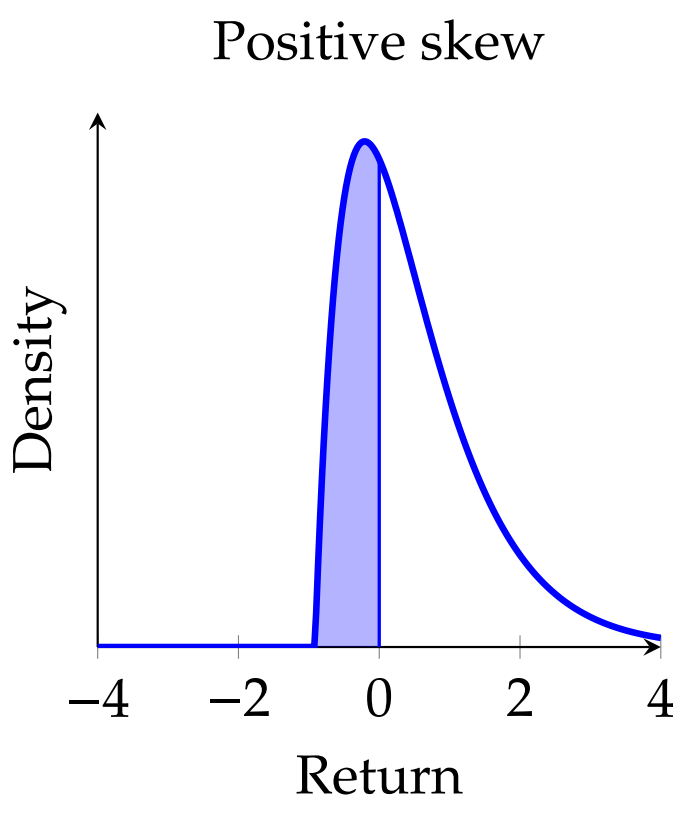
Same Sharpe Ratio, Different Sortino Ratios



- Common Sharpe ratio: 0.5
- Sortino ratios: 1.09 vs. 0.83

Same Sharpe Ratio, Different Sortino Ratios





- Common Sharpe ratio: 0.5
- Sortino ratios: 1.09 vs. 1.87

Outline

1. Motivation
2. The (Generalized) Sortino Ratio
3. Difference of Two Sortino Ratios
4. Single Sortino Ratio
5. Multiple Testing
6. Empirical Applications
   - Difference of Two Ratios
   - Multiple Testing
7. Conclusions

General Setup & Notation

We use the same notation as Ledoit and Wolf (2008, JEF):

- There are two investment strategies i and n
- Their excess returns are r_{ti} and r_{tn}, for $t = 1, \ldots, T$
- Bivariate return series is assumed stationary

Strategy i: for a threshold τ_i and a power $p \geq 1$, define:

$$e_{ti} := r_{ti} - \tau_i \quad \text{and} \quad h_{ti} := \max(0, -e_{ti})^2$$

with means

$$\mu_i := \mathbb{E}(e_{ti}) \quad \text{and} \quad \gamma_i := \mathbb{E}(h_{ti})$$

Do the same for strategy n.

Then, the parameter of interest is

$$\Delta := \frac{\mu_i}{\sqrt{\gamma_i}} - \frac{\mu_n}{\sqrt{\gamma_n}} \quad \text{with} \quad \hat{\Delta} := \frac{\hat{\mu}_i}{\sqrt{\hat{\gamma}_i}} - \frac{\hat{\mu}_n}{\sqrt{\hat{\gamma}_n}}$$

(The estimators of μ and γ are the corresponding sample means.)

Smooth Function of Means

Define $v := (\mu_i, \mu_n, \gamma_i, \gamma_n)'$ and $\hat{v} := (\hat{\mu}_i, \hat{\mu}_n, \hat{\gamma}_i, \hat{\gamma}_n)'$.

Then: $\Delta = f(v)$ and $\hat{\Delta} = f(\hat{v})$ with

$$f(a, b, c, d) := \frac{a}{\sqrt{c}} - \frac{b}{\sqrt{d}}$$

The gradient of $f(\cdot)$ is given by

$$\nabla f(a, b, c, d) := \left(\frac{1}{c^{1/2}}, -\frac{1}{d^{1/2}}, -\frac{a}{2c^{3/2}}, \frac{b}{2d^{3/2}} \right)'$$

HAC Inference

Under moment and weak-dependence conditions:

$$\sqrt{T}(\hat{v} - v) \xrightarrow{d} N(0, \Psi)$$

for some symmetric and positive-definite matrix Ψ.

Get a consistent estimator $\hat{\Psi}$ via standard HAC methodology.

Basis for Inference:

$$\frac{\hat{\Delta} - \Delta}{s(\hat{\Delta})} \dot{\sim} N(0,1) \quad \text{with} \quad s(\hat{\Delta}) := \sqrt{\frac{\nabla f(\hat{v})' \hat{\Psi} \nabla f(\hat{v})}{T}}$$

Throughout, we use the prewhitened Parzen kernel with automatic choice of bandwidth of Andrews and Monahan (1992, ECMA).

Bootstrap Inference

In a nutshell:

- Generate bootstrap data: $\{(r^*_{ti}, r^*_{tn})\}_{t=1}^{T}$
- Compute estimator from bootstrap data: $\hat{\Delta}^*$
- Compute corresponding standard error: $s(\hat{\Delta}^*)$

Basis for Inference:

$$\frac{\hat{\Delta} - \Delta}{s(\hat{\Delta})} \overset{\cdot}{\sim} \frac{\hat{\Delta}^* - \hat{\Delta}}{s(\hat{\Delta}^*)}$$

Motivation:

- More reliable than HAC inference in small samples under strong dependence of return data

Relevant Details

Use the circular block bootstrap to generate bootstrap data; resample pairs of data to account for cross-sectional dependence.

The "natural" bootstrap standard error makes use of the particular dependence structure (independent blocks of block size b).

Have calibration method to pick a data-dependent block size.

Outline

1. Motivation
2. The (Generalized) Sortino Ratio
3. Difference of Two Sortino Ratios
4. Single Sortino Ratio
5. Multiple Testing
6. Empirical Applications
   - Difference of Two Ratios
   - Multiple Testing
7. Conclusions

General Setup & Notation

Simplify to a single strategy:

- There is one investment strategy
- Its excess returns are r_t, for $t = 1, \ldots, T$
- Return series is assumed stationary

For a threshold τ and a power $p \geq 1$, define:

$$e_t := r_t - \tau \quad \text{and} \quad h_t := \max(0, -e_t)^2$$

with means

$$\mu := \mathbb{E}(e_t) \quad \text{and} \quad \gamma := \mathbb{E}(h_t)$$

Then, the parameter of interest is

$$So := \frac{\mu}{\sqrt{\gamma}} \quad \text{with} \quad \widehat{So} := \frac{\hat{\mu}}{\sqrt{\hat{\gamma}}}$$

Smooth Function of Means

Define $v := (\mu, \gamma)'$ and $\hat{v} := (\hat{\mu}, \hat{\gamma})'$.

Then: $So = f(v)$ and $\widehat{So} = f(\hat{v})$ with

$$f(a, b) := \frac{a}{\sqrt{b}}$$

The gradient of $f(\cdot)$ is given by

$$\nabla f(a, b) := \left(\frac{1}{b^{1/2}}, -\frac{a}{2b^{3/2}} \right)'$$

HAC Inference

Under moment and weak-dependence conditions:

$$\sqrt{T}(\hat{v} - v) \xrightarrow{d} N(0, \Psi)$$

for some symmetric and positive-definite matrix Ψ.

Get a consistent estimator $\hat{\Psi}$ via standard HAC methodology.

Basis for Inference:

$$\frac{\widehat{So} - So}{s(\widehat{So})} \dot{\sim} N(0,1) \quad \text{with} \quad s(\widehat{So}) := \sqrt{\frac{\nabla f(\hat{v})' \hat{\Psi} \nabla f(\hat{v})}{T}}$$

Bootstrap Inference

In a nutshell:

- Generate bootstrap data: $\{r_t^*\}_{t=1}^T$
- Compute estimator from bootstrap data: $\widehat{So}^*$
- Compute corresponding standard error: $s(\widehat{So}^*)$

Basis for Inference:

$$\frac{\widehat{So} - So}{s(\widehat{So})} \overset{.}{\sim} \frac{\widehat{So^*} - \widehat{So}}{s(\widehat{So^*})}$$

Motivation:

- More reliable than HAC inference in small samples under strong dependence of return data

Outline

1. Motivation
2. The (Generalized) Sortino Ratio
3. Difference of Two Sortino Ratios
4. Single Sortino Ratio
5. Multiple Testing
6. Empirical Applications
   - Difference of Two Ratios
   - Multiple Testing
7. Conclusions

General Setup & Notation

Data:

- Observe $T \times (S+1)$ excess return matrix
- Last column is a common benchmark
- Multivariate return series is assumed stationary

Let $\Delta_s := So_s - So_{S+1}$, for $s = 1, \ldots, S$.

Interested in the multiple testing problem

$$H_s : \Delta_s = 0 \quad \text{vs} \quad H_s' : \Delta_s \neq 0 \,, \qquad \text{for } s = 1, \ldots, S$$

Can consider many variations of the theme.

High-level Methodology

Compute adjusted p-values following Romano and Wolf (2016, SPL).

Only need to specify how to compute:

- Individual test statistics t_s
- Null bootstrap test statistics t_s^*

This bootstrap method is of stepwise nature and accounts for dependence of individual test statistics t_s.

It is more powerful than the Holm method, which is also of stepwise nature but does not account for such dependence.

Relevant Details

Individual test statistics:

$$t_s := \frac{|\hat{\Delta}_s|}{s(\hat{\Delta}_s)}$$

Use the circular block bootstrap to generate bootstrap data.
Resample rows of data to account for cross-sectional dependence.

Null bootstrap test statistics:

$$t_s^* := \frac{|\hat{\Delta}_s^* - \hat{\Delta}_s|}{s(\hat{\Delta}_s^*)}$$

Have calibration method to pick a data-dependent block size.

Outline

1. Motivation
2. The (Generalized) Sortino Ratio
3. Difference of Two Sortino Ratios
4. Single Sortino Ratio
5. Multiple Testing
6. Empirical Applications
   - Difference of Two Ratios
   - Multiple Testing
7. Conclusions

Outline

1. Motivation
2. The (Generalized) Sortino Ratio
3. Difference of Two Sortino Ratios
4. Single Sortino Ratio
5. Multiple Testing
6. Empirical Applications
   - Difference of Two Ratios
   - Multiple Testing
7. Conclusions

Put-writing Strategy vs. Factor Portfolio

Strategy 1:

- CBOE S&P 500 PutWrite Index (PUT)
- Sells one-month, at-the-money SPX put options
- Subtract risk-free rate
- Pronounced negative skewness

Strategy 2:

- "Value and Momentum Everywhere" portfolios (by AQR)
- VAL2JP: mid-value Japanese equities
- Already reported in excess of the risk-free rate
- Mild positive skewness

Data:

- Monthly returns, 01/1996–12/2025 ($T = 360$)

Put-writing Strategy vs. Factor Portfolio

Strategy	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\gamma}$	$\widehat{Sh}$	$\widehat{So}$	*skew*
CBOE PUT	0.559	3.011	2.260	0.186	0.247	−1.706
VAL2JP	0.255	4.469	3.013	0.057	0.085	0.023

Table: Summary statistics for monthly excess returns

H_0	HAC	Boot
$\Delta_{\text{Sh}} = 0$	0.041	0.044
$\Delta_{\text{So}} = 0$	0.094	0.095

Table: *p*-values for testing equality of Sharpe and Sortino ratios

Stock Index vs. Alternative Investments

Strategy 1:

- S&P 500 Total Return Index (SP500-TR)
- Subtract risk-free rate
- Moderate negative skewness

Strategy 2:

- 80% Gold and 20% Bitcoin (GOLD-BTC)
- Subtract risk-free rate
- Moderate positive skewness

Data:

- Monthly returns, 10/2014–12/2025 ($T = 135$)

Stock Index vs. Alternative Investments

Strategy	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\gamma}$	$\widehat{Sh}$	$\widehat{So}$	*skew*
SP500-TR	1.003	4.270	2.721	0.235	0.369	−0.408
GOLD-BTC	1.883	5.375	2.385	0.350	0.790	0.513

Table: Summary statistics for monthly excess returns

H_0	HAC	Boot
$\Delta_{\text{Sh}} = 0$	0.193	0.186
$\Delta_{\text{So}} = 0$	0.078	0.074

Table: p-values for testing equality of Sharpe and Sortino ratios

Outline

1. Motivation
2. The (Generalized) Sortino Ratio
3. Difference of Two Sortino Ratios
4. Single Sortino Ratio
5. Multiple Testing
6. Empirical Applications
   - Difference of Two Ratios
   - Multiple Testing
7. Conclusions

ETFs vs. SPY

Strategies:

- Consider 40 ETFs
- Common benchmark is SPY
- Subtract risk-free rate
- Moderate negative to moderate positive skewness

Asset classes:

- Equities: 21
- Fixed income: 8
- Alternative assets: 12

Data:

- Monthly returns, 12/2007–12/2025 ($T = 216$)

ETFs vs. SPY

How many ETFs have a ratio significantly different from SPY?

Significance level is $\alpha = 5\%$.

Sharpe ratio:

- Marginal p-values: 24
- Holm adjustment: 5
- Bootstrap adjustment: 6

Sortino ratio:

- Marginal p-values: 22
- Holm adjustment: 3
- Bootstrap adjustment: 8

Note:

- In all significant cases, $\hat{\Delta} < 0$

Outline

1. Motivation
2. The (Generalized) Sortino Ratio
3. Difference of Two Sortino Ratios
4. Single Sortino Ratio
5. Multiple Testing
6. Empirical Applications
   - Difference of Two Ratios
   - Multiple Testing
7. Conclusions

Conclusions

The Sortino ratio is widely used in practice and has attracted growing interest in the academic literature.

By focusing on downside risk, it can lead to different conclusions than the Sharpe ratio when return distributions are skewed.

Proper performance evaluation requires statistical inference, not just sample-based point estimates.

We provide a toolbox of inference methods for:

- Difference of two Sortino ratios
- Single Sortino ratio
- Multiple testing

References

Andrews, D. W. K. and Monahan, J. C. (1992). An improved heteroskedasticity and autocorrelation consistent covariance matrix estimator. *Econometrica*, 60(4):953–966.

Ledoit, O. and Wolf, M. (2008). Robust performance hypothesis testing with the Sharpe ratio. *Journal of Empirical Finance*, 15:850–859.

Markowitz, H. M. (1959). *Portfolio Selection: Efficient Diversification of Investments*. Yale University Press, New Haven. Cowles Foundation Monograph 16.

Romano, J. P. and Wolf, M. (2016). Efficient computation of adjusted p-values for resampling-based stepdown multiple testing. *Statistics & Probability Letters*, 113:38–40.