David Florysiak, VTAD Meeting, 14. September 2022

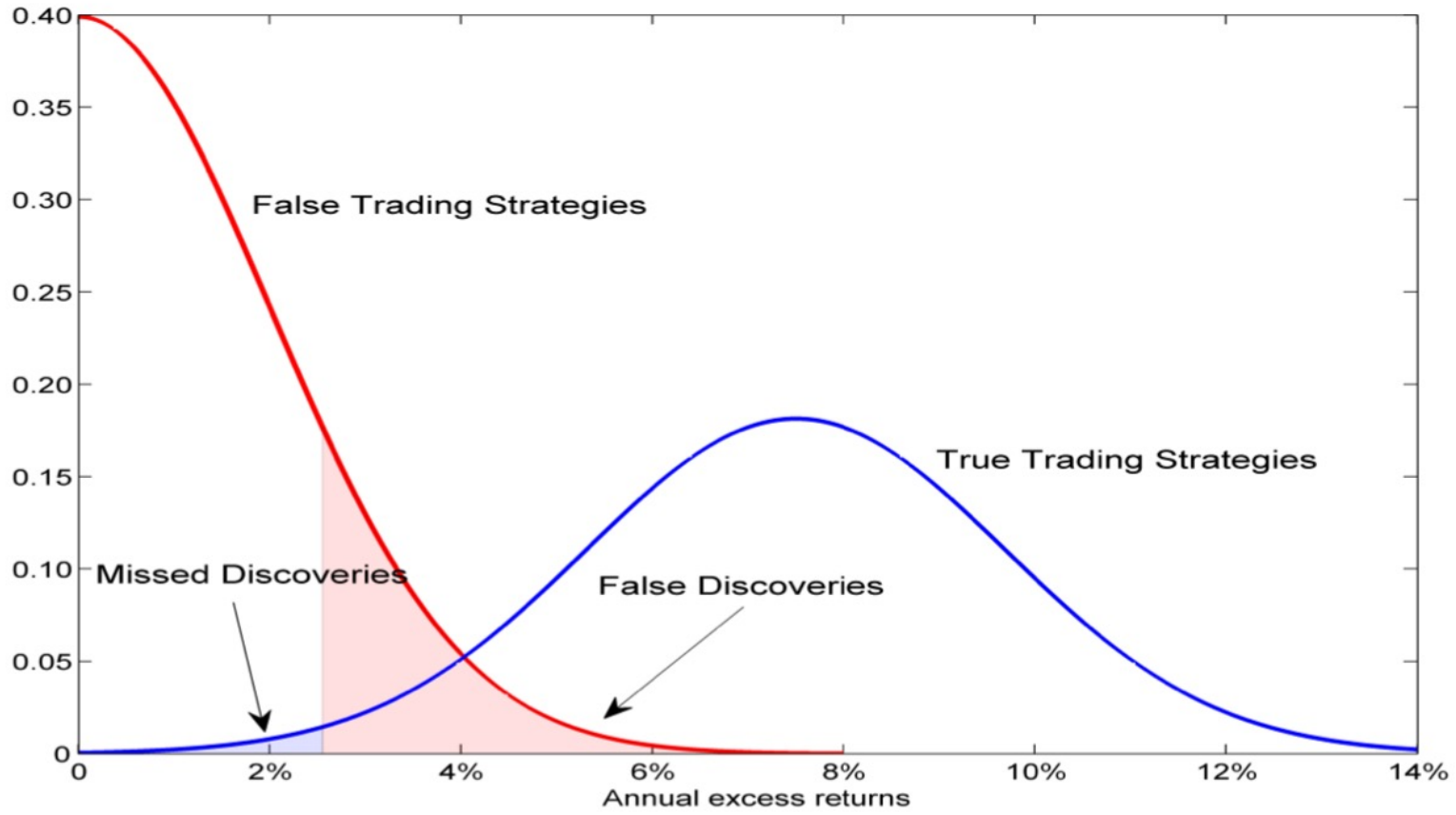
Backtesting von Trading Strategien

Motivation

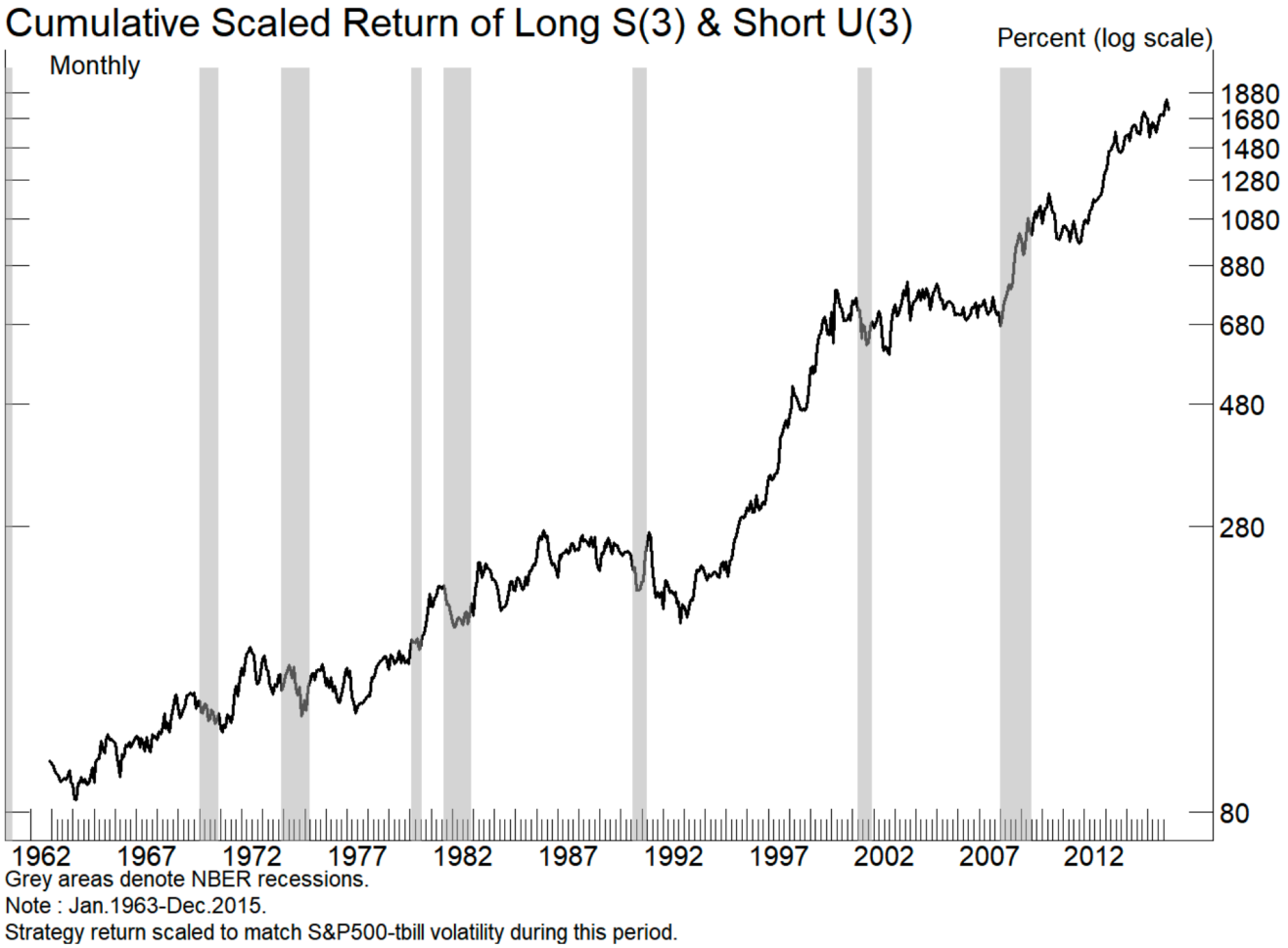
- Diskussion zu Backtesting auf dem Börsentag München

- Zu meiner Person:
  - https://www.linkedin.com/in/florysiak/
  - https://www.florysiak.com

Finding Profitable Trading Strategies



Finding Profitable Trading Strategies



[Arnott, Harvey, and Markowitz (2018)]

Sharpe Ratio and T-Statistic

- Candidate strategy has an annualized Sharpe ratio of 0.92 and t-statistic of 2.91.

$$\widehat{SR} = \frac{\hat{\mu}}{\hat{\sigma}}, \qquad t\text{-ratio} = \frac{\hat{\mu}}{\hat{\sigma}/\sqrt{T}}.$$

- The observed profitability is about three standard deviations (t-distribution, T = 10 years) from the null hypothesis of zero profitability.

$$\begin{aligned} p^S &= Pr(|r| > t\text{-ratio}) \\ &= Pr(|r| > \widehat{SR} \cdot \sqrt{T}) \end{aligned}$$

p = (1 - t (2.91, 10*12-1)) * 2 = 0.004

- This means that the chance that our trading strategy is a false discovery (Type I error) is 0.4%.

Multiple Testing

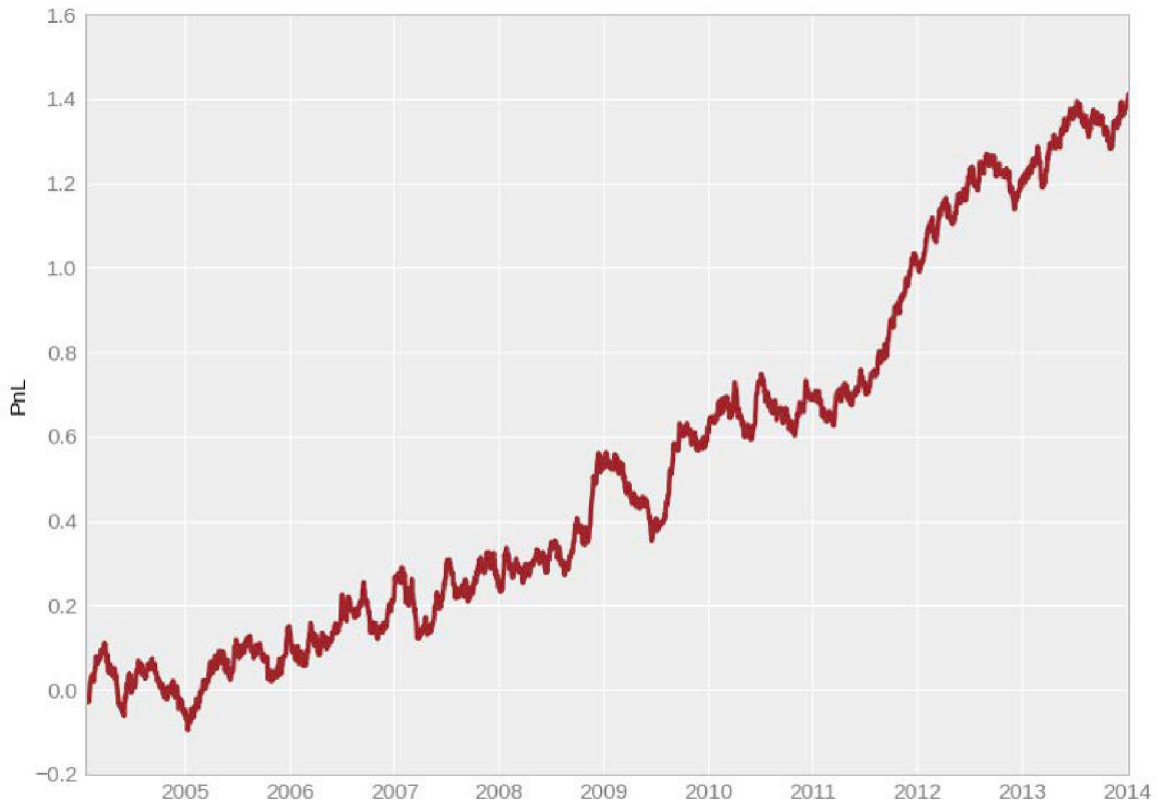


Exhibit 1: A candidate trading strategy

Source: Harvey, C. R., & Liu, Y. (2014). Evaluating Trading Strategies. https://doi.org/10.2139/ssrn.2474755

Multiple Testing

- If you test (“draw”) one strategy, what is the probability to find a profitable strategy if the significance level alpha = 0.05?
- If you test (“draw”) 10 strategies, what is the probability to find a profitable strategy if the significance level alpha = 0.05?

$$\begin{aligned} p^S &= Pr(|r| > t\text{-ratio}) \\ &= Pr(|r| > \widehat{SR} \cdot \sqrt{T}) \end{aligned}$$

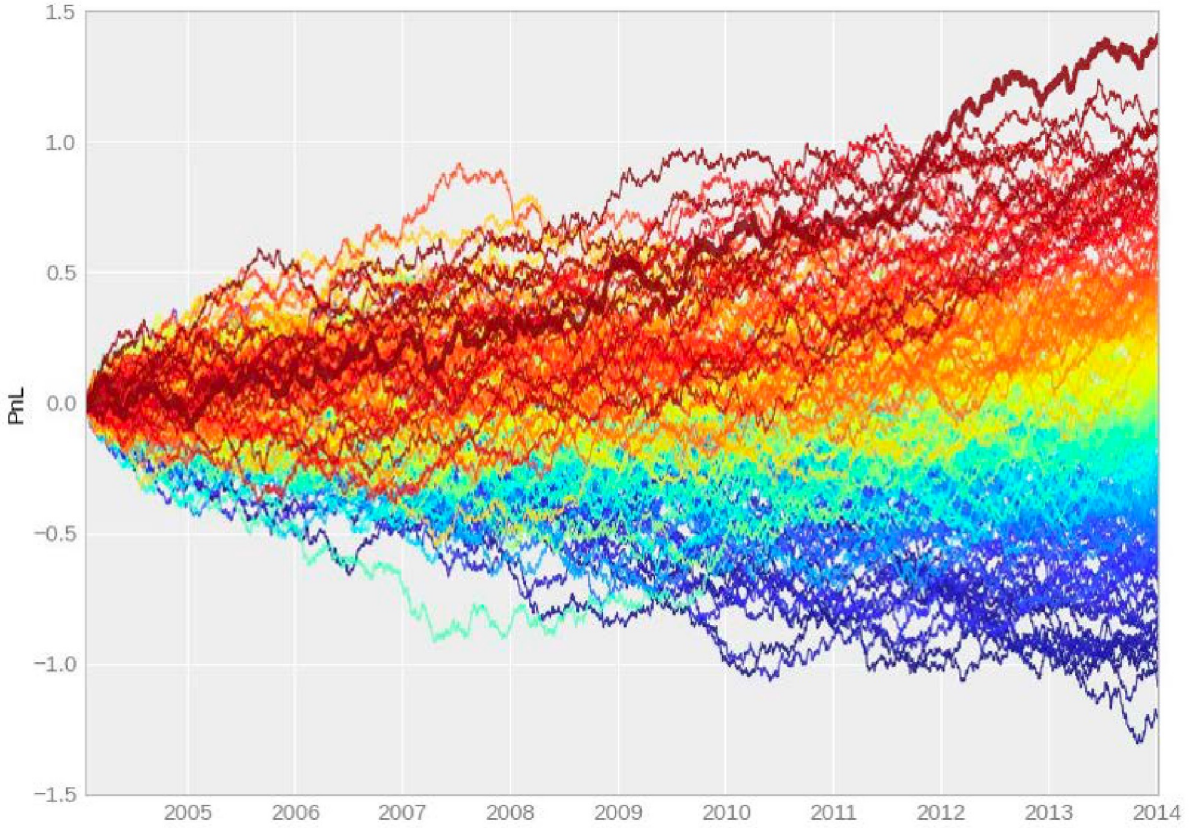


Exhibit 2. 200 randomly generated trading strategies

$$\begin{aligned} p^M &= Pr(\max\{|r_i|, i = 1, \ldots, N\} > t\text{-ratio}) \\ &= 1 - \prod_{i=1}^{N} Pr(|r_i| \leq t\text{-ratio}) \\ &= 1 - (1 - p^S)^N. \end{aligned}$$

Solution: “Haircut”/Adjusted Sharpe Ratio

$$p^M = Pr(|r| > \widehat{HSR} \cdot \sqrt{T})$$

- Assume a Sharpe ratio of 0.75 and a p-value of 0.0008 for a single test.
- When N = 200, p^M = 0.15.
- The adjusted Sharpe ratio is HSR = 0.32.
- Hence, multiple testing with 200 tests reduces the original Sharpe ratio by approximately 60% (=(0.75-0.32)/0.75).

More Solutions: Adjusting p-values

- Multiple Testing Framework (e.g. family-wise error rate and false-discovery rate corrections): Bonferroni; Holm; Benjamini, Hochberg and Yekutieli (BHY)

- Multiple testing methods are designed to limit incorrectly "discovering" a profitable trading strategy.

- The simplest one

$$Bonferroni: \qquad p_{(i)}^{Bonferroni} = \min[Mp_{(i)}, 1], i = 1, \ldots, M.$$

- If M = 200, $p^S = 0.004$ => $p^{Bonferroni} = 0.80$.

More Solutions

- Multiple testing and cross-validation
  - To quantify the degreee of backtest overfitting, Lopéz de Prado (2013) propose the calculation of the probability of backtest overfitting (PBO) that measures the relative performance of a particular backtest among a basket of strategies using cross-validation techniques.

- In-sample multiple testing vs. out-of-sample validation

Limitations of Using the Sharpe Ratio as a Measure of Performance

- SR not normally distributed.
  - Two trading strategies might have identical Sharpe Ratios but one of them might be preferred because it has less severe downside risk.
- Current collection of strategies.
  - For example, a strategy with a lower Sharpe might be preferred because the strategy is relatively uncorrelated with current strategies in production.
- Further measures (e.g. information ratio, VaR, …) and corresponding adjustments for multiple testing.
- In sample vs. out of sample testing.

Exhibit 2. Seven-Point Protocol for Research in Quantitative Finance

1. **Research Motivation**
   a) Does the model have a solid economic foundation?
   b) Did the economic foundation or hypothesis exist *before* the research was conducted?

2. **Multiple Testing and Statistical Methods**
   a) Did the researcher keep track of all models and variables that were tried (both successful and unsuccessful) and are the researchers aware of the multiple-testing issue?
   b) Is there a full accounting of all possible interaction variables if interaction variables are used?
   c) Did the researchers investigate all variables set out in the research agenda or did they cut the research as soon as they found a good model?

3. **Data and Sample Choice**
   a) Do the data chosen for examination make sense? And, if other data are available, does it make sense to exclude these data?
   b) Did the researchers take steps to ensure the integrity of the data?
   c) Do the data transformations, such as scaling, make sense? Were they selected in advance? And are the results robust to minor changes in these transformations?
   d) If outliers are excluded, are the exclusion rules reasonable?
   e) If the data are winsorized, was there a good reason to do it? Was the winsorization rule chosen before the research was started? Was only one winsorization rule tried (as opposed to many)?

4. **Cross-Validation**
   a) Are the researchers aware that true out-of-sample tests are only possible in live trading?
   b) Are steps in place to eliminate the risk of out-of-sample "iterations" (i.e., an in-sample model that is later modified to fit out-of-sample data)?
   c) Is the out-of-sample analysis representative of live trading? For example, are trading costs and data revisions taken into account?

5. **Model Dynamics**
   a) Is the model resilient to structural change and have the researchers taken steps to minimize the overfitting of the model dynamics?
   b) Does the analysis take into account the risk/likelihood of overcrowding in live trading?
   c) Do researchers take steps to minimize the tweaking of a live model?

6. **Complexity**
   a) Does the model avoid the curse of dimensionality?
   b) Have the researchers taken steps to produce the simplest practicable model specification?
   c) Is an attempt made to interpret the predictions of the machine learning model rather than using it as a black box?

7. **Research Culture**
   a) Does the research culture reward quality of the science rather than finding the winning strategy?
   b) Do the researchers and management understand that most tests will fail?
   c) Are expectations clear (that researchers should seek the truth not just something that works) when research is delegated?

Multiple Testing: Conclusion

"Most of the empirical research in finance, whether published in academic journals or put into production as an active trading strategy by an investment manager, is likely false. Second, this implies that half the financial products (promising outperformance) that companies are selling to clients are false."

"It is also clear that investment managers want to promote products that are most likely to outperform in the future. That is, there is a strong incentive to get the testing right. No one wants to disappoint a client and no one wants to lose their bonus – or their job. Employing the statistical tools of multiple testing in the evaluation of trading strategies reduces the number of false discoveries."

Source: Harvey, C. R., & Liu, Y. (2013). Backtesting. https://doi.org/10.2139/ssrn.2345489