Backtesting Review - Cheat Sheet.

By GoshawkTrades, Founder of Unbiased Trading.

▼ Drawdowns

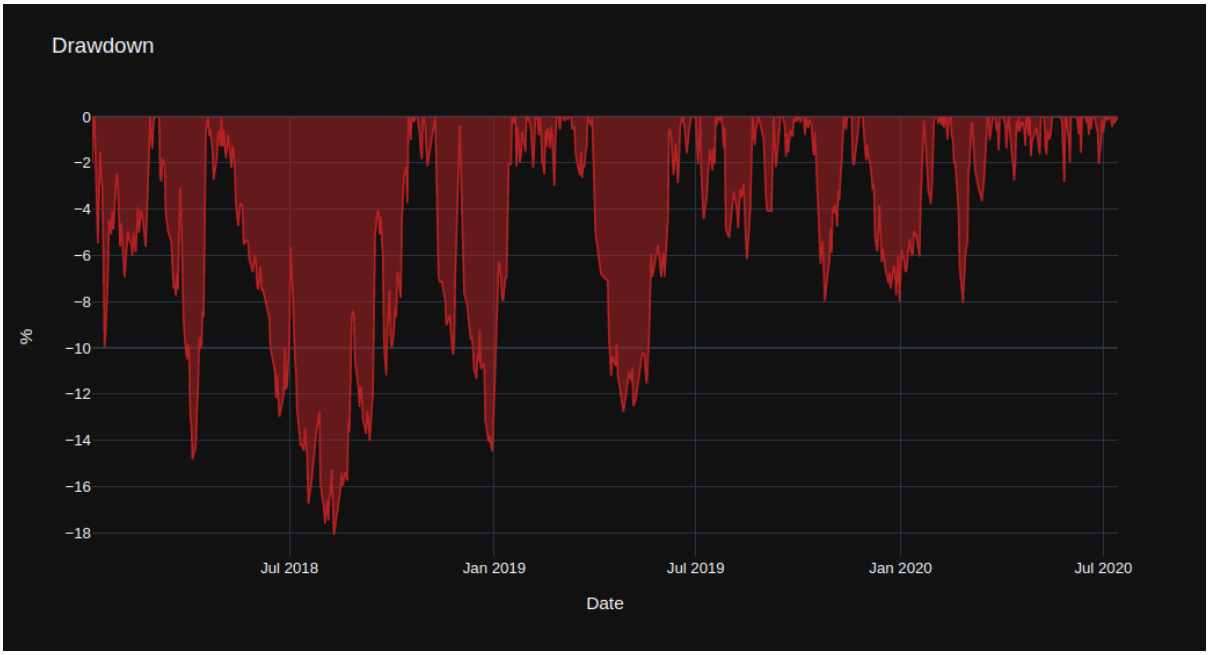
Drawdowns occur with every strategy and serve as one of the most significant metrics for evaluation, often surpassing the returns metric.

Formula:

$$\text{Maximum Drawdown (\%)} = \frac{\text{Peak Value} - \text{Trough Value}}{\text{Peak Value}}$$

- Drawdowns help determine how much capital is needed to run a strategy, but margin/leverage is also needed on top depending on the market and strategy.
- Look out for excessively long drawdowns as they require more resilience to not interfere with the strategy whether its execution or parameters or stopping the algorithm.
- How long does your average drawdown last? What was the longest? These metrics will help you understand when your strategy is potentially no longer working and keep you confident in it during a drawdown.
- Keep an eye out for recent drawdowns in the strategy backtest. This could be a good opportunity to take a strategy live if the backtest has been conducted properly. In general, it's still better to take a strategy live as soon as possible after thorough backtesting rather than trying to time it, but starting in a drawdown can be to your advantage sometimes.

- The rule of thumb is to normally 1.5x or 2x the Monte Carlos max drawdown for capital requirements, however, a max drawdown on a singular strategy can still be a good starting place.
- A good drawdown is usually below 15%, but it can increase if returns are substantial. However, risk management, such as sizing, must then be essential.
- A terrible drawdown is anything over 50%, but it's also important to consider it in relation to returns. For instance, a 0.25% drawdown for 5% return isn't bad.



▼ Year On Year Performance

Absolute return isn't always the full story, to have a realistic expectation of your strategy you need to gauge yearly performance.

Formula:

$$CAGR = \left(\frac{Final\ Value}{Starting\ Value}\right)^{\frac{1}{N}} - 1$$

- For most strategies to be worthwhile, they need to beat the risk-free rate. This rate typically ranges from 7% to 9% annually, based on the SPY 500 or you can use other market benchmarks like BTC or ETH.
- The annualized rate of return is not the best way to measure yearly returns. Instead, use CAGR (Compound Annual Growth Rate) as it includes compounding, leading to more accurate results.
- Yearly returns provide a clearer picture of your expected returns and allow you to analyze if one year made up for the majority of the returns, which could indicate inconsistent performance.
- The CAGR is a great metric for comparing strategies as for example, a strategy that yielded 20% over 10 years will have the same absolute return as a strategy that yielded 20% in 1 year making CAGR a more accurate metric compared to absolute return.
- A strong compound annual growth rate (CAGR) would exceed that of the SPY 500. The only exception would be if the strategy considerably minimized drawdowns and volatility. This creates a less risky strategy that can benefit select traders and individuals.

▼ Win Rate

Win Rate is often overlooked as a simple metric, but it often tells a bigger story about your potential experience trading the strategy.

Formula:

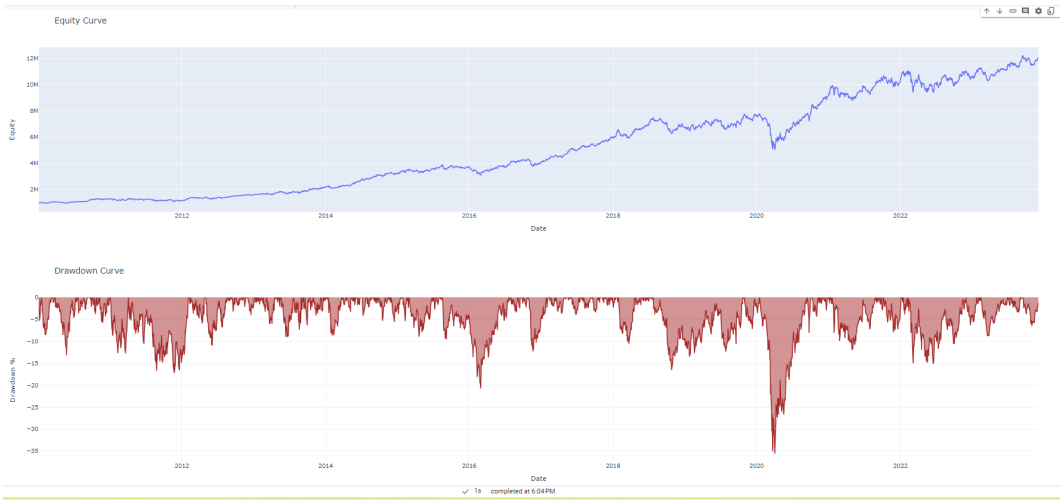
- Higher win rate:
  - Easier compounding, plus you can use bigger size on any trade as there is a higher chance of being profitable.
  - Shorter drawdown periods as losses are recouped faster.
  - More conviction and confidence due to more time winning, meaning less resilience is needed to execute well.
- Lower win rate:
  - Longer drawdown periods due to higher amounts of consecutive losses as well as losing trades.
  - More reliant on taking every single trade, as missing one big winning trade can highly impact performance.
  - More exponential returns capability (with no compounding), due to bigger wins.
  - More resilience is needed to withstand drawdowns, depleting more mental capital.
- There is no good or bad win rate; the risk to reward just needs to be high enough to compensate for the win rate for the strategy to remain profitable. However, both approaches have their advantages and disadvantages, and you must align with what best suits your trading personality.

▼ Equity Curve

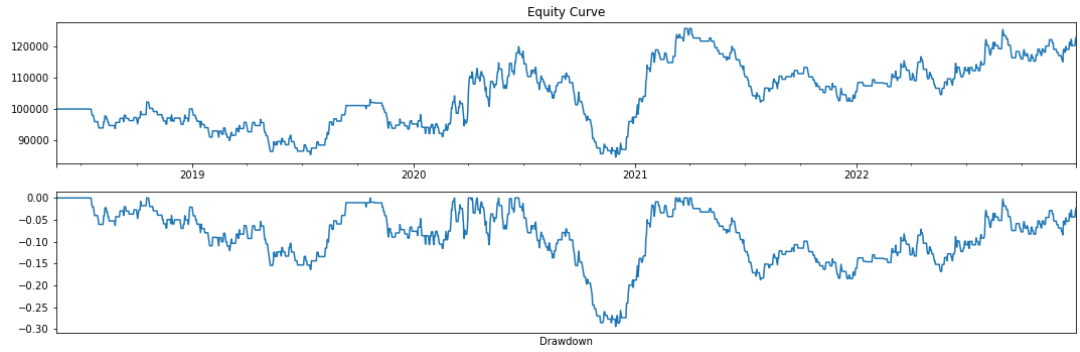
Equity curves represent a standard backtest output. However, many traders fail to extract its full information, which can reveal more than just

performance metrics.

Example:

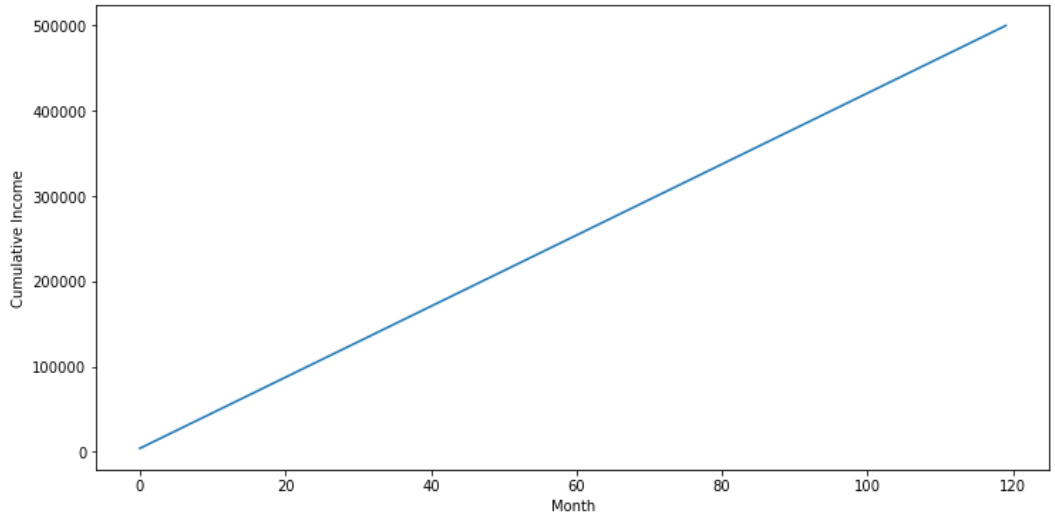


- You are ideally looking for a stable and consistent equity curve with no huge drawdowns or high volatility.
    - What you don’t want:

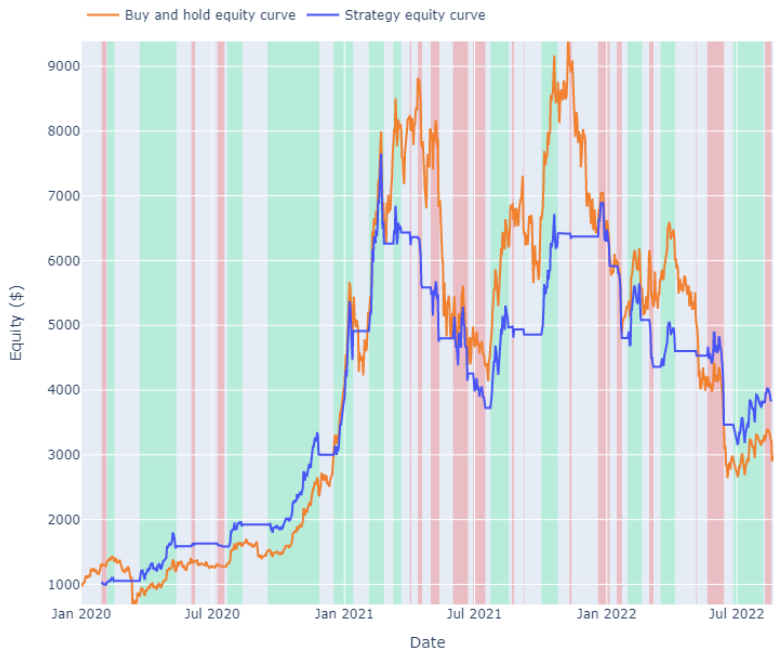


- Profitability should not be sporadic or only work in certain years, as this can indicate a 'lucky' market environment or timing resulting in the majority of the returns. Keep in mind, if a strategy was working in 2018 but in 2020 the returns went exponentially higher, this is a good sign as it shows it produces even more profits in a higher opportunity environment.
- Equity curves are also one of the easiest ways to spot overfit or potential code errors in a backtest.

- If your results look anything like this (no drawdowns of any kind or volatility) this usually indicates an issue:



- Equity curves plotted using days often don't show the full Intra-trade drawdown picture. To compensate for this, be sure to plot a daily and trade-based equity curve to see the actual volatility you may experience on a trade-by-trade basis.
- Plotting a buy-and-hold curve on your equity curve is a quick and easy way to compare your strategy and see how it reacts to the general market. It can also help identify potential drawdown correlations.



▼ Parameters

We all start developing strategies thinking we need hundreds of parameters to find a unique edge, but the profits lie in simplicity.

- Too many parameters lead to overfitting (occurs when a model becomes too closely tailored to historical data) making it unprofitable on live (general) data.
- Too few parameters lead to underfitting causing underperformance making the strategy less worthwhile to trade.
- In most cases, a good parameter count typically ranges from 2 to 6, although this may vary by market. This conclusion is based on my own testing and wider discussion with other quants.
- Double-check all parameters and their calculations to make sure none are using data unavailable at the time of the historical trade (meaning the results are useless and inaccurate), this is a common shortfall when using highs and lows of candles yet to be formed.
- Use only parameters that significantly impact the strategy's results, such as returns or drawdowns. Adding more parameters increases the risk of overfitting, so use them sparingly.

- Understanding parameters is important for fully evaluating and accurately interpreting backtest results. To research this further, look up the ‘black box theory’.

▼ Amount Of Trades

The more trades you have in your backtest, the more confidence you can have in your strategy. This means that your strategy's performance is a true reflection of its effectiveness, rather than just a result of chance.

- The more trades you can have in your backtest, the more statistically significant your results will be.
- You ideally want a minimum of 200 samples however this can depend on the market, timeframe, etc.
- Prioritizing statistical significance over returns/performance is normally a good practice.
- Testing your strategy over a longer historical period allows you to analyze its performance across various market conditions, providing a more comprehensive assessment of its robustness due to a higher sample count.
- If you can't increase your trade count, make sure to run a forward test for a considerable amount of time to check if the results remain steady on the live market.

▼ Consecutive Losses

Without statistics, if you stumble upon a string of losing trades, that is when you start to doubt your strategy the most.

- It is crucial to know the average amount of losing trades you can encounter for realistic expectations.
- You need to know the highest number of losing trades you could encounter using a large historical backtest, as well as a Monte Carlo test.
    - This gives you perspective on when a strategy might no longer be working if it breaches the maximum number of losing trades in a row.
- Knowing your consecutive loss statistics gives you mental resilience when going through it in live performance because you were able to expect it

beforehand.

- Consecutive losses can also help gauge your capital requirements by calculating your risk per trade and then multiplying it by the highest number of losing trades in a row.

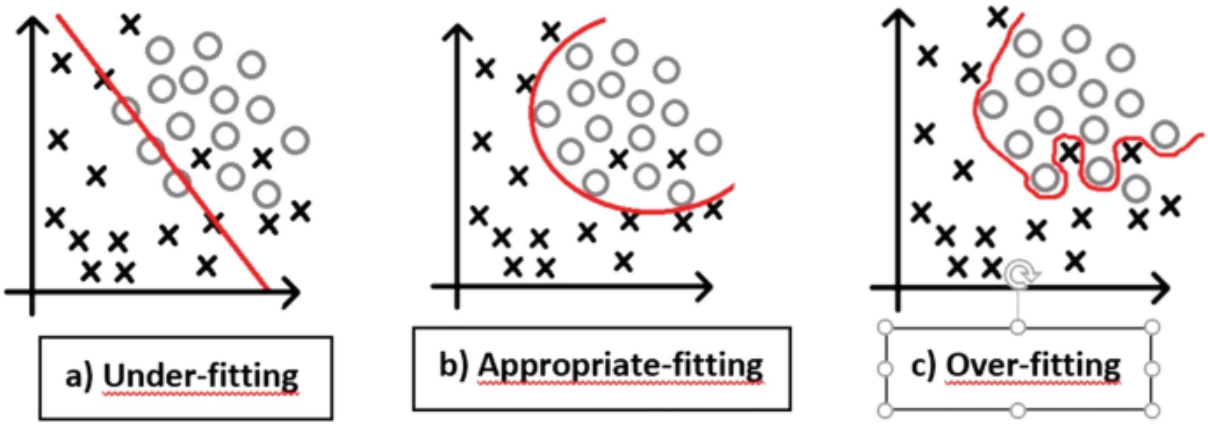
	Probability of X consecutive losing trades within a 100 trade sequence									
Win rate	1	2	3	4	5	6	7	8	9	10
5%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
10%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
15%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
20%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
25%	100%	100%	100%	100%	100%	100%	100%	100%	100%	99%
30%	100%	100%	100%	100%	100%	100%	100%	100%	98%	93%
35%	100%	100%	100%	100%	100%	100%	99%	95%	85%	71%
40%	100%	100%	100%	100%	100%	99%	93%	79%	61%	42%
45%	100%	100%	100%	100%	99%	93%	76%	54%	35%	21%
50%	100%	100%	100%	100%	95%	78%	52%	31%	16%	9%
55%	100%	100%	100%	98%	83%	55%	30%	14%	7%	3%
60%	100%	100%	100%	92%	63%	32%	14%	6%	2%	1%
65%	100%	100%	99%	77%	40%	16%	6%	2%	1%	0%
70%	100%	100%	93%	55%	21%	7%	2%	1%	0%	0%
75%	100%	100%	79%	32%	9%	2%	1%	0%	0%	0%
80%	100%	98%	54%	14%	3%	1%	0%	0%	0%	0%
85%	100%	89%	28%	5%	1%	0%	0%	0%	0%	0%
90%	100%	63%	9%	1%	0%	0%	0%	0%	0%	0%
95%	99%	22%	1%	0%	0%	0%	0%	0%	0%	0%

▼ Amount of Iterations

Overfitting is every backtester's nightmare, and the number of iterations has a strong correlation to how likely it will become your reality.

- Overfitting occurs when you fit your parameters too closely to the historical data, leaving no room for generalization and making it very sensitive to minor changes in the data and prone to failure in the future.
    - The perfect recipe to do this is by running hundreds of iterations/optimizations to find the best results.
- In my trading, I try not to exceed 2 to 3 optimizations for any strategy, treating iterations like a Mario life on a level. Otherwise, there is a high risk of overfitting to the historical data.
- Don't strive for the best backtesting results, but for ones you believe could survive the live markets.

- This applies to anything optimization related, including walk-forward optimizations, which are often treated as robustness tests but can actually lead to overfitting.



▼ Manual or Coded

"Manual backtests" are a great way to get started, but they're hard to rely on and trust. Code is the new standard in backtesting for all retail traders.

- Manual backtests have no barrier to entry and can be done in a notepad.
  - But, they are extremely time-intensive and not scalable enough to achieve statistical significance.
  - Can take time away from what is most important in trading, which is screen time or creating strategy ideas.
- Manual backtests can be filled with inaccuracies and biases due to tiredness during data entry or ignorance of biases affecting historical testing.
  - Biases: Look-ahead Bias - Survivorship Bias - Neglecting Market Impacts - Analysis errors - Not accounting for costs.
- Coded backtests have a higher barrier to entry but are the ultimate tool for testing and building conviction in a strategy.
  - Provides full scalability to run backtests as far back as 15 years, resulting in ample data and samples for statistical significance.
  - The ability to run more complex tests, such as parameter sensitivity, walk-forward optimizations, and much more, is available.

    - 99% accuracy is due to a computer handling historical executions.
    - There are no human biases occurring such as changing individual trades during the backtest.
    - The ability to automate after.
    - The ability to add new data / expand to new market testing with very low time requirement to achieve.
- I will always prefer a coded backtest to a manual one, and I would recommend always doing thorough backtests using code instead of doing it manually.

▼ Realistic Simulations

Every trade automatically loses you money, but most backtests don't account for this, leaving you with profitable historical equity curves and losing live strategies.

- A backtest should account for three main costs:
    - Commissions
    - Slippage
    - Locates / Other
- To account for commissions, use your expected or current broker's commission structure pricing in your backtests.
- Slippage is difficult to measure accurately, but here are two easy methods:
    - Applying a standard amount of slippage on each trade that you expect based on live testing and volatility.
    - Applying a variable amount of slippage at random on each trade. For example, you could have $0.05 to $1 of slippage, and the range is applied at random on each trade.
    - Advanced methods:

The Almgren-Chriss (A-C) framework:

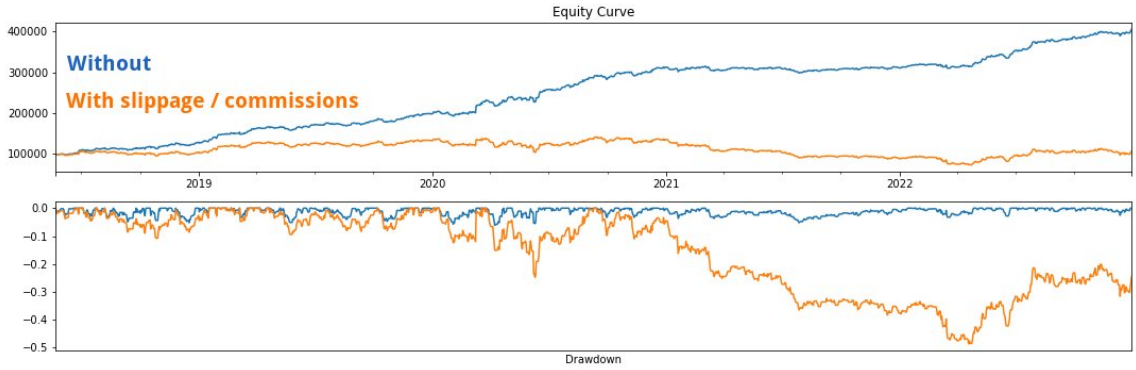
- **Volatility (Vol):** A higher volatility environment generally leads to higher slippage, as executing large orders becomes more challenging due to wider bid-ask spreads.
- **Participation Rate (Participation):** This reflects the proportion of the Average Daily Volume (ADV) your order represents. A higher participation rate indicates a larger order, which can incur greater slippage as it takes more time and effort to execute without significantly impacting the market price.

The A-C formula incorporates these factors in a way that is:

- **Linear in Vol:** Slippage increases proportionally with volatility.
- **Square Root in Participation:** Slippage increases with the square root of the participation rate.
- A great article on it: https://medium.com/@anboto_labs/deep-dive-into-is-the-almgren-chriss-framework-be45a1bde831
  (Suren Markosov (Anboto Labs), 'Deep Dive into IS: The Almgren-Chriss Framework', April 12 2024)

You can also simulate slippage by using Impact models, Liquidity based models, and machine learning techniques.

- Sometimes, there are additional costs associated with trading, such as locates, gas (crypto), and others. These costs can surprise traders when they implement their strategy live. Therefore, it is important to always conduct a forward test using a small size on a live account to identify these mistakes and errors.
- This guide is aimed at newer traders. However, if you are trading with a large size, be sure to account for liquidity and float to ensure that the live market can handle your size without causing huge slippage and profitability issues. This is something we often help with in Unbiased Trading for larger sized traders.



▼ Backtest Length

In order to ensure that a backtest is statistically significant, it is important to have a sufficient amount of historical data to test the strategy for a high enough sample count as well as various market conditions.

Length:

- **Strategy Type:**
  - **Short-term (Intraday):** 3 to 5 years of data is sufficient.
  - **Long-term (Swing/Positional):** Ideally, 5-10 years of data is better to capture broader market cycles.
  - **High-frequency:** This highly depends but 1 year of data would be enough if you are having 30+ samples a day.
- **Sample Size:** Aim for at least 200 to 1000 samples for statistical significance.

Other Considerations:

- **Market Conditions:** Include periods of bull, bear, and volatile markets to assess strategy robustness.
  - **Market Regimes:** Consider segmenting the backtest results by distinct market regimes (e.g., bull vs bear markets) to see how the strategy performs in different environments, this is just a potential point of analysis.

- **Survivorship Bias:** Use data including delisted securities to avoid overestimating returns.

▼ Sharpe Ratio

The Sharpe Ratio is a metric used to assess the risk-adjusted performance of an investment. It compares the average return of an investment (like a portfolio or strategy) to the risk-free rate of return (typically government bonds) relative to the investment's volatility (standard deviation of returns).

- **Formula:** Sharpe Ratio = (Return on Investment - Risk-Free Rate) / Standard Deviation of Investment Returns
- **Interpretation:** A higher Sharpe Ratio indicates better risk-adjusted performance. You're getting more return for each unit of risk taken.

Good vs. Bad Sharpe Ratio:

- **Generally good:** Above 0.75
- **Very good:** Between 1 and 2
- **Excellent:** Above 2
- **Negative:** Less than 0 (investment underperforms risk-free rate)
- Suspicious: Above 3 (Once above 3 it is more likely the strategy has a bias such as look-ahead bias)

Important Considerations:

- **Limitations:**
  - Assumes normal distribution of returns (not always true in reality).
  - Doesn't capture the potential for extreme events ("Black Swan" events).
- **Nassim Taleb's Critique:** High Sharpe Ratios might indicate strategies prone to sudden blowups. Volatility can be a sign of resilience, not just risk.

Sharpe Ratio for Different Strategies:

- **Long-Term Strategies (3+ years):** Focus on Sharpe Ratio as a good risk-adjusted performance indicator.

- **Short-Term Strategies (Intraday/HFT):** Consider other metrics alongside Sharpe Ratio due to potential limitations for short timeframes.

▼ Advanced Techniques

While traditional backtesting provides valuable insights, advanced techniques can offer a more nuanced understanding of your strategy's robustness and potential pitfalls. Here's an introduction into them:

1. Monte Carlo Simulations:

This technique simulates thousands of possible equity curves based on the historical trades of your backtest. By running your strategy on these simulations, you can assess its performance distribution (how often it performs well, poorly, etc.) and gain insights into potential risks under various trade sequencing.

- **Benefits:**
  - Provides a broader picture of strategy performance beyond the original backtest.
  - Quantifies the uncertainty and potential range of future returns and drawdown.
- **Considerations:**
  - It can't account for bear markets, recessions, or any other kind of financial crisis that might impact potential results.
  - The outcomes from running a Monte Carlo simulation may vary due to statistical factors. These simulations offer approximations and chances, rather than exact forecasts. As more simulations are conducted, the precision of the approximations tends to enhance, yet there will persist a certain level of unpredictability in the results.
  - It must be coded for the highest accuracy and flexibility to apply to any strategy.

2. Walk-Forward Optimization:

This approach involves progressively splitting your historical data into training (In-sample) and validation (Out of sample) sets. The strategy is optimized on the training data and then tested on the validation data. This process is repeated by moving the validation window forward, mimicking a real-world scenario.

- **Benefits:**
  - Reduces the risk of overfitting to historical data.
  - Tests the strategy's ability to adapt to changing market conditions.
- **Considerations:**
  - Requires more data compared to traditional backtesting.
  - The choice of window size can impact results.

3. Stress Testing:

This involves testing your strategy under extreme market conditions not seen in the original historical backtest. This could involve simulating historical crashes, sudden volatility spikes, or liquidity droughts.

- **Benefits:**
  - Identifies potential weaknesses in the strategy during extreme events.
  - Helps assess how the strategy would handle unexpected market disruptions and any potential position sizing or risk changes required.
- **Considerations:**
  - Requires additional assumptions about the nature of extreme events.

4. Machine Learning Integration:

Machine learning algorithms can be used to identify patterns in historical data and potentially improve backtesting accuracy. These algorithms can be trained on historical data and then used to predict future market movements or optimize trading strategies.

- **Benefits:**

    - Can potentially identify complex patterns missed by traditional methods.
    - Offers a data-driven approach to strategy optimization.
    - Can be incredibly good at strategy portfolio rebalancing and optimization.
- **Considerations:**
    - Requires more technical skills to accomplish.
    - Risk of overfitting the model to historical data.

If you would like to perform these advanced techniques to improve the robustness of your trading performance,

Book a 30-minute call to discuss how our quant developers can help backtest and automate your trading:

https://www.unbiasedtrading.info/