CHARTA LEARNING · LESSON NOTES

Backtesting a strategy: aims, stages, metrics and flaws

Portfolio Management

1 A story: the shop that never tried the recipe

- Mira's idea: buy the cheapest shares each month
- Question: would it have worked over ten years?
- A backtest runs the rules on past data
- Like a small test batch before opening the shop

IN THE LESSON

- Meet Mira Castellan, an analyst at a fictional firm called Larkspur Asset Management.
- She has an idea.
- Buy the cheapest shares, measured by price compared with company book value, hold them for a month, then do it again.
- Before the committee puts money behind it, someone must ask a fair question.
- Would this idea have worked, had we used it in real time over the last ten years?
- Think of a baker who tests a new recipe on a small batch before opening a shop.
- A backtest is that small batch.
- It is cheap, fast and honest, as long as it is done carefully.

2 What a backtest aims to achieve

- Estimate return and risk
- Compare with a benchmark and alternatives
- See when and why it works
- Measure drawdowns, turnover and sensitivity

IN THE LESSON

- A backtest applies a rules based strategy to past data as if it had been run in real time.
- The word rules matters.
- Every decision must be written down so a computer can follow it, with no hindsight and no gut feeling.
- What does the exercise aim to do?
- It estimates the return and the risk the strategy would have had.
- It compares the result with a benchmark and with other ideas.
- It helps us understand when and why the strategy works, and when it does not.
- It measures drawdowns, which are the painful falls, and turnover, which is how much trading the strategy needs.
- And it shows how sensitive the answer is to choices such as the look back period.

3 Mira's first backtest report

Strategy backtest, ten years, before costs

Measure	Value
Average return a year	9.6%
Benchmark return a year	7.2%
Maximum drawdown	−32%
Turnover a year	120%

IN THE LESSON

- Here is what Mira's first run produced, ten years of monthly data.
- All figures are made up for the lesson.
- The strategy earned an average of nine point six per cent a year before costs.
- A broad benchmark earned seven point two.
- The worst fall from a peak, the maximum drawdown, was thirty two per cent.
- The strategy turned over one hundred and twenty per cent of the portfolio every year.
- Turnover matters because every trade costs money, and the first figure, nine point six, is before any of that cost.
- So how much is left after trading?

4 Net return after trading costs

Net return = Gross return − Turnover × Cost per unit traded

Net return	What the investor keeps
Gross return	Return before any costs
Turnover	Share of portfolio traded a year
Cost per unit traded	Loss for each unit traded

IN THE LESSON

- Here is the formula that answers it.
- Net return equals gross return, minus turnover times the cost per unit traded.
- Gross return is what the backtest earned before costs.
- Turnover is the share of the portfolio traded in a year, so one hundred and twenty per cent means every dollar of the portfolio is traded about one and a fifth times.
- Cost per unit traded is what each dollar traded loses to commissions and to the gap between buying and selling prices.
- Multiply turnover by that cost and you have the annual drag.
- Subtract it from the gross return, and what remains is the return an investor could really have kept.

5 Example 1: from gross to net

Gross 9.6%, turnover 120%, cost 0.5% per unit traded

Gross return	from the backtest	**9.6%**
Cost drag	1.20 × 0.5%	**0.6%**
Net return	9.6% − 0.6%	**9.0%**
Benchmark		**7.2%**
Active return	net minus benchmark	**1.8%**

IN THE LESSON

- Example one.
- Mira assumes it costs half a per cent of the amount traded, to buy or sell.
- Turnover is one hundred and twenty per cent.
- The drag is one point two times zero point five, which is zero point six per cent a year.
- Take that from the gross nine point six, and net return is nine point zero per cent.
- Compare with the benchmark at seven point two.
- The strategy still wins, by one point eight per cent a year.
- That one point eight is the active return, and it is the number the committee will argue about.
- That is why we always look at net figures.

6 From gross to net, drawn

Return (% a year)

Gross	**9.6%**
Cost drag	**0.6%**
Net	**9.0%**
Benchmark	**7.2%**

Costs shrink the edge; the benchmark is the bar to beat

IN THE LESSON

- Here are the same figures as bars.
- The tall first bar is the gross return, nine point six per cent.
- The short gold bar beside it is the cost drag, only zero point six, but it is the price of all that trading.
- The third bar is the net return, nine point zero.
- And the last bar is the benchmark at seven point two.
- The gap between the third bar and the fourth is the active return, one point eight.
- Notice how a strategy with much higher turnover would have a much taller cost bar, and would eat into the net bar very quickly.
- We will see exactly that in the last part.

7 A filter, not a proof

- Failing the backtest: usually reject the idea
- Passing it: only a reason to look further
- Good past results may be luck or tuning
- Or a regime that has passed

IN THE LESSON

- Now the trap in this first outcome.
- Suppose the backtest looks wonderful.
- It is tempting to say, great, this strategy will earn the same in future.
- Resist that.
- A backtest is a necessary filter for an idea.
- If a sensible strategy fails on past data, it rarely deserves real money.
- But passing is not proof.
- The good result might be pure luck, especially if many ideas were tried.
- It might come from rules tuned to fit that one sample.
- Or it might be a real effect from a regime that has now passed, like an era of falling interest rates.
- So a strong backtest is a reason to look further, not a reason to invest.

8 A story: the forecaster who read tomorrow's paper

- A forecaster who reads tomorrow's paper
- Perfect record, no value
- Use only what was known on each date
- Ask at every step: could we have known?

IN THE LESSON

- Imagine a weather forecaster who is judged on last week's forecasts, but who is allowed to read last week's newspapers before making them.
- Her record would be perfect, and completely useless.
- A good backtest must prevent exactly this.
- At every date in the test, the strategy may use only what an investor could have known on that date.
- Not a day later, not a number revised later.
- This is the idea behind the stages of a backtest, and behind the rolling window design.
- Whenever you read about a backtest procedure, ask one question.
- At this step, could a real investor have known this?

9 The stages of a backtest

The four stages

Specify the strategy → Gather point in time data → Run the simulation → Report the statistics

Inside the simulation: one rolling step

Use data up to the date → Form the portfolio → Record next period return → Roll the window forward

IN THE LESSON

- Here are the stages, in order.
- Along the top row, first, specify the strategy.
- Write down the universe of shares you may buy, the signal that ranks them, how often you rebalance, and the benchmark.
- Second, gather point in time data, meaning the figures exactly as they stood on each date, including the reporting delays.
- Third, run the simulation to get the portfolio returns.
- Fourth, produce the statistics: return, risk, drawdown and turnover.
- The bottom row shows what happens inside the third stage.
- At each rebalancing date, estimate the signal using only past data, form the portfolio, record the return over the next period, then roll forward.

10 Stage one: specify before you look

- Universe: which shares may be held
- Signal: the number that ranks them
- Rebalancing: how often to reset
- Benchmark and cost assumptions, written first

IN THE LESSON

- The first stage is the one people rush.
- Mira must write down the universe, the signal, the rebalancing frequency and the benchmark, before she sees a single result.
- The universe is the list of shares she is allowed to hold, such as the largest one thousand companies.
- The signal is the number that ranks them, here price compared with book value.
- Rebalancing frequency is how often she resets the portfolio, here monthly.
- The benchmark is what she compares against.
- She should also write down her cost assumptions and any limit on shorting.
- Why?
- Because choices made after seeing results tend to flatter the results, as we will see in the last part.

11 Example 2: point in time data

Annual accounts for a December year end

Fiscal year ends	**31 Dec**
Accounts published	**31 Mar**
Reporting lag	**3 months**
Earliest use in test	**1 Apr**
Wrong use in test	**1 Jan**

IN THE LESSON

- Example two is about the data stage.
- Mira's signal uses company accounts.
- A firm's financial year ends on the thirty first of December.
- But the annual accounts are not published until the thirty first of March, three months later.
- A real investor in January could not have seen them.
- So in a proper backtest, the accounts for that year may be used only from the first of April.
- If the backtest simply files the accounts under their year end date and uses them on the first of January, it uses information three months before it existed.
- That is a built in knowledge of the future, and it makes the results look better than any real investor could have achieved.

12 Example 3: the rolling window

120 months of data, 60 month rolling window

Total history		**120 months**
First estimation window		**60 months**
Out of sample months	120 − 60	**60**
In sample Sharpe		**1.10**
Out of sample Sharpe	half of 1.10	**0.55**

IN THE LESSON

- Example three counts the months.
- Mira has one hundred and twenty months of data.
- She uses a rolling window of sixty months.
- At the end of month sixty, she estimates the model on those sixty months only, forms the portfolio, and records the return in month sixty one.
- Then the window moves forward one month, and she repeats.
- That leaves sixty months of results, every one of them out of sample, meaning the model had never seen that month when it was built.
- Suppose the in sample Sharpe ratio was one point one, but the out of sample ratio is zero point five five.
- It fell by half.
- That shrinkage is normal, and it is why only out of sample results are believed.

13 Mistakes: the traps so far

- Strong backtest does not mean future success
- Fitting and testing on the same period
- Fix: rolling window, out of sample results
- Forgetting the reporting lag on accounts

IN THE LESSON

- Let us stop and collect the traps.
- The first trap is treating a strong backtest as proof the strategy will deliver the same in future.
- It is a filter, not a promise.
- The second trap is the one the stages exist to prevent.
- It is estimating the signal using the whole sample, and then testing it on the very same period.
- The model has then seen the answers.
- The result is built on knowledge of the future, and any strategy will look good.
- The honest way is the rolling window, where each month is predicted only from earlier months.
- A third, smaller trap is forgetting the reporting lag.
- Remember Example two, where accounts were used three months too early.

14 A story: the dashboard of a car

- Cumulative return chart
- Average return, volatility, Sharpe ratio
- Maximum drawdown and turnover
- Signal scores: information coefficient, hit rate

IN THE LESSON

- A backtest report is like the dashboard of a car.
- One dial, the speedometer, tells you how fast you are going, but nobody drives on that dial alone.
- You also need the fuel gauge, the warning lights and the temperature.
- In the same way, Mira's report has a chart of cumulative return, which is the growth of one unit of money.
- It has the average return and the volatility.
- It has the Sharpe ratio, the maximum drawdown and the turnover.
- And it has two scores for the signal itself, the information coefficient and the hit rate.
- Each dial answers a different question, and the danger is reading only the flattering ones.

15 Mira's monthly numbers

Monthly backtest statistics

Measure	Value
Average return a month	0.80%
Volatility a month	3.00%
Risk free rate a month	0.20%
Hit rate	54%

IN THE LESSON

- Here are the monthly numbers behind the annual figures.
- Average monthly return is zero point eight per cent.
- Monthly volatility is three per cent.
- The risk free rate is zero point two per cent a month.
- And the hit rate, which is the share of months in which the signal picked the right direction, is fifty four per cent.
- Notice these are monthly.
- The Sharpe ratio is the ratio of reward to risk, and if you compute it from monthly data, you get a monthly Sharpe ratio.
- Reports usually quote annual figures, so we must know how to convert.
- First, the formula itself.

16 The Sharpe ratio, annualised

Sharpe = (Average return − Rf) ÷ Volatility
Annual = Monthly × √12

Sharpe ratio	Reward for each unit of risk
Average return	Mean return of the strategy
Risk free rate	Return with no risk
Volatility	Standard deviation of returns
Annualise	Monthly ratio × 3.46

IN THE LESSON

- The Sharpe ratio is average return above the risk free rate, divided by volatility.
- It tells you how much extra return you earned for each unit of risk.
- For monthly data, the annual Sharpe ratio is the monthly one multiplied by the square root of twelve.
- Why the square root?
- Average excess return grows twelve times in a year, but volatility grows only by the square root of twelve, because monthly ups and downs partly cancel out.
- So the ratio grows by twelve divided by the square root of twelve, which is the square root of twelve again.
- The common error is multiplying by twelve, which overstates the ratio badly.

17 Example 4: annualising the Sharpe ratio

Monthly mean 0.80%, risk free 0.20%, volatility 3.00%

Monthly excess return	0.80% − 0.20%	**0.60%**
Monthly Sharpe	0.60 ÷ 3.00	**0.20**
Multiplier	√12	**3.46**
Annual Sharpe	0.20 × 3.46	**0.69**

IN THE LESSON

- Example four.
- Excess return each month is zero point eight, minus zero point two, which is zero point six per cent.
- Divide by monthly volatility of three per cent, and the monthly Sharpe ratio is zero point two.
- Now multiply by the square root of twelve, which is about three point four six.
- The annual Sharpe ratio is zero point six nine.
- As a check, annual excess return is twelve times zero point six, seven point two per cent, and annual volatility is three times three point four six, ten point four per cent.
- Seven point two divided by ten point four also gives zero point six nine.
- Good, the two routes agree.

18 Maximum drawdown

Maximum drawdown = (Trough value − Peak value) ÷ Peak value

Maximum drawdown	Biggest fall, as a percentage
Trough value	Lowest value after the peak
Peak value	Highest value before the fall

IN THE LESSON

- Next, the fuel gauge: maximum drawdown.
- It is the largest fall from a previous peak to a later trough, measured as a percentage of the peak.
- Picture a hiker who climbs to a high point, then descends into a valley before climbing again.
- The drawdown is how far down she went, compared with how high she had climbed.
- The Sharpe ratio uses average ups and downs, but investors feel the single worst fall, the time the portfolio stood far below its high.
- Drawdown measures that pain.
- A strategy with a decent Sharpe ratio but a very deep drawdown may be impossible to hold, because clients and managers lose heart and sell at the bottom.

19 The cumulative return chart

Chart: Portfolio value vs Time

- Peak 160
- Trough 108.8

IN THE LESSON

- Here is the cumulative return chart of Mira's strategy, the most common picture in a backtest report.
- Along the bottom is time.
- Up the side is the value of the portfolio, starting at one hundred.
- Follow the curve.
- It climbs to a peak of one hundred and sixty.
- Then it falls to a trough of one hundred and eight point eight.
- Then it recovers and ends near one hundred and seventy five.
- The dashed line marks the old peak.
- The vertical gap between the peak and the trough is the drawdown.
- A chart like this shows what an average hides, namely how long the investor sat below the old high, and how it felt on the way.

20 Example 5: the size of the drawdown

Peak 160, trough 108.8

Peak value		**160**
Trough value		**108.8**
Fall	160 − 108.8	**51.2**
Maximum drawdown	51.2 ÷ 160	**−32%**
Gain needed to recover	160 ÷ 108.8 − 1	**47.1%**

IN THE LESSON

- Example five reads the chart in numbers.
- The peak is one hundred and sixty, the trough one hundred and eight point eight.
- The fall is fifty one point two.
- Divided by the peak, that is thirty two per cent, so maximum drawdown is minus thirty two per cent.
- Now an exam favourite: how much must the portfolio gain to get back to its peak?
- It is not thirty two.
- From one hundred and eight point eight to one hundred and sixty is a gain of forty seven point one per cent.
- The deeper the hole, the harder the climb out, because the gain is measured on a smaller base.

21 When averages mislead: tails and skew

- Skewness: are the big moves gains or losses?
- Excess kurtosis: fat tails
- Downside deviation, VaR, conditional VaR
- Negative skew and fat tails: Sharpe flatters

IN THE LESSON

- Strategy returns are rarely a neat bell curve, so the report may add more dials.
- Skewness shows whether big surprises are mostly good or mostly bad.
- Negative skewness means the large moves are losses.
- Excess kurtosis shows fat tails, meaning extreme outcomes happen more often than a bell curve says.
- Downside deviation measures volatility using only bad months.
- Value at risk, called VaR, is a loss level that should be exceeded only rarely, and conditional VaR is the average loss when it is exceeded.
- The key lesson is this.
- With negative skewness and fat tails, the standard deviation understates the chance of large losses, so the Sharpe ratio flatters the strategy.

22 Same Sharpe ratio, very different strategies

Two strategies, annual statistics

Measure	Value
Sharpe ratio, A and B	0.70
Strategy A max drawdown	−18%
Strategy B max drawdown	−45%
Strategy B excess kurtosis	6.0

IN THE LESSON

- Here is why we never judge on the Sharpe ratio alone.
- Two strategies, A and B, both have an annual Sharpe ratio of zero point seven.
- On that dial they are twins.
- Now look at the warning lights.
- Strategy A has a maximum drawdown of eighteen per cent.
- Strategy B has a drawdown of forty five per cent and excess kurtosis of six.
- Strategy B earns its steady gains by quietly selling insurance against rare disasters, and then, once in a while, the disaster arrives.
- A committee that looked at the average return and the Sharpe ratio would call them equal, and would be badly wrong.

23 Example 6: is the hit rate luck?

Hit rate 54%, 120 months, null hypothesis 50%

Excess over a coin toss	54% − 50%	**4.0%**
Standard error	√(0.25 ÷ 120)	**4.56%**
Test statistic	4.0 ÷ 4.56	**0.88**
Critical value	5% level	**1.96**
Verdict		**not significant**

IN THE LESSON

- The last dials score the signal.
- The information coefficient is the correlation between the signal and the returns that followed.
- The hit rate is the share of periods when the signal pointed the right way.
- Example six asks whether Mira's fifty four per cent hit rate over one hundred and twenty months could be luck.
- If the signal were useless, the hit rate would be fifty per cent, with a standard error of the square root of one quarter divided by one hundred and twenty, about four point six per cent.
- Fifty four minus fifty is four, and four divided by four point six is only zero point eight eight, well below one point nine six.
- So this hit rate is not convincing on its own.

24 A story: the recipe book with only the winners

- Survivorship bias
- Look ahead bias
- Data snooping
- Ignored costs, shorting limits and capacity

IN THE LESSON

- Picture a cookbook of recipes that have won prizes.
- Every dish looks delicious.
- But nobody mentions the hundred recipes that were tried and thrown away.
- Judging cooking by that book would be a mistake, and judging a strategy by a flawed backtest is the same mistake.
- Four flaws spoil most backtests.
- Survivorship bias, look ahead bias, data snooping, and ignoring the real costs and limits of trading.
- Each one makes the result look better than anything an investor could have achieved.
- A good analyst learns to hunt for them, one at a time, every time she reads a backtest.

25 The four flaws, defined

- Survivorship: failed firms missing from data
- Look ahead: information not yet known
- Snooping: many trials, keep the best
- Costs, shorting limits, capacity ignored

IN THE LESSON

- Here are the four flaws.
- Survivorship bias arises when the data contain only firms that still exist, because the failures have been deleted, and that lifts returns.
- Look ahead bias means using information that was not available on the date, such as accounts published weeks after the period end, or figures later restated.
- Data snooping means testing many variations and keeping the best one, so the winner is partly luck.
- And ignoring transaction costs, shorting constraints and capacity inflates the result, most of all for strategies that trade a lot.
- Capacity means how much money a strategy can manage before its own trading moves prices against it.
- These are the flaws to look for in every report.

26 Example 7: survivorship bias

100 firms at the start, 20 later failed

Survivors (80 firms)	average	**+10%**
Failed firms (20)	average	**−50%**
Backtest sees only survivors		**+10%**
True starting group	0.8×10 + 0.2×(−50)	**−2%**

IN THE LESSON

- Example seven.
- Mira's data set holds one hundred firms that existed ten years ago.
- Twenty of them have since failed and been dropped from the database.
- The eighty survivors returned an average of ten per cent a year.
- The twenty failures returned an average of minus fifty per cent.
- A backtest on today's database sees only the survivors and reports ten per cent.
- The truth for the whole starting group is eighty per cent times ten, plus twenty per cent times minus fifty.
- That is eight, minus ten, which is minus two per cent a year.
- The backtest turned a loss into a gain simply by deleting the losers.

27 Data snooping: how many tries?

P(at least one false winner) = 1 − (1 − α)^N

Chance of a false winner	Luck alone finds a winner
Significance level	Chance of a lucky pass each time
Number of trials	Variations tried and compared

IN THE LESSON

- Now data snooping, with a formula.
- Suppose each trial of a useless strategy has a five per cent chance of looking significant by luck.
- That is the significance level.
- If you try N variations, the chance that at least one looks significant is one minus the quantity one minus alpha, raised to the power N.
- Each variation might change the look back period, the rebalancing day or the signal definition.
- With one trial, the chance is five per cent.
- With twenty trials, it is sixty four per cent.
- The more trials, the higher the hurdle must be.
- That is why for a heavily searched strategy, analysts ask for a t statistic above three rather than two.
- That is a sensible rule for any strategy that was searched hard.

28 Example 8: twenty variations

20 variations tried, 5% level each

Variations tried		**20**
Chance of a false winner	1 − 0.95^20	**64%**
Best t statistic		**2.6**
Hurdle after many trials		**3.0**
Verdict		**not convincing**

IN THE LESSON

- Example eight.
- Mira tried twenty variations of her strategy, changing the look back period and the rebalancing day, and kept the best one.
- Its t statistic is two point six.
- At first glance that beats the usual hurdle of two.
- But the chance that at least one of twenty useless variations passes by luck is one minus zero point nine five, to the power twenty, which is about sixty four per cent.
- So the best of twenty must clear a higher hurdle, say three.
- Two point six does not.
- The honest remedy is to test the chosen variation on fresh data it has never seen, and to report how many variations were tried.

29 Example 9: costs and turnover

Gross 9.6%, cost 0.5% per unit, benchmark 7.2%

Turnover		**600%**
Cost drag	6.00 × 0.5%	**3.0%**
Net return	9.6% − 3.0%	**6.6%**
Against benchmark	6.6% − 7.2%	**−0.6%**

IN THE LESSON

- Example nine returns to costs.
- A second strategy, a short term version of Mira's idea, also earns a gross return of nine point six per cent, but it turns over six hundred per cent a year.
- At the same half per cent per unit traded, the drag is six times zero point five, which is three per cent.
- Net return falls to six point six per cent.
- That is below the benchmark at seven point two, so the edge has gone, and turned negative by zero point six.
- High turnover is where costs bite.

30 Flaws and their remedies

Match each flaw to its fix

Flaw	Remedy
Survivorship bias	include failed firms
Look ahead bias	point in time, lags
Data snooping	out of sample, higher hurdle
Ignored costs	cost model, capacity

IN THE LESSON

- Every flaw has a remedy, and the exam likes to pair them.
- For survivorship bias, use a database that includes firms that failed or were delisted.
- For look ahead bias, use point in time data with realistic reporting lags.
- For data snooping, test out of sample, count your trials and raise the hurdle.
- For ignored costs, use a cost model, test with higher cost assumptions and check capacity.
- One trap is worth repeating.
- Look ahead bias is easy to miss when data are dated by the fiscal period they describe rather than the date they were published.
- Always ask, on what date could an investor have actually seen this number?

31 Recap

- Backtest: rules on past data, a filter not proof
- Specify, point in time data, rolling window
- Sharpe times √12; check drawdown and tails
- Hunt survivorship, look ahead, snooping, costs

IN THE LESSON

- Let us pull it together.
- A backtest applies written rules to past data as if run in real time.
- It estimates return and risk, and compares with a benchmark, but passing it is only a filter.
- Its stages are specify the strategy, gather point in time data, and run a rolling window simulation that gives out of sample results.
- In the report, annualise a monthly Sharpe ratio by the square root of twelve, read the maximum drawdown and the tails, and never judge on average return and Sharpe alone.
- And hunt for survivorship, look ahead, snooping and cost flaws, with their remedies.

32 Try three quick questions

1. Monthly Sharpe ratio is 0.25. Annual Sharpe ratio?
   Answer: 0.25 × √12 = 0.87
2. Peak 200, trough 140. Maximum drawdown?
   Answer: (140 − 200) ÷ 200 = −30%
3. 10 useless variations, 5% each. Chance of a lucky winner?
   Answer: 1 − 0.95^10 = 40%

IN THE LESSON

- Now try three questions yourself, pausing after each.
- Question one.
- A monthly Sharpe ratio is zero point two five.
- What is the annual figure?
- Multiply by the square root of twelve, which is three point four six, giving about zero point eight seven.
- Question two.
- A portfolio peaks at two hundred and falls to one hundred and forty.
- What is the maximum drawdown?
- Sixty divided by two hundred, minus thirty per cent.
- Question three.
- An analyst tries ten useless variations, each with a five per cent chance of a lucky pass.
- What is the chance at least one passes?
- One minus zero point nine five to the power ten, about forty per cent.
- So a strong result from many trials may simply be luck.