ĐỀ CƯƠNG ÔN TẬP

TRÍ TUỆ NHÂN TẠO TRONG GIAO DỊCH ĐỊNH LƯỢNG

Dành cho sinh viên đại học ngành công nghệ tài chính, tài chính – ngân hàng

Artificial Intelligence in Quantitative Trading – a bilingual study guide

Dữ liệu tài chính · Mô hình chuỗi thời gian · Học máy dự báo lợi suất
Tối ưu danh mục · Ngang bằng rủi ro · Giao dịch cặp · Kiểm định ngược

Bài tập có lời giải · Tự luận có dàn ý · Trắc nghiệm có đáp án · Mã Python chạy thật

Song ngữ Việt – Anh · Vietnamese – English

Phiên bản 1.0 – Tháng 11 năm 2026
https://tailieuso.vin

MỤC LỤC

CONTENTS

Chương 1. Giao dịch định lượng và học máy tài chính: tổng quan – *Chapter 1. Quantitative trading and financial machine learning: an overview* 7

1.1. Giao dịch định lượng là gì – *1.1. What quantitative trading is* 7

1.2. Bức tranh lớn: từ dữ liệu đến giám sát – *1.2. The big picture: from data to monitoring* 9

1.3. Giá là lời dự báo, thông tin thì nhiều, dạng hàm thì chưa biết – *1.3. A price is a forecast, information is plentiful, the functional form is unknown* 11

1.4. Học máy và kinh tế lượng: khác mục tiêu, khác cách đánh giá – *1.4. Machine learning and econometrics: different goals, different ways of judging* 13

1.5. Thách thức khi dùng học máy trong tài chính – *1.5. Challenges of using machine learning in finance* 15

1.6. Lợi suất đơn, lợi suất log và lợi suất danh mục – *1.6. Simple returns, log returns and portfolio returns* 17

1.7. Các chiến lược định lượng phổ biến – *1.7. Common quantitative strategies* 20

1.8. Sản phẩm tài chính – ngân hàng dùng mô hình định lượng – *1.8. Financial and banking products that use quantitative models* 22

1.9. Quy tắc của sàn ảnh hưởng thế nào đến mô hình – *1.9. How exchange rules affect a model* 23

1.10. Pháp lý và vị trí việc làm – *1.10. Law and jobs* 26

1.11. Thực hành bằng Python – *1.11. Practising with Python* 27

1.12. Bài tập vận dụng có lời giải – *Applied exercises with solutions* 33

1.13. Câu hỏi tự luận và dàn ý trả lời – *Essay questions and answer outlines* 35

1.14. Câu hỏi trắc nghiệm – *Multiple-choice questions* 36

Chương 2. Dữ liệu tài chính và các đặc điểm điển hình – *Chapter 2. Financial data and stylised facts* 39

2.1. Giá, lợi suất và giá điều chỉnh – *2.1. Prices, returns and adjusted prices* 39

2.2. Các đặc điểm điển hình của lợi suất – *2.2. Stylised facts of returns* 44

2.3. Độ lệch, độ nhọn và phân phối chuẩn – *2.3. Skewness, kurtosis and the normal distribution* 46

2.4. Tự tương quan và cụm biến động – *2.4. Autocorrelation and volatility clustering* 50

2.5. Cấu trúc giữa các tài sản: tương quan và nhân tố thị trường – *2.5. Structure across assets: correlation and the market factor* 54

2.6. Dữ liệu cho học máy: đặc trưng, tần suất và hai sai lệch – *2.6. Data for machine learning: features, frequency and two biases* 58

2.7. Làm sạch dữ liệu giá – *2.7. Cleaning price data* 61

2.8. Ứng dụng ở quỹ, công ty chứng khoán và ngân hàng – *2.8. Applications at funds, securities firms and banks* 64

2.9. Thực hành bằng Python – *2.9. Practising with Python* 66

2.10. Bài tập vận dụng có lời giải – *Applied exercises with solutions* 70

2.11. Câu hỏi tự luận và dàn ý trả lời – *Essay questions and answer outlines* 71
2.12. Câu hỏi trắc nghiệm – *Multiple-choice questions* 73

Chương 3. Ước lượng tham số, co rút và mô hình nhân tố – *Chapter 3. Parameter estimation, shrinkage and factor models* 76
3.1. Mô hình độc lập cùng phân phối – *3.1. The independent and identically distributed model* 76
3.2. Ước lượng mẫu: trung bình, phương sai, hiệp phương sai – *3.2. Sample estimators: mean, variance, covariance* 78
3.3. Ước lượng vị trí vững: trung vị và trung bình cắt – *3.3. Robust location estimators: median and trimmed mean* 82
3.4. Ước lượng hợp lý cực đại: dữ liệu chuẩn và đuôi dày – *3.4. Maximum likelihood estimation: normal and heavy-tailed data* 84
3.5. Co rút: kéo ước lượng về một mục tiêu đơn giản – *3.5. Shrinkage: pulling an estimate towards a simple target* 86
3.6. Mô hình nhân tố và Black–Litterman – *3.6. Factor models and Black–Litterman* 89
3.7. Lý thuyết định giá kinh doanh chênh lệch, phần bù rủi ro và nhân tố từ phân tích thành phần chính – *3.7. Arbitrage pricing theory, risk premia and principal component analysis factors* 92
3.8. Nhân tố có điều kiện và phân tích thành phần chính có điều kiện – *3.8. Conditional factors and instrumented principal component analysis* 96
3.9. Ứng dụng ở quỹ, công ty chứng khoán và ngân hàng – *3.9. Applications at funds, securities firms and banks* 98
3.10. Thực hành bằng Python – *3.10. Practice with Python* 99
3.11. Bài tập vận dụng có lời giải – *Applied exercises with solutions* 103
3.12. Câu hỏi tự luận và dàn ý trả lời – *Essay questions and answer outlines* 105
3.13. Câu hỏi trắc nghiệm – *Multiple-choice questions* 107

Chương 4. Mô hình chuỗi thời gian và đồ thị tài chính – *Chapter 4. Time-series models and financial graphs* 109
4.1. Trung bình và biến động đổi theo thời gian – *4.1. Mean and volatility change over time* 109
4.2. Mô hình trung bình theo thời gian – *4.2. Mean models over time* 112
4.3. Mô hình biến động theo thời gian – *4.3. Volatility models over time* 114
4.4. Không gian trạng thái và lọc Kalman một biến – *4.4. State-space models and the one-variable Kalman filter* 118
4.5. Đồ thị tài chính: nút, cạnh và ma trận kề – *4.5. Financial graphs: nodes, edges and the adjacency matrix* 121
4.6. Học đồ thị từ dữ liệu – *4.6. Learning a graph from data* 122
4.7. Ứng dụng: phân nhóm mã và phát hiện chế độ thị trường – *4.7. Applications: grouping stocks and detecting market regimes* 125
4.8. Thực hành bằng Python – *4.8. Practising with Python* 127
4.9. Bài tập vận dụng có lời giải – *Applied exercises with solutions* 135
4.10. Câu hỏi tự luận và dàn ý trả lời – *Essay questions and answer outlines* 136
4.11. Câu hỏi trắc nghiệm – *Multiple-choice questions* 138

Chương 5. Dự báo lợi suất bằng học máy – *Chapter 5. Return prediction with machine learning* 140

5.1. Bài toán dự báo lợi suất – *5.1. The return-forecasting problem* 140

5.2. Thiết kế thực nghiệm: chia dữ liệu theo thời gian – *5.2. Experimental design: splitting data in time order* 144

5.3. Mô hình tuyến tính làm mốc và tuyến tính có phạt – *5.3. The linear benchmark and penalised linear models* 147

5.4. Giảm chiều: thành phần chính và hồi quy từng phần – *5.4. Dimension reduction: principal components and partial least squares* 151

5.5. Cây quyết định, rừng ngẫu nhiên và tăng cường – *5.5. Decision trees, random forests and boosting* 153

5.6. Mạng nơ-ron – *5.6. Neural networks* 156

5.7. So sánh các mô hình và nêm phức tạp – *5.7. Comparing models and the complexity wedge* 159

5.8. Từ dự báo đến danh mục – *5.8. From forecasts to a portfolio* 161

5.9. Alpha, kiểm định alpha và kiểm định bội – *5.9. Alpha, testing alpha and multiple testing* 163

5.10. Ứng dụng ở quỹ, công ty chứng khoán và ngân hàng – *5.10. Applications at a fund, a securities firm and a bank* 164

5.11. Thực hành bằng Python – *5.11. Practice in Python* 166

5.12. Bài tập vận dụng có lời giải – *Applied exercises with solutions* 172

5.13. Câu hỏi tự luận và dàn ý trả lời – *Essay questions and answer outlines* 174

5.14. Câu hỏi trắc nghiệm – *Multiple-choice questions* 176

Chương 6. Cơ bản danh mục và lý thuyết danh mục hiện đại – *Chapter 6. Portfolio basics and modern portfolio theory* 178

6.1. Danh mục, lợi suất và rủi ro của danh mục – *6.1. Portfolios, portfolio return and portfolio risk* 178

6.2. Ràng buộc khi lập danh mục – *6.2. Constraints on a portfolio* 183

6.3. Đo hiệu quả của danh mục – *6.3. Measuring portfolio performance* 186

6.4. Danh mục theo quy tắc đơn giản – *6.4. Simple rule-based portfolios* 189

6.5. Tối ưu lồi trong một trang – *6.5. Convex optimisation on one page* 192

6.6. Danh mục dựa trên rủi ro – *6.6. Risk-based portfolios* 194

6.7. Markowitz, đường biên hiệu quả và các danh mục theo hàm mục tiêu – *6.7. Markowitz, the efficient frontier and objective-based portfolios* 197

6.8. Nhược điểm của Markowitz – *6.8. The drawbacks of Markowitz* 201

6.9. Ứng dụng ở quỹ và ngân hàng – *6.9. Applications at funds and banks* 203

6.10. Thực hành bằng Python – *6.10. Practice with Python* 204

6.11. Bài tập vận dụng có lời giải – *Applied exercises with solutions* 208

6.12. Câu hỏi tự luận và dàn ý trả lời – *Essay questions and answer outlines* 210

6.13. Câu hỏi trắc nghiệm – *Multiple-choice questions* 212

Chương 7. Danh mục nâng cao: rủi ro phía dưới, ngang bằng rủi ro và đồ thị – *Chapter 7. Advanced portfolios: downside risk, risk parity and graphs* 214

7.1. Khi phương sai chưa đủ: mô-men bậc cao – *7.1. When variance is not enough: higher moments* 214
7.2. Đo rủi ro phía lỗ: bán phương sai và tổn thất ở đuôi – *7.2. Measuring the downside: semi-variance and tail losses* 218
7.3. Sụt giảm: thước đo nhìn vào đường đi – *7.3. Drawdown: a measure that looks at the path* 221
7.4. Danh mục giảm rủi ro phía dưới, danh mục CVaR và danh mục kiểm soát sụt giảm – *7.4. Downside risk, CVaR and drawdown-control portfolios* 224
7.5. Từ chia đều vốn sang chia đều rủi ro – *7.5. From splitting capital equally to splitting risk equally* 226
7.6. Ngang bằng rủi ro cho nhiều mã: bài toán tổng quát – *7.6. Risk parity for many stocks: the general problem* 229
7.7. Đồ thị của các mã: khoảng cách và phân cụm thứ bậc – *7.7. Graphs of stocks: distances and hierarchical clustering* 232
7.8. Phân bổ rủi ro theo thứ bậc – *7.8. Hierarchical risk parity* 234
7.9. Ứng dụng ở quỹ, công ty chứng khoán và ngân hàng – *7.9. Applications at funds, securities companies and banks* 238
7.10. Thực hành bằng Python: so sánh các danh mục trên dữ liệu tự sinh – *7.10. Practice in Python: comparing portfolios on self-generated data* 241
7.11. Bài tập vận dụng có lời giải – *Applied exercises with solutions* 247
7.12. Câu hỏi tự luận và dàn ý trả lời – *Essay questions and answer outlines* 249
7.13. Câu hỏi trắc nghiệm – *Multiple-choice questions* 251

Chương 8. Theo dõi chỉ số, danh mục vững và học sâu cho danh mục – *Chapter 8. Index tracking, robust portfolios and deep learning for portfolios* 253

8.1. Chủ động và thụ động, chỉ số là gì – *8.1. Active and passive investing, and what an index is* 253
8.2. Theo dõi chỉ số và sai số theo dõi – *8.2. Index tracking and tracking error* 255
8.3. Theo dõi thưa: dùng ít mã – *8.3. Sparse tracking: using few stocks* 257
8.4. Theo dõi nâng cao – *8.4. Enhanced index tracking* 259
8.5. Danh mục vững – *8.5. Robust portfolios* 260
8.6. Lấy mẫu lại danh mục – *8.6. Portfolio resampling* 262
8.7. Học máy và học sâu cho thiết kế danh mục – *8.7. Machine learning and deep learning for portfolio design* 263
8.8. Ước lượng và tối ưu tích hợp, hồi quy Sharpe tối đa, nhân tố chiết khấu ngẫu nhiên – *8.8. Integrated estimation and optimisation, maximum Sharpe regression, stochastic discount factor* 265
8.9. Chi phí giao dịch, vòng quay và học tăng cường – *8.9. Trading costs, turnover and reinforcement learning* 267
8.10. Ứng dụng ở quỹ, công ty chứng khoán và ngân hàng – *8.10. Applications at funds, brokerages and banks* 269
8.11. Thực hành bằng Python – *8.11. Practice in Python* 270
8.12. Bài tập vận dụng có lời giải – *Applied exercises with solutions* 279
8.13. Câu hỏi tự luận và dàn ý trả lời – *Essay questions and answer outlines* 280
8.14. Câu hỏi trắc nghiệm – *Multiple-choice questions* 282

Chương 9. Giao dịch cặp và kinh doanh chênh lệch thống kê – *Chapter 9. Pairs trading and statistical arbitrage* 284

9.1. Hồi quy về trung bình và chu kỳ bán rã – *9.1. Mean reversion and half-life* 284

9.2. Đồng liên kết khác tương quan – *9.2. Cointegration is not correlation* 288

9.3. Giao dịch cặp: từ chênh lệch đến lệnh – *9.3. Pairs trading: from the spread to an order* 291

9.4. Tìm cặp đồng liên kết – *9.4. Finding a cointegrated pair* 294

9.5. Giao dịch chênh lệch: ngưỡng, quy mô và chi phí – *9.5. Trading the spread: thresholds, size and costs* 297

9.6. Lọc Kalman cho cặp: hệ số β thay đổi theo thời gian – *9.6. A Kalman filter for pairs: a β that changes over time* 301

9.7. Kinh doanh chênh lệch thống kê nhiều mã – *9.7. Statistical arbitrage across many stocks* 304

9.8. Rủi ro: cặp gãy, chi phí và hạn chế bán khống – *9.8. Risks: broken pairs, costs and short-selling limits* 307

9.9. Ứng dụng tại quỹ, công ty chứng khoán và ngân hàng – *9.9. Applications at funds, brokerages and banks* 309

9.10. Thực hành bằng Python – *9.10. Hands-on with Python* 310

9.11. Bài tập vận dụng có lời giải – *Applied exercises with solutions* 314

9.12. Câu hỏi tự luận và dàn ý trả lời – *Essay questions and answer outlines* 316

9.13. Câu hỏi trắc nghiệm – *Multiple-choice questions* 317

Chương 10. Kiểm định ngược, triển khai và quản trị hệ thống giao dịch AI – *Chapter 10. Backtesting, deployment and governance of AI trading systems* 320

10.1. Một lần kiểm định ngược điển hình – *10.1. A typical backtest* 321

10.2. Bảy 'tội' của đầu tư định lượng – *10.2. The seven 'sins' of quantitative investing* 325

10.3. Quá khớp và Sharpe bị thổi phồng – *10.3. Overfitting and the inflated Sharpe ratio* 327

10.4. Kiểm định ngược trên dữ liệu lịch sử – *10.4. Backtesting on historical data* 332

10.5. Kiểm định ngược trên dữ liệu tổng hợp – *10.5. Backtesting on synthetic data* 335

10.6. Triển khai: từ kiểm định ngược đến giao dịch thật – *10.6. Deployment: from backtest to live trading* 339

10.7. Quản trị rủi ro mô hình – *10.7. Model risk management* 341

10.8. Pháp lý và đạo đức – *10.8. Legal and ethical points* 342

10.9. Điều học máy làm được và chưa làm được – *10.9. What machine learning can and cannot do* 343

10.10. Ứng dụng ở quỹ, công ty chứng khoán và ngân hàng – *10.10. Applications at funds, securities firms and banks* 344

10.11. Thực hành bằng Python: dự án của Hoàng Nam – *10.11. Practice in Python: Hoàng Nam's project* 345

10.12. Bài tập vận dụng có lời giải – *Applied exercises with solutions* 350

10.13. Câu hỏi tự luận và dàn ý trả lời – *Essay questions and answer outlines* 352

10.14. Câu hỏi trắc nghiệm – *Multiple-choice questions* 353

Phụ lục A. Đáp án câu hỏi trắc nghiệm – *Appendix A. Answer key* 356

Phụ lục B. Công thức – *Appendix B. Formulas* **361**

Phụ lục C. Thuật ngữ Việt – Anh – *Appendix C. Vietnamese–English glossary* **368**

Phụ lục D. Tài liệu tham khảo – *Appendix D. References* **375**

Chương 1. Giao dịch định lượng và học máy tài chính: tổng quan – *Chapter 1. Quantitative trading and financial machine learning: an overview*

Mục tiêu chương – *Chapter objectives*

- Nói được giao dịch định lượng khác giao dịch theo cảm tính ở đâu và ai làm nghề này.
  Say how quantitative trading differs from discretionary trading and who does it.
- Vẽ được bảy bước từ dữ liệu đến giám sát và nói mỗi bước tạo ra sản phẩm gì.
  Draw the seven steps from data to monitoring and say what each step produces.
- Giải thích vì sao giá là dự báo, vì sao học máy hữu ích, và bốn thách thức khi dùng nó ở tài chính.
  Explain why a price is a forecast, why machine learning helps, and the four challenges of using it in finance.
- Tính được lợi suất đơn, lợi suất log và lợi suất danh mục từ giá, rồi tránh lỗi cộng lợi suất đơn qua nhiều ngày.
  Compute simple, log and portfolio returns from prices and avoid the error of adding simple returns over several days.
- Đổi trọng số của mô hình thành lệnh hợp lệ theo biên độ, lô, bước giá và thời hạn thanh toán của sàn.
  Turn a model weight into an order that is valid under the price band, lot, price step and settlement period of the exchange.

1.1. Giao dịch định lượng là gì – *1.1. What quantitative trading is*

Hoàng Nam vào thực tập ở bộ phận Đầu tư định lượng của Công ty Quản lý quỹ Cánh Diều. Buổi đầu, chị Thu Hà, trưởng nhóm nghiên cứu, hỏi cậu: 'Giả sử em quản lý 10 tỷ đồng. Em mua mã nào, mua lúc nào, bán lúc nào, và dựa vào đâu?'. Nam nghĩ một lúc rồi kể ra ba căn cứ: đọc tin, nhìn đồ thị giá, hỏi người quen trong ngành. Chị Hà gật đầu và nói ở nhóm chị, chữ 'dựa vào đâu' phải đổi thành 'dựa vào con số nào'.

Hoàng Nam starts an internship in the quantitative investment unit of Cánh Diều Asset Management Company. On the first day Thu Hà, the head of research, asks him: 'Suppose you manage 10 billion dong. Which stock do you buy, when do you buy it, when do you sell, and on what basis?' Nam thinks for a moment and lists three grounds: reading the news, looking at the price chart, asking people he knows in the industry. Hà nods and says that in her team 'on what basis' has to become 'on which numbers'.

Giao dịch định lượng (quantitative trading) là cách đầu tư mà mọi quyết định mua, bán và chia tỷ trọng đều đi ra từ một quy tắc viết rõ bằng số và chạy được bằng mã. Hai người chạy cùng quy tắc trên cùng dữ liệu phải ra cùng một lệnh. Quy tắc đó có thể kiểm lại trên dữ liệu quá khứ trước khi đem tiền thật vào. Con người vẫn quyết định dữ liệu nào đưa vào, quy tắc nào đáng thử và lúc nào tắt hệ thống.

Quantitative trading is investing in which every decision to buy, sell and size a position comes out of a rule written in numbers that code can run. Two people running the same rule on the

same data must produce the same order. The rule can be re-checked on past data before real money goes in. People still decide which data to feed in, which rules are worth testing and when to switch the system off.

Cách làm ngược lại gọi là giao dịch theo cảm tính (discretionary trading): người giao dịch tự đọc tình hình rồi quyết định. Hai cách không loại trừ nhau, nhiều quỹ dùng cả hai. Bảng sau so sánh theo năm tiêu chí.

The opposite way is called discretionary trading: the trader reads the situation and decides personally. The two ways do not exclude each other and many funds use both. The table compares them on five criteria.

Tiêu chí *Criterion*	**Giao dịch định lượng** *Quantitative*	**Giao dịch theo cảm tính** *Discretionary*
Căn cứ ra quyết định *Basis of decisions*	Quy tắc bằng số, chạy bằng mã *Numeric rules run by code*	Kinh nghiệm, trực giác, đọc tin *Experience, intuition, news reading*
Lặp lại được *Repeatable*	Có, cùng dữ liệu cùng lệnh *Yes, same data gives same order*	Khó, mỗi người một cách đọc *Hard, each person reads differently*
Kiểm lại trên quá khứ *Test on the past*	Làm được, nhưng dễ quá khớp, tức học thuộc cả nhiễu *Possible, but prone to overfitting, which is learning the noise too*	Khó, ít khi viết thành quy tắc *Hard, rarely written as a rule*
Số mã theo dõi cùng lúc *Stocks followed at once*	Hàng trăm mã *Hundreds*	Vài chục mã *A few dozen*
Điểm yếu điển hình *Typical weakness*	Mô hình sai khi thị trường đổi chế độ *Model fails when the market changes regime*	Thiên kiến cá nhân, khó mở rộng *Personal bias, hard to scale*

Hàng 3 và hàng 5 đáng đọc kỹ: kiểm lại trên quá khứ là điểm mạnh của giao dịch định lượng, và cũng là chỗ sinh ra lỗi quá khớp mà chương 10 sẽ mổ xẻ.

Rows 3 and 5 deserve a careful read: re-testing on the past is the strength of quantitative trading, and it is also where the overfitting error comes from, which Chapter 10 takes apart.

Ai làm giao dịch định lượng – *Who does quantitative trading*

Ba nhóm tổ chức dùng cách này nhiều nhất ở Việt Nam. Mỗi nhóm giao dịch bằng nguồn vốn khác nhau, nên câu hỏi mà người kiểm tra rủi ro đặt ra cũng khác nhau.

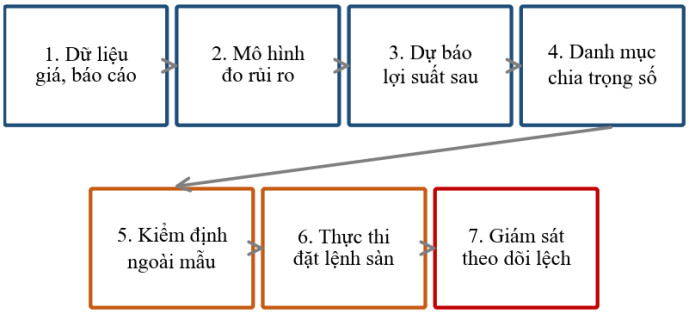
Three groups of organisations use this approach most in Vietnam. Each trades with a different source of capital, so the questions the risk reviewer asks are different too.

Tổ chức *Organisation*	**Vốn của ai** *Whose capital*	**Dùng mô hình để làm gì** *What the models are for*	**Người kiểm tra hay hỏi** *What the reviewer asks*
Công ty quản lý quỹ *Fund manager*	Tiền của nhà đầu tư vào quỹ *Investors' money in the fund*	Chọn mã, chia tỷ trọng cho quỹ *Pick stocks, size positions for the fund*	Có đúng chính sách đầu tư đã công bố không? *Does it follow the published investment policy?*
Tự doanh của công ty chứng khoán *Brokerage proprietary desk*	Vốn của chính công ty *The firm's own capital*	Giao dịch bằng vốn công ty, tạo lệnh tự động *Trade the firm's capital, generate orders automatically*	Lỗ tối đa trong ngày là bao nhiêu? *What is the maximum loss in a day?*
Ngân hàng *Bank*	Vốn ngân hàng, quỹ hưu trí nội bộ *Bank capital, internal pension fund*	Quản lý danh mục trái phiếu, cổ phiếu *Manage bond and equity portfolios*	Danh mục có hợp khẩu vị rủi ro của ngân hàng không? *Does the portfolio fit the bank's risk appetite?*

1.2. Bức tranh lớn: từ dữ liệu đến giám sát – *1.2. The big picture: from data to monitoring*

Một hệ thống giao dịch định lượng giống một dây chuyền nhà máy. Nguyên liệu vào đầu dây chuyền là dữ liệu, sản phẩm ở cuối là các lệnh mua bán, và mỗi trạm làm đúng một việc. Palomar (2025) gom phần đầu của dây chuyền thành hai khối: mô hình hóa dữ liệu rồi tối ưu danh mục. Sách này kéo dài dây chuyền thêm ba trạm ở cuối, vì tiền thật chỉ vào lúc đó.

A quantitative trading system looks like a factory line. Raw material at the head of the line is data, the product at the end is buy and sell orders, and each station does exactly one job. Palomar (2025) groups the first part of the line into two blocks: modelling the data, then optimising the portfolio. This book extends the line by three stations at the end, because real money enters there.



Hình 1.1. Bảy trạm của dây chuyền. Nhìn hàng trên: bốn hộp xanh là việc làm trên máy. Hàng dưới: hai hộp cam đưa kết quả ra thị trường, hộp đỏ cuối cùng phát hiện lúc mô hình lệch.
Figure 1.1. *The seven stations of the line. On the top row the four blue boxes are desk work on the computer. On the bottom row two orange boxes take the result to the market and the last, red box detects when the model drifts.*

1. Dữ liệu (data): thu giá, khối lượng, báo cáo tài chính, làm sạch ngày nghỉ và giá trị thiếu.
   Data: collect prices, volumes and financial statements, clean holidays and missing values.
2. Mô hình hóa (modelling): đo mức biến động, tương quan giữa các mã, độ nhạy với thị trường chung.
   Modelling: measure volatility, the correlation between stocks, the sensitivity to the overall market.
3. Dự báo (forecasting): ước lượng lợi suất kỳ sau của từng mã, kèm độ tin cậy.
   Forecasting: estimate each stock's next-period return together with its reliability.
4. Xây dựng danh mục (portfolio construction): đổi dự báo và rủi ro thành trọng số vốn cho từng mã.
   Portfolio construction: turn forecasts and risks into a capital weight for each stock.
5. Kiểm định (validation, còn gọi là kiểm định ngược hay backtesting): chạy quy tắc trên dữ liệu quá khứ chưa dùng khi xây mô hình, có tính chi phí giao dịch.
   Validation (also called back-testing): run the rule on past data not used when the model was built, with transaction costs included.
6. Thực thi (execution): đổi trọng số thành lệnh đúng lô, đúng bước giá (nấc nhảy nhỏ nhất của giá), đúng giờ của sàn.
   Execution: turn weights into orders of the right lot size, the right price step (the smallest jump of a price) and the right time of day on the exchange.
7. Giám sát (monitoring): so lợi suất và rủi ro thực tế với kỳ vọng, dừng hệ thống khi lệch quá ngưỡng.
   Monitoring: compare actual return and risk with expectations and stop the system when the gap passes a threshold.

Bước *Step*	**Sản phẩm của bước** *Output of the step*	**Nơi học trong sách** *Where to learn it*
1. Dữ liệu *1. Data*	Bảng giá sạch, bộ đặc trưng *Clean price table, feature set*	Chương 2 *Chapter 2*
2. Mô hình hóa *2. Modelling*	Trung bình, độ biến động, tương quan *Means, volatilities, correlations*	Chương 3, 4 *Chapters 3, 4*
3. Dự báo *3. Forecasting*	Dự báo lợi suất từng mã *Return forecast for each stock*	Chương 5 *Chapter 5*
4. Danh mục *4. Portfolio*	Vectơ trọng số *Weight vector*	Chương 6, 7, 8, 9 *Chapters 6, 7, 8, 9*
5. Kiểm định *5. Validation*	Lợi suất và lỗ tối đa ngoài mẫu *Out-of-sample return and maximum loss*	Chương 10 *Chapter 10*
6. Thực thi *6. Execution*	Danh sách lệnh đúng quy tắc sàn *Order list that obeys exchange rules*	Mục 1.9, chương 8 *Section 1.9, Chapter 8*
7. Giám sát *7. Monitoring*	Báo cáo rủi ro, nhật ký lệnh *Risk report, order log*	Chương 10 *Chapter 10*

Điều anh Phong, giám đốc quản trị rủi ro, hay nhắc: dây chuyền chỉ tốt bằng trạm yếu nhất. Dự báo giỏi mà thực thi sai lô thì danh mục thật khác danh mục trên giấy. Kiểm định đẹp mà dùng dữ liệu tương lai khi dựng đặc trưng thì con số đẹp đó vô nghĩa.

What Phong, the director of risk management, keeps repeating is that the line is only as good as its weakest station. A good forecast with an order of the wrong lot size leaves the real portfolio different from the paper one. A fine validation that used future data to build a feature gives a fine number that means nothing.

1.3. Giá là lời dự báo, thông tin thì nhiều, dạng hàm thì chưa biết – *1.3. A price is a forecast, information is plentiful, the functional form is unknown*

Giá là một lời dự báo – *A price is a forecast*

Giá vé máy bay dịp Tết tăng từ tháng trước, trước cả khi ai đặt vé. Hãng đã đoán nhu cầu và định giá theo cái đoán đó. Giá cổ phiếu hoạt động tương tự. Giá hôm nay là số tiền thị trường sẵn lòng trả cho các khoản thu trong tương lai của doanh nghiệp, sau khi trừ đi mức ngại rủi ro.

An air fare at Tết rises a month earlier, before anyone has booked. The airline has guessed demand and priced on that guess. A share price works in the same way. Today's price is what the market is willing to pay for the firm's future receipts, after allowing for the dislike of risk.

$$P_t = E(m_{t+1} * x_{t+1} \mid I_t) \qquad (1.1)$$

Ở đây P_t là giá hôm nay, x_{t+1} là khoản thu kỳ sau của tài sản, m_{t+1} là hệ số chiết khấu phản ánh mức ngại rủi ro, và I_t là toàn bộ thông tin có trong tay nhà đầu tư lúc này. Ký hiệu E là kỳ vọng, tức giá trị trung bình nếu tình huống lặp lại nhiều lần. Công thức nói một

điều: giá đã gói sẵn dự đoán của cả thị trường. Tin mới đến thì dự đoán đổi, giá nhảy. Vì vậy dự báo lợi suất là việc khó, vì bạn phải đoán điều mà cả thị trường chưa đoán ra.

Here P_t is today's price, x_{t+1} is the asset's receipt next period, m_{t+1} is a discount factor that reflects the dislike of risk, and I_t is all the information investors hold right now. The symbol E is the expectation, the average if the situation repeated many times. The formula says one thing: the price already packs in the guess of the whole market. When news arrives the guess changes and the price jumps. That is why forecasting returns is hard: you must guess what the whole market has not yet guessed.

Tập thông tin lớn – *A large information set*

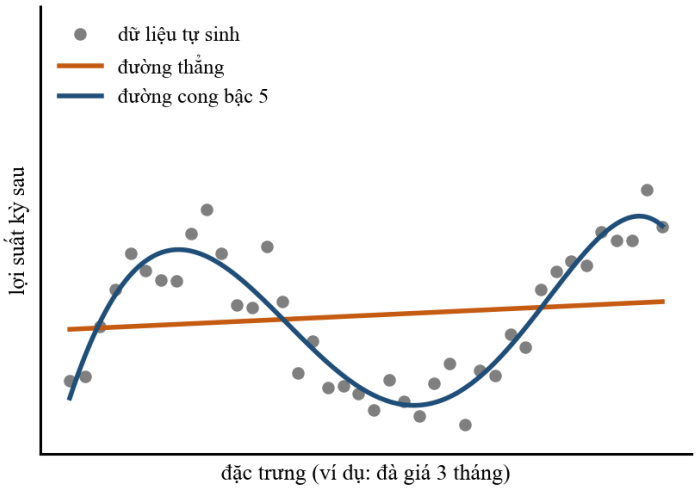
Thông tin ảnh hưởng đến giá có ở khắp nơi. Có số liệu vĩ mô, báo cáo doanh nghiệp, tin tức, mức độ nhộn nhịp của sổ lệnh. Dữ liệu tài chính còn có hai chiều. Chiều thời gian cho biết một mã biến đổi thế nào qua các ngày. Chiều chéo cho biết các mã khác nhau ra sao trong cùng một ngày. Một mô hình tốt phải xử lý được cả điểm chung của thị trường lẫn nét riêng của từng mã.

Information that moves prices is everywhere: macro figures, company reports, news, how busy the order book is. Financial data also has two dimensions. The time dimension shows how one stock changes across days. The cross-sectional dimension shows how different stocks compare on one day. A good model must handle both what the market has in common and what is particular to each stock.

Dạng hàm chưa biết trước – *An unknown functional form*

Giả sử bạn muốn nối 'đà giá 3 tháng' với 'lợi suất tháng sau'. Dạng hàm (functional form) là hình dáng của quan hệ đó: đường thẳng, đường cong, hay bậc thang. Lý thuyết kinh tế hiếm khi cho biết hình dáng đúng. Cách truyền thống là chọn trước một dạng từ lý thuyết, rồi ước lượng các hệ số. Theo Kelly và Xiu (2023), cách thứ hai là để dữ liệu góp ý, bằng cách thử nhiều dạng và chọn dạng dự báo tốt nhất trên dữ liệu chưa thấy.

Suppose you want to link 'three-month price momentum' to 'next month's return'. The functional form is the shape of that link: a straight line, a curve or a staircase. Economic theory rarely tells you the right shape. The traditional route is to pick a form in advance from theory and then estimate its coefficients. According to Kelly and Xiu (2023), the second route is to let the data advise, by trying many forms and keeping the one that forecasts best on unseen data.



Hình 1.2. Cùng 40 điểm dữ liệu tự sinh, hai dạng hàm khác nhau. Nhìn vào đường thẳng cam: nó bỏ sót chỗ dữ liệu cong. Đường cong xanh bám sát hơn, nhưng đuôi bên phải bắt đầu uốn theo nhiễu.

Figure 1.2. *The same 40 synthetic data points with two different functional forms. Look at the orange line: it misses the part where the data bends. The blue curve follows more closely, but its right-hand end begins to bend with the noise.*

Hình 1.2 cho thấy cả hai mặt. Đường thẳng đơn giản nên ước lượng ổn định, nhưng sai về hình dáng. Đường cong mềm hơn nên bám dữ liệu, nhưng dễ học cả phần nhiễu. Học máy là bộ công cụ để chọn mức mềm vừa phải.

Figure 1.2 shows both sides. The straight line is simple so its estimate is stable, but its shape is wrong. The curve is looser so it follows the data, but it can also learn the noise. Machine learning is the toolkit for choosing a moderate looseness.

> **PHẢI THUỘC / MEMORISE –** Giá là dự báo có điều kiện theo thông tin hiện có. Thông tin thì nhiều, dạng hàm nối thông tin với giá thì chưa biết. Hai điều đó khiến học máy hữu ích: nó xử lý nhiều biến cùng lúc và để dữ liệu chọn hình dáng quan hệ.
> *A price is a forecast conditional on current information. Information is plentiful and the functional form linking it to the price is unknown. Those two facts make machine learning useful: it handles many variables at once and lets the data choose the shape of the relationship.*

1.4. Học máy và kinh tế lượng: khác mục tiêu, khác cách đánh giá – *1.4. Machine learning and econometrics: different goals, different ways of judging*

Học máy (machine learning) là nhóm phương pháp để máy tự rút quan hệ từ dữ liệu. Trí tuệ nhân tạo (AI, artificial intelligence) là khái niệm rộng hơn, trong đó học máy là phần được dùng nhiều nhất ở tài chính. Kinh tế lượng (econometrics) là ngành dùng thống kê để kiểm tra các giả thuyết kinh tế trên số liệu. Hai bên dùng chung nhiều công cụ, ví dụ hồi quy. Chúng khác nhau ở câu hỏi đặt ra.

Machine learning is a group of methods in which the machine draws relationships out of data. Artificial intelligence (AI) is the broader idea, and machine learning is the part used most in finance. Econometrics is the field that uses statistics to test economic hypotheses on data. The two share many tools, regression for example. They differ in the question asked.

Kinh tế lượng hỏi 'quan hệ này có thật không, hệ số bằng bao nhiêu?'. Nó chấm bằng kiểm định giả thuyết trên chính dữ liệu đã dùng để ước lượng. Học máy hỏi 'dự báo kỳ sau chính xác đến đâu?'. Nó chấm bằng sai số trên dữ liệu chưa thấy. Bảng sau tóm các khác biệt.

Econometrics asks 'is this relationship real, and what is the coefficient?'. It grades by hypothesis tests on the same data used to estimate. Machine learning asks 'how accurate is the forecast for the next period?'. It grades by error on unseen data. The table summarises the differences.

Tiêu chí *Criterion*	**Kinh tế lượng** *Econometrics*	**Học máy** *Machine learning*
Mục tiêu chính *Main goal*	Hiểu cơ chế, kiểm giả thuyết *Understand a mechanism, test a hypothesis*	Dự báo chính xác *Forecast accurately*
Cách đánh giá *How it is judged*	Ý nghĩa thống kê của hệ số *Statistical significance of coefficients*	Sai số trên dữ liệu chưa thấy *Error on unseen data*
Cỡ mô hình *Model size*	Nhỏ, vài biến chọn từ lý thuyết *Small, a few variables chosen by theory*	Lớn, nhiều biến và nhiều dạng hàm *Large, many variables and functional forms*
Chọn mô hình *Model choice*	Chọn trước, kiểm sau *Chosen first, tested afterwards*	Thử nhiều mô hình, chọn bằng dữ liệu *Many models tried, chosen with data*
Chống quá khớp *Guarding against overfitting*	Ít cần, vì mô hình nhỏ *Seldom needed, the model is small*	Bắt buộc: phạt độ phức tạp, tách dữ liệu *Compulsory: penalise complexity, split the data*

Hàng cuối cho thấy một điểm của Kelly và Xiu (2023): vì học máy thử nhiều mô hình, nó mặc nhiên bị quá khớp trên dữ liệu đã thấy. Để giữ mô hình ổn định ngoài mẫu, người làm dùng kỹ thuật chính quy hóa (regularisation), tức cộng một khoản phạt vào mô hình phức tạp. Mô hình lớn chỉ được chọn khi nó thật sự dự báo tốt hơn trên dữ liệu chưa thấy. Kinh tế lượng cũng từng dùng ý này dưới tên chọn mô hình, nên ranh giới giữa hai ngành khá mờ.

The last row shows a point of Kelly and Xiu (2023): because machine learning tries many models, it overfits the data it has seen by default. To keep a model stable out of sample, practitioners use regularisation, which adds a penalty on complex models. A large model is chosen only when it forecasts better on unseen data. Econometrics has long used this idea under the name of model selection, so the border between the two fields is blurred.

Hai cách tiếp cận: dựa lý thuyết và dựa dữ liệu – *Two approaches: theory-led and data-led*

Một người dựng mô hình có thể xuất phát từ lý thuyết. Lý thuyết cho biết dạng hàm và biến cần dùng, mô hình nhỏ, giải thích được, nhưng hay sai ngoài phạm vi nó được thiết kế. Người khác xuất phát từ dữ liệu. Anh ta chấp nhận mô hình không có lời giải thích

kinh tế sẵn, miễn là dự báo chính xác hơn mốc so sánh. Phần lớn học máy tài chính hiện nay thuộc nhóm thứ hai.

A model builder can start from theory. Theory supplies the functional form and the variables, so the model is small and explainable, but it often fails outside the range it was designed for. Another builder starts from the data. He accepts a model with no ready economic explanation as long as it forecasts better than the benchmark. Most financial machine learning today belongs to the second group.

Nhóm thứ hai hay bị chê là 'không có kinh tế học'. Lời chê này chưa chắc đúng. Một cách cân bằng hơn là dùng lý thuyết cho phần đã biết rõ, và dùng học máy cho phần lý thuyết chưa nói tới. Ví dụ Nam giữ phần 'beta' (mức mã tăng giảm theo thị trường chung) theo lý thuyết, rồi để mô hình học thêm phần còn lại từ đặc trưng doanh nghiệp. Lý thuyết giảm số hệ số phải ước lượng, nên dữ liệu ít cũng chịu nổi.

The second group is often accused of having 'no economics'. The charge is not always fair. A more balanced way is to use theory for what is well understood and machine learning for what theory is silent about. For example Nam keeps the 'beta' part (how far a stock moves with the overall market) from theory and lets the model learn the rest from company features. Theory reduces the number of coefficients to estimate, so a small dataset can cope.

1.5. Thách thức khi dùng học máy trong tài chính – *1.5. Challenges of using machine learning in finance*

Học máy cho nhận diện ảnh làm tốt vì ảnh có hàng triệu mẫu, và con mèo hôm nay trông giống con mèo hôm qua. Tài chính khác ở bốn điểm. Cả bốn đều làm kết quả đẹp trên giấy thành kết quả xấu ngoài thị trường.

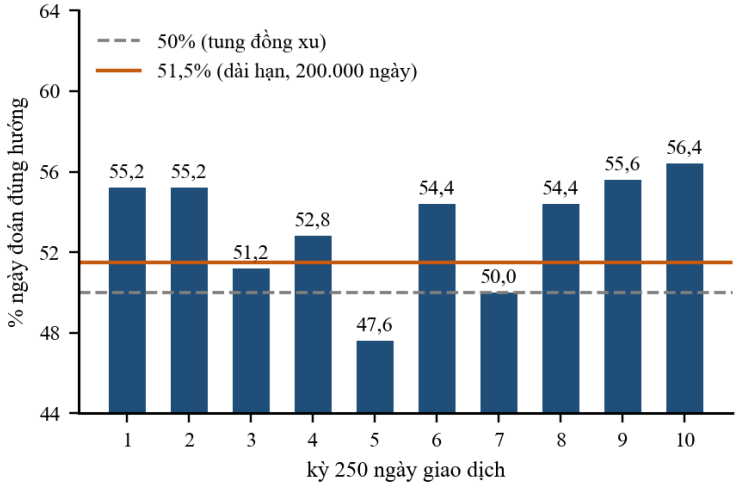
Machine learning for image recognition works well because there are millions of samples and today's cat looks like yesterday's cat. Finance differs on four points. All four turn a good result on paper into a poor one in the market.

1. Tín hiệu yếu, nhiễu lớn. Giá biến động chủ yếu do tin bất ngờ, mà tin bất ngờ thì không dự báo được. Phần dự báo được còn lại nhỏ. Giống như nghe một người thì thầm giữa chợ ồn.
   Weak signal, loud noise. Prices move mostly on surprise news, and surprises cannot be forecast. The forecastable part that remains is small. It is like hearing someone whisper in a noisy market.
2. Dữ liệu ít. Ba năm giá ngày cho khoảng 750 điểm, và chỉ có thêm khi thời gian trôi. Bạn không thể 'sinh thêm' lịch sử như sinh thêm ảnh con mèo.
   Little data. Three years of daily prices give about 750 points, and more only arrive as time passes. You cannot 'generate more' history the way you generate more cat pictures.
3. Thị trường đổi chế độ. Quy tắc đúng thời lãi suất thấp có thể sai khi lãi suất tăng. Luật, công nghệ, thói quen nhà đầu tư thay đổi và mô hình phải theo kịp.
   Changing regimes. A rule that held when interest rates were low may fail when they rise. Rules, technology and investor habits change and the model must keep up.
4. Đối thủ học theo. Một quy luật càng nhiều người dùng thì lợi nhuận từ nó càng bị chia nhỏ rồi mất. Khác với ảnh con mèo, giá phản ứng lại khi bạn khai thác nó.

Rivals learn too. The more people use a pattern, the more its profit is shared out and then lost. Unlike a cat picture, the price reacts when you exploit it.

Điểm 1 và điểm 2 cộng lại nguy hiểm hơn từng điểm riêng. Hãy xem một thí nghiệm trên dữ liệu tự sinh: tín hiệu có tác dụng thật nhưng nhỏ, và nhiễu lớn gấp 20 lần. Mỗi kỳ 250 ngày, ta đếm số ngày tín hiệu đoán đúng hướng lợi suất hôm sau. Mã nguồn ở mục 1.11, kết quả vẽ ở Hình 1.3.

Points 1 and 2 together are more dangerous than either alone. Consider an experiment on synthetic data: the signal has a real but small effect, and the noise is 20 times larger. In each 250-day period we count the days on which the signal guessed the direction of the next day's return. The code is in section 1.11 and the result is plotted in Figure 1.3.



Hình 1.3. Tỷ lệ đoán đúng hướng của một tín hiệu có thật, qua mười kỳ 250 ngày. Nhìn vào cột thấp nhất: 47,6%, thấp hơn tung đồng xu. Cột cao nhất 56,4% trông như tài năng, dù mọi kỳ dùng chính tín hiệu đó.

Figure 1.3. *The hit rate of a signal that is real, over ten periods of 250 days. Look at the lowest bar: 47.6%, below a coin toss. The highest bar, 56.4%, looks like skill even though every period uses the same signal.*

Trong dài hạn tín hiệu này chỉ đoán đúng khoảng 51,5% số ngày. Nhưng qua một kỳ 250 ngày, tỷ lệ nhảy từ 47,6% đến 56,4%. Nếu bạn chỉ có đúng một kỳ để kiểm, bạn không phân biệt nổi tín hiệu có thật với một lần may. Đây là lý do mọi kết luận ở các chương sau đều kèm kiểm tra ngoài mẫu.

In the long run this signal guesses right on only about 51.5% of days. Yet across single 250-day periods the rate swings from 47.6% to 56.4%. If you have just one period to test on, you cannot tell a real signal from one lucky stretch. That is why every conclusion in the later chapters comes with an out-of-sample check.

Một cách nhìn gọn: lấy lợi suất trung bình chia cho độ lệch chuẩn, tức mức dao động quanh trung bình. Thương số này là tỷ số Sharpe (Sharpe ratio, Sharpe 1966, khi bỏ qua lãi suất phi rủi ro). Tỷ số thấp nghĩa là đường tài sản có đi lên nhưng giật nhiều, và có thể chạm đáy trước khi lên. Học máy tài chính phải sống với tỷ số thấp.

A compact way to see it: divide the average return by the standard deviation, the size of the swings around the average. The quotient is the Sharpe ratio (Sharpe 1966, ignoring the risk-free rate). A low ratio means the wealth line rises but jerks about, and may hit bottom before it climbs. Financial machine learning has to live with low ratios.

> **LỖI HAY GẶP / COMMON MISTAKE –** Thắng một kỳ rồi kết luận mô hình tốt. Với tín hiệu yếu, một kỳ 250 ngày có thể cho tỷ lệ đúng từ dưới 48% đến trên 56%. Một kỳ không đủ để phân biệt kỹ năng với may mắn.
> *Winning one period and concluding the model is good. With a weak signal, one 250-day period can give a hit rate anywhere from below 48% to above 56%. One period is not enough to tell skill from luck.*

1.6. Lợi suất đơn, lợi suất log và lợi suất danh mục – *1.6. Simple returns, log returns and portfolio returns*

Mọi mô hình trong sách này làm việc với lợi suất (return), tức tỷ lệ tăng hoặc giảm của giá, thay cho giá thô. Lý do đơn giản: lợi suất so sánh được giữa mã giá 20.000 đồng và mã giá 200.000 đồng, còn giá thô thì không. Có hai cách tính. Cách thứ nhất là lợi suất đơn (simple return), cách thứ hai là lợi suất log (log return).

Every model in this book works with returns, the proportional rise or fall of the price, instead of raw prices. The reason is simple: returns are comparable between a stock priced at 20,000 dong and one at 200,000 dong, while raw prices are not. There are two ways to compute them. The first is the simple return, the second is the log return.

$$r_t = P_t / P_{t-1} - 1 \tag{1.2}$$

$$\ell_t = \ln(P_t / P_{t-1}) \tag{1.3}$$

P_t là giá đóng cửa ngày t, ln là logarit tự nhiên. Hai số gần nhau khi biến động nhỏ, và lệch nhau khi giá nhảy mạnh. Hãy tính tay trên năm ngày giá của mã XQZ, một mã tự đặt của Công ty cổ phần XQZ Vật liệu.

P_t is the closing price on day t and ln is the natural logarithm. The two numbers are close when moves are small and drift apart when the price jumps. Compute by hand on five days of prices for XQZ, an invented stock of XQZ Materials Joint Stock Company.

1. Ghi giá đóng cửa: 20.000; 20.400; 20.200; 20.800; 20.600; 21.000 đồng (ngày 0 đến ngày 5).
   Write the closing prices: 20,000; 20,400; 20,200; 20,800; 20,600; 21,000 dong (day 0 to day 5).
2. Lợi suất đơn ngày 1 = 20.400 ÷ 20.000 − 1 = 0,0200, tức 2,00%. Ngày 2 = 20.200 ÷ 20.400 − 1 = −0,0098. Làm tiếp ngày 3, 4, 5 theo cùng cách.
   Simple return on day 1 = 20,400 ÷ 20,000 − 1 = 0.0200, that is 2.00%. Day 2 = 20,200 ÷ 20,400 − 1 = −0.0098. Continue for days 3, 4, 5 in the same way.
3. Lợi suất log ngày 1 = ln(20.400 ÷ 20.000) = ln(1,02) = 0,0198, tức 1,98%. Làm tiếp các ngày còn lại.
   Log return on day 1 = ln(20,400 ÷ 20,000) = ln(1.02) = 0.0198, that is 1.98%. Continue for the remaining days.

4. Cộng cả năm số của mỗi cột, rồi so với kết quả thật: giá đi từ 20.000 lên 21.000, tăng 5,00%.
Add the five numbers in each column, then compare with the true result: the price went from 20,000 to 21,000, a rise of 5.00%.

Ngày *Day*	**Giá (đồng)** *Price (dong)*	**Lợi suất đơn** *Simple return*	**Lợi suất log** *Log return*
0	20.000 *20,000*		
1	20.400 *20,400*	2,00% *2.00%*	1,98% *1.98%*
2	20.200 *20,200*	−0,98% *−0.98%*	−0,99% *−0.99%*
3	20.800 *20,800*	2,97% *2.97%*	2,93% *2.93%*
4	20.600 *20,600*	−0,96% *−0.96%*	−0,97% *−0.97%*
5	21.000 *21,000*	1,94% *1.94%*	1,92% *1.92%*
Cộng *Sum*		4,97% *4.97%*	4,88% *4.88%*

Tổng lợi suất đơn là 4,97%, thấp hơn thực tế 5,00%. Tổng lợi suất log là 4,88%, và đúng bằng ln(21.000 ÷ 20.000) = ln(1,05). Đó là tính chất cần nhớ: lợi suất log cộng được qua thời gian, lợi suất đơn thì không. Muốn ghép lợi suất đơn qua nhiều ngày, bạn phải nhân từng thừa số (1 + r).
The sum of simple returns is 4.97%, short of the true 5.00%. The sum of log returns is 4.88% and equals exactly ln(21,000 ÷ 20,000) = ln(1.05). That is the property to remember: log returns add across time, simple returns do not. To chain simple returns over several days you must multiply the factors (1 + r).

$$\Sigma_t \ell_t = \ln(P_n / P_0) \tag{1.4}$$

$$1 + R = (1 + r_1) * (1 + r_2) * \ldots * (1 + r_n) \tag{1.5}$$

Một ví dụ cực đoan cho thấy vì sao. Giá đi từ 100 lên 110 rồi về 100. Lợi suất đơn là +10% rồi −9,09%, cộng lại +0,91% dù bạn không lãi đồng nào. Lợi suất log là +9,53% rồi −9,53%, cộng lại đúng 0.
An extreme example shows why. A price goes from 100 up to 110 and back to 100. The simple returns are +10% then −9.09%, which add to +0.91% although you gained nothing. The log returns are +9.53% then −9.53%, which add to exactly 0.

Lợi suất danh mục – *Portfolio return*

Danh mục (portfolio) là rổ nhiều mã, mỗi mã chiếm một phần vốn gọi là trọng số (weight). Lợi suất của danh mục trong một ngày là trung bình có trọng số của lợi suất đơn từng mã. Chú ý: với danh mục, ta cộng lợi suất đơn theo chiều ngang giữa các mã. Lợi suất log thì không cộng như vậy được, vì logarit của một tổng không bằng tổng các logarit.

A portfolio is a basket of several stocks, each taking a share of capital called its weight. A portfolio's return in one day is the weighted average of the simple returns of its stocks. Note that for a portfolio we add simple returns across stocks. Log returns cannot be added that way, because the logarithm of a sum is not the sum of the logarithms.

$$r_p = \Sigma_i\, w_i * r_i\,, \quad \Sigma_i\, w_i = 1 \tag{1.6}$$

Thêm mã JWK (giá 40.000; 39.600; 40.200; 40.800; 40.400; 41.200 đồng) và giữ trọng số 60% XQZ, 40% JWK, cân bằng lại mỗi ngày. Ngày 1: 0,6 ∗ 2,00% + 0,4 ∗ (−1,00%) = 1,20% − 0,40% = 0,80%. Bảng làm tiếp cho năm ngày.

Add the stock JWK (prices 40,000; 39,600; 40,200; 40,800; 40,400; 41,200 dong) and hold weights of 60% XQZ and 40% JWK, rebalanced every day. Day 1: 0.6 ∗ 2.00% + 0.4 ∗ (−1.00%) = 1.20% − 0.40% = 0.80%. The table continues for five days.

Ngày *Day*	**XQZ**	**JWK**	**Danh mục 60/40** *Portfolio 60/40*
1	2,00% *2.00%*	−1,00% *−1.00%*	0,80%
2	−0,98% *−0.98%*	1,52% *1.52%*	0,02% *0.02%*
3	2,97% *2.97%*	1,49% *1.49%*	2,38% *2.38%*
4	−0,96% *−0.96%*	−0,98% *−0.98%*	−0,97% *−0.97%*
5	1,94% *1.94%*	1,98% *1.98%*	1,96% *1.96%*

Nhân năm thừa số 1,0080; 1,0002; 1,0238; 0,9903; 1,0196, vốn của danh mục tăng 1,0422 lần, tức +4,22% sau năm ngày. Ngày 4 là ngày cả hai mã cùng giảm, và đó là ngày duy nhất danh mục lỗ.

Multiplying the five factors 1.0080; 1.0002; 1.0238; 0.9903; 1.0196, the portfolio's capital grows 1.0422 times, which is +4.22% after five days. Day 4 is the day both stocks fell, and it is the only losing day for the portfolio.

Bài mẫu 1: anh Phong hỏi quỹ hơn mua giữ bao nhiêu – ***Worked example 1: Phong asks how far the fund beat buy and hold***

Quỹ cổ phiếu định lượng Cánh Diều chia 60% vốn cho XQZ và 40% cho JWK, cân bằng lại mỗi ngày, và thu +4,22% trong năm ngày. Anh Phong hỏi: nếu mua rồi cứ giữ nguyên thì được bao nhiêu?

The Cánh Diều Quantitative Equity Fund puts 60% of capital in XQZ and 40% in JWK, rebalances every day and earns +4.22% in five days. Phong asks: if we had bought and simply held, what would we have made?

1. Lợi suất 5 ngày của XQZ = 21.000 ÷ 20.000 − 1 = 5,00%.
   XQZ's five-day return = 21,000 ÷ 20,000 − 1 = 5.00%.
2. Lợi suất 5 ngày của JWK = 41.200 ÷ 40.000 − 1 = 3,00%.
   JWK's five-day return = 41,200 ÷ 40,000 − 1 = 3.00%.
3. Mua giữ 60/40 từ đầu: 0,6 ∗ 5,00% + 0,4 ∗ 3,00% = 3,00% + 1,20% = 4,20%.
   Buy and hold 60/40 from the start: 0.6 ∗ 5.00% + 0.4 ∗ 3.00% = 3.00% + 1.20% = 4.20%.
4. So với quỹ: 4,22% − 4,20% = 0,02 điểm phần trăm. Cân bằng lại mỗi ngày gần như không đổi kết quả trong năm ngày, nhưng đòi hỏi mua bán hằng ngày.
   Against the fund: 4.22% − 4.20% = 0.02 percentage points. Daily rebalancing hardly changes the result over five days, yet it requires trading every day.

Nghĩa là chênh lệch 0,02 điểm phần trăm sẽ biến mất ngay khi trừ phí giao dịch của việc cân bằng lại. Chi phí giao dịch là điều anh Phong sẽ hỏi ngay sau câu này.

In other words the 0.02 percentage point gap vanishes as soon as the trading cost of rebalancing is deducted. Transaction cost is what Phong will ask about right after this question.

> **LỖI HAY GẶP / COMMON MISTAKE –** Cộng lợi suất đơn của các ngày để ra lợi suất cả kỳ. Ở bảng trên tổng là 4,97% trong khi lợi suất thật là 5,00%. Chênh lệch lớn dần khi chuỗi dài và biến động mạnh.
> *Adding the simple returns of the days to get the return of the whole period. In the table above the sum is 4.97% while the true return is 5.00%. The gap grows as the series gets longer and the swings larger.*

1.7. Các chiến lược định lượng phổ biến – *1.7. Common quantitative strategies*

Một chiến lược là một quy tắc biến dữ liệu thành lệnh. Sáu họ chiến lược hay gặp nhất. Mỗi họ đứng trên một niềm tin khác về cách giá vận động, nên cũng chịu rủi ro khác nhau.

A strategy is a rule that turns data into orders. Six families of strategy are met most often. Each stands on a different belief about how prices move, so each carries different risks.

Họ chiến lược *Family*	**Ý tưởng** *Idea*	**Dữ liệu chính** *Main data*	**Rủi ro điển hình** *Typical risk*
Theo đà *Momentum*	Mã tăng gần đây tiếp tục tăng một thời gian *Stocks that rose lately keep rising for a while*	Giá nhiều tháng *Prices over months*	Đảo chiều đột ngột làm lỗ nhanh *A sudden reversal loses quickly*
Đảo chiều *Reversal*	Mã giảm quá tay sẽ bật lại *Stocks that fell too far bounce back*	Giá vài ngày đến vài tuần *Prices over days to weeks*	Mua đúng lúc giá đang sụp tiếp *Buying just as the price keeps falling*
Giá trị *Value*	Mã rẻ so với lợi nhuận, tài sản sẽ được định giá lại *Stocks cheap against earnings or assets get re-priced*	Báo cáo tài chính *Financial statements*	Mã rẻ vì có lý do chính đáng *The stock is cheap for a good reason*
Nhân tố *Factor*	Gom mã theo đặc điểm (quy mô, giá trị, chất lượng) đem lại phần bù ổn định *Group stocks by trait (size, value, quality) for a steady premium*	Giá và báo cáo *Prices and statements*	Nhân tố nhiều năm không sinh lời *A factor earns nothing for years*
Chênh lệch thống kê *Statistical arbitrage*	Hai mã thường đi cùng nhau, lệch nhau thì chờ họp lại *Two stocks usually move together, bet on the gap closing*	Giá cặp mã *Prices of pairs*	Cặp 'gãy', hai mã tách hẳn *The pair 'breaks' and the stocks drift apart*
Theo dõi chỉ số *Index tracking*	Dựng danh mục ít mã bám sát một chỉ số *Build a few-stock portfolio that follows an index*	Giá và quyền số của chỉ số *Index prices and weights*	Sai số theo dõi, chi phí cân bằng lại *Tracking error, rebalancing cost*

Theo đà có bằng chứng sớm từ Jegadeesh và Titman (1993): các mã thắng trong vài tháng qua có xu hướng thắng tiếp vài tháng sau. Chiến lược nhân tố nối với Fama và French (1993), người gom mã theo quy mô và tỷ lệ giá trên giá trị sổ sách. Chênh lệch thống kê sẽ được dựng chi tiết ở chương 9, còn theo dõi chỉ số ở chương 8.

Momentum has early evidence from Jegadeesh and Titman (1993): stocks that won over the past few months tend to keep winning for a few months after. Factor strategies connect to Fama and French (1993), who grouped stocks by size and by the ratio of price to book value. Statistical arbitrage is built in detail in Chapter 9 and index tracking in Chapter 8.

Cặp theo đà và đảo chiều cho thấy một điều đáng nhớ: hai niềm tin ngược nhau đều có thể đúng, ở hai khung thời gian khác nhau. Vì thế, câu hỏi đúng trước khi dùng chiến lược là 'chiến lược này kiếm tiền từ ai, và vì sao họ chịu để mất?'.

The pair of momentum and reversal shows a point worth remembering: two opposite beliefs can both be right, on two different time frames. So the right question before using a strategy is 'who does this strategy earn from, and why do they accept the loss?'.

1.8. Sản phẩm tài chính – ngân hàng dùng mô hình định lượng – *1.8. Financial and banking products that use quantitative models*

Mô hình định lượng không nằm riêng trong phòng nghiên cứu. Nó đi vào nhiều sản phẩm mà khách hàng mua hằng ngày. Bảng sau liệt kê năm sản phẩm và chỗ mô hình tham gia.
Quantitative models do not sit alone in the research room. They enter many products customers buy every day. The table lists five products and where the models take part.

Sản phẩm *Product*	**Mô hình làm gì** *What the model does*	**Ví dụ ở Cánh Diều** *Example at Cánh Diều*
Quỹ chỉ số, quỹ hoán đổi danh mục (ETF, exchange-traded fund) *Index fund, exchange-traded fund (ETF)*	Chọn ít mã để bám sát chỉ số, ước lượng sai số theo dõi *Pick few stocks to follow the index, estimate tracking error*	Quỹ bám một chỉ số cổ phiếu *A fund following an equity index*
Quỹ định lượng *Quantitative fund*	Dự báo lợi suất, chia tỷ trọng theo rủi ro *Forecast returns, size weights by risk*	Quỹ cổ phiếu định lượng Cánh Diều *Cánh Diều Quantitative Equity Fund*
Danh mục ủy thác *Discretionary mandate*	Dựng danh mục riêng cho mỗi khách theo ràng buộc của họ *Build a separate portfolio per client under their constraints*	Khách không muốn giữ mã ngành năng lượng *A client who refuses energy stocks*
Quỹ hưu trí *Pension fund*	Cân đối tài sản với nghĩa vụ chi trả dài hạn *Match assets to long-dated payment obligations*	Quỹ hưu trí nội bộ của Ngân hàng TMCP Cánh Diều *Internal pension fund of Cánh Diều Commercial Bank*
Danh mục trái phiếu của ngân hàng *Bank bond portfolio*	Đo rủi ro lãi suất, chọn kỳ hạn *Measure interest-rate risk, choose maturities*	Danh mục trái phiếu của Ngân hàng TMCP Cánh Diều *Bond portfolio of Cánh Diều Commercial Bank*

Ứng dụng ở quỹ, công ty chứng khoán và ngân hàng – *Applications at funds, brokerages and banks*

Ở quỹ, nhóm của chị Hà dựng mô hình dự báo, chọn trọng số, rồi báo cho anh Phong số liệu rủi ro mỗi sáng. Ở công ty chứng khoán, bộ phận tự doanh dùng quy tắc tự động để đặt lệnh bằng vốn công ty, nên cần giới hạn lỗ trong ngày. Ở ngân hàng, phòng đầu tư dùng mô hình để chọn kỳ hạn trái phiếu cho danh mục và để chia vốn của quỹ hưu trí nội bộ.

At the fund, Hà's team builds the forecasting models, chooses the weights and sends Phong risk figures each morning. At a brokerage, the proprietary desk uses automatic rules to place orders with the firm's own capital, so it needs a daily loss limit. At a bank, the investment department uses models to pick bond maturities for its portfolio and to split the capital of the internal pension fund.

Anh Phong có ba câu hỏi cố định cho mọi mô hình: có quá khớp không, lỗ tối đa bao nhiêu, chi phí giao dịch tính chưa. Câu 1 sẽ được trả lời bằng kiểm định ngoài mẫu ở chương 5 và 10. Câu 2 cần thước đo mức lỗ tối đa ở chương 6 và 7. Câu 3 cần phần chi phí ở chương 8.

Phong has three fixed questions for every model: does it overfit, what is the maximum loss, has the transaction cost been counted. Question 1 is answered by out-of-sample validation in Chapters 5 and 10. Question 2 needs the maximum-loss measures of Chapters 6 and 7. Question 3 needs the cost section of Chapter 8.

Lập luận ủng hộ *Arguments for*	**Lập luận phản đối** *Arguments against*
Xử lý nhiều mã và nhiều loại dữ liệu cùng lúc *Handles many stocks and many types of data at once*	Tín hiệu yếu nên dễ mất trắng nếu quá khớp *The signal is weak, so an overfitted model can lose everything*
Kiểm lại được quy tắc trên quá khứ trước khi bỏ tiền *Rules can be tested on the past before money is committed*	Quá khứ không lặp lại khi thị trường đổi chế độ *The past does not repeat when the market changes regime*
Giảm thiên kiến cảm xúc của từng người giao dịch *Reduces the emotional bias of individual traders*	Nhiều quỹ cùng dùng một ý thì lợi nhuận bị chia nhỏ *When many funds use one idea the profit is shared out*

Cột trái không loại bỏ được cột phải. Người dùng mô hình phải ghi cả hai vào hồ sơ trước khi dùng vốn thật.

The left column does not cancel the right one. The people who use the model must record both in the file before committing real capital.

1.9. Quy tắc của sàn ảnh hưởng thế nào đến mô hình – *1.9. How exchange rules affect a model*

Mô hình trên giấy cho phép mua 3.738,3 cổ phiếu ở giá 26.750 đồng. Sàn thì không. Các quy tắc sau lấy từ tài liệu hướng dẫn của Sở Giao dịch Chứng khoán Thành phố Hồ Chí Minh (HOSE, Ho Chi Minh Stock Exchange) công bố tháng 2 năm 2026. Theo Điều 4 của Thông tư giao dịch chứng khoán (hợp nhất 08/VBHN-BTC năm 2026), biên độ dao động giá do Sở Giao dịch Chứng khoán Việt Nam quyết định sau khi Ủy ban Chứng khoán Nhà nước (UBCK) chấp thuận. Vì quy tắc có thể được điều chỉnh, bạn nên kiểm văn bản hiện hành trước khi dùng cho hệ thống thật.

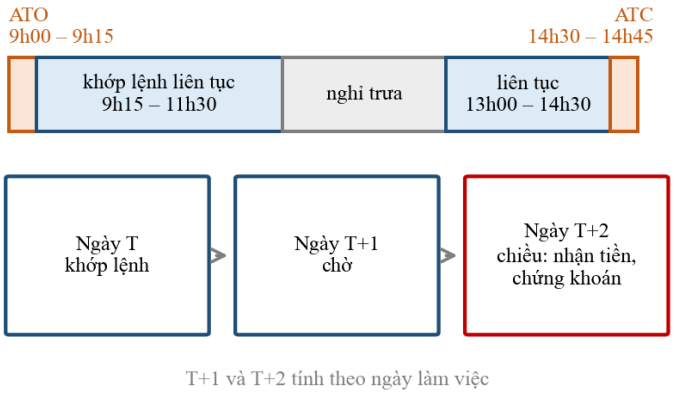
A model on paper may buy 3,738.3 shares at 26,750 dong. The exchange will not allow that. The rules that follow come from the guidance document of the Ho Chi Minh Stock Exchange (HOSE)

published in February 2026. Under Article 4 of the Securities Trading Circular (consolidated 08/VBHN-BTC, 2026), the price band is decided by the Vietnam Stock Exchange after the State Securities Commission (SSC) approves. Because the rules may be adjusted, check the current text before using them in a real system.

Các chữ viết tắt dùng ở bảng sau: ATO là lệnh xác định giá mở cửa (at-the-open), ATC là lệnh xác định giá đóng cửa (at-the-close), LO là lệnh giới hạn (limit order, chỉ khớp ở giá bạn đặt hoặc tốt hơn), MTL là lệnh thị trường chuyển về lệnh giới hạn (market-to-limit). T+2 nghĩa là hai ngày làm việc sau ngày giao dịch T.

The abbreviations used in the next table: ATO is the at-the-open order, ATC is the at-the-close order, LO is the limit order (it executes only at your price or better) and MTL is the market-to-limit order. T+2 means two working days after the trading day T.

Quy tắc *Rule*	**Nội dung tại HOSE** *Content at HOSE*	**Ảnh hưởng đến mô hình** *Effect on the model*
Biên độ giá *Price band*	±7% (cổ phiếu, chứng chỉ quỹ đóng, ETF); ±20% ngày đầu hoặc sau tạm ngừng từ 25 ngày *±7% (shares, closed-end fund certificates, ETFs); ±20% on the first day or after a halt of 25 days or more*	Lợi suất ngày bị chặn, đuôi phân phối bị cắt *Daily returns are capped and the tail is cut*
Đơn vị giao dịch *Trading unit*	Lô chẵn 100 cổ phiếu, tối đa 500.000 cổ phiếu mỗi lệnh; lô lẻ 1 đến 99 *Board lot 100 shares, at most 500,000 shares per order; odd lot 1 to 99*	Trọng số thật lệch trọng số mô hình *Actual weights differ from model weights*
Bước giá *Price step*	Dưới 10.000 đồng: 10 đồng; 10.000 đến 49.950: 50 đồng; từ 50.000: 100 đồng *Below 10,000 dong: 10 dong; 10,000 to 49,950: 50 dong; from 50,000: 100 dong*	Giá đặt phải là bội của bước giá *An order price must be a multiple of the step*
Phiên giao dịch *Sessions*	ATO 9h00 – 9h15; liên tục 9h15 – 11h30 và 13h00 – 14h30; ATC 14h30 – 14h45 *ATO 9:00 – 9:15; continuous 9:15 – 11:30 and 13:00 – 14:30; ATC 14:30 – 14:45*	Giá đóng cửa hình thành ở phiên ATC *The closing price forms in the ATC session*
Thanh toán *Settlement*	Nhận tiền và chứng khoán vào chiều ngày T+2 *Cash and securities received on the afternoon of day T+2*	Vốn và cổ phiếu về chậm so với lệnh *Cash and shares arrive later than the order*
Loại lệnh *Order types*	LO, ATO, ATC, MTL *LO, ATO, ATC, MTL*	Quyết định giá khớp thực tế *Decide the real execution price*



Hình 1.4. Một ngày giao dịch ở HOSE và vòng thanh toán T+2. Nhìn hàng trên: hai ô cam là phiên định kỳ, hai ô xanh là khớp liên tục. Hàng dưới: tiền và chứng khoán chỉ về chiều ngày T+2.

Figure 1.4. *A trading day at HOSE and the T+2 settlement cycle. On the top row the two orange boxes are the periodic sessions and the two blue boxes are continuous trading. On the bottom row, cash and securities arrive only on the afternoon of day T+2.*

$$\text{Giá trần} = \text{giá tham chiếu} * (1 + b)\,, \quad \text{Giá sàn} = \text{giá tham chiếu} * (1 - b) \qquad (1.7)$$

Ở đây b là biên độ, ví dụ 0,07. Với tham chiếu 25.000 đồng, giá trần là 25.000 ∗ 1,07 = 26.750 và giá sàn là 25.000 ∗ 0,93 = 23.250. Cả hai đều là bội của bước giá 50 đồng, nên đặt lệnh được.

Here b is the band, for example 0.07. With a reference price of 25,000 dong, the ceiling is 25,000 ∗ 1.07 = 26,750 and the floor is 25,000 ∗ 0.93 = 23,250. Both are multiples of the 50 dong step, so orders can be placed there.

Bài mẫu 2: đổi trọng số của mô hình thành lệnh thật – ***Worked example 2: turning a model weight into a real order***

Quỹ có vốn 2 tỷ đồng. Mô hình muốn dành 5% cho mã ZVB, giá đang là 26.750 đồng (đúng giá trần). Nam phải đổi trọng số thành lệnh.

The fund has 2 billion dong of capital. The model wants 5% for the stock ZVB, whose price is 26,750 dong (exactly at the ceiling). Nam must turn the weight into an order.

1. Số tiền mục tiêu = 2.000.000.000 ∗ 5% = 100.000.000 đồng.
   Target amount = 2,000,000,000 ∗ 5% = 100,000,000 dong.
2. Số cổ phiếu mong muốn = 100.000.000 ÷ 26.750 = 3.738,3. Lấy phần nguyên: 3.738.
   Shares wanted = 100,000,000 ÷ 26,750 = 3,738.3. Take the whole part: 3,738.
3. Tách lô: 3.738 gồm 3.700 (37 lô chẵn) và 38 lô lẻ.
   Split into lots: 3,738 is 3,700 (37 board lots) plus 38 odd-lot shares.
4. Nếu chỉ đặt phần lô chẵn: 3.700 ∗ 26.750 = 98.975.000 đồng, tức 4,949% vốn, thay vì 5%.
   If only the board-lot part is ordered: 3,700 ∗ 26,750 = 98,975,000 dong, which is 4.949% of capital instead of 5%.

Nghĩa là mô hình bị lệch 0,051 điểm phần trăm chỉ vì lô. Với danh mục 2 tỷ đồng sai lệch này nhỏ. Với danh mục 20 triệu đồng thì cùng mã đó không mua nổi một lô chẵn, và mô hình phải đổi cách chia vốn.

In other words the model is off by 0.051 percentage points just because of lots. For a 2 billion dong portfolio the gap is small. For a 20 million dong portfolio the same stock cannot buy even one board lot, and the model must change how it splits the capital.

> **LỖI HAY GẶP / COMMON MISTAKE –** Bỏ qua biên độ, lô và bước giá khi kiểm định. Lệnh đặt 26.800 đồng khi trần là 26.750 đồng bị từ chối, nên kết quả kiểm định tính khớp lệnh ở giá đó là ảo.
> *Ignoring the band, lot and price step when testing. An order at 26,800 dong when the ceiling is 26,750 dong is rejected, so a back-test that counts a fill at that price is illusory.*

1.10. Pháp lý và vị trí việc làm – *1.10. Law and jobs*

Điều 12 Luật Chứng khoán liệt kê các hành vi bị cấm, trong đó có gian lận và công bố thông tin sai, giao dịch nội gián, thao túng thị trường, dùng tài khoản của khách hàng trái phép. Hành vi đó vẫn bị cấm khi lệnh do chương trình tự động phát ra. Một chương trình đặt rồi hủy hàng loạt lệnh chỉ để tạo ấn tượng cung cầu giả có nguy cơ thuộc nhóm thao túng, nên người viết chiến lược phải hỏi bộ phận pháp chế từ khi thiết kế. Sách này không kết luận cho từng trường hợp.

Article 12 of the Securities Law lists prohibited acts, including fraud and false disclosure, insider trading, market manipulation and unauthorised use of a client's account. The acts remain prohibited when the orders are issued by an automatic program. A program that places and cancels orders in bulk only to create a false impression of supply and demand risks falling into manipulation, so whoever writes the strategy must ask the legal department from the design stage. This book does not decide individual cases.

Luật Trí tuệ nhân tạo số 134/2025 có hiệu lực từ ngày 01 tháng 3 năm 2026. Điều 9 chia hệ thống AI thành ba mức rủi ro: cao, trung bình và thấp. Hệ thống giao dịch của quỹ thuộc mức nào phụ thuộc tiêu chí trong Điều 9 và các văn bản hướng dẫn. Bộ phận pháp chế của quỹ cần xác định, và kết quả quyết định nghĩa vụ đi kèm như minh bạch hay báo cáo sự cố.

The Law on Artificial Intelligence No. 134/2025 takes effect on 1 March 2026. Article 9 divides AI systems into three risk levels: high, medium and low. Which level a fund's trading system falls into depends on the criteria in Article 9 and its guidance. The fund's legal department must determine it, and the result decides the accompanying duties such as transparency or incident reporting.

Về nghề nghiệp, bốn vị trí sau xuất hiện ở hầu hết tổ chức làm giao dịch định lượng.

As for careers, the four positions that follow appear in most organisations that do quantitative trading.

Vị trí *Position*	**Việc hằng ngày** *Daily work*	**Kỹ năng cần** *Skills needed*	**Học ở** *Studied in*
Phân tích định lượng *Quantitative analyst*	Dựng và kiểm tra mô hình dự báo, đo rủi ro *Build and test forecasting models, measure risk*	Thống kê, Python *Statistics, Python*	Chương 3 đến 5 *Chapters 3 to 5*
Phát triển chiến lược *Strategy developer*	Biến ý tưởng thành quy tắc, chạy kiểm định ngược *Turn ideas into rules, run back-tests*	Danh mục, kiểm định *Portfolios, validation*	Chương 6 đến 10 *Chapters 6 to 10*
Kỹ sư dữ liệu *Data engineer*	Thu thập, làm sạch, lưu và cung cấp dữ liệu giá *Collect, clean, store and serve price data*	Lập trình, cơ sở dữ liệu *Programming, databases*	Chương 2 *Chapter 2*
Quản trị rủi ro mô hình *Model risk manager*	Thẩm định độc lập, đặt giới hạn, theo dõi mô hình *Validate independently, set limits, monitor models*	Thống kê, pháp lý *Statistics, regulation*	Chương 10 *Chapter 10*

1.11. Thực hành bằng Python – *1.11. Practising with Python*

Các đoạn mã sau chỉ dùng thư viện chuẩn của Python, nên chạy trên máy nào cũng được. Dữ liệu là dữ liệu tự sinh, không phải giá thật.

The code blocks that follow use only Python's standard library, so they run on any machine. The data is synthetic, not real prices.

Đoạn mã 1: lợi suất đơn, log và danh mục – *Code 1: simple, log and portfolio returns*

Đoạn mã làm lại đúng phép tính tay của mục 1.6 trên XQZ và JWK. Số in ra phải khớp bảng.

The code repeats the hand calculation of section 1.6 on XQZ and JWK. The numbers printed must match the tables.

```
import math

xqz = [20000, 20400, 20200, 20800, 20600, 21000]
jwk = [40000, 39600, 40200, 40800, 40400, 41200]

def simple(p):
    return [p[i] / p[i - 1] - 1 for i in range(1, len(p))]

def logret(p):
    return [math.log(p[i] / p[i - 1]) for i in range(1, len(p))]

s, g = simple(xqz), logret(xqz)
print("simple %:", [round(x * 100, 2) for x in s])
print("log %   :", [round(x * 100, 2) for x in g])
grow = 1.0
for x in s:
    grow *= 1 + x
```

```
print("sum simple %.4f | compounded %.4f" % (sum(s), grow - 1))
print("sum log %.4f | ln(P5/P0) %.4f" % (sum(g), math.log(21000 / 20000)))


w = [0.6, 0.4]
r2 = simple(jwk)
port = [w[0] * a + w[1] * b for a, b in zip(s, r2)]
print("portfolio %:", [round(x * 100, 2) for x in port])
wealth = 1.0
for x in port:
    wealth *= 1 + x
print("wealth after 5 days: %.4f" % wealth)
```

Kết quả chạy / *Output*

```
simple %: [2.0, -0.98, 2.97, -0.96, 1.94]
log %   : [1.98, -0.99, 2.93, -0.97, 1.92]
sum simple 0.0497 | compounded 0.0500
sum log 0.0488 | ln(P5/P0) 0.0488
portfolio %: [0.8, 0.02, 2.38, -0.97, 1.96]
wealth after 5 days: 1.0422
```

Hai dòng cuối của khối lợi suất cho thấy cộng lợi suất đơn ra 0,0497 còn nhân dồn ra 0,0500. Cộng lợi suất log ra 0,0488, trùng ln(P5 ÷ P0). Vốn danh mục sau năm ngày là 1,0422.

The last lines of the return block show that adding simple returns gives 0.0497 while compounding gives 0.0500. Adding log returns gives 0.0488, matching ln(P5 ÷ P0). The portfolio's capital after five days is 1.0422.

Đoạn mã 2: ba mã, 750 ngày, đa dạng hóa – ***Code 2: three stocks, 750 days, diversification***

Đoạn mã sinh lợi suất ngày cho ba mã có chung một biến động thị trường, mỗi mã có độ nhạy (beta) khác nhau. Sau đó nó năm hóa: lợi suất ngày ∗ 252 và độ lệch chuẩn ngày ∗ √252, với 252 là số ngày giao dịch ước tính trong năm.

The code generates daily returns for three stocks that share one market move, each with a different sensitivity (beta). It then annualises: daily return ∗ 252 and daily standard deviation ∗ √252, where 252 is the estimated number of trading days in a year.

```
import math
import random


random.seed(7)
names = ["XQZ", "JWK", "ZVB"]
beta = [1.0, 0.8, 1.2]
days = 750
rets = [[], [], []]
for _ in range(days):
    m = random.gauss(0.0004, 0.008)          # common market move
    for i in range(3):
        e = random.gauss(0.0, 0.010 + 0.002 * i)
        rets[i].append(beta[i] * m + e)


def mean(x):
    return sum(x) / len(x)


def std(x):
    mu = mean(x)
    return math.sqrt(sum((v - mu) ** 2 for v in x) / (len(x) - 1))


def show(tag, x):
    print(tag, "mean %.2f%%/yr" % (mean(x) * 252 * 100),
```

```
          "vol %.1f%%/yr" % (std(x) * math.sqrt(252) * 100))

for i, k in enumerate(names):
    show(k, rets[i])
port = [sum(rets[i][t] for i in range(3)) / 3 for t in range(days)]
show("EW ", port)
wealth = 1.0
for x in port:
    wealth *= 1 + x
print("equal-weight wealth after %d days: %.3f" % (days, wealth))
```

Kết quả chạy / *Output*

```
XQZ mean 9.58%/yr vol 20.7%/yr
JWK mean 13.52%/yr vol 21.1%/yr
ZVB mean 14.41%/yr vol 27.2%/yr
EW  mean 12.50%/yr vol 17.2%/yr
equal-weight wealth after 750 days: 1.389
```

Danh mục chia đều có biến động 17,2% một năm, thấp hơn cả mã ít biến động nhất (XQZ, 20,7%) và thấp hơn mức trung bình 23,0% của ba mã. Đây là đa dạng hóa: các mã không tăng giảm đúng cùng lúc nên dao động triệt tiêu một phần. Dữ liệu tự sinh với hạt giống 7, kết quả thật sẽ khác, và con số lợi suất trung bình ở đây không dự báo gì.

The equal-weight portfolio has a volatility of 17.2% a year, below even the least volatile stock (XQZ, 20.7%) and below the 23.0% average of the three. That is diversification: the stocks do not rise and fall at exactly the same time, so part of the swings cancels. The data is synthetic with seed 7, real results will differ, and the average return figures here forecast nothing.

Đoạn mã 3: tín hiệu yếu trong nhiễu lớn – *Code 3: a weak signal in loud noise*

Đoạn mã tạo tín hiệu x và lợi suất hôm sau r = 0,0005 ∗ x + nhiễu, nhiễu có độ lệch chuẩn 0,010, tức gấp 20 lần hệ số tín hiệu. Nó đếm số ngày dấu của x trùng dấu của r trong mười kỳ 250 ngày, rồi trong một kỳ dài 200.000 ngày.

The code creates a signal x and the next-day return r = 0.0005 ∗ x + noise, with noise of standard deviation 0.010, which is 20 times the signal coefficient. It counts the days on which the sign of x matches the sign of r over ten periods of 250 days, then over one long period of 200,000 days.

```
import random

random.seed(11)

def one_period(days=250):
    hits = 0
    sxy = sxx = syy = 0.0
    for _ in range(days):
        x = random.gauss(0, 1)                      # signal
        r = 0.0005 * x + random.gauss(0, 0.010)     # next-day return
        hits += (x > 0) == (r > 0)
        sxy += x * r
        sxx += x * x
        syy += r * r
    return hits / days, sxy * sxy / (sxx * syy)

res = [one_period() for _ in range(10)]
hit = [round(h * 100, 1) for h, _ in res]
print("hit rate per 250-day period (%):")
print(hit)
print("min %.1f  max %.1f" % (min(hit), max(hit)))
```

```
h, r2 = one_period(200000)
print("200000 days: hit rate %.1f%%, R2 %.2f%%" % (h * 100, r2 * 100))
```

Kết quả chạy / *Output*

```
hit rate per 250-day period (%):
[55.2, 55.2, 51.2, 52.8, 47.6, 54.4, 50.0, 54.4, 55.6, 56.4]
min 47.6  max 56.4
200000 days: hit rate 51.5%, R2 0.24%
```

Mười kỳ cho tỷ lệ đúng từ 47,6% đến 56,4%, còn kỳ dài cho 51,5% với R^2 (phần biến động của lợi suất mà tín hiệu giải thích được) chỉ 0,24%. Nghĩa là tín hiệu có thật nhưng chỉ hơn đồng xu khoảng 1,5 điểm phần trăm. Ngoài đời, mức đó dễ biến mất sau chi phí giao dịch.

The ten periods give hit rates from 47.6% to 56.4%, while the long run gives 51.5% with an R^2 (the share of the return's variation that the signal explains) of only 0.24%. So the signal is real but beats a coin by only about 1.5 percentage points. In real markets that margin can disappear after transaction costs.

Đoạn mã 4: quy tắc sàn – *Code 4: exchange rules*

Đoạn mã tính giá trần, sàn, bước giá, tách lô cho bài mẫu 2 và tìm ngày nhận tiền khi bỏ qua ngày nghỉ lễ. Hàm settle chỉ nhảy qua thứ Bảy và Chủ nhật.

The code computes ceiling, floor and price step, splits lots for worked example 2 and finds the cash date, ignoring public holidays. The function settle only skips Saturday and Sunday.

```
import datetime

def tick(p):
    return 10 if p < 10000 else 50 if p < 50000 else 100

for ref in (25000, 10000, 60000):
    floor, ceil = ref * 93 // 100, ref * 107 // 100
    print(ref, "floor", floor, "ceiling", ceil, "tick", tick(ref))

capital, weight, price = 2_000_000_000, 0.05, 26750
shares = int(capital * weight // price)
board = shares // 100 * 100
print("wanted", shares, "board lot", board, "odd lot", shares - board)
print("weight reached: %.3f%%" % (board * price / capital * 100))

def settle(day, n=2):
    while n:
        day += datetime.timedelta(days=1)
        if day.weekday() < 5:
            n -= 1
    return day

t = datetime.date(2026, 11, 5)
print(t.strftime("%a %d/%m"), "->", settle(t).strftime("%a %d/%m"))
```

Kết quả chạy / *Output*

```
25000 floor 23250 ceiling 26750 tick 50
10000 floor 9300 ceiling 10700 tick 50
60000 floor 55800 ceiling 64200 tick 100
wanted 3738 board lot 3700 odd lot 38
weight reached: 4.949%
Thu 05/11 -> Mon 09/11
```

Ba dòng đầu khớp công thức (1.7) với tham chiếu 25.000 đồng. Hai dòng sau khớp bài mẫu 2: 3.738 cổ phiếu mong muốn, 3.700 lô chẵn, 38 lô lẻ, đạt 4,949% vốn. Dòng cuối cho thấy lệnh mua thứ Năm 05/11 được nhận chiều thứ Hai 09/11.

The first three lines match formula (1.7) with a reference price of 25,000 dong. The next two match worked example 2: 3,738 shares wanted, 3,700 in board lots, 38 odd lots, reaching 4.949% of capital. The last line shows that an order bought on Thursday 05/11 is received on the afternoon of Monday 09/11.

> **PHẢI THUỘC / MEMORISE –** Lợi suất log cộng được qua thời gian, lợi suất đơn cộng được giữa các mã trong danh mục. Giá thô không so sánh được giữa các mã. Tín hiệu yếu cần nhiều kỳ kiểm, và mô hình phải qua được lô, bước giá, biên độ, T+2 của sàn.
> *Log returns add across time, simple returns add across the stocks of a portfolio. Raw prices cannot be compared between stocks. A weak signal needs many test periods, and a model must pass the lot, price step, band and T+2 of the exchange.*

Bảng thuật ngữ – *Glossary*

Thuật ngữ *Term (Vietnamese)*	**Tiếng Anh** *English*	**Giải thích ngắn** *Short definition*
giao dịch định lượng	quantitative trading	Giao dịch theo quy tắc bằng số, chạy bằng mã, kiểm lại được trên dữ liệu. *Trading by numeric rules run by code and testable on data.*
giao dịch theo cảm tính	discretionary trading	Người giao dịch tự đọc tình hình rồi quyết định. *The trader reads the situation and decides personally.*
trí tuệ nhân tạo (AI)	artificial intelligence	Khái niệm rộng về máy làm việc cần trí thông minh; học máy là một phần. *The broad idea of machines doing work that needs intelligence; machine learning is a part.*
học máy	machine learning	Phương pháp để máy tự rút quan hệ từ dữ liệu. *Methods in which a machine draws relationships from data.*
kinh tế lượng	econometrics	Ngành dùng thống kê để kiểm giả thuyết kinh tế. *The field that uses statistics to test economic hypotheses.*
dạng hàm	functional form	Hình dáng của quan hệ giữa biến vào và biến ra. *The shape of the link between input and output variables.*
chính quy hóa	regularisation	Phạt mô hình phức tạp để giữ ổn định ngoài mẫu. *Penalising complex models to keep them stable out of sample.*

Thuật ngữ *Term (Vietnamese)*	**Tiếng Anh** *English*	**Giải thích ngắn** *Short definition*
quá khớp	overfitting	Mô hình học cả nhiễu nên tốt trên dữ liệu cũ, kém trên dữ liệu mới. *A model that learns the noise, good on old data, poor on new data.*
ngoài mẫu	out of sample	Dữ liệu không dùng khi xây mô hình. *Data not used when the model was built.*
tín hiệu	signal	Phần của dữ liệu giúp dự báo được lợi suất. *The part of the data that helps forecast returns.*
nhiễu	noise	Phần biến động không dự báo được. *The part of the variation that cannot be forecast.*
độ lệch chuẩn	standard deviation	Mức dao động quanh giá trị trung bình. *The size of swings around the average.*
tỷ số Sharpe	Sharpe ratio	Lợi suất trung bình chia cho độ lệch chuẩn. *Average return divided by standard deviation.*
lợi suất đơn	simple return	Giá mới chia giá cũ rồi trừ 1. *New price divided by old price, minus 1.*
lợi suất log	log return	Logarit tự nhiên của giá mới chia giá cũ. *Natural logarithm of new price divided by old price.*
danh mục	portfolio	Rổ nhiều mã, mỗi mã một phần vốn. *A basket of several stocks, each with a share of capital.*
trọng số	weight	Phần vốn dành cho một mã trong danh mục. *The share of capital placed in one stock of the portfolio.*
chế độ thị trường	market regime	Giai đoạn thị trường có cách vận động riêng. *A market period with its own way of moving.*
kiểm định ngược	backtesting	Chạy quy tắc giao dịch trên dữ liệu quá khứ. *Running a trading rule on past data.*
đa dạng hóa	diversification	Chia vốn cho nhiều mã để dao động triệt tiêu một phần. *Spreading capital over many stocks so swings partly cancel.*
biên độ giá	price band	Mức tăng giảm tối đa của giá trong một ngày so với giá tham chiếu. *The largest daily rise or fall from the reference price.*
lô chẵn	board lot	Đơn vị giao dịch 100 cổ phiếu ở HOSE. *The trading unit of 100 shares at HOSE.*
bước giá	price step	Khoảng nhảy nhỏ nhất của giá đặt lệnh. *The smallest jump in an order price.*

Thuật ngữ *Term (Vietnamese)*	**Tiếng Anh** *English*	**Giải thích ngắn** *Short definition*
T+2	T+2 settlement	Nhận tiền và chứng khoán sau hai ngày làm việc. *Receiving cash and securities two working days after the trade.*

1.12. Bài tập vận dụng có lời giải – *Applied exercises with solutions*

Bài 1. Lợi suất đơn và lợi suất log của mã ZVB – ***Exercise 1. Simple and log returns of ZVB***

Mã ZVB đóng cửa ngày 0 là 30.000 đồng, ngày 1 là 30.600 đồng, ngày 2 là 30.000 đồng, ngày 3 là 31.500 đồng. (a) Tính lợi suất đơn từng ngày. (b) Cộng ba lợi suất đó và so với lợi suất thật của cả kỳ. (c) Tính lợi suất log từng ngày và tổng của chúng. (d) Nhận xét về ngày 1 và ngày 2.

The stock ZVB closes at 30,000 dong on day 0, 30,600 dong on day 1, 30,000 dong on day 2 and 31,500 dong on day 3. (a) Compute the simple return for each day. (b) Add the three returns and compare with the true return of the whole period. (c) Compute the log return for each day and their sum. (d) Comment on days 1 and 2.

1. Lợi suất đơn: ngày 1 = 30.600 ÷ 30.000 − 1 = 2,00%; ngày 2 = 30.000 ÷ 30.600 − 1 = −1,96%; ngày 3 = 31.500 ÷ 30.000 − 1 = 5,00%.
   Simple returns: day 1 = 30,600 ÷ 30,000 − 1 = 2.00%; day 2 = 30,000 ÷ 30,600 − 1 = −1.96%; day 3 = 31,500 ÷ 30,000 − 1 = 5.00%.
2. Tổng ba số = 2,00% − 1,96% + 5,00% = 5,04%. Lợi suất thật cả kỳ = 31.500 ÷ 30.000 − 1 = 5,00%. Cộng đơn lệch 0,04 điểm phần trăm.
   The sum = 2.00% − 1.96% + 5.00% = 5.04%. The true return of the period = 31,500 ÷ 30,000 − 1 = 5.00%. Adding simple returns is off by 0.04 percentage points.
3. Lợi suất log: ngày 1 = ln(1,02) = 1,98%; ngày 2 = ln(30.000 ÷ 30.600) = −1,98%; ngày 3 = ln(1,05) = 4,88%.
   Log returns: day 1 = ln(1.02) = 1.98%; day 2 = ln(30,000 ÷ 30,600) = −1.98%; day 3 = ln(1.05) = 4.88%.
4. Tổng log = 1,98% − 1,98% + 4,88% = 4,88% = ln(31.500 ÷ 30.000). Cộng log khớp lợi suất cả kỳ.
   The log sum = 1.98% − 1.98% + 4.88% = 4.88% = ln(31,500 ÷ 30,000). Adding logs matches the return of the period.
5. Đọc kết quả: ngày 1 giá lên rồi ngày 2 về chỗ cũ. Lợi suất log đối nhau (+1,98% và −1,98%), lợi suất đơn thì không (+2,00% và −1,96%).
   Reading the result: the price rises on day 1 and returns to its old level on day 2. The log returns are opposite (+1.98% and −1.98%), the simple returns are not (+2.00% and −1.96%).

Bài 2. Danh mục 70/30 trong hai ngày không cân bằng lại – ***Exercise 2. A 70/30 portfolio over two days without rebalancing***

Quỹ chia vốn 70% cho mã A và 30% cho mã B. Ngày 1: A tăng 3,00%, B giảm 2,00%. Ngày 2: A giảm 1,00%, B tăng 2,00%. Quỹ không cân bằng lại giữa hai ngày. Tính: (a) lợi suất danh mục ngày 1; (b) tỷ trọng mới của A và B cuối ngày 1; (c) lợi suất danh mục ngày 2; (d) vốn cuối ngày 2 nếu vốn đầu là 1; (e) lợi suất ngày 2 nếu cân bằng lại về 70/30.

The fund puts 70% of capital in stock A and 30% in stock B. Day 1: A rises 3.00%, B falls 2.00%. Day 2: A falls 1.00%, B rises 2.00%. The fund does not rebalance between the two days. Compute: (a) the portfolio return on day 1; (b) the new weights of A and B at the end of day 1; (c) the portfolio return on day 2; (d) the capital at the end of day 2 if the initial capital is 1; (e) the day 2 return if the fund rebalances to 70/30.

1. Ngày 1: $0{,}7 * 3{,}00\% + 0{,}3 * (-2{,}00\%) = 2{,}10\% - 0{,}60\% = 1{,}50\%$.
   Day 1: $0.7 * 3.00\% + 0.3 * (-2.00\%) = 2.10\% - 0.60\% = 1.50\%$.
2. Giá trị cuối ngày 1: A = $0{,}7 * 1{,}03 = 0{,}7210$; B = $0{,}3 * 0{,}98 = 0{,}2940$; tổng 1,0150. Tỷ trọng: A = $0{,}7210 \div 1{,}0150 = 71{,}03\%$; B = 28,97%.
   Values at the end of day 1: A = $0.7 * 1.03 = 0.7210$; B = $0.3 * 0.98 = 0.2940$; total 1.0150. Weights: A = $0.7210 \div 1.0150 = 71.03\%$; B = 28.97%.
3. Ngày 2: $0{,}7103 * (-1{,}00\%) + 0{,}2897 * 2{,}00\% = -0{,}7103\% + 0{,}5794\% = -0{,}13\%$.
   Day 2: $0.7103 * (-1.00\%) + 0.2897 * 2.00\% = -0.7103\% + 0.5794\% = -0.13\%$.
4. Vốn cuối: A = $0{,}7210 * 0{,}99 = 0{,}71379$; B = $0{,}2940 * 1{,}02 = 0{,}29988$; tổng 1,01367, tức +1,37% sau hai ngày.
   Final capital: A = $0.7210 * 0.99 = 0.71379$; B = $0.2940 * 1.02 = 0.29988$; total 1.01367, that is +1.37% after two days.
5. Nếu cân bằng lại về 70/30 trước ngày 2: $0{,}7 * (-1{,}00\%) + 0{,}3 * 2{,}00\% = -0{,}70\% + 0{,}60\% = -0{,}10\%$.
   If the fund rebalances to 70/30 before day 2: $0.7 * (-1.00\%) + 0.3 * 2.00\% = -0.70\% + 0.60\% = -0.10\%$.
6. Đọc kết quả: A tăng nên A chiếm phần lớn hơn, và ngày 2 A giảm nên danh mục lỗ −0,13% thay vì −0,10%. Không cân bằng lại, tỷ trọng tự trôi theo giá.
   Reading the result: A rose so A holds a bigger share, and A fell on day 2 so the portfolio loses −0.13% instead of −0.10%. Without rebalancing, weights drift with prices.

Bài 3. Giá trần, sàn và lệnh mua của quỹ – ***Exercise 3. Ceiling, floor and the fund's buy order***

Mã MNK có giá tham chiếu 20.000 đồng, biên độ ±7%, bước giá 50 đồng ở khoảng giá này, lô chẵn 100 cổ phiếu. (a) Tính giá trần và giá sàn. (b) Các giá đặt 20.075; 21.450; 19.250 đồng có hợp lệ không? (c) Quỹ muốn mua 2.350 cổ phiếu ở giá 20.000 đồng. Tách lô chẵn và lô lẻ, tính giá trị phần lô chẵn. (d) Lệnh mua vào thứ Sáu 06/11, bỏ qua ngày nghỉ lễ: chứng khoán về tài khoản chiều ngày nào?

The stock MNK has a reference price of 20,000 dong, a band of ±7%, a price step of 50 dong in this range and a board lot of 100 shares. (a) Compute the ceiling and the floor. (b) Are order

prices of 20,075; 21,450; 19,250 dong valid? (c) The fund wants to buy 2,350 shares at 20,000 dong. Split board lots from odd lots and compute the value of the board-lot part. (d) The order is bought on Friday 06/11, ignoring public holidays: on the afternoon of which day do the securities arrive?

1. Giá trần = 20.000 ∗ 1,07 = 21.400. Giá sàn = 20.000 ∗ 0,93 = 18.600.
   Ceiling = 20,000 ∗ 1.07 = 21,400. Floor = 20,000 ∗ 0.93 = 18,600.
2. 20.075: nằm giữa sàn và trần nhưng không là bội của 50, nên không hợp lệ.
   20,075: between floor and ceiling but not a multiple of 50, so invalid.
3. 21.450: cao hơn trần 21.400, không hợp lệ. 19.250: trong khoảng 18.600 đến 21.400 và là bội của 50, hợp lệ.
   21,450: above the ceiling of 21,400, invalid. 19,250: within 18,600 to 21,400 and a multiple of 50, valid.
4. 2.350 cổ phiếu = 23 lô chẵn (2.300 cổ phiếu) + 50 cổ phiếu lô lẻ.
   2,350 shares = 23 board lots (2,300 shares) + 50 odd-lot shares.
5. Giá trị phần lô chẵn = 2.300 ∗ 20.000 = 46.000.000 đồng.
   The board-lot part is worth 2,300 ∗ 20,000 = 46,000,000 dong.
6. Thứ Sáu 06/11 là ngày T. T+1 là thứ Hai 09/11, T+2 là thứ Ba 10/11, nên chứng khoán về chiều thứ Ba 10/11.
   Friday 06/11 is day T. T+1 is Monday 09/11 and T+2 is Tuesday 10/11, so the securities arrive on the afternoon of Tuesday 10/11.

1.13. Câu hỏi tự luận và dàn ý trả lời – *Essay questions and answer outlines*

Câu 1. So sánh giao dịch định lượng với giao dịch theo cảm tính, và nêu ba nhóm tổ chức làm giao dịch định lượng cùng nguồn vốn của họ.

Question 1. *Compare quantitative with discretionary trading, and name three groups of organisations that do quantitative trading together with their sources of capital.*

- Định nghĩa giao dịch định lượng: quy tắc bằng số, chạy bằng mã, kiểm lại được.
  Define quantitative trading: numeric rules, run by code, testable.
- Nêu ít nhất ba khác biệt về căn cứ quyết định, khả năng lặp lại, kiểm lại trên quá khứ.
  State at least three differences in basis of decision, repeatability and testing on the past.
- Nêu điểm yếu của mỗi cách.
  State the weakness of each way.
- Ba nhóm tổ chức: công ty quản lý quỹ, tự doanh công ty chứng khoán, ngân hàng, kèm nguồn vốn.
  Three groups: fund managers, brokerage proprietary desks, banks, each with its source of capital.
- Cho biết con người còn quyết định việc gì trong hệ thống định lượng.
  Say what people still decide in a quantitative system.

Câu 2. Giải thích vì sao 'giá là dự báo' và vì sao học máy hữu ích khi tập thông tin lớn và dạng hàm chưa biết. Học máy khác kinh tế lượng ở đâu?

Question 2. *Explain why 'a price is a forecast' and why machine learning is useful when the*

information set is large and the functional form is unknown. Where does machine learning differ from econometrics?

- Giải thích giá là kỳ vọng có điều kiện theo thông tin hiện có, và vì sao tin mới làm giá nhảy.
  Explain that a price is an expectation conditional on current information and why news makes it jump.
- Nêu hai chiều của dữ liệu tài chính: thời gian và chéo giữa các mã.
  State the two dimensions of financial data: time and cross-section.
- Giải thích dạng hàm chưa biết và rủi ro của mô hình quá mềm hoặc quá cứng.
  Explain an unknown functional form and the risk of a model that is too loose or too rigid.
- Nêu khác biệt về mục tiêu và cách đánh giá giữa hai ngành.
  State the differences in goal and in how the two fields judge.
- Giải thích chính quy hóa và vai trò của dữ liệu chưa thấy.
  Explain regularisation and the role of unseen data.

Câu 3. Nêu bốn thách thức khi dùng học máy trong tài chính. Vì sao một mô hình thắng ở một kỳ vẫn chưa đủ để tin? Quy tắc của sàn làm mô hình sai lệch thế nào?
Question 3. *State four challenges of using machine learning in finance. Why is a model that wins in one period still not enough to trust? How do exchange rules distort a model?*

- Bốn thách thức: tín hiệu yếu nhiễu lớn, dữ liệu ít, đổi chế độ, đối thủ học theo.
  Four challenges: weak signal with loud noise, little data, changing regimes, rivals learning.
- Dùng thí nghiệm tỷ lệ đoán đúng qua mười kỳ để giải thích vì sao một kỳ không đủ.
  Use the ten-period hit-rate experiment to explain why one period is not enough.
- Nêu các quy tắc sàn ảnh hưởng mô hình: biên độ, lô, bước giá, T+2.
  Name the exchange rules that affect a model: band, lot, price step, T+2.
- Cho ví dụ bằng số về lệch trọng số do lô chẵn.
  Give a numeric example of a weight gap caused by board lots.
- Nêu vì sao hành vi thao túng thị trường vẫn bị cấm khi dùng chương trình tự động (Luật Chứng khoán Điều 12).
  State why market manipulation stays prohibited when an automatic program is used (Securities Law, Article 12).

1.14. Câu hỏi trắc nghiệm – *Multiple-choice questions*

1. Giá một mã đi từ 40.000 đồng lên 42.000 đồng. Lợi suất đơn là bao nhiêu?
A stock's price goes from 40,000 dong to 42,000 dong. What is the simple return?

A. 2,0% / *2.0%*
B. 4,76% / *4.76%*
C. 5,0% / *5.0%*
D. 5,5% / *5.5%*

2. Cùng mã đó, giá đi từ 40.000 lên 42.000 đồng. Lợi suất log gần nhất với số nào?
For the same stock the price goes from 40,000 to 42,000 dong. Which number is closest to the log return?

A. 5,00% / *5.00%*
B. 4,88% / *4.88%*
C. 5,13% / *5.13%*
D. 2,44% / *2.44%*

3. Danh mục giữ 60% mã A và 40% mã B. Trong ngày A tăng 2% và B giảm 1%. Lợi suất danh mục là bao nhiêu?
A portfolio holds 60% of stock A and 40% of stock B. In one day A rises 2% and B falls 1%. What is the portfolio return?

A. 0,5% / *0.5%*
B. 1,0% / *1.0%*
C. 1,2% / *1.2%*
D. 0,8% / *0.8%*

4. Cổ phiếu có giá tham chiếu 30.000 đồng, biên độ ±7%. Giá trần là bao nhiêu?
A share has a reference price of 30,000 dong and a band of ±7%. What is the ceiling price?

A. 32.100 đồng / *32,100 dong*
B. 32.000 đồng / *32,000 dong*
C. 36.000 đồng / *36,000 dong*
D. 30.700 đồng / *30,700 dong*

5. Điều nào giải thích đúng nhất vì sao học máy hữu ích cho dự báo giá?
Which statement best explains why machine learning helps forecast prices?

A. Thông tin liên quan đến giá lớn và dạng hàm nối nó với giá chưa biết trước / *The information bearing on prices is large and the form linking it to prices is unknown in advance*
B. Dữ liệu tài chính nhiều hơn dữ liệu ở mọi lĩnh vực khác / *Financial data is more plentiful than data in any other field*
C. Tín hiệu dự báo trong lợi suất vốn đã mạnh và dễ thấy / *The forecasting signal in returns is already strong and easy to see*
D. Thị trường giữ nguyên quy luật qua nhiều thập kỷ / *The market keeps the same rules for decades*

6. Một tín hiệu có R^2 ngoài mẫu là 0,24%. Nhận định nào hợp lý nhất?
A signal has an out-of-sample R^2 of 0.24%. Which judgement is most reasonable?

A. Tín hiệu vô dụng và phải bỏ ngay vì R^2 nhỏ hơn 1% / *The signal is useless and must be dropped because R^2 is below 1%*

B. Tín hiệu chắc chắn sinh lời sau mọi chi phí giao dịch / *The signal will surely earn a profit after all trading costs*

C. Tín hiệu có thể có thật nhưng yếu, cần kiểm nhiều kỳ và tính chi phí / *The signal may be real but weak, so test over many periods and count costs*

D. R^2 nhỏ như vậy chứng tỏ mô hình đã quá khớp / *An R^2 this small proves that the model has overfitted*

7. Mô hình muốn mua 1.270 cổ phiếu. Phần mua được theo lô chẵn 100 cổ phiếu là bao nhiêu?
The model wants to buy 1,270 shares. How many can be bought in board lots of 100 shares?

A. 1.270 cổ phiếu / *1,270 shares*

B. 1.200 cổ phiếu / *1,200 shares*

C. 1.300 cổ phiếu / *1,300 shares*

D. 1.000 cổ phiếu / *1,000 shares*

8. Một chương trình tự động đặt và hủy hàng loạt lệnh để tạo cung cầu giả. Theo Điều 12 Luật Chứng khoán, điều gì đúng?
An automatic program places and cancels orders in bulk to fake supply and demand. Under Article 12 of the Securities Law, which is right?

A. Hợp pháp vì lệnh do máy phát ra, không do người / *Lawful because the orders are issued by a machine, not a person*

B. Hợp pháp nếu quỹ báo cáo số lệnh cho nhà đầu tư / *Lawful if the fund reports the order count to investors*

C. Chỉ bị cấm khi chương trình gây lỗ cho quỹ / *Prohibited only when the program causes the fund a loss*

D. Thao túng thị trường vẫn bị cấm, dù người hay chương trình thực hiện / *Market manipulation stays prohibited whether a person or a program does it*

Chương 2. Dữ liệu tài chính và các đặc điểm điển hình – *Chapter 2. Financial data and stylised facts*

Mục tiêu chương – *Chapter objectives*

- Tính lợi suất đơn, lợi suất log và điều chỉnh giá cho cổ tức, chia tách.
  Compute simple and log returns and adjust prices for dividends and splits.
- Nhận ra và đo các đặc điểm điển hình: đuôi dày, bất đối xứng, cụm biến động, tương quan tăng lúc căng thẳng.
  Recognise and measure the stylised facts: heavy tails, asymmetry, volatility clustering, correlation rising under stress.
- Tính độ lệch, độ nhọn, tự tương quan và ma trận tương quan bằng tay và bằng Python.
  Compute skewness, kurtosis, autocorrelation and a correlation matrix by hand and with Python.
- Dựng bảng dữ liệu cho học máy mà không dính sai lệch sống sót và sai lệch nhìn trước.
  Build a data table for machine learning without survivorship or look-ahead bias.
- Làm sạch dữ liệu giá bằng một quy trình có nhật ký.
  Clean price data with a logged procedure.

2.1. Giá, lợi suất và giá điều chỉnh – *2.1. Prices, returns and adjusted prices*

Tuần đầu thực tập ở Công ty Quản lý quỹ Cánh Diều, Hoàng Nam nhận từ chị Thu Hà một tệp giá đóng cửa hằng ngày của cổ phiếu QTK. Chị dặn: trước khi dựng mô hình nào, em phải hiểu từng con số trong tệp này. Bạn cũng nên làm như Nam. Mọi mô hình ở các chương sau đều bắt đầu từ cách đọc giá và lợi suất.

In his first week as an intern at Cánh Diều Fund Management, Hoàng Nam receives from Ms Thu Hà a file of daily closing prices for share QTK. She tells him that before building any model he must understand every number in it. You should do the same. Every model in later chapters starts from how prices and returns are read.

Giá đóng cửa là mức giá khớp cuối cùng của một phiên giao dịch. Lợi suất cho biết giá đã đổi bao nhiêu so với phiên trước, tính theo phần trăm. Có hai cách tính. Lợi suất đơn (simple return) là giá mới chia giá cũ rồi trừ 1. Lợi suất log (log return) là logarit tự nhiên của giá mới chia giá cũ.

The closing price is the last matched price of a trading session. The return tells you how much the price has changed against the previous session, as a percentage. There are two ways to compute it. The simple return is the new price divided by the old price, minus 1. The log return is the natural logarithm of the new price divided by the old price.

$$R(t) = P(t) / P(t-1) - 1 \tag{2.1}$$

$$r(t) = \ln P(t) - \ln P(t-1) \tag{2.2}$$

Hai số này gần nhau khi giá đổi ít: giá tăng 1% thì lợi suất đơn là 0,01 và lợi suất log là 0,00995. Chúng khác nhau khi bạn cộng nhiều kỳ. Lợi suất log cộng được: tổng lợi suất

log của nhiều ngày bằng lợi suất log của cả giai đoạn. Lợi suất đơn không cộng được. Muốn gộp lợi suất đơn bạn phải nhân (1 + R) của từng ngày rồi trừ 1.

The two numbers are close when the price moves little: a 1% rise gives a simple return of 0.01 and a log return of 0.00995. They differ when you add several periods. Log returns add up: the sum of the log returns of many days equals the log return of the whole period. Simple returns do not add. To combine simple returns you must multiply the (1 + R) of each day and subtract 1.

Nhiều tài liệu mô hình hóa log giá bằng bước ngẫu nhiên có trôi: log giá hôm nay bằng log giá hôm qua cộng một hằng số μ (độ trôi) cộng một nhiễu ngẫu nhiên. Theo Palomar (2025), đó là mô hình đơn giản nhất cho giá, và các mục sau sẽ kiểm xem nó sai ở chỗ nào.

Many texts model the log price as a random walk with drift: today's log price equals yesterday's log price plus a constant μ (the drift) plus a random noise term. According to Palomar (2025), that is the simplest model for prices, and the next sections test where it goes wrong.

$$\ln P(t) = \mu + \ln P(t-1) + \varepsilon(t) \qquad (2.3)$$

Cách tính lợi suất từ một cột giá – ***How to compute returns from a column of prices***

1. Lấy cột giá đóng cửa đã điều chỉnh (xem phần sau), xếp theo thứ tự ngày tăng dần.
   Take the column of adjusted closing prices (see the next steps), sorted by date in ascending order.
2. Với mỗi ngày từ ngày thứ hai, chia giá hôm nay cho giá hôm qua. Ví dụ 41.200 / 40.000 = 1,03.
   For each day from the second, divide today's price by yesterday's. For example 41,200 / 40,000 = 1.03.
3. Trừ 1 để có lợi suất đơn: 1,03 − 1 = 0,03, tức 3%.
   Subtract 1 to get the simple return: 1.03 − 1 = 0.03, which is 3%.
4. Lấy logarit tự nhiên của kết quả bước 2 để có lợi suất log: ln(1,03) = 0,02956, tức 2,956%.
   Take the natural logarithm of the result of step 2 to get the log return: ln(1.03) = 0.02956, which is 2.956%.
5. Muốn lợi suất cả giai đoạn, cộng các lợi suất log rồi đổi lại bằng e mũ tổng, trừ 1.
   For the return of the whole period, add the log returns and convert back with e to the power of the sum, minus 1.

Bài mẫu 1: năm lợi suất của cổ phiếu QTK – ***Worked example 1: five returns of share QTK***

Nam có sáu giá đóng cửa của QTK. Hãy tính lợi suất đơn và lợi suất log của từng ngày, rồi so tổng với lợi suất thật cả kỳ.

Nam has six closing prices of QTK. Compute the simple and log return of each day, then compare their sums with the true return of the whole period.

Ngày *Day*	**Giá đóng cửa (đồng)** *Close (dong)*	**Lợi suất đơn (%)** *Simple return (%)*	**Lợi suất log (%)** *Log return (%)*
1	40.000 *40,000*	–	–
2	41.200 *41,200*	3,000 *3.000*	2,956 *2.956*
3	40.600 *40,600*	−1,456 *−1.456*	−1,467 *−1.467*
4	42.100 *42,100*	3,695 *3.695*	3,628 *3.628*
5	41.500 *41,500*	−1,425 *−1.425*	−1,435 *−1.435*
6	43.000 *43,000*	3,614 *3.614*	3,551 *3.551*

Tổng năm lợi suất đơn là 7,428%. Tổng năm lợi suất log là 7,232%. Lợi suất thật của cả kỳ là 43.000 / 40.000 − 1 = 7,500%. Đổi từ log: e mũ 0,07232 trừ 1 cũng ra 7,500%. Nghĩa là cộng lợi suất đơn làm lệch −0,072 điểm phần trăm chỉ sau năm ngày.

The sum of the five simple returns is 7.428%. The sum of the five log returns is 7.232%. The true return of the whole period is 43,000 / 40,000 − 1 = 7.500%. Converting from logs, e to the power 0.07232 minus 1 also gives 7.500%. So adding simple returns is off by −0.072 percentage points after only five days.

```
import math, random
random.seed(3)
gia = [40000.0]
for _ in range(500):
    gia.append(gia[-1] * math.exp(random.gauss(0.0003, 0.012)))
don = [b / a - 1 for a, b in zip(gia, gia[1:])]
log = [math.log(b / a) for a, b in zip(gia, gia[1:])]
print("Giá đầu, giá cuối:", round(gia[0]), round(gia[-1]))
print("Cộng lợi suất đơn (%):", round(100 * sum(don), 2))
print("Cộng lợi suất log (%):", round(100 * sum(log), 2))
print("Lợi suất thực cả kỳ (%):", round(100 * (gia[-1] / gia[0] - 1), 2))
print("Đổi từ log (%):", round(100 * (math.exp(sum(log)) - 1), 2))
```

Kết quả chạy / *Output*

```
Giá đầu, giá cuối: 40000 32926
Cộng lợi suất đơn (%): -15.96
Cộng lợi suất log (%): -19.46
Lợi suất thực cả kỳ (%): -17.69
Đổi từ log (%): -17.69
```

Mã mô phỏng 500 ngày cho giá cuối 32.926 đồng từ 40.000 đồng. Tổng lợi suất đơn là −15,96%, trong khi lợi suất thật là −17,69%: lệch 1,73 điểm. Tổng lợi suất log là −19,46%, đổi lại thành −17,69%, khớp với lợi suất thật. Chuỗi giá này tự sinh, nên con số chỉ minh họa cách cộng.

The simulated 500-day series ends at 32,926 dong from 40,000 dong. The simple returns add to −15.96% while the true return is −17.69%: a gap of 1.73 points. The log returns add to

−19.46%, which converts back to −17.69%, matching the true return. The price series is self-generated, so the figures only show how the addition works.

> **LỖI HAY GẶP / COMMON MISTAKE –** Cộng lợi suất đơn của nhiều kỳ rồi coi đó là lợi suất cả giai đoạn. Chuỗi 500 ngày ở trên cho hai con số khác hẳn nhau. Hãy cộng lợi suất log, hoặc nhân (1 + R).
> *Adding the simple returns of several periods and treating the sum as the return of the whole period. The 500-day series above gives two clearly different figures. Add log returns, or multiply the (1 + R) terms.*

Còn một điều về giá cần xử lý trước khi tính lợi suất. Giá điều chỉnh (adjusted price) là giá đã sửa theo các sự kiện làm giá rơi mà tài sản của cổ đông không giảm. Có hai sự kiện chính. Cổ tức tiền mặt: công ty trả tiền cho cổ đông, và từ ngày giao dịch không hưởng quyền (ex-dividend date) giá cổ phiếu thấp đi đúng bằng phần tiền đó. Chia tách (stock split): một cổ phiếu cũ đổi thành vài cổ phiếu mới, giá mỗi cổ phiếu giảm tương ứng. Thưởng cổ phiếu cho kết quả tương tự.

One more thing about prices needs handling before returns are computed. The adjusted price is the price corrected for events that make the price fall while the shareholder's wealth does not. There are two main events. Cash dividend: the company pays money to shareholders, and from the ex-dividend date the share price is lower by about that amount. Stock split: one old share becomes several new shares, and the price per share falls in proportion. A bonus share issue has a similar effect.

Nếu bạn bỏ qua điều chỉnh, lợi suất của ngày xảy ra sự kiện sai, và chuỗi lợi suất xuất hiện những ngày lỗ nửa giá chưa từng có. Mô hình sẽ học nhầm các cú rơi giả đó thành rủi ro thật. Cách sửa là nhân mọi giá trước ngày sự kiện với một hệ số.

If you skip the adjustment, the return of the event day is wrong and the return series shows days of a half-price loss that never happened. A model will learn those false drops as real risk. The fix is to multiply every price before the event date by a factor.

1. Cổ tức tiền mặt: hệ số f = (giá đóng cửa ngày trước − cổ tức) / giá đóng cửa ngày trước.
   Cash dividend: factor f = (previous close − dividend) / previous close.
2. Chia tách một thành k: hệ số f = 1 / k. Ví dụ một thành hai thì f = 0,5.
   Split of one into k: factor f = 1 / k. For example one into two gives f = 0.5.
3. Nhân mọi giá TRƯỚC ngày sự kiện với f. Giá từ ngày sự kiện trở đi giữ nguyên.
   Multiply every price BEFORE the event date by f. Prices from the event date onward stay unchanged.
4. Tính lợi suất từ chuỗi giá đã nhân. Với nhiều sự kiện, nhân các hệ số lại với nhau.
   Compute returns from the adjusted series. With several events, multiply the factors together.

$$f = (P(\text{trước}) - D) \,/\, P(\text{trước}) \tag{2.4}$$

Bài mẫu 2: hai sự kiện của cổ phiếu QTK – *Worked example 2: two events of share QTK*

Tình huống 1: QTK đóng cửa 50.000 đồng, hôm sau là ngày không hưởng quyền cổ tức 2.000 đồng, giá đóng cửa 48.200 đồng. Tình huống 2: một mã khác đóng cửa 90.000 đồng, hôm sau chia tách một thành hai và đóng cửa 45.500 đồng.

Case 1: QTK closes at 50,000 dong, and the next day is the ex-dividend day for 2,000 dong, with a close of 48,200 dong. Case 2: another share closes at 90,000 dong, and the next day splits one into two and closes at 45,500 dong.

1. Cổ tức: hệ số f = (50.000 − 2.000) / 50.000 = 0,96. Giá ngày trước điều chỉnh = 50.000 ∗ 0,96 = 48.000 đồng.
   Dividend: factor f = (50,000 − 2,000) / 50,000 = 0.96. Adjusted previous price = 50,000 ∗ 0.96 = 48,000 dong.
2. Lợi suất thô = 48.200 / 50.000 − 1 = −3,60%. Lợi suất điều chỉnh = 48.200 / 48.000 − 1 = 0,42%.
   Raw return = 48,200 / 50,000 − 1 = −3.60%. Adjusted return = 48,200 / 48,000 − 1 = 0.42%.
3. Chia tách: f = 0,5. Giá ngày trước điều chỉnh = 90.000 ∗ 0,5 = 45.000 đồng.
   Split: f = 0.5. Adjusted previous price = 90,000 ∗ 0.5 = 45,000 dong.
4. Lợi suất thô = 45.500 / 90.000 − 1 = −49,44%. Lợi suất điều chỉnh = 45.500 / 45.000 − 1 = 1,11%.
   Raw return = 45,500 / 90,000 − 1 = −49.44%. Adjusted return = 45,500 / 45,000 − 1 = 1.11%.

Đọc kết quả: giá thô báo lỗ 3,60% ở tình huống 1 và gần một nửa ở tình huống 2, nhưng cổ đông thực tế lãi 0,42% và 1,11%. Chỉ lợi suất điều chỉnh mới đưa vào mô hình.

Reading the result: raw prices report a loss of 3.60% in case 1 and almost half in case 2, yet shareholders actually gain 0.42% and 1.11%. Only the adjusted return belongs in a model.

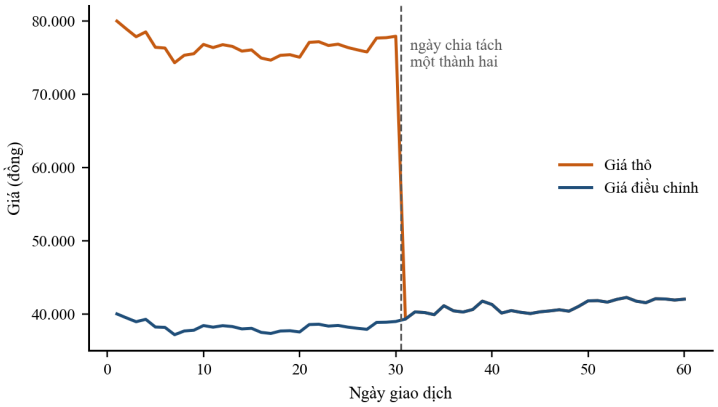
```
def dieu_chinh(gia, ngay, he_so):
    return [g * he_so if i < ngay else g for i, g in enumerate(gia)]
def loi_suat(gia, i):
    return gia[i] / gia[i - 1] - 1
cotuc = [49000, 50000, 48200]
he_so = (50000 - 2000) / 50000
dc = dieu_chinh(cotuc, 2, he_so)
print("Cổ tức hệ số:", round(he_so, 4))
print("Cổ tức thô (%):", round(100 * loi_suat(cotuc, 2), 3))
print("Cổ tức điều chỉnh (%):", round(100 * loi_suat(dc, 2), 3))
tach = [88000, 90000, 45500]
dc2 = dieu_chinh(tach, 2, 0.5)
print("Tách thô (%):", round(100 * loi_suat(tach, 2), 3))
print("Tách điều chỉnh (%):", round(100 * loi_suat(dc2, 2), 3))
print("Giá sau điều chỉnh:", [round(g) for g in dc2])
```

Kết quả chạy / *Output*

```
Cổ tức hệ số: 0.96
Cổ tức thô (%): -3.6
Cổ tức điều chỉnh (%): 0.417
Tách thô (%): -49.444
Tách điều chỉnh (%): 1.111
Giá sau điều chỉnh: [44000, 45000, 45500]
```

Mã trên cho đúng bốn số của bài mẫu: −3,600%, 0,417%, −49,444% và 1,111%. Hàm điều chỉnh chỉ nhân các giá có chỉ số nhỏ hơn ngày sự kiện, đúng như bước 3.

The code gives exactly the four numbers of the worked example: −3.600%, 0.417%, −49.444% and 1.111%. The function multiplies only prices whose index is smaller than the event day, as in step 3.



Hình 2.1. Hai đường giá của cùng một cổ phiếu tự sinh. Nhìn đường cam: giá thô tụt một nửa ngay ngày chia tách. Đường xanh liền mạch vì giá trước đó đã nhân 0,5.

Figure 2.1. *Two price lines for the same simulated share. Look at the orange line: the raw price drops by half on the split day. The blue line is continuous because earlier prices were multiplied by 0.5.*

> **PHẢI THUỘC / MEMORISE –** Lợi suất đơn R = P(t) / P(t − 1) − 1 và lợi suất log r = ln P(t) − ln P(t − 1). Log cộng được qua các kỳ, đơn thì không. Luôn tính từ giá điều chỉnh: nhân giá trước ngày sự kiện với f.
>
> *Simple return R = P(t) / P(t − 1) − 1 and log return r = ln P(t) − ln P(t − 1). Log returns add across periods, simple ones do not. Always compute from adjusted prices: multiply prices before the event day by f.*

2.2. Các đặc điểm điển hình của lợi suất – *2.2. Stylised facts of returns*

Đặc điểm điển hình (stylised fact) là một tính chất của dữ liệu đã được nhiều nghiên cứu độc lập quan sát thấy ở nhiều tài sản, nhiều thị trường và nhiều thời kỳ. Biết các đặc điểm này giống như biết thời tiết trước khi lên kế hoạch dã ngoại: bạn chưa biết ngày mai mưa hay nắng, nhưng biết mùa nào hay mưa. Theo Palomar (2025), có sáu đặc điểm đáng nhớ.

A stylised fact is a property of the data that many independent studies have observed across many assets, markets and periods. Knowing these facts is like knowing the climate before planning a picnic: you do not know whether tomorrow is wet or dry, but you know which season is rainy. According to Palomar (2025), six facts are worth remembering.

Đặc điểm *Fact*	**Nghĩa là** *What it means*	**Ví dụ đời thường** *Everyday example*	**Hệ quả khi dựng mô hình** *Consequence for models*
Không dừng *Non-stationary*	Trung bình và biến động đổi theo thời gian *Mean and volatility change over time*	Nhiệt độ trung bình tháng Giêng khác tháng Bảy *The average temperature in January differs from July*	Mô hình chạy tốt năm trước có thể hỏng năm sau *A model that worked last year may fail next year*
Gần như không tự tương quan *Almost no autocorrelation*	Lợi suất hôm nay ít cho biết lợi suất ngày mai *Today's return says little about tomorrow's*	Tung đồng xu: mặt sấp vừa ra không báo trước lần tới *A coin toss: the face just shown does not predict the next*	Tín hiệu trong lợi suất mỏng, khó dự báo *The signal in returns is thin and hard to forecast*
Đuôi dày *Heavy tails*	Ngày cực đoan thường hơn phân phối chuẩn nói *Extreme days are more frequent than the normal says*	Đa số ngày êm, thỉnh thoảng có một cơn bão lớn *Most days are calm, now and then a big storm*	Đo rủi ro bằng mô hình chuẩn sẽ đánh giá thấp *Measuring risk with a normal model understates it*
Bất đối xứng lãi lỗ *Gain and loss asymmetry*	Phân phối lệch, lỗ lớn nhiều hơn lãi lớn *The distribution is skewed, large losses outnumber large gains*	Leo cầu thang từng bước, ngã thì chỉ một khoảnh khắc *Climbing stairs step by step, falling is quick*	Phải đo độ lệch, đừng chỉ nhìn trung bình *Measure skewness, do not look only at the mean*
Cụm biến động *Volatility clustering*	Ngày lớn nối ngày lớn, ngày nhỏ nối ngày nhỏ *Large days follow large days, small days follow small*	Mưa bão đến theo từng đợt *Storms come in spells*	Biến động dự báo được dù lợi suất khó dự báo *Volatility is forecastable even if returns are not*
Tương quan cùng chiều *Positive correlation*	Các mã thường đi cùng nhau; lúc thị trường xấu càng sát nhau *Shares tend to move together, and closer still in bad markets*	Ngày nắng mỗi người đi một hướng, ngày bão cả nhà chạy vào một chỗ *On sunny days people scatter, in a storm everyone runs to one place*	Đa dạng hóa yếu đi đúng lúc cần nhất *Diversification weakens just when it is needed*

Cột cuối của bảng là điều nên nhớ: mỗi đặc điểm đòi một việc cụ thể khi bạn dựng dữ liệu hoặc mô hình. Ba mục sau đo lần lượt đuôi dày và bất đối xứng (mục 2.3), tự tương quan và cụm biến động (mục 2.4), tương quan giữa các mã (mục 2.5).

The last column is the part to remember: each fact calls for a specific action when you build data or models. The next three sections measure, in turn, heavy tails and asymmetry (section 2.3), autocorrelation and volatility clustering (section 2.4), and correlation between shares (section 2.5).

Một con số cho thấy đuôi dày nghĩa là gì. Nếu lợi suất ngày theo phân phối chuẩn, xác suất một ngày lệch xa trung bình hơn 5 độ lệch chuẩn là 5,73 phần mười triệu. Tức khoảng một lần trong 7.000 năm giao dịch, tính 250 ngày mỗi năm. Trong dữ liệu thật những ngày như vậy xuất hiện trong vài năm, không phải vài nghìn năm.

One number shows what heavy tails mean. If daily returns followed the normal distribution, the probability of a day more than 5 standard deviations from the mean would be 5.73 in ten million. That is about once in 7,000 trading years of 250 days. In real data such days appear within a few years, not a few thousand.

Tần suất dữ liệu cũng đổi cách các đặc điểm biểu hiện. Dữ liệu tuần, tháng gần chuẩn hơn sau khi hiệu chỉnh cụm biến động, nhưng có quá ít quan sát. Dữ liệu ngày có đuôi dày rõ và đủ quan sát nếu mô hình không có quá nhiều tham số. Dữ liệu trong ngày có hàng triệu quan sát, hợp với học máy, nhưng nhiễu của cơ chế khớp lệnh lớn dần.

Data frequency also changes how the facts show. Weekly and monthly data are closer to normal once volatility clustering is corrected, but there are too few observations. Daily data have clear heavy tails and enough observations if the model does not have too many parameters. Intraday data have millions of observations, which suits machine learning, but the noise of the order-matching mechanism grows.

Tần suất *Frequency*	**Số quan sát trong 10 năm** *Observations in 10 years*	**Điều cần cẩn thận** *What to watch*
Tháng *Monthly*	khoảng 120 *about 120*	Quá ít dữ liệu cho nhiều tham số *Too little data for many parameters*
Ngày *Daily*	khoảng 2.500 *about 2,500*	Đuôi dày, cụm biến động rõ *Clear heavy tails and clustering*
Trong ngày *Intraday*	hàng triệu *millions*	Nhiễu khớp lệnh, cần mô hình riêng *Matching noise, needs its own models*

Nam hỏi chị Thu Hà nên bắt đầu từ tần suất nào. Chị trả lời: dữ liệu ngày, vì đủ dài để ước lượng và đủ gọn để kiểm tra bằng tay.

Nam asks Ms Hà which frequency to start with. She answers: daily data, because it is long enough to estimate from and compact enough to check by hand.

2.3. Độ lệch, độ nhọn và phân phối chuẩn – *2.3. Skewness, kurtosis and the normal distribution*

Phân phối chuẩn (normal distribution) là đường chuông đối xứng, chỉ cần hai số để mô tả: trung bình μ và phương sai σ². Nó dễ tính toán nên được dùng nhiều. Với tiếng ồn của mạch điện, giả định đó khá hợp lý. Với lợi suất cổ phiếu thì thiếu hai mô tả: mức bất đối xứng và độ dày của đuôi.

The normal distribution is a symmetric bell curve described by just two numbers: the mean μ and the variance σ². It is easy to calculate with, so it is widely used. For electrical circuit noise the assumption is fairly reasonable. For share returns two descriptions are missing: the degree of asymmetry and the thickness of the tails.

Độ lệch (skewness) đo bất đối xứng. Bằng 0 là đối xứng. Âm nghĩa là đuôi bên lỗ dài hơn đuôi bên lãi. Độ nhọn (kurtosis) đo độ dày của đuôi so với đường chuông. Phân phối chuẩn có độ nhọn bằng 3. Lớn hơn 3 là đuôi dày: ngày cực đoan thường hơn. Lợi suất tài chính thường có độ lệch âm và độ nhọn lớn, nghĩa là có khả năng lỗ lớn ở đuôi trái.

Skewness measures asymmetry. Zero means symmetric. Negative means the loss side has a longer tail than the gain side. Kurtosis measures the thickness of the tails compared with the bell curve. The normal distribution has kurtosis 3. Above 3 the tails are fat: extreme days are more frequent. Financial returns usually show negative skewness and large kurtosis, which means a chance of very large losses in the left tail.

$$S = [\Sigma\,(r_i - \mu)^3 / n] / \sigma^3 \qquad (2.5)$$

$$K = [\Sigma\,(r_i - \mu)^4 / n] / \sigma^4 \qquad (2.6)$$

Cách đo độ lệch và độ nhọn – ***How to measure skewness and kurtosis***

1. Tính trung bình μ = tổng lợi suất / n.
   Compute the mean μ = sum of returns / n.
2. Với mỗi ngày, tính độ lệch so với trung bình: $r_i - \mu$.
   For each day compute the deviation from the mean: $r_i - \mu$.
3. Tính phương sai σ^2 = trung bình của $(r_i - \mu)^2$, rồi $\sigma = \sqrt{\sigma^2}$.
   Compute the variance σ^2 = the average of $(r_i - \mu)^2$, then $\sigma = \sqrt{\sigma^2}$.
4. Tính trung bình của $(r_i - \mu)^3$ rồi chia cho σ^3. Đó là độ lệch S.
   Compute the average of $(r_i - \mu)^3$ and divide by σ^3. That is the skewness S.
5. Tính trung bình của $(r_i - \mu)^4$ rồi chia cho σ^4. Đó là độ nhọn K. So K với 3.
   Compute the average of $(r_i - \mu)^4$ and divide by σ^4. That is the kurtosis K. Compare K with 3.

Bài mẫu 3: tám lợi suất ngày của mã HLM – ***Worked example 3: eight daily returns of share HLM***

Tám lợi suất ngày của HLM (%) gồm một ngày giảm mạnh. Chia cho n = 8, làm đủ năm bước.

Eight daily returns of HLM (%) include one sharp fall. Divide by n = 8 and do all five steps.

Ngày *Day*	**r (%)**	**r − μ**	**(r − μ)²**	**(r − μ)³**	**(r − μ)⁴**
1	1,2 *1.2*	1,375 *1.375*	1,891 *1.891*	2,60 *2.60*	3,6 *3.6*
2	0,6 *0.6*	0,775 *0.775*	0,601 *0.601*	0,47 *0.47*	0,4 *0.4*
3	−0,4 *−0.4*	−0,225 *−0.225*	0,051 *0.051*	−0,01 *−0.01*	0,0 *0.0*
4	0,9 *0.9*	1,075 *1.075*	1,156 *1.156*	1,24 *1.24*	1,3 *1.3*

Ngày *Day*	**r (%)**	**r − μ**	**(r − μ)²**	**(r − μ)³**	**(r − μ)⁴**
5	−5,8 *−5.8*	−5,625 *−5.625*	31,641 *31.641*	−177,98 *−177.98*	1.001,1 *1,001.1*
6	0,7 *0.7*	0,875 *0.875*	0,766 *0.766*	0,67 *0.67*	0,6 *0.6*
7	1,1 *1.1*	1,275 *1.275*	1,626 *1.626*	2,07 *2.07*	2,6 *2.6*
8	0,3 *0.3*	0,475 *0.475*	0,226 *0.226*	0,11 *0.11*	0,1 *0.1*

Trung bình μ = −1,4 / 8 = −0,175. Phương sai σ² = 37,955 / 8 = 4,744, độ lệch chuẩn σ = 2,178. Độ lệch S = (−170,83 / 8) / 10,334 = −2,07. Độ nhọn K = (1.009,7 / 8) / 22,509 = 5,61.

Mean μ = −1.4 / 8 = −0.175. Variance σ² = 37.955 / 8 = 4.744, standard deviation σ = 2.178. Skewness S = (−170.83 / 8) / 10.334 = −2.07. Kurtosis K = (1,009.7 / 8) / 22.509 = 5.61.

Đọc kết quả: độ lệch −2,07 âm mạnh và độ nhọn 5,61 vượt xa 3, vì một ngày giảm 5,8% chi phối cả bảng. Tám ngày quá ít để kết luận về thị trường, nhưng đủ để thấy cách một ngày xấu kéo hai số này.

Reading the result: skewness −2.07 is strongly negative and kurtosis 5.61 is far above 3, because one day of a 5.8% fall dominates the table. Eight days are too few to conclude about the market, but enough to see how one bad day pulls both numbers.

Độ lệch và độ nhọn cho bạn hai con số. Để thấy cả hình dạng, nhà phân tích dùng đồ thị Q–Q: xếp lợi suất từ nhỏ đến lớn rồi vẽ từng giá trị với giá trị tương ứng của phân phối đang thử. Nếu các điểm nằm trên một đường thẳng thì phân phối hợp, nếu hai đầu cong ra xa thì đuôi dày hơn. Để có kết luận bằng số, người ta dùng kiểm định thống kê, ví dụ kiểm định Anderson–Darling. Giá trị p (p-value) là xác suất thấy dữ liệu lệch như vậy nếu phân phối đang thử là đúng. Giá trị p nhỏ hơn 0,05 thường được coi là bằng chứng bác phân phối đó.

Skewness and kurtosis give you two numbers. To see the whole shape, analysts use the Q–Q plot: sort the returns from small to large and plot each against the matching value of the distribution under test. If the points lie on a straight line the distribution fits, and if the two ends curve away the tails are fatter. For a verdict in numbers, a statistical test is used, for example the Anderson–Darling test. The p-value is the probability of seeing data this far off if the tested distribution were true. A p-value below 0.05 is usually taken as evidence against that distribution.

Với chuỗi lợi suất ngày của một chỉ số lớn, các kiểm định kiểu này bác phân phối chuẩn dứt khoát. Phân phối Student t (có đuôi dày) khá hơn, và phân phối t lệch (vừa dày đuôi vừa bất đối xứng) hợp nhất. Bạn chưa cần ước lượng các phân phối ấy ở đây; điều cần nhớ là chuẩn thường không đủ.

On daily return series of a large index, tests of this kind reject the normal distribution firmly. The Student t distribution (with fat tails) fits better, and the skewed t distribution (fat-tailed and

asymmetric) fits best. You do not need to estimate those distributions here; the point to remember is that the normal is usually not enough.

```
import math, random
def sinh(n, seed):
    random.seed(seed)
    h, out = 0.00005, []
    for _ in range(n):
        chi = sum(random.gauss(0, 1) ** 2 for _ in range(5))
        z = random.gauss(0, 1) / math.sqrt(chi / 3)
        r = 0.0003 + math.sqrt(h) * z
        if random.random() < 0.004:
            r -= abs(random.gauss(0.04, 0.01))
        out.append(r)
        h = 0.000002 + 0.10 * (r - 0.0003) ** 2 + 0.86 * h
    return out
def mom(x):
    n = len(x)
    m = sum(x) / n
    s2 = sum((v - m) ** 2 for v in x) / n
    sk = sum((v - m) ** 3 for v in x) / n / s2 ** 1.5
    ku = sum((v - m) ** 4 for v in x) / n / s2 ** 2
    return m, math.sqrt(s2), sk, ku

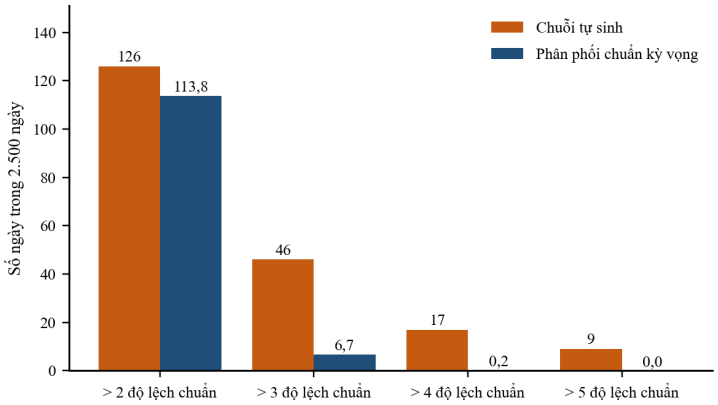
r = sinh(2500, 7)
m, s, sk, ku = mom(r)
print("Trung bình (%/ngày):", round(100 * m, 4))
print("Độ lệch chuẩn (%/ngày):", round(100 * s, 4))
print("Độ lệch:", round(sk, 3))
print("Độ nhọn:", round(ku, 3))
print("Ngày tệ nhất (%):", round(100 * min(r), 2))
print("Ngày tệ nhất (số σ):", round((min(r) - m) / s, 2))
ngoai = sum(1 for v in r if abs(v - m) > 3 * s)
print("Số ngày ngoài 3σ:", ngoai)
print("Kỳ vọng nếu chuẩn:", round(2500 * math.erfc(3 / 2 ** 0.5), 1))
jb = len(r) / 6 * (sk ** 2 + (ku - 3) ** 2 / 4)
print("Thống kê Jarque-Bera:", round(jb, 1))
```

Kết quả chạy / *Output*

```
Trung bình (%/ngày): 0.0075
Độ lệch chuẩn (%/ngày): 0.8739
Độ lệch: -0.208
Độ nhọn: 9.959
Ngày tệ nhất (%): -5.49
Ngày tệ nhất (số σ): -6.29
Số ngày ngoài 3σ: 46
Kỳ vọng nếu chuẩn: 6.7
Thống kê Jarque-Bera: 5061.8
```

Trên 2.500 ngày tự sinh, độ lệch là −0,208 và độ nhọn là 9,959. Ngày tệ nhất lỗ 5,49%, tức 6,29 độ lệch chuẩn dưới trung bình. Có 46 ngày ngoài 3 độ lệch chuẩn, trong khi phân phối chuẩn chỉ kỳ vọng 6,7. Thống kê Jarque–Bera, kết hợp độ lệch và độ nhọn, bằng 5.061,8, lớn hơn hẳn ngưỡng 5,99 ở mức 5%. Chuỗi này tự sinh với đuôi dày có chủ đích, nên đây là cách kiểm tra, chưa phải phát hiện về thị trường.

On 2,500 simulated days, skewness is −0.208 and kurtosis is 9.959. The worst day loses 5.49%, which is 6.29 standard deviations below the mean. There are 46 days beyond 3 standard deviations, whereas the normal distribution expects only 6.7. The Jarque–Bera statistic, which combines skewness and kurtosis, is 5,061.8, very large against the 5% threshold of 5.99. The series is simulated with deliberately fat tails, so this shows how to test, not a finding about the market.



Hình 2.2. Số ngày vượt k độ lệch chuẩn trong 2.500 ngày tự sinh. Nhìn cặp cột ở k = 4 và 5: cột cam cao vài ngày, cột xanh gần bằng 0.

Figure 2.2. *Number of days beyond k standard deviations in 2,500 simulated days. Look at the pairs at k = 4 and 5: the orange bar is several days high while the blue bar is almost zero.*

LỖI HAY GẶP / COMMON MISTAKE – Tin trung bình mẫu ngắn hạn. Trung bình lợi suất ngày chỉ là một phần nhỏ của độ lệch chuẩn, nên ước lượng từ vài tháng dao động mạnh. Hãy báo kèm khoảng sai số, đừng coi nó là lợi suất kỳ vọng.
Trusting a short-sample mean. The mean daily return is only a small fraction of the standard deviation, so an estimate from a few months swings widely. Report an error band with it and do not treat it as the expected return.

PHẢI THUỘC / MEMORISE – Phân phối chuẩn có độ lệch 0 và độ nhọn 3. Lợi suất tài chính thường có độ lệch âm và độ nhọn lớn hơn 3: đuôi dày, lỗ lớn thường hơn mô hình chuẩn nói.
The normal distribution has skewness 0 and kurtosis 3. Financial returns usually have negative skewness and kurtosis above 3: fat tails, with large losses more frequent than a normal model says.

2.4. Tự tương quan và cụm biến động – *2.4. Autocorrelation and volatility clustering*

Câu hỏi lớn nhất của dự báo là: quá khứ của một chuỗi có cho biết gì về tương lai của nó không? Giả thuyết thị trường hiệu quả (efficient-market hypothesis, EMH) nói rằng giá đã phản ánh mọi thông tin sẵn có, nên không ai có thể thắng thị trường đều đặn bằng cách chọn mã hay chọn thời điểm, trừ khi nhận thêm rủi ro. Giả thuyết này được tranh luận nhiều thập kỷ, và nhiều nhà đầu tư chủ động vẫn tin có thể thiết kế danh mục thắng thị trường.

The biggest question in forecasting is whether the past of a series says anything about its future. The efficient-market hypothesis (EMH) states that prices already reflect all available information, so nobody can beat the market steadily by picking shares or timing, unless taking more risk. The hypothesis has been debated for decades, and many active investors still believe portfolios can be designed to beat the market.

Dữ liệu cho một cách đo cụ thể. Tự tương quan (autocorrelation) là hệ số tương quan giữa chuỗi và chính nó lùi k ngày, gọi k là độ trễ (lag). Hàm tự tương quan

(autocorrelation function, ACF) là danh sách các giá trị này theo k = 1, 2, 3… Nếu lợi suất hôm nay không cho biết gì về ngày mai, ACF ở mọi k gần 0.

The data give a concrete measure. Autocorrelation is the correlation coefficient between a series and itself shifted back by k days, with k called the lag. The autocorrelation function (ACF) is the list of these values for k = 1, 2, 3… If today's return says nothing about tomorrow's, the ACF is near 0 at every k.

$$\rho_k = \Sigma\,(r_t - \mu)(r_{t-k} - \mu) \,/\, \Sigma\,(r_t - \mu)^2 \qquad (2.7)$$

Cách đo tự tương quan độ trễ 1 – *How to measure lag-1 autocorrelation*

1. Tính trung bình μ của chuỗi và độ lệch $r_t - \mu$ của từng ngày.
   Compute the mean μ of the series and the deviation $r_t - \mu$ of each day.
2. Với mỗi ngày từ ngày thứ hai, nhân độ lệch của ngày đó với độ lệch của ngày liền trước.
   For each day from the second, multiply that day's deviation by the previous day's.
3. Cộng các tích ở bước 2. Cộng bình phương độ lệch của mọi ngày.
   Add the products from step 2. Add the squared deviations of all days.
4. Chia tổng tích cho tổng bình phương: đó là ρ_1. Số gần 0 là không tự tương quan.
   Divide the sum of products by the sum of squares: that is ρ_1. A number near 0 means no autocorrelation.
5. Lặp lại với chuỗi bình phương lợi suất r_t^2. Nếu ρ_1 của r^2 dương và lớn thì biến động có tính cụm.
   Repeat with the series of squared returns r_t^2. If ρ_1 of r^2 is positive and large, volatility clusters.

Bài mẫu 4: mười ngày của mã VZP – *Worked example 4: ten days of share VZP*

Mười lợi suất ngày của VZP (%) có hai ngày yên rồi một đợt biến động. Tính ρ_1 của r và của r^2.

Ten daily returns of VZP (%) show quiet days and then a spell of large moves. Compute ρ_1 of r and of r^2.

Ngày *Day*	**r (%)**	**r^2**	**$r - \mu$**	**Tích r** *r product*	**$r^2 - \mu_2$**	**Tích r^2** *r^2 product*
1	0,3 *0.3*	0,09 *0.09*	0,21 *0.21*	–	−2,62 *−2.62*	–
2	−0,5 *−0.5*	0,25 *0.25*	−0,59 *−0.59*	−0,12 *−0.12*	−2,46 *−2.46*	6,42 *6.42*
3	0,4 *0.4*	0,16 *0.16*	0,31 *0.31*	−0,18 *−0.18*	−2,54 *−2.54*	6,25 *6.25*
4	0,2 *0.2*	0,04 *0.04*	0,11 *0.11*	0,03 *0.03*	−2,67 *−2.67*	6,78 *6.78*

Ngày *Day*	**r (%)**	**r^2**	**$r - \mu$**	**Tích r** *r product*	**$r^2 - \mu_2$**	**Tích r^2** *r^2 product*
5	−2,6 *−2.6*	6,76 *6.76*	−2,69 *−2.69*	−0,30 *−0.30*	4,06 *4.06*	−10,81 *−10.81*
6	−2,2 *−2.2*	4,84 *4.84*	−2,29 *−2.29*	6,16 *6.16*	2,14 *2.14*	8,66 *8.66*
7	2,9 *2.9*	8,41 *8.41*	2,81 *2.81*	−6,43 *−6.43*	5,71 *5.71*	12,18 *12.18*
8	2,5 *2.5*	6,25 *6.25*	2,41 *2.41*	6,77 *6.77*	3,54 *3.54*	20,22 *20.22*
9	0,3 *0.3*	0,09 *0.09*	0,21 *0.21*	0,51 *0.51*	−2,62 *−2.62*	−9,27 *−9.27*
10	−0,4 *−0.4*	0,16 *0.16*	−0,49 *−0.49*	−0,10 *−0.10*	−2,54 *−2.54*	6,66 *6.66*

Với r: trung bình $\mu = 0{,}090$, tổng tích = 6,33, tổng bình phương = 26,97, nên $\rho_1 = 0{,}235$. Với r^2: trung bình $\mu_2 = 2{,}705$, tổng tích = 47,09, tổng bình phương = 105,88, nên $\rho_1 = 0{,}445$.

For r: mean $\mu = 0.090$, sum of products = 6.33, sum of squares = 26.97, so $\rho_1 = 0.235$. For r^2: mean $\mu_2 = 2.705$, sum of products = 47.09, sum of squares = 105.88, so $\rho_1 = 0.445$.

Đọc kết quả: ρ_1 của r là 0,23, nhỏ, còn ρ_1 của r^2 là 0,44, lớn hơn rõ. Với chỉ 10 ngày, ngưỡng ý nghĩa $2 / \sqrt{n}$ là 0,63, nên cả hai chưa vượt ngưỡng; bài này chỉ để bạn luyện cách tính. Chuỗi dài ở phần sau mới cho kết luận.

Reading the result: ρ_1 of r is 0.23, small, while ρ_1 of r^2 is 0.44, clearly larger. With only 10 days the significance threshold $2 / \sqrt{n}$ is 0.63, so neither passes it; this exercise is only for practising the calculation. The long series in the next code block gives the conclusion.

```
import math, random
def sinh(n, seed):
    random.seed(seed)
    h, out = 0.00005, []
    for _ in range(n):
        chi = sum(random.gauss(0, 1) ** 2 for _ in range(5))
        z = random.gauss(0, 1) / math.sqrt(chi / 3)
        r = 0.0003 + math.sqrt(h) * z
        if random.random() < 0.004:
            r -= abs(random.gauss(0.04, 0.01))
        out.append(r)
        h = 0.000002 + 0.10 * (r - 0.0003) ** 2 + 0.86 * h
    return out
def acf(x, k):
    n = len(x)
    m = sum(x) / n
    c0 = sum((v - m) ** 2 for v in x)
    ck = sum((x[i] - m) * (x[i + k] - m) for i in range(n - k))
    return ck / c0

r = sinh(2500, 7)
q = [v * v for v in r]
for k in range(1, 6):
    print(f"k={k}:", round(acf(r, k), 3), round(acf(q, k), 3))
```

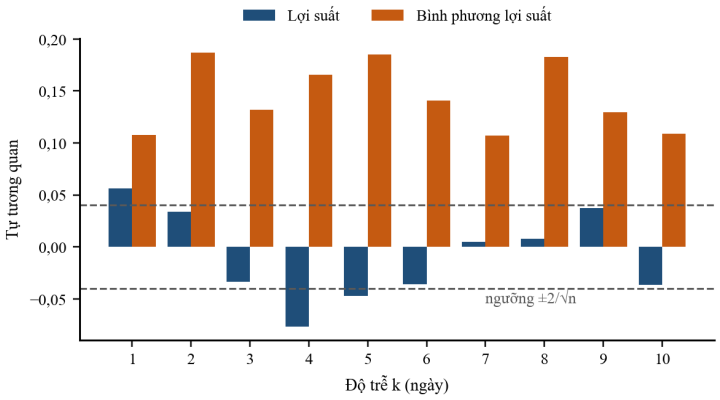
```
print("Ngưỡng 2/√n:", round(2 / math.sqrt(len(r)), 3))
print("ACF r² ở k=20:", round(acf(q, 20), 3))
```

Kết quả chạy / *Output*

```
k=1: 0.056 0.108
k=2: 0.034 0.187
k=3: -0.034 0.132
k=4: -0.076 0.165
k=5: -0.047 0.185
Ngưỡng 2/√n: 0.04
ACF r² ở k=20: 0.085
```

Trên 2.500 ngày tự sinh, tự tương quan của lợi suất ở độ trễ 1 là 0,056, chỉ cùng cỡ ngưỡng 0,040. Tự tương quan của bình phương lợi suất ở độ trễ 1 là 0,108, gấp 2,7 lần ngưỡng. Ở độ trễ 20 ngày nó vẫn là 0,085, còn cao hơn ngưỡng. Biến động có trí nhớ dài, còn hướng đi của giá thì không.

On 2,500 simulated days, the lag-1 autocorrelation of returns is 0.056, only of the order of the threshold 0.040. The lag-1 autocorrelation of squared returns is 0.108, 2.7 times the threshold. At a lag of 20 days it is still 0.085, above the threshold. Volatility has a long memory, whereas the direction of prices has none.



Hình 2.3. Tự tương quan của lợi suất và của bình phương lợi suất, độ trễ 1 đến 10. Nhìn cột xanh: thấp, sát hai đường đứt. Nhìn cột cam: cao hơn hẳn các cột xanh.

Figure 2.3. *Autocorrelation of returns and of squared returns, lags 1 to 10. Look at the blue bars: low, close to the two dashed lines. Look at the orange bars: clearly taller than the blue ones.*

Hai kết quả cùng đúng và không mâu thuẫn. Hướng đi ngày mai khó đoán từ hướng đi hôm nay, nên dự báo lợi suất là việc khó. Mức dao động ngày mai lại đoán được khá tốt từ mức dao động gần đây, nên dự báo biến động là việc làm được. Nhà phân tích gọi đường bao của dao động theo thời gian là đường bao biến động. Nếu bạn chia từng lợi suất cho biến động ước lượng của ngày đó, chuỗi chuẩn hóa nhìn đều hơn nhiều, và cụm biến động gần như biến mất.

Both results hold and they do not conflict. The direction of tomorrow is hard to guess from the direction of today, so forecasting returns is difficult. The size of tomorrow's swing can be guessed fairly well from recent swings, so forecasting volatility is doable. Analysts call the envelope of the swings over time the volatility envelope. If you divide each return by that day's estimated volatility, the standardised series looks much more even and the clustering almost disappears.

Một công cụ đi kèm là tự tương quan riêng phần (partial autocorrelation function, PACF). Nó cũng đo quan hệ giữa chuỗi và chính nó lùi k ngày, nhưng loại phần ảnh hưởng đi qua các ngày ở giữa. Ví dụ PACF ở độ trễ 3 chỉ giữ quan hệ trực tiếp ngày t với ngày t − 3. Các mô hình chuỗi thời gian ở chương sau dùng ACF và PACF để chọn cấu trúc.

A companion tool is the partial autocorrelation function (PACF). It also measures the link between the series and itself shifted back by k days, but removes the part that passes through the days in between. For example the PACF at lag 3 keeps only the direct link between day t and day t − 3. The time-series models of a later chapter use the ACF and the PACF to choose a structure.

> **PHẢI THUỘC / MEMORISE –** Lợi suất gần như không tự tương quan, bình phương lợi suất thì tự tương quan dương và kéo dài. Hướng giá khó dự báo, mức biến động dự báo được.
> *Returns are almost uncorrelated whereas squared returns are positively autocorrelated for long. The direction of prices is hard to forecast, the size of swings is forecastable.*

2.5. Cấu trúc giữa các tài sản: tương quan và nhân tố thị trường – *2.5. Structure across assets: correlation and the market factor*

Đến đây bạn nhìn từng mã một. Nhưng danh mục có nhiều mã, và rủi ro của cả danh mục phụ thuộc vào việc các mã đi cùng hay đi ngược nhau. Hệ số tương quan (correlation coefficient) đo điều đó bằng một số từ −1 đến 1. Bằng 1 là hai mã luôn cùng chiều theo tỷ lệ cố định. Bằng 0 là chưa thấy quan hệ tuyến tính. Bằng −1 là luôn ngược chiều.

Until now you have looked at one share at a time. But a portfolio holds many shares, and its risk depends on whether they move together or against each other. The correlation coefficient measures that with one number from −1 to 1. A value of 1 means the two shares always move in the same direction in a fixed proportion. A value of 0 means no linear link has been seen. A value of −1 means they always move in opposite directions.

$$\rho = \mathrm{Cov}(A, B) \,/\, (\sigma A * \sigma B) \qquad (2.8)$$

Cách tính hệ số tương quan của hai mã – ***How to compute the correlation of two shares***

1. Tính trung bình của mỗi mã, rồi độ lệch của từng ngày so với trung bình của mã đó.
   Compute the mean of each share, then each day's deviation from that share's mean.
2. Nhân hai độ lệch của cùng một ngày. Cộng các tích rồi chia cho n: đó là hiệp phương sai.
   Multiply the two deviations of the same day. Add the products and divide by n: that is the covariance.
3. Tính độ lệch chuẩn của từng mã: căn bậc hai của trung bình bình phương độ lệch.
   Compute the standard deviation of each share: the square root of the average squared deviation.
4. Chia hiệp phương sai cho tích hai độ lệch chuẩn. Kết quả luôn nằm trong [−1, 1].
   Divide the covariance by the product of the two standard deviations. The result always lies in [−1, 1].

Bài mẫu 5: ba mã KDX, MNQ, TVK trong sáu ngày – ***Worked example 5: shares KDX, MNQ and TVK over six days***

Lợi suất ngày (%) của ba mã nằm trong bảng. Hoàng Nam làm tay cặp KDX và MNQ, rồi chị Thu Hà bảo thêm hai cặp còn lại.

The daily returns (%) of three shares are in the table. Hoàng Nam works out the pair KDX and MNQ by hand, then Ms Hà asks him to add the other two pairs.

Ngày *Day*	KDX (%)	MNQ (%)	a − μa	b − μb	Tích *Product*	(a − μa)²	(b − μb)²
1	1,0 *1.0*	0,8 *0.8*	0,600 *0.600*	0,467 *0.467*	0,280 *0.280*	0,360 *0.360*	0,218 *0.218*
2	−0,6 *−0.6*	−0,9 *−0.9*	−1,000 *−1.000*	−1,233 *−1.233*	1,233 *1.233*	1,000 *1.000*	1,521 *1.521*
3	1,8 *1.8*	1,5 *1.5*	1,400 *1.400*	1,167 *1.167*	1,633 *1.633*	1,960 *1.960*	1,361 *1.361*
4	−1,2 *−1.2*	−0,7 *−0.7*	−1,600 *−1.600*	−1,033 *−1.033*	1,653 *1.653*	2,560 *2.560*	1,068 *1.068*
5	0,5 *0.5*	0,2 *0.2*	0,100 *0.100*	−0,133 *−0.133*	−0,013 *−0.013*	0,010 *0.010*	0,018 *0.018*
6	0,9 *0.9*	1,1 *1.1*	0,500 *0.500*	0,767 *0.767*	0,383 *0.383*	0,250 *0.250*	0,588 *0.588*

Trung bình KDX là 0,400, trung bình MNQ là 0,333. Tổng tích là 5,170, tổng bình phương của KDX là 6,140, của MNQ là 4,773. Tương quan = 5,170 / √(6,140 ∗ 4,773) = 0,955. Hai cặp còn lại, tính cùng cách, cho KDX và TVK là −0,249, MNQ và TVK là −0,257.

The mean of KDX is 0.400 and the mean of MNQ is 0.333. The sum of products is 5.170, the sum of squares of KDX is 6.140 and of MNQ is 4.773. Correlation = 5.170 / √(6.140 ∗ 4.773) = 0.955. The other two pairs, by the same method, give −0.249 for KDX and TVK and −0.257 for MNQ and TVK.

Ma trận tương quan *Correlation matrix*	KDX	MNQ	TVK
KDX	1	0,955 *0.955*	−0,249 *−0.249*
MNQ	0,955 *0.955*	1	−0,257 *−0.257*
TVK	−0,249 *−0.249*	−0,257 *−0.257*	1

Đọc kết quả: KDX và MNQ gần như một đôi sinh đôi (tương quan 0,95), còn TVK đi gần như riêng (tương quan −0,25 và −0,26). Chia vốn cho KDX và MNQ giống chia vốn