








Modelling Financial Markets

Chapter 19: Review and Project Presentations



Daniel Traian PELE

Bucharest University of Economic Studies
IDA Institute Digital Assets
Blockchain Research Center
AI4EFin Artificial Intelligence for Energy Finance
Romanian Academy, Institute for Economic Forecasting
MSCA Digital Finance

When Does an Empirical Result Deserve Trust?



Albert J. Menkveld. **Photo: Me Judice (2015); CC BY 3.0; Wikimedia Commons**

"In science, evidence is generated to test hypotheses in an evidence-generating process (EGP). We claim that EGP variation across researchers adds uncertainty—nonstandard errors (NSEs)."

— *Menkveld et al. (2024)*

- Menkveld et al. (2024): 164 research teams test the same six hypotheses on the same data
  - for market efficiency, the interquartile range (IQR) of the estimates across teams is 6.7 percentage points (pp)
  - the median standard error (SE) implies a Gaussian IQR of only 3.4%, half the 6.7 pp
- On course data, a mini-multiverse of 27 specifications per index shows the same phenomenon
  - IQR of the estimated annual trend: 64.6 pp for the Euro Stoxx 50 and 82.6 pp for the BET (Bucharest Exchange Trading), the benchmark index of the Bucharest Stock Exchange
  - benchmark due to sampling alone: 41.3 pp and 43.4 pp, respectively
- The choice of risk model also changes the conclusion: VaR 1% (Value-at-Risk) for the BET over 5 432 evaluation days
  - the Normal model is breached on 2.25% of days, more than twice as often as it should
  - only FHS (filtered historical simulation), at 1.23%, lies in the 95% binomial range [0.74%, 1.27%]
- The chapter reviews Chapters 0–18 through seven pitfalls and prepares the team project
  - replication, pre-registration and a specification grid

▸ Central chart: A Mini-Multiverse on Course Data: 27 Paths per Index

What Turns a Result into a Finding? (1/2)

The Chapter in Brief

- The question: when does an empirical result about markets deserve trust?
  - and when does it only reflect the data, the design or luck?
- Part I: the course in one picture, a review of Chapters 0–18
  - seven pitfalls
  - one integrated case study, from raw prices of the BET to a backtested value-at-risk (VaR) at 1%
  - BET: the benchmark index of the Bucharest Stock Exchange, BVB
- The main thread: GARCH forecasts volatility, DCC the correlations, VaR and ES turn them into a loss
  - GARCH: generalised autoregressive conditional heteroskedasticity
  - DCC: dynamic conditional correlation
  - ES: expected shortfall, the mean loss beyond VaR
- Part II: the team project and artificial intelligence (AI) in the project
  - a research case study on nonstandard errors
  - the AI discovery loop
  - the project step by step: ▸ Appendix

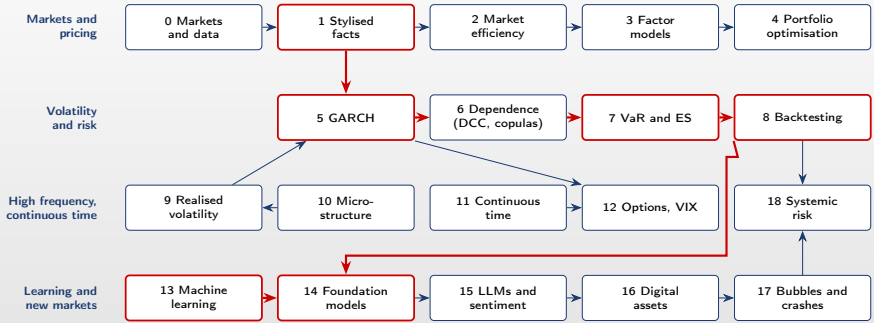
What Turns a Result into a Finding? (2/2)

What you should be able to do

- Compute a backtested risk forecast from the data, with valid inference
- Interpret an empirical result after checking the seven pitfalls in your own code
- Compute the power of the main test within a project plan with replication, identification and a specification grid
- Further Topics and Appendix: reading for familiarity

- Detailed examples and advanced topics are in Further Topics, after the Conclusions
  - the footer button leads there; the Back button returns to the starting slide

Structure of the Course



What Connects the Steps of the Analysis

- Red: the main thread, from stylised facts to risk forecasts we can test
  - Chapter 1: heavy tails, volatility clustering, almost no linear predictability
  - Chapters 5–8: GARCH gives σ_{t+1}, VaR and ES the loss
  - backtesting says whether to trust it
- Chapters 9–12 measure volatility better and model it in continuous time
  - intraday data, option prices, the VIX
  - VIX: the implied-volatility index of the Standard & Poor's 500 index (S&P 500)
- Chapters 13–15 bring flexible models, including large language models (LLMs)
  - judged with the tools of Chapter 8
- Chapters 16–18 apply the same toolkit to crypto, bubbles and the whole financial system
- Special Chapter 20: realised volatility with path signatures
  - on the VOLARE data (VOLatility Archive for Realized Estimates)
- Further inference formulas: the Shanken covariance, quasi-maximum likelihood (QMLE), the empirical quantile:
  ▸ Appendix

Five Questions for Every Model

Before estimating

- **Data**: which series, which calendar, which adjustments (dividends, splits, holidays)?
- **Model**
  - which stylised fact does it capture?
  - which stylised fact does it ignore?
- **Estimation**: on which window, with which information at time t?

After estimating

- **Evaluation**: out of sample, with a loss that matches the target?
  - QLIKE (the quasi-likelihood loss) for variance
  - FZ0 (the Fissler–Ziegel score) for a (VaR, ES) pair
- **Uncertainty**: robust standard errors, bootstrap, a test against a simple benchmark?
- If one answer is missing, the result is not yet a finding

Chapter 0: Financial Markets in 2026

- Markets combine exchanges, over-the-counter dealers, funds, algorithms and blockchains
  - exchange-traded funds (ETFs)
- SPY, 2015–2026: CAGR 13.7%, annualised volatility 17.6%
  - SPY: S&P 500 ETF, dividends reinvested
  - CAGR: compound annual growth rate
- Bitcoin, same period: CAGR 60.6%, annualised volatility 66.7%
- BET, July 2007 to February 2009: a drawdown of -82.5%
- Standard errors of the mean and the volatility from i.i.d. log returns
  - Y: sample length in years; T: number of daily returns; K: population kurtosis

$$\mathrm{SE}(\hat{\mu}_{\mathbf{ann}}) = \frac{\sigma_{\mathbf{ann}}}{\sqrt{Y}}, \qquad \mathrm{SE}(\hat{\sigma}) \approx \sigma \sqrt{\frac{K-1}{4T}}$$

  - $\mathrm{SE}(\hat{\sigma})$ requires a finite fourth moment, which Chapter 1 questions
- SPY: $\mathrm{SE}(\hat{\mu}_{\mathbf{ann}}) \approx 5.1\%$ for an annualised mean log return of 12.8%
  - means are barely estimable, volatilities are

Chapter 1: Stylised Facts and Returns

- Returns are uncorrelated but not independent Cont (2001)
  - their magnitude is predictable, their sign largely is not
- Tail index α: the probability of a loss above x decays like $x^{-\alpha}$
- The Hill estimator uses the k largest losses, ordered $L_{(\mathbf{1})} \geq \cdots \geq L_{(k+\mathbf{1})}$ Hill (1975)

$$\hat{\alpha} = \Big[\frac{1}{k}\sum_{i=\mathbf{1}}^{k} \ln \frac{L_{(i)}}{L_{(k+\mathbf{1})}}\Big]^{-\mathbf{1}}, \qquad \mathrm{SE}(\hat{\alpha}) \approx \frac{\hat{\alpha}}{\sqrt{k}} \text{ for i.i.d. losses}$$

- BET, $k = 153$: $\hat{\alpha} = 2.39$, block-bootstrap 95% CI $[2.12, 3.03]$
  - CI: confidence interval
- The i.i.d. interval $\hat{\alpha} \pm 1.96\,\hat{\alpha}/\sqrt{k} = [2.01, 2.77]$ is narrower
  - large losses come in clusters
- $\alpha < 4$: the fourth moment is likely infinite
- Days beyond 5σ are more than 1000 times more frequent than under the Normal distribution
  - the Normal distribution gives $2\Phi(-5) \approx 5.7 \times 10^{-\mathbf{7}}$ a day
  - that is, one such day in about 7 000 years

Chapter 2: Market Efficiency

- The efficient market hypothesis: prices fully reflect the information set
- A test of abnormal-return predictability is a joint test Fama (1970)
  - of efficiency and of the expected-return model
- The q-period variance ratio and its heteroskedasticity-robust statistic Lo & MacKinlay (1988)

$$VR(q) = \frac{\mathrm{Var}(r_t + \cdots + r_{t-q+1})}{q\,\mathrm{Var}(r_t)}, \qquad z^*(q) = \frac{\widehat{VR}(q) - 1}{\widehat{\mathrm{SE}}^*} \Rightarrow N(0,1)$$

  - $z^*(q)$ requires a finite fourth moment, which Chapter 1 questions
- Several horizons at once: $\max_q |z^*(q)|$ with the Chow–Denning critical values Chow & Denning (1993)
- Share of rolling windows that reject the random walk
  - S&P 500 6.7%, BET 22.4%, Bitcoin 0.0%
- Statistical rejection is not exploitable profit

Chapter 3: Factor Models and Asset Pricing

- Expected returns should reward only exposure to systematic factors
- The GRS (Gibbons–Ross–Shanken) test that all intercepts (pricing errors) $\hat{\alpha}$ are zero Gibbons, Ross & Shanken (1989)
  - T periods, N assets and K factors

$$\mathrm{GRS} = \frac{T-N-K}{N}\,\frac{\hat{\alpha}^{\top}\hat{\boldsymbol{\Sigma}}^{-1}\hat{\alpha}}{1+\hat{\mu}_f^{\top}\hat{\boldsymbol{\Omega}}^{-1}\hat{\mu}_f} \;\sim\; F(N, T-N-K)$$

  - $\hat{\mu}_f$: factor means; $\hat{\boldsymbol{\Sigma}}$, $\hat{\boldsymbol{\Omega}}$: residual and factor covariances
  - the F is exact only under i.i.d. Normal errors
- 25 size and book-to-market portfolios
  - SML (security market line) slope -4.4%
- GRS $= 4.18$, against a 5% critical value below 1.6
  - the CAPM (capital asset pricing model) is rejected ($p < 0.01$)
- Sector ETFs: the first principal component (the market) explains 66% of the variance

Chapter 4: Portfolio Optimisation

- Markowitz asks the right question, but estimation error dominates the answer Markowitz (1952)
- Standard error of the per-period Sharpe ratio, i.i.d. Normal returns over T periods Lo (2002)

$$\mathrm{SE}(\widehat{SR}) \approx \sqrt{(1 + \widehat{SR}^2/2)/T}$$

  - $\mathrm{SE}(\widehat{SR})$ requires a finite fourth moment, which Chapter 1 questions
- 60-month samples: median in-sample Sharpe 1.56
  - median true Sharpe only 0.30
- 96% of the estimated tangency portfolios are truly worse than 1/N DeMiguel, Garlappi & Uppal (2009)
  - 1/N is a hard benchmark
- Portfolio A against 1/N: test $SR_A - SR_B = 0$ with HAC (heteroskedasticity- and autocorrelation-consistent) standard errors or studentised bootstrap inference Ledoit & Wolf (2008)
  - not by comparing two point estimates

Chapter 5: Conditional Volatility: GARCH

- Volatility clusters, so today's shock predicts tomorrow's variance Engle (1982), Bollerslev (1986)
- The GARCH(1,1) model and its h-day forecast
  - ε_t: the shock; α, β: the weights of the last shock and of the last variance

$$\sigma^2_{t+1} = \omega + \alpha\,\varepsilon^2_t + \beta\,\sigma^2_t, \qquad \bar{\sigma}^2 = \frac{\omega}{1-\alpha-\beta}$$

$$\sigma^2_{T+h|T} = \bar{\sigma}^2 + (\alpha+\beta)^{h-1}\left(\sigma^2_{T+1|T} - \bar{\sigma}^2\right)$$

- The persistence $\alpha+\beta$ is the rate at which the forecast returns to the long-run level $\bar{\sigma}^2$
  - the half-life is $h_{1/2} = \ln 0.5 / \ln(\alpha+\beta)$
- S&P 500, 2000–2026: $\alpha+\beta = 0.9948$, so $h_{1/2} = 132$ days
- The BET persistence cannot be bounded away from 1
  - pitfall "Persistence Near 1"

Chapter 6: Multivariate Volatility and Dependence

- Correlations move with the macro regime
  - diversification is not guaranteed
- Crisis correlation rises mechanically with volatility, so test on the adjusted correlation Forbes & Rigobon (2002)

$$\rho^{*} = \frac{\rho}{\sqrt{1 + \delta(1 - \rho^{2})}}$$

  - ρ: crisis correlation
  - δ: relative rise in the variance of the source market from calm to crisis
  - the test on ρ^{*} requires a finite fourth moment, which Chapter 1 questions
- SPY–TLT: ρ rose from -0.40 to 0.11 after 2022
  - TLT: long US (United States) Treasury bond ETF
- Over the same time a 60/40 portfolio went from 10.4% to 13.2% volatility
- Bitcoin–SPY DCC correlation: 0.04 before 2020, 0.31 since

Chapter 7: Value-at-Risk and Expected Shortfall

- VaR is a quantile of the loss
- ES, the average loss beyond it, is coherent
- Definitions, with q_α the α-quantile of the returns r (continuous distribution)

$$\mathrm{VaR}_\alpha = -q_\alpha, \qquad \mathrm{ES}_\alpha = -\mathbb{E}\big[r \mid r \le q_\alpha\big] = \frac{1}{\alpha}\int_0^\alpha \mathrm{VaR}_u \, du$$

- S&P 500, daily: historical $\mathrm{VaR}_{1\%} = 3.45\%$, $\mathrm{ES}_{1\%} = 5.03\%$
- The Normal distribution gives only 2.80%
  - heavy tails matter
- Filtered historical simulation and conditional EVT (Extreme Value Theory) react to volatility within days McNeil & Frey (2000)

Chapter 8: Backtesting and Evaluating Risk Forecasts

- Test the number and the timing of breaches: coverage and independence Kupiec (1995), Christoffersen (1998)
- Kupiec, with x breaches in T days and the observed rate $\hat{\pi} = x/T$

$$LR_{uc} = -2 \ln \frac{(1-\alpha)^{T-x}\alpha^{x}}{(1-\hat{\pi})^{T-x}\hat{\pi}^{x}} \sim \chi^2_1$$

- Christoffersen: $LR_{ind} \sim \chi^2_1$ compares the breach rate after a quiet day with the rate after a breach
  - formula: Step 4 of the case study
- S&P 500, VaR 1%, 2007–2026: breach rate (target 1%)
  - Normal 2.70%
  - HS (historical simulation) 1.67%
  - FHS (filtered historical simulation) 1.23%
  - GARCH-EVT (GARCH filter with a tail from extreme value theory) 1.15%
- On the S&P 500 and the BET, a GARCH filter plus a flexible tail wins
  - FHS, GARCH-EVT

Chapter 9: Realised Volatility

- Intraday prices measure daily variance (RV, realised variance)
- The log-HAR (heterogeneous autoregressive) model approximates its observed persistence over finite horizons Corsi (2009)

$$\ln RV_{t+\mathbf{1}} = b_{\mathbf{0}} + b_d \ln RV_t + b_w \overline{\ln RV}_t^{(\mathbf{5})} + b_m \overline{\ln RV}_t^{(\mathbf{22})} + u_{t+\mathbf{1}}$$

  - $\overline{\ln RV}_t^{(\mathbf{5})}$, $\overline{\ln RV}_t^{(\mathbf{22})}$: the mean of $\ln RV$ over the last 5 and 22 days
  - a stationary HAR (Heterogeneous AutoRegressive, model) is a finite autoregressive (AR) model: short memory asymptotically
- SPY: kurtosis 15.73 for open-to-close returns
  - 2.66 after dividing by $\sqrt{RV_t}$
- log-HAR beats GARCH(1,1)-t in QLIKE
  - by 10% for SPY ($p = 0.021$)
  - by 22% for Bitcoin
  - the Diebold–Mariano (DM) test on QLIKE requires finite moments and weak dependence of the loss differential d_t, not of the returns

Chapter 10: Market Microstructure

- Prices form in limit order books
- The spread pays for processing, inventory and adverse selection
- The Roll spread: the bid–ask bounce
  - makes consecutive changes of the transaction price, Δp_t, negatively correlated

$$s_{\mathbf{Roll}} = 2\sqrt{-\mathrm{Cov}(\Delta p_t, \Delta p_{t-\mathbf{1}})}$$

- When the sample covariance is positive the estimator is undefined
  - report how often
- Amihud ratio (basis points, bp, per 1 million US dollars, USD)
  - SPY 0.0016, Banca Transilvania 24.7, Transelectrica 1024.0
- SPY: the opening 5-minute bar moves 2.06 times as much as a midday bar

Chapter 11: Continuous-Time Models

- Brownian motion is the limit of random walks with ever smaller steps
- Itô's lemma solves the stochastic differential equations of prices
- Normality of the log returns of geometric Brownian motion (GBM)
  - the Jarque–Bera test, with S the sample skewness and K the sample kurtosis

$$JB = \frac{T}{6}\left(S^2 + \frac{(K-3)^2}{4}\right) \sim \chi^2_2$$

- S&P 500 against GBM: 62 days beyond ± 4 standard deviations, against 0.59 expected
- The Hill index of S&P 500 losses is below 4 (Chapters 1 and 16)
  - the fourth moment is likely infinite
- Jumps: $(RV - BV)$ standardised $\Rightarrow N(0,1)$ Barndorff-Nielsen & Shephard (2006)
  - BV: the bipower variation, robust to jumps

Chapter 12: Options and the Volatility Surface

- Black–Scholes prices options by hedging Black & Scholes (1973)
  - implied volatility: the value of σ extracted from option prices, a risk-neutral quantity, not a forecast of physical volatility
- The (ex-ante) variance risk premium: $E_t^Q[RV_{t,t+\mathbf{21}}] - E_t^P[RV_{t,t+\mathbf{21}}]$
  - $VIX_t^{\mathbf{2}}$ estimates the first term, the risk-neutral expected variance over the next month; the second needs a forecast
  - $RV_{t,t+\mathbf{21}}$: realised variance over the next 21 trading days
- The ex-post realised gap, known only at $t + 21$:

$$VIX_t^{\mathbf{2}} - RV_{t,t+\mathbf{21}}$$

- Consecutive values overlap, so the mean needs Newey–West standard errors with at least 21 lags
  - the standard error of the mean requires a finite fourth moment, which Chapter 1 questions
- 1990–2026: mean VIX 19.4
  - mean realised volatility over the next month 15.3
- The ex-post realised gap is positive on 86% of days, but not in crashes
  - its sign does not identify the sign of the ex-ante premium

Chapter 13: Machine Learning in Finance

- Flexible models, but little signal and much noise López de Prado (2018)
  - the finance-specific part is the validation process
- Shuffled K-fold reports 69.4% accuracy on pure noise
  - leakage: test information leaking into training
- Remedy: purging and an embargo after the test fold
  - purging: dropping training labels that overlap the test fold
- Selection among N strategies: the deflated Sharpe ratio tests $SR > \mathbb{E}[\max_N \widehat{SR}]$ Bailey & López de Prado (2014)
  - the threshold: the expected best estimated Sharpe ratio of N no-skill strategies
  - the threshold grows with N
- S&P 500, fractional differencing of order d: a unit-root test with one lag rejects from $d^* = 0.25$
  - a rejection by the ADF (Augmented Dickey–Fuller) test does not prove stationarity: the finite filter has weights summing to a nonzero value (0.178 at $d = 0.25$), so the series keeps a component proportional to the log price
  - its correlation with the log price, 0.98, only illustrates this component
- No model beats the naive rule "always up"

Chapter 14: Deep Learning and Time-Series Foundation Models

- LSTMs (long short-term memory networks) and Transformers learn from sequences
- Foundation models forecast zero-shot
  - without training on the series at hand
- The Diebold–Mariano test on loss differences d_t between two models Diebold & Mariano (1995)
  - $\hat{\Omega}$: the HAC long-run variance of d_t

$$t_{DM} = \frac{\bar{d}}{\sqrt{\hat{\Omega}/T}} \Rightarrow N(0,1)$$

- SPY realised volatility, QLIKE (lower is better)
  - Chronos-2 0.223, log-HAR 0.226, GARCH-t 0.273, on the exact grid of 78 five-minute bars
- Chronos-2 and log-HAR are not significantly different (DM $p = 0.539$)
  - the DM test on QLIKE requires finite moments and weak dependence of the loss differential d_t, not of the returns
- Chronos-2 VaR 1% without recalibration on the S&P 500: 1.67% breaches
- With many candidate models: the model confidence set (MCS) Hansen, Lunde & Nason (2011)

Chapter 15: LLMs and Sentiment Analysis

- From dictionaries to FinBERT and LLMs: text becomes a number
- Predictive regression $r_{t+\mathbf{1}} = a + b\,S_t + \varepsilon_{t+\mathbf{1}}$, S_t the sentiment score Stambaugh (1999)
  - a persistent S_t whose innovations correlate with returns biases $\hat{b}$ and its t-test
- Classifiers scored on the same texts are compared as pairs, with the exact McNemar test
- Qwen 7B vs 1.5B on financial headlines: $+21.5$ percentage points in accuracy
  - McNemar $p < 0.001$
- A headline-sentiment strategy: Sharpe $+0.63$ before costs
  - and -1.48 after costs
- Break-even cost: 1.5 bp per trade

Chapter 16: Digital Assets and DeFi

- The chapter case study (Pele et al. (2023)): cryptocurrencies form a distinct asset class
  - in a feature space where a tail factor separates them
- DeFi (decentralised finance) runs on smart contracts
- Bitcoin–S&P 500 link: sup-Wald 20.1 Andrews (1993)
  - sup-Wald: the largest Wald statistic over candidate break dates
  - 5% critical value 11.6
  - the Wald statistic requires $\mathbb{E}[x_t^2 u_t^2] < \infty$ (regressor $\times$ error), a fourth-order condition that Chapter 1 questions
- Estimated break: 9 Mar 2020, not the 2024 ETF launch
- Hill tail index of losses
  - Bitcoin 2.73, 95% CI $[2.12, 3.33]$
  - S&P 500 2.87, 95% CI $[2.07, 3.67]$
- The Bitcoin $-$ S&P 500 difference, from a joint block bootstrap: -0.14, 95% CI $[-1.68, 0.71]$
  - the interval contains 0: no evidence of heavier Bitcoin tails; a non-rejection is not equality
  - the overlap of the two intervals is not a test of the difference

Chapter 17: Bubbles and Crashes

- A bubble is a price path that grows faster than fundamentals can justify
- GSADF: the largest right-tailed ADF statistic Phillips, Shi & Yu (2015)
  - over all sample fractions $[r_1, r_2]$ of length at least r_0

$$GSADF = \sup_{r_2 \in [r_0, 1]} \sup_{r_1 \in [0, r_2 - r_0]} ADF_{r_1}^{r_2}$$

- Volatility shifts distort the size
  - use wild-bootstrap critical values
- Bitcoin 2014–2026 (weekly): GSADF 3.85 against a 95% wild-bootstrap critical value of 2.82 (2.27 Monte Carlo)
  - explosive episodes are detected
- One vs two Markov regimes for Bitcoin: $LR = 123.2$
  - against a parametric-bootstrap 95% critical value of 11.8 ($p = 0.005$)

Chapter 18: Systemic Risk and Networks

- Systemic risk: losses of one institution that spread to the whole system
- ΔCoVaR: the change in the system's VaR when institution i moves from its median to distress
  - X^i: the return of institution i

$$\Delta\text{CoVaR}^i = \text{CoVaR}_\alpha^{sys|X^i=-\text{VaR}_\alpha^i} - \text{CoVaR}_\alpha^{sys|X^i=\textbf{median}}$$

- ΔCoVaR 1% of eleven banks: 4 of 25 within-region pairs differ at 5%
  - 0 after the Holm correction
- Overlap of intervals alone would have ranked none
- LRMES, simulated from GJR–GARCH (Glosten–Jagannathan–Runkle asymmetric GARCH) and DCC Brownlees & Engle (2017)
  - LRMES: the expected loss of a bank when the market falls by more than 10% over 22 days
  - JPMorgan 10.6% today, 23.7% with the information of 31 March 2020

Design Errors: Look-Ahead Bias and Overfitting

- **Look-ahead bias**: a full-sample estimate leaks the future
  - any parameter, threshold or scaling estimated on the full sample
  - Examples: full-sample GARCH parameters or standardisation, labels that overlap the test set (Chapter 13; the effect on the BET: case study)
  - Remedy: rolling or expanding windows, purging and embargo, the information set $\mathcal{F}_{t-1}$ written next to every forecast
- **Overfitting**: the in-sample fit rewards noise (Chapter 4)
  - the noise will not repeat out of sample
  - Backtest overfitting: trying many strategies and keeping the best Bailey et al. (2017)
  - The deflated Sharpe ratio corrects for the number of trials and for non-normal returns Bailey & López de Prado (2014)
  - Remedy: the design fixed before the test period, every model tried reported, a simple benchmark (1/N, historical simulation, random walk, log-HAR)

Design Error: Multiple Testing (1/2)

- Chapter 3: with 300 useless factors, on average 15.0 pass $|t| > 1.96$
  - the best reaches $|t| \approx 3.06$
  - hence the literature asks for $|t| > 3$ Harvey, Liu & Zhu (2016)
- Calendar anomalies, trading rules, specifications
  - each extra test adds a chance of a false discovery
- Remedy: count your tests and correct

Design Error: Multiple Testing (2/2)

- With m tests, ordered p-values $p_{(1)} \leq \cdots \leq p_{(m)}$ and level a, reject $H_{(i)}$ if

Bonferroni:	$p_{(i)} \leq a/m$
Holm:	$p_{(j)} \leq a/(m-j+1)$ for every $j \leq i$
Benjamini–Hochberg:	$i \leq \max\{k : p_{(k)} \leq k\, a/m\}$

  - Bonferroni and Holm Holm (1979) control the chance of any false discovery (FWER)
  - Holm rejects at least everything Bonferroni rejects
  - Benjamini–Hochberg (BH) Benjamini & Hochberg (1995) controls the expected share of false discoveries among the rejections (FDR)
  - For forecasts: the model confidence set (MCS) instead of many pairwise tests Hansen, Lunde & Nason (2011)

Data Error: Aligning Returns Instead of Prices

- Bitcoin trades every day, SPY only on weekdays
  - Wrong: compute each return on its own calendar, then keep the common days: Bitcoin's weekend moves disappear
  - Right: keep the common days of the **prices**, then compute returns: Monday covers Friday to Monday for both
- Bitcoin, 2020–2026, on the common days: cumulative return 1 025% (right) against 680% (wrong)
  - Monday standard deviation 5.04% (right) against 3.83% (wrong): the wrong alignment hides weekend risk
- Remedy: align prices on common days first
  - one series alone stays on its own calendar

Inference Error: Heavy Tails Ignored

- Classical standard errors assume independence and finite moments of the quantity analysed
  - variances, autocorrelations and Sharpe ratios: the fourth moment of returns; DM: the moments of the loss differential; Gaussian QMLE: the fourth moment of the standardised innovations
  - BET: $\hat{\alpha} = 2.39$, 95% CI $[2.12, 3.03]$ (Chapter 1): under a Pareto-type tail, consistent with an infinite kurtosis, not a proof
  - The sample excess kurtosis (11.1) estimates no finite parameter if $\alpha < 4$, and a bootstrap interval for it is not valid Athreya (1987): report the tail index instead
- Remedy: HAC errors, block bootstrap, intervals that respect the parameter boundary
  - block bootstrap where it is valid
- The Newey–West HAC standard error of a mean $\bar{d}$ Newey & West (1987): $\mathrm{SE}(\bar{d}) = \sqrt{\hat{\Omega}/T}$, with

$$\hat{\Omega} = \hat{\gamma}_0 + 2\sum_{j=1}^{L}\Big(1 - \frac{j}{L+1}\Big)\hat{\gamma}_j$$

  - $\hat{\gamma}_j$: the sample autocovariance at lag j; L: the number of lags; the Bartlett weights guarantee $\hat{\Omega} \geq 0$
  - The same long-run variance enters the Diebold–Mariano test (Step 5)

Persistence Near 1: Which Boundary? (1/2)

- $p = \alpha + \beta \leq 1$ is a restriction we choose, not a property of GARCH
  - its interior gives a finite unconditional variance and a finite half-life
  - equality is the IGARCH (integrated GARCH) boundary
- Without it, p slightly above 1 is admissible Nelson (1990)
  - the process can still be strictly stationary, when $E \ln(\alpha z_t^2 + \beta) < 0$
  - but it has no finite variance and no half-life
- BET: $\hat{p} = 0.9907$ is inside the region
  - the difficulty is the **true** p, which may equal 1
- At $p = 1$ the estimator is not approximately Normal Andrews (1999)
  - so the Wald interval [0.974, 1.007] is uninformative under our restriction, rather than "wrong"

Persistence Near 1: Which Boundary? (2/2)

- The likelihood-ratio test of H_0: $p = 1$ has the boundary distribution $\frac{1}{2}\chi^2_0 + \frac{1}{2}\chi^2_1$
  - the 5% critical value is 2.71 instead of 3.84
  - BET: $LR = 1.01$, p-value 0.16, so IGARCH is not rejected
- Profile-likelihood CI within $p \leq 1$: $[0.975, 1]$
  - so the half-life is at least 27 days, with no finite upper bound
- Boundary critical values apply only when the true parameter can sit on the edge of the space you imposed
  - not whenever an estimate is close to some bound

Evaluation Error: a Loss That Cannot Rank the Target

- A variance forecast f_t is compared with a noisy proxy $\hat{\sigma}_t^2$ (r_t^2 or RV_t)
  - the MSE (mean squared error) loss: $(\hat{\sigma}_t^2 - f_t)^2$
  - QLIKE: $\hat{\sigma}_t^2/f_t - \ln(\hat{\sigma}_t^2/f_t) - 1$
- With a conditionally unbiased proxy, MSE and QLIKE rank forecasts as the true variance would Patton (2011)
- QLIKE depends only on the ratio proxy/forecast, so volatile days weigh less than under MSE
- Chapter 9: log-HAR beats GARCH in QLIKE
  - the MSE comparison is inconclusive
- ES alone has no strictly consistent scoring function (one minimised in expectation only by the correct forecast) on the usual broad classes of distributions
- The pair (VaR, ES) has one Fissler & Ziegel (2016): the FZ0 score (Step 5)
  - with Diebold–Mariano tests Diebold & Mariano (1995)
  - their precision is limited in finite samples Pele & Mazurencu-Marinescu-Pele (2026)
- Remedy: choose the loss before the comparison, and state it in the project report

Integrated Case Study: Methods from Four Chapters (1/2)

The question

- How much can a BET index position lose tomorrow, with probability 1%?
- Which forecast of that loss would we have trusted in real time, 2005–2026?

Data

- BET index, official daily closing values, 6 Jan 2000 – 18 Sep 2026
  - 6 670 daily log returns
- Log return in %: $r_t = 100 \ln(P_t / P_{t-1})$, loss $L_t = -r_t$
- Removed: duplicate records and weekend-dated rows
  - days with an unchanged close kept

Integrated Case Study: Methods from Four Chapters (2/2)

The five steps

- Chapter 1: stylised facts, with robust inference
- Chapter 5: GARCH(1,1) with Student-t innovations
- Chapter 7: one-day VaR 1% from four models
- Chapter 8: coverage, independence and Basel zones
- Formal comparison of the four models
  - FZ0 loss, Diebold–Mariano (DM) tests, model confidence set

Uses of the Symbol α in the Case Study

Where	Meaning of α	Value	Enters
Step 1, Hill	tail index of losses	2.39	$P(L > x)$ decays like $x^{-\alpha}$
Step 2, GARCH	weight of yesterday's squared shock	0.192	$\sigma_t^2 = \omega + \alpha\varepsilon_{t-1}^2 + \beta\sigma_{t-1}^2$
Steps 3–4, VaR and Kupiec	VaR level: promised breach probability	1%	LR_{uc}, DQ
Step 5, FZ0 and ES	ES level used to score the pair (VaR, ES)	2.5%	$S(v, e; y)$

- Before substituting a number, ask which α the formula means
  - the Kupiec test needs $\alpha = 0.01$, never 2.39 or 0.192
- The letter p also has two meanings here
  - the GARCH persistence $p = \alpha + \beta$ and a p-value

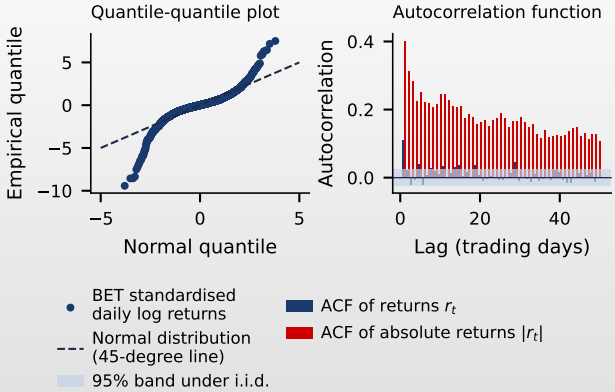
Step 1: What We Measure and How We Test It

- Tails: the Hill tail index $\hat{\alpha}$ of losses, with a block-bootstrap 95% CI
  - formula: the Chapter 1 review
- Volatility clustering: the autocorrelations of $|r_t|$ and the Ljung–Box test on r_t^2

$$Q(m) = T(T+2)\sum_{k=1}^{m} \frac{\hat{\rho}_k^2}{T-k} \;\sim\; \chi^2_m \text{ under } H_0$$

  - H_0: the first m autocorrelations are zero; $\hat{\rho}_k$: the sample autocorrelation at lag k
  - here $m = 10$, so the 5% critical value is 18.3
- Next chart
  - left, quantiles of the standardised returns against those of the Normal distribution
  - right, the ACF (autocorrelation function) of r_t and of $|r_t|$

Step 1: Stylised Facts of the BET



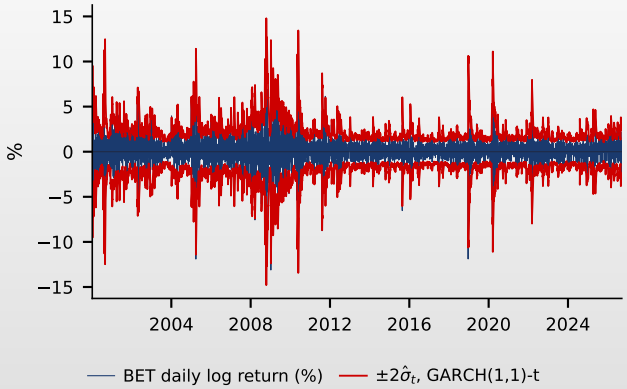
MFM_ch19_case_study ▶ Interactive

Step 1: Reading the Chart

- **Heavy tails** (left)
  - Hill tail index of losses $\hat{\alpha} = 2.39$, 95% CI [2.12, 3.03]
  - the kurtosis is not reported, because $\hat{\alpha} < 4$
  - 51 days beyond ± 4 standard deviations, against 0.42 expected under the Normal distribution
  - worst daily loss -13.1%, on 7 Jan 2009
- **Volatility clustering** (right)
  - ACF of $|r_t|$ at lag 1: 0.40, slow decay (the band assumes i.i.d. returns)
  - Ljung–Box $Q(10)$ on r_t^2: 2 440, against the critical value 18.3
  - the test on r_t^2 requires a finite eighth moment for the usual χ^2 limit
- First-order autocorrelation of the returns: 0.11
  - typical of a less liquid market (Chapters 2, 10)
- A Normal, constant-volatility VaR is therefore too low in stress and too high in calm periods

MFM_ch19_case_study

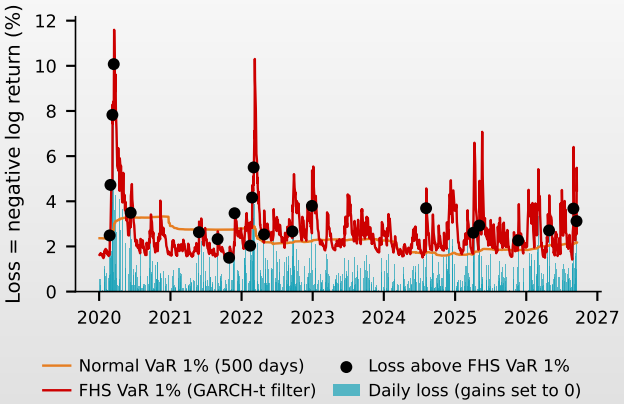
Step 2: GARCH(1,1)-t Volatility



- BET daily log returns with $\pm 2\hat{\sigma}_t$ from a GARCH(1,1) with Student-t innovations, 2000–2026
  - Robust standard errors: $\hat{\alpha} = 0.192$ (0.025), $\hat{\beta} = 0.799$ (0.024), $\hat{\nu} = 5.01$
  - $\hat{\nu}$, the degrees of freedom of the t innovations, is low: heavy tails even after filtering
- Persistence $\hat{\alpha} + \hat{\beta} = 0.9907$
  - a half-life of about 74 trading days
  - it cannot be bounded away from 1 (pitfall "Persistence Near 1")
- Peak daily volatility 7.39% on 15 Oct 2008
  - at the end of the sample (18 Sep 2026), $\hat{\sigma}_T = 1.67\%$

MFM_ch19_case_study ▶ Interactive

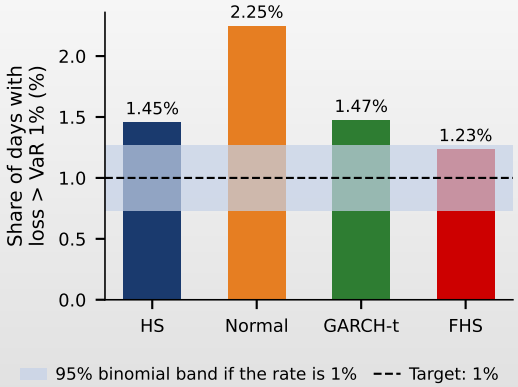
Step 3: Losses and VaR 1%, 2020–2026



- Daily BET losses (gains set to 0), Normal and FHS VaR 1%
  - black dots: losses above the FHS VaR
- FHS reacts within days
  - it peaked at 11.6% on 18 Mar 2020, during the COVID-19 crash
  - The Normal VaR on a 500-day window moves only slowly
- 1 675 days, about 17 breaches expected
  - FHS 22, Normal 41
  - Normal breaches cluster in the first weeks of every stress episode
- Last VaR forecast: FHS 4.48%, Normal 2.18%

MFM_ch19_case_study ▶ Interactive

Step 4: Breach Rates



- Share of days with a loss above the VaR 1% forecast, over the $T = 5\,432$ evaluation days of Step 3
- Light-blue band: the 95% binomial range [0.74%, 1.27%] if the true rate is 1%
- Only FHS (1.23%) lies inside the band
  - the Normal model (2.25%) breaches more than twice as often as it should

MFM_ch19_case_study ▶ Interactive

Step 5: Results

Model	DQ	p_{DQ}	FZ0	p_{MCS}	ES zone
HS	192.1	<0.001	1.422	0.00	standard
Normal	364.6	<0.001	1.495	0.00	red
GARCH-t	17.9	0.007	1.175	0.18	green
FHS	9.9	0.129	1.157	1.00	green

- Only FHS passes the DQ (Dynamic Quantile) test at 5% ($p = 0.129$)
  - GARCH-t fails it ($p = 0.007$) through its excess of breaches
- FZ0 (lower is better): FHS and GARCH-t are not significantly different (DM $t = 1.39$)
  - both beat HS and Normal ($|t| > 4.9$)
  - The 10% model confidence set is {GARCH-t, FHS}
- Comparative ES backtest against HS as the standard model: Normal red, GARCH-t and FHS green
- The GARCH-filtered models are superior
  - FHS remains the only model rejected by no test (coverage, independence, DQ)
- Type of evidence: comparison of procedures (it shows which method works better on the same data, not why)
  - FZ0 scores the (VaR 2.5%, ES 2.5%) pair, not the VaR 1% of Steps 3–4

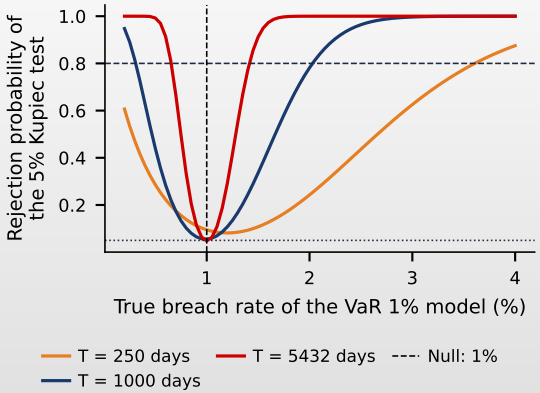
The Power of a Backtest: Definition

- The question: if the VaR 1% model is wrong, how often does the Kupiec test notice?
- Power at the true breach rate π: an exact binomial probability
  - the probability of the rejection region R_T of the 5% test

$$\text{Power}(\pi) = \sum_{x \in R_T} \binom{T}{x} \pi^x (1-\pi)^{T-x}, \qquad R_T = \{x : LR_{uc}(x) > 3.84\}$$

  - At $\pi = 1\%$ the same sum gives the actual size (the rejection rate of a correct model) of the nominal 5% test: 9.48% at $T = 250$
  - At $T = 250$, R_T holds 0 breaches and any number from 7 up
- Minimum detectable effect (MDE): the smallest true rate $\pi > 1\%$ with $\text{Power}(\pi) \geq 80\%$
- Power grows with T
  - the standard error of the observed rate $\hat{\pi}$ is $\sqrt{\pi(1-\pi)/T}$

How Much Can a Backtest Detect? The Power of Kupiec



- Exact binomial probability that the 5% Kupiec test rejects a VaR 1% model
  - by its true breach rate
- $T = 250$ (one Basel year): a true rate of 2% is detected only 24% of the time
- MDE, searched on a grid with steps of 0.05 percentage points
  - 3.65% at $T = 250$, 2.05% at $T = 1\,000$, 1.45% at $T = 5\,432$
  - Even at $T = 5\,432$, a true 1.3% is detected with probability 55% only
- A non-rejection with low power is "no evidence", not "no problem"
  - state the power before the test

MFM_ch19_case_study ▶ Interactive

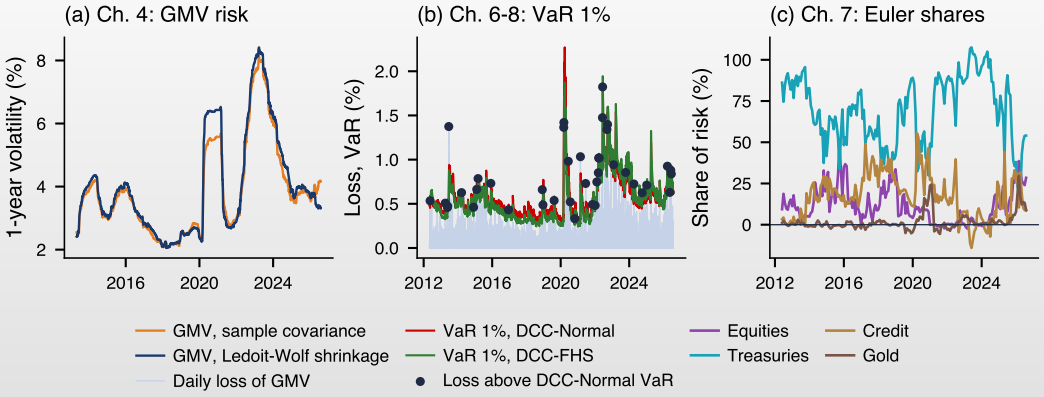
Multi-Asset Portfolio: the Question and the Decision Rule

- The question: which global minimum-variance (GMV) portfolio of eight ETFs do we hold, and which one-day VaR 1% forecast do we rely on?
- The multi-asset universe of Chapter 4, daily adjusted prices, Apr 2007 – Jul 2026
  - equities: SPY (US), EFA (developed markets ex-US), EEM (emerging markets)
  - US Treasuries: TLT (20+ years), IEF (7–10 years)
  - corporate bonds: LQD (investment grade), HYG (high yield); gold: GLD (SPDR Gold Shares)
- The evaluation: 1 May 2012 – 31 Jul 2026, 3 583 days
- Four chapters
  - Chapter 4: monthly GMV weights from the last 60 months, with the sample covariance or with Ledoit–Wolf shrinkage Ledoit & Wolf (2004)
  - Chapter 6: conditional covariance H_t from GARCH(1,1) and DCC Engle (2002), re-estimated every January
  - Chapter 7: portfolio VaR 1% and Euler risk contributions
  - Chapter 8: the Kupiec (coverage) and Christoffersen (independence, conditional coverage) tests

The decision rule

- Weights: the GMV variant with the lower volatility to 2018; from 2019 we test the choice
- Forecast: a VaR model rejected at 5% by none of the three tests
- On a tie: the model with the lower mean VaR

Multi-Asset Portfolio: Weights, VaR and Risk Contributions



MFM_ch19_integrated_portfolio ▶ Interactive

Multi-Asset Portfolio: the Steps in Numbers

- **(a) Chapter 4**: daily out-of-sample volatility, annualised (in the chart: over a rolling year)
  - GMV with the sample covariance: 4.13% a year
  - GMV with Ledoit–Wolf shrinkage: 4.28% a year, with a mean shrinkage intensity $\delta = 0.17$
  - monthly turnover: 0.085 with Ledoit–Wolf, against 0.135
  - gross exposure $\sum_i |w_i|$ of the target weights: mean 1.80 with Ledoit–Wolf, against 2.42; maximum 2.30, against 3.78
  - to 2018: 3.08% against 3.15%: the rule picks the sample covariance; from 2019: 4.87% against 5.07%, so the choice holds
- **(b) Chapters 6–8**: $\sigma_{p,t} = \sqrt{w_t' H_t w_t}$, the forecast for day t, with w_t the weights at the start of the day (rebalanced monthly, then drifting with the prices)
  - DCC at the last re-estimation (Jan 2026, 4 712 days): $\hat{a} = 0.024$, $\hat{b} = 0.968$
  - DCC-Normal: VaR $= 2.326\, \sigma_{p,t}$
  - DCC-FHS: $\sigma_{p,t}$ times the 99% quantile of the standardised loss $-r_{p,s}/\sigma_{p,s}$ over the last 1,000 days
  - HS: the loss quantile over the last 500 days, with the weights w_t
  - 35.8 breaches of the VaR 1% expected
- **(c) Chapter 7**: the Euler share $w_i (H_t w_t)_i / w_t' H_t w_t$ of each asset class in the variance
  - it is also the share in VaR and ES under an elliptical distribution
  - mean over the evaluation: Treasuries 69%, credit 17%, equities 12%, gold 2%
  - 27 Mar 2020, the day with the highest $\sigma_{p,t}$ (0.98%): credit carries 63% of the risk with a weight of 25%

Multi-Asset Portfolio: the Result and the Decision

VaR 1%	Breaches	Rate	Kupiec p	indep. p	cond. p	mean VaR
DCC-Normal	40	1.12%	0.492	0.083	0.175	0.58%
DCC-FHS	52	1.45%	0.011	0.046	0.005	0.55%
HS, 500 days	48	1.34%	0.052	<0.001	<0.001	0.63%

- Type of evidence: a retrospective comparison of procedures, on one portfolio and one period
  - only DCC-Normal passes all three tests; the VaR model is chosen on the same period on which it is evaluated
  - HS fails on independence, DCC-FHS on coverage and on independence
  - the DCC-FHS history applies today's weights to the data they were estimated on (the in-sample understatement of risk of Chapter 4)
  - standard deviation of $r_{p,s}/\sigma_{p,s}$: 0.91 inside the FHS window, 0.95 outside it
- Limits
  - the independence p-value of DCC-Normal is 0.083, below 10%
  - the Kupiec MDE at $T = 3\,583$: 1.55%; no transaction costs
- **What I choose**: GMV with the sample covariance and the DCC-Normal VaR 1% (0.64% for 31 Jul 2026)
- **Under which conditions**: monthly weights and DCC re-estimated yearly, with only the information available on the forecast date
- **What would make me drop it**
  - a rejection of independence on new data
  - a gross-exposure limit c: imposed in the optimisation at every rebalancing and checked daily; GMV breaches it for $c < 3.78$ (daily: 3.86), Ledoit–Wolf for $c < 2.30$ (2.34)

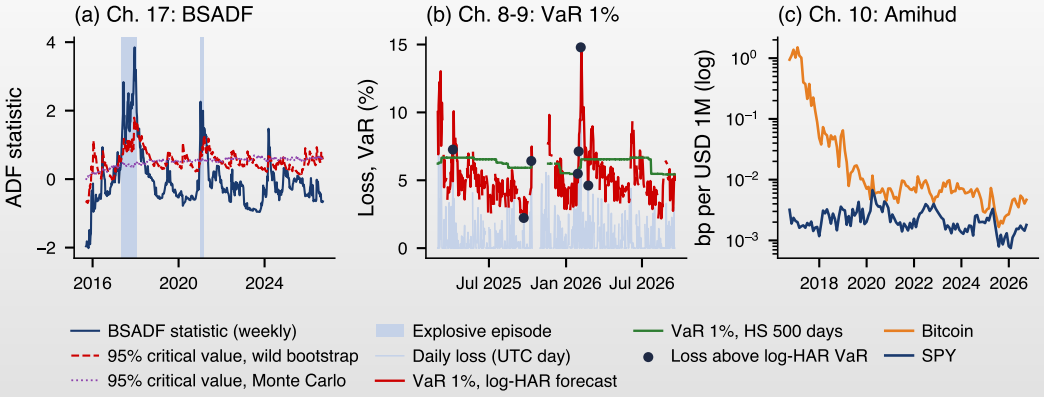
Bitcoin: the Question and the Decision Rule

- The question: a risk manager holds Bitcoin; which one-day loss limit does she set, and which signal would make her cut the position?
- Data
  - daily closes, 18 Sep 2014 – 18 Sep 2026: 4 384 returns
  - 5-minute bars on UTC (Coordinated Universal Time) days, 1 Sep 2024 – 17 Sep 2026: 659 complete days
- Five chapters
  - Chapter 1: stylised facts
  - Chapter 9: realised variance RV and log-HAR forecasts Corsi (2009)
  - Chapter 8: backtesting the VaR 1%
  - Chapter 10: spreads and Amihud illiquidity
  - Chapter 17: the GSADF bubble test and date-stamping with the BSADF (Backward Sup Augmented Dickey–Fuller) statistic Phillips, Shi & Yu (2015)

The decision rule

- Daily limit: the log-HAR VaR 1%, if it is not rejected at 5% and needs less capital on average than HS; otherwise the HS VaR
- On days without a log-HAR forecast (incomplete 5-minute bars): the HS VaR
- Cutting the position: BSADF above its 95% wild-bootstrap critical value for at least 7 weeks in a row
  - the minimum episode length of Chapter 17: $\lceil \ln T \rceil = 7$ weeks

Bitcoin: Bubbles, Realised-Volatility VaR and Liquidity



MFM_ch19_integrated_bitcoin ▶ Interactive

Bitcoin: the Steps in Numbers

- Chapter 1: daily returns (no panel in the figure)
  - Hill tail index of losses: 2.71 ($k = 109$, 2.5% of the days)
  - annualised volatility: 66.6%
  - ACF of $|r_t|$ at lag 1: 0.21
  - 29 days beyond ± 4 standard deviations, against 0.28 under the Normal distribution
- (b) Chapter 9: log-HAR on $\ln RV$, with the 1-, 7- and 30-day means, $R^2 = 0.36$
  - coefficients, Newey–West errors in brackets: $\hat{b}_d = 0.38$ (0.04), $\hat{b}_w = 0.47$ (0.08), $\hat{b}_m = -0.02$ (0.10)
  - expanding-window forecasts from 1 Mar 2025: 422 days, the same as in Chapter 9
- (b) Chapter 8: VaR 1% $= 2.326\sqrt{\widehat{RV}_t}$, against HS on the last 500 days
  - log-HAR VaR: mean 5.29%, between 1.99% and 14.45%; HS VaR: mean 6.24%
  - 144 of the 566 calendar days have no log-HAR forecast: the evaluation covers the days selected by data availability
  - $r_t / \sqrt{\widehat{RV}_t}$: standard deviation 1.00, excess kurtosis 2.54
  - 1% quantile of $r_t / \sqrt{\widehat{RV}_t}$: -2.52, against -2.33 for the Normal distribution
- (c) Chapter 10: Amihud illiquidity, in bp per USD 1 million traded
  - last two years: 0.0037 for Bitcoin, 0.0016 for SPY
  - monthly mean: 0.80 over the first 12 months from 2016, 0.0040 over the last 12 months
  - spreads from daily data disagree: Roll 85 bp, Corwin–Schultz 81 bp, EDGE (Efficient Discrete Generalized Estimator) -8 bp (EDGE estimates the squared spread, which comes out negative when price variance dominates)
- (a) Chapter 17: GSADF 3.85, against a 95% wild-bootstrap critical value of 2.82 (homoskedastic Monte Carlo: 2.27)
  - explosive episodes, with the wild-bootstrap critical values: May 2017 – Jan 2018 and Dec 2020 – Feb 2021
  - since 2022: BSADF above the wild critical value in 6 weeks, at most 3 in a row (peak 1.47, on 8 Mar 2024)
  - last value: -0.65, against a wild critical value of 0.58

Bitcoin: the Result and the Decision

VaR 1%	Breaches	Expected	Kupiec p	indep. p	cond. p	mean VaR
log-HAR	7	4.2	0.214	0.604	0.404	5.29%
HS, 500 days	7	4.2	0.214	0.604	0.404	6.24%

- Type of evidence: a retrospective comparison of procedures, on 422 days with complete bars (1 Mar 2025 – 17 Sep 2026)
  - the same coverage, with a mean VaR lower by 0.95 percentage points
  - indep. p: only the 372 pairs of consecutive calendar days
  - the full rule (HS on days without a forecast), over all 566 days: 9 breaches, against 5.7 expected; Kupiec p 0.194, indep. p 0.589
  - the power is low: MDE 2.70%; a true rate of 2% is detected with probability 47% (power of the Kupiec test)
- Limits
  - only two years of 5-minute bars
  - the Normal quantile understates the tail: -2.52, against -2.33
  - Amihud uses the volume reported by aggregated exchanges
- **What I choose**: a daily limit equal to the log-HAR VaR 1% (5.04% for 17 Sep 2026), recomputed every day
- **Under which conditions**: the 5-minute bars cover at least 95% of the day; illiquidity stays of the order of that of SPY
- **What would make me drop it**
  - 5 or more breaches in 250 days (the Basel yellow zone, Chapter 8): I switch to the empirical quantile of $r_t / \sqrt{\widehat{RV}_t}$
  - BSADF above the wild critical value for 7 weeks in a row: I cut the position
  - weekly monitoring repeats the test: a pointwise 95% threshold does not control false alarms over the whole horizon, so I confirm with the family-wise threshold of Chapter 17

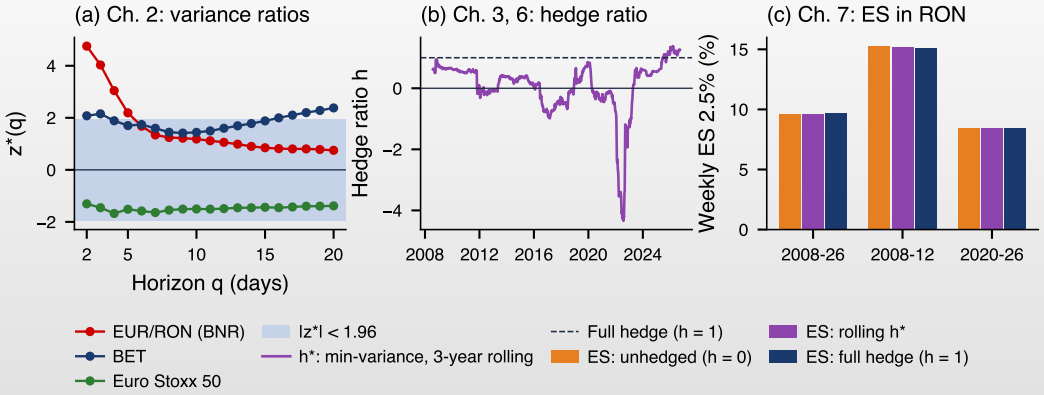
A Romanian Investor: the Question and the Decision Rule

- The question: an investor who measures her wealth in lei holds 50% BET and 50% Euro Stoxx 50, bought in euro; does she hedge the currency risk of the euro leg, that is, of the EUR/RON rate, the number of Romanian lei (RON) per euro (EUR)?
- Data: BET and Euro Stoxx 50 (daily closes), the BNR (Banca Națională a României, the National Bank of Romania) EUR/RON reference rate, Jul 2005 – Sep 2026
  - weekly (Friday) returns: the BNR fixes the rate at 13:00, the exchanges close at other times
- Five chapters
  - Chapter 2: variance ratios Lo & MacKinlay (1988)
  - Chapter 3: exposures with Newey–West errors
  - Chapter 5: GARCH(1,1)-t volatility
  - Chapter 6: the minimum-variance hedge ratio h^* over a rolling three years
  - Chapter 7: VaR 1% and ES 2.5% by historical simulation
- The benchmark Campbell, Serfaty-de Medeiros & Viceira (2010): a currency that is negatively correlated with equities lowers risk, so a full hedge does not minimise risk

The decision rule

- Three choices, with h the share of the euro leg sold forward: $h = 0$ (no hedge), rolling h^*, $h = 1$ (full hedge)
- We pick the choice with the lowest weekly ES 2.5%
- If the ES values differ by less than 0.25 percentage points in every period: $h = 0$, which needs no forward contracts; we switch to $h = 1$ after six months in a row of positive three-year correlation between equities and EUR/RON

A Romanian Investor: Efficiency, Currency Hedging and ES in Lei



MFM_ch19_integrated_romania ▶ Interactive

A Romanian Investor: the Steps in Numbers

- **(a) Chapter 2**: the robust statistic $z^*(q)$ on daily returns; $|z^*| > 1.96$ rejects the random walk at 5%
  - EUR/RON: $z^*(2) = 4.76$, with VR(2) $= 1.17$ and a first-order autocorrelation of 0.17
  - BET: $z^*(2) = 2.08$; the Chow–Denning test over $q = 2, 5, 10, 20$: $p = 0.067$
  - Euro Stoxx 50: $|z^*(q)| \leq 1.68$; Chow–Denning: $p = 0.43$
  - the daily EUR/RON risk times $\sqrt{h}$ understates the h-day risk, so we work on weeks
- **Chapter 3**: exposures over 1 101 weeks, with $f = \Delta$EUR/RON ($f > 0$: the leu depreciates)
  - Euro Stoxx 50 in euro on f: $\hat{\beta} = -0.52$ ($t = -2.61$)
  - BET on Euro Stoxx 50 and f: 0.56 ($t = 7.4$) and -0.54 ($t = -2.11$), $R^2 = 0.30$
  - the leu depreciates in the weeks in which equities fall
- **Chapter 5**: GARCH(1,1)-t on daily returns
  - EUR/RON: persistence 1.000, at the boundary of stationarity
  - EUR/RON: volatility today 0.96% a year, against a mean of 4.4%
  - BET: 25.9% today, against a mean of 21.2%
  - Euro Stoxx 50: 13.8% today, against a mean of 21.0%
- **(b) Chapter 6**: $h^* = \mathrm{Cov}(r^{RON}, f)/\mathrm{Var}(f)$, with $r^{RON} = r^{EUR} + f$ the return in lei of the euro leg
  - full sample: $h^* = 0.48$
  - over a rolling three years: between -4.34 (Jul 2022) and 1.38 (Mar 2026); last value 1.27
  - the three-year correlation between the Euro Stoxx 50 and f is negative in 93% of the windows
- **(c) Chapter 7**: weekly ES 2.5% of the 50/50 portfolio, rebalanced weekly, for the three choices (table on the next slide)
  - the exact simple return $R_p = \frac{1}{2}(e^b - 1) + \frac{1}{2}[e^{s+f} - 1 - h(e^f - 1)]$, with b, s the log returns of the BET and the Euro Stoxx 50; a forward on the notional at the start of the week, with zero carry
  - the loss, in log points: $L = -100 \ln(1 + R_p)$

A Romanian Investor: the Result and the Decision

Weekly ES 2.5%	Weeks	$h = 0$	rolling h^*	$h = 1$
Aug 2008 – Sep 2026	944	9.60%	9.62%	9.67%
Aug 2008 – Dec 2012	228	15.23%	15.20%	15.12%
Jan 2020 – Sep 2026	351	8.43%	8.43%	8.47%

- The retrospective conclusion: the ES values of the three choices differ by at most 0.11 percentage points, so the choice of h mattered little for the risk
  - annualised weekly volatility: EUR/RON 4.6%, Euro Stoxx 50 20.9%
  - the negative correlation lets the depreciation of the leu offset part of the losses of the euro leg
- Type of evidence: comparison of hedging choices, for one portfolio
- Limits
  - the cost of the hedge, the RON–EUR interest differential, is not in the course data
  - the depreciation of the leu added 1.55% a year to the euro leg (Jul 2005 – Sep 2026); a forward replaces it with that differential
  - h^* is unstable when $\mathrm{Var}(f)$ is small, as in 2019–2022: the estimation risk of Chapter 4
  - h^* hedges the euro leg; for the whole portfolio, $\mathrm{Cov}(b + s + f, f)/\mathrm{Var}(f) = -0.35$ (linear approximation)
- **What I choose now**: the full hedge ($h = 1$)
  - the three-year correlation is positive after 12 Sep 2025 (0.037 at the end): the rule for dropping $h = 0$ was triggered on 13 Mar 2026
  - the rolling h^* (1.27) would sell forward more than the euro leg
- **Under which conditions**: the three-year correlation between equities and EUR/RON stays positive
- **What would make me drop it**: a negative three-year correlation for six months in a row (back to $h = 0$)

The Project in Four Steps

- An empirical study of one question about financial markets, answered with the methods of this course
  - **70%** of the final grade
  - quizzes and seminar activity 20%, attendance 10%, no final exam
- **1. Replicate**: one full table or figure of a published paper (or of the lecture)
  - cell by cell (standard: next slide)
- **2. Extend**: new data, a new market (the BVB, crypto) or a more recent sample
  - under a **pre-analysis plan** recorded in a dated commit before the test period is opened
  - The plan states the null, the test, the loss, the sample split, the identification strategy, the power of the main test and the specification grid with its correction
- **3. Deliver on GitHub**: a repository that regenerates every number
  - with `AI_USE.md` and `AI_ERRORS.md` mandatory
- **4. Present and defend**: team presentation, then a 10-minute oral defence
  - Each member explains one result and answers a "what changes if..." question
  - Grades can differ between members

Replication Standard

- Target: one full table or figure, with its DOI (Digital Object Identifier), page and table number, not a single coefficient
  - Report a cell-by-cell comparison: published value, your value, difference, and the reason for each gap (sample, filter, data vendor, definition)
- Large-scale replications show why
  - many anomalies shrink out of sample or with better data Hou, Xue & Zhang (2020)
  - most factors replicate once the design is right Jensen, Kelly & Pedersen (2023)
- Reproducibility
  - a lockfile of package versions, fixed random seeds, series names and dates of the data
  - one command that regenerates every number
  - Different teams on the same data disagree more than their standard errors suggest Menkveld et al. (2024)
  - more reproducible work is associated with a smaller gap (research case study, this chapter)

Choosing the Question

A good question is

- **Specific**: one market, one period, one quantity
- **Testable**: a null hypothesis and a benchmark
- **Feasible**: data you can obtain and clean in a few weeks
  - the course data sets in the course GitHub repository
  - the National Bank of Romania, BNR; Federal Reserve Economic Data, FRED; Kenneth French Data Library; DefiLlama
- **Relevant**: an answer matters to an investor, a bank or a regulator

Examples

- Too vague: "Is Bitcoin risky?"
  - Better: "Did spot ETFs change the one-day VaR 1% of Bitcoin?"
  - Then: "Which model forecasts it best since 2024?"
- Too broad: "Can machine learning predict the market?"
  - Better: "Does a boosting model beat log-HAR for next-day BET volatility, in QLIKE, 2018–2026?"

Project Grading Criteria

Component	Grading criteria
Replication (step 1)	A published table or figure reproduced cell by cell; every gap explained
Question, data, extension (step 2)	Specific, testable question; documented data and calendar alignment
Methods and inference	Plan, identification, power, real-time design, benchmark, specification curve with a correction
Interpretation and report	Economic meaning, limitations, honest conclusion; 8–12 pages, 10–15 references
GitHub repository (step 3)	One command reproduces every result (lockfile, seeds); `AI_USE.md`, `AI_ERRORS.md`
Presentation and defence (step 4)	Individual: one result explained, answers to "what changes if..."

Deliverables and Timeline

- Week 1 after this chapter: team, question, one-page proposal
  - proposal: Seminar 19, C2
  - GitHub repository
- Weeks 2–3: data, replication (step 1)
  - extension plan recorded in a dated commit
- Weeks 4–5: extension, evaluation, robustness (step 2)
- Final weeks: the 8–12-page report
  - repository with README and the two AI files (step 3)
  - presentation and oral defence (step 4)
- Exact dates are announced at the course and seminar

The AI Policy and the Two Mandatory Files

Policy

- **Allowed**: AI assistants for code, debugging, literature search and writing
- **Declared**: undeclared use of AI counts as plagiarism
- **Checked**: you are responsible for every line
  - a reference that does not exist counts as fabricated data
- **No AI during the oral defence**: "the AI wrote it" is not an explanation

The two files

- `AI_USE.md`: tools and versions
  - who used them for what
  - where AI-written code or text was kept
  - how you verified the replicated table, data, formulas and references
- `AI_ERRORS.md`: at least **three** places where the assistant was wrong
  - what you asked, what it answered, what was wrong, how you found out, the correct version
- Errors you caught are a strength of the project, not a weakness

How Much Do Researchers' Choices Matter? (1/2)

- Menkveld et al. (2024): 164 research teams test the same six hypotheses on the same data
  - *Journal of Finance* 79(3)
  - Finance Crowd Analysis Project, #fincap
- **Standard error**
  - Uncertainty from sampling, the data-generating process
  - **Nonstandard error**: uncertainty from the analysis path each researcher picks, the evidence-generating process
  - Model: team j reports $X_j = e_j + \varepsilon_j$, e_j the researcher fixed effect of its path, ε_j the sampling error
  - Nonstandard error $:= Q_{0.75}(x) - Q_{0.25}(x)$, the interquartile range (IQR) of the estimates across teams, $Q_p(x)$ being the p-quantile of the team estimates x
  - The IQR, not the standard deviation, because the e_j can have fat tails

How Much Do Researchers' Choices Matter? (2/2)

- **Data**
  - All trades in the Euro Stoxx 50 index futures, 2002–2018, 720 million records
  - Six hypotheses of "no annual trend" (market efficiency, realised spread, client share of volume, ...)
  - Each team reports the average yearly change in % with its standard error (SE)
  - Four stages in 2021: independent analysis, peer feedback, the five best papers shared, final estimates
  - design details in the Appendix
- Link to this chapter: every team project is one path through the "garden of forking paths"
  - Pre-registration, the specification grid and a repository that reruns every number are the remedies

Key Results: Size and Quality

- **Size**
  - For market efficiency the median estimate is -1.1% per year, with an IQR of 6.7 percentage points (pp)
  - Even for the client share of volume, a simple ratio, the IQR is 1.2 pp around -3.3% (Table I, Panel C)
  - The median SE of 2.5% implies a Gaussian IQR of $1.35 \times 2.5 = 3.4\%$, half the 6.7 pp
  - The no-dispersion null is rejected for all six hypotheses, $p < 0.0001$ (Table II)
- **Quality** (Table III, Fig. 2)
  - a rise of one standard deviation in reproducibility is associated with a 25.0% lower IQR
  - the same rise in peer rating, with a 33.3% lower IQR
  - Code that reruns end to end is associated with answers closer to the consensus
- Type of evidence: description (it shows what happened in the data, without testing a cause)
  - of the dispersion across teams; the link with quality is only an association

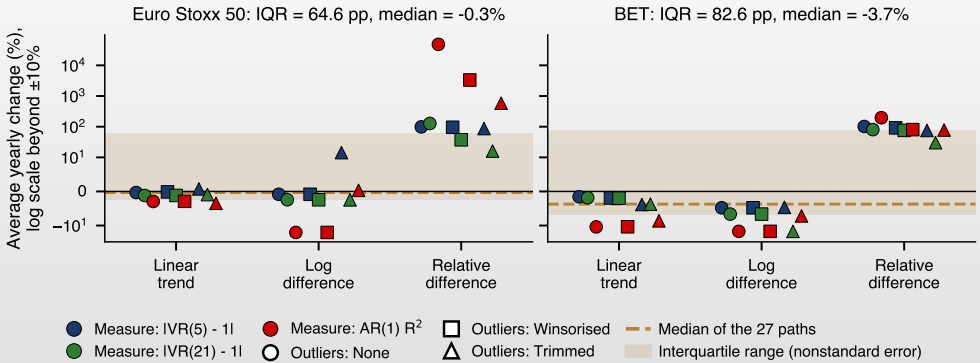
A Mini-Multiverse on Course Data: the Design

- Daily closes 2002–2018 from the course data, without the artificial non-trading-day fills (the rule of Chapter 2)
  - Forks: outliers × measure × model, 27 paths
  - The band is the IQR, the nonstandard error of the multiverse
- Measures of inefficiency
  - $|VR(q) - 1|$ ($VR(q)$: the variance ratio of Chapter 2, with the Lo–MacKinlay corrections, $q = 5, 21$ days) and the R^2 of an autoregressive AR(1) model: departures from a random walk
  - A positive trend means declining efficiency

MFM_ch19_nonstandard_errors

A Mini-Multiverse on Course Data: 27 Paths per Index

Legend: $VR(q)$, the q-day variance ratio; AR(1), the first-order autoregressive model; IQR, the interquartile range; pp, percentage points.



MFM_ch19_nonstandard_errors ▶ Interactive

A Mini-Multiverse on Course Data: Results

- IQR of the estimates: 64.6 pp (Euro Stoxx 50) and 82.6 pp (BET)
  - around medians of -0.3% and -3.7% a year
- The model fork creates the dispersion
  - Over the nine linear-trend paths the IQR is only 2.6 pp for the Euro Stoxx 50 and 6.7 pp for the BET
  - Relative differences explode: medians $+98.7\%$ (Euro Stoxx 50) and $+80.3\%$ (BET), up to 48 415%
- Sampling benchmark: the Gaussian IQR, $1.35\times$ the median SE
  - 41.3 pp (Euro Stoxx 50) and 43.4 pp (BET)
  - The IQR exceeds this benchmark for both indices
- Type of evidence: description
  - of the dispersion over the 27 paths of the grid, not across real teams

Limits and an Open Extension

- **Authors**: the multiverse covers a few forks only and explains the reported estimates poorly
- **Authors**: beliefs were asked as standard deviations, not IQRs
- **Authors**: the test of dispersion assumes consistent reported SEs
- **Our replication**: a multiverse spans only the options in its grid
  - so its IQR neither estimates nor bounds from below the dispersion across real teams
- **Our replication**: daily closes only
  - so the paper's intraday frequency pairs cannot be run
- **Open extension**: does a pre-registered grid shrink the nonstandard error of a risk question?
  - Measure the IQR of VaR 1% conclusions for the BET across defensible paths

An Open Question

- **Open question**: how much of a result is the data, and how much the researcher?
  - Replication is not hopeless: in a Bayesian model of replication most asset-pricing factors replicate and work out of sample in 93 countries Jensen, Kelly & Pedersen (2023)
  - AI assistants make every extra specification almost free: the forking paths multiply, but so does the ability to report all of them
  - Formal version: the conclusion must hold over a pre-registered set $\mathcal{S}$ of defensible specifications, with the family corrected (FWER or FDR)
  - Report its distribution over $\mathcal{S}$, not the best cell

The Discovery Loop and Its Checks (1/2)

Loop

- Question and H_0/H_1 (week 1)
- Literature and the replication target seen in the PDF (weeks 1–2)
- Pre-registration of the extension, recorded in a dated commit before the first extension result (week 2)
- Replication as a passing test in the repository (weeks 2–3)
- The whole grid (weeks 4–5)
- An AI examiner before the defence

The Discovery Loop and Its Checks (2/2)

- **Checks by the researcher**
  - The target number is in the paper
  - The plan precedes the first results file in the commit history
  - Every run is logged and counted
  - A fresh clone regenerates every number
  - Each member can derive the key formula without AI
  - Prompts for every step and the test for each check: Appendix

Mini-Case: Replication and Extension (1/2)

- **Replicate**: FHS VaR 1% for the BET
  - Chapter 8 reports 59 breaches in 4 937 days (1.20%, Kupiec $p = 0.181$)
  - The case study of this chapter: 67 in 5 432 days (1.23%, $p = 0.095$)
  - Same model, different evaluation start (2007 vs 2005) and settings: document every gap before calling a replication failed
- A typical assistant proposal: "only GARCH-EVT is correctly calibrated on both markets"
  - the assistant reads each p-value alone at 5%

Mini-Case: Replication and Extension (2/2)

- **Extend**
  - The 12 conditional-coverage tests of Chapter 8 (S&P 500 and BET $\times$ 6 models) form one family
  - Bitcoin and EUR/RON stay outside the family: on these markets some GARCH forecasts are missing (Chapter 8)
  - Rejections at 5%: 8 without correction, 6 with Holm (FWER), 7 with Benjamini–Hochberg, BH (FDR 5%)
  - FHS fails only on the S&P 500 ($p = 0.040$); within the family of 12 this rejection is no longer significant, so after Holm FHS and GARCH-EVT pass on both markets
- Interpretation: "FHS fails for the S&P 500" depends on whether the test is read alone or in its family
- Type of evidence: comparison of procedures
- In the project, fix the family in the plan before seeing results and report both readings

Your Turn: the Capstone Plan

- **Question**: your project idea from Part C, written as H_0/H_1, with the observation that would falsify it
- **Replicate first**: one published table seen in the paper, turned into a passing test in the repository
  - Then extend over a pre-registered grid, with the family corrected and the full distribution reported
- **Deliverables** along the timeline
  - proposal and replication target, pre-registration commit
  - passing replication test, grid of results
  - mock defence, report, oral defence
- **AI log** (templates are in the project folder)
  - the prompts that changed the design, in `AI_USE.md`
  - every caught error, with the test that caught it, in `AI_ERRORS.md`

Key Takeaways

- Returns are hard to predict, their risk is not
  - volatility, tails and dependence can be forecast
- Every model is judged out of sample, against a simple benchmark, with a loss that fits the target
- Heavy tails change inference
  - use robust errors and the bootstrap
  - report intervals, state the power
- Most wrong results come from the design, not from the model
  - look-ahead, overfitting, multiple testing, calendars
- The analysis path is itself a source of uncertainty
  - pre-register, replicate, report the whole grid
- New tools (ML, foundation models, LLMs) raise the bar for evaluation, not lower it

Check Your Understanding (1/2)

Questions

- Why does the Kupiec test need $\alpha = 0.01$, not 2.39 or 0.192?
- Why is the sample excess kurtosis of BET returns not an estimate of anything?
- Why is the Wald interval for GARCH persistence uninformative once we restrict to $p \leq 1$, and what replaces it?
- What does a non-rejection of the Kupiec test at $T = 250$ tell you?
- Why is the IQR of a 27-path multiverse not the nonstandard error of 164 teams?

Check Your Understanding (2/2)

Further study

- Seminar 19: the exercises A1–A12, B1–B8 and C1–C3 apply this chapter to the BET
- Textbooks: Franke, Härdle & Hafner (2019); Härdle, Chen & Overbeck (2017); McNeil, Frey & Embrechts (2015); Tsay (2010)
- Advanced courses on quantinar.com: Statistics of Financial Markets
  - Measuring Statistical Risk
- Code and data: Quantlets/Ch_19 (GitHub)

Two Computations on New Numbers (1/2)

Try it

- **1.** A VaR 1% model has $x = 6$ breaches in $T = 400$ days
  - Compute LR_{uc} with the Kupiec formula of the Chapter 8 review and decide at 5%
- **2.** A GARCH(1,1) has $\alpha + \beta = 0.97$
  - Compute the half-life $h_{1/2}$
  - How does it change if $\alpha + \beta = 0.985$?

Two Computations on New Numbers (2/2)

Check your answer

- **1.** $\hat{\pi} = 6/400 = 1.5\%$

$$LR_{uc} = 2\big[(394 \ln 0.985 + 6 \ln 0.015) - (394 \ln 0.99 + 6 \ln 0.01)\big] = 0.88$$

  - Below 3.84: not rejected
  - At $T = 400$ the test has little power, so this is "no evidence", not "no problem"
- **2.** $\ln 0.5 / \ln 0.97 = 22.8$ days and $\ln 0.5 / \ln 0.985 = 45.9$ days
  - Halving $1 - \alpha - \beta$ roughly doubles the half-life

The Practical Decision

- **What I choose**: for a risk project, a GARCH filter with a flexible tail (FHS), against historical simulation as the simple benchmark
  - on the BET, FHS is the only model rejected by no test (coverage, independence, DQ)
  - the 10% model confidence set on FZ0 is {GARCH-t, FHS}
- **Under which conditions**
  - real-time forecasts, with the information set $\mathcal{F}_{t-1}$ and parameters re-estimated on a rolling window
  - an evaluation with enough power: the Kupiec MDE is 1.45% at $T = 5\,432$, against 3.65% at $T = 250$
  - the loss function and the test family fixed in the plan by a dated commit
- **What result would make me drop it**
  - a rejection that survives the Holm correction over the family of conditional-coverage tests
  - a conclusion that changes across the pre-registered specification grid
  - a replication of the target table with deviations I cannot explain cell by cell

Step 2: the GARCH(1,1)-t Model

- The GARCH(1,1) of the Chapter 5 review, with Student-t innovations Bollerslev (1986)

$$r_t = \mu + \varepsilon_t, \qquad \varepsilon_t = \sigma_t z_t, \qquad \sigma_t^2 = \omega + \alpha\, \varepsilon_{t-1}^2 + \beta\, \sigma_{t-1}^2$$

  - σ_t: conditional volatility, known at $t-1$
  - z_t: i.i.d. Student-t with ν degrees of freedom, scaled to variance 1
  - $\omega > 0$, $\alpha, \beta \geq 0$: α measures the reaction to news, β the memory
  - a small ν means heavy tails
- Persistence $\alpha + \beta$ and half-life $h_{1/2}$: the formulas of the Chapter 5 review
- Estimated by maximum likelihood, with robust (sandwich) standard errors

Step 2: Today's Volatility and the Long-Run Level (1/2)

- Two numbers at the end of the sample
  - $\hat{\sigma}_T$ is known at $T-1$
  - $\hat{\sigma}_{T+1}$ also uses the last return $r_T = 2.17\%$ and is the input of tomorrow's VaR

$$\hat{\sigma}^2_{T+1} = \hat{\omega} + \hat{\alpha}\,(r_T - \hat{\mu})^2 + \hat{\beta}\,\hat{\sigma}^2_T = 0.0448 + 0.192\,(2.17 - 0.08)^2 + 0.799 \times 1.67^2 = 3.11$$

  - So $\hat{\sigma}_{T+1} = \sqrt{3.11} = 1.76\%$: a large last return lifts the forecast above $\hat{\sigma}_T = 1.67\%$
- Long-run variance implied by the model

$$\bar{\sigma}^2 = \frac{\omega}{1-\alpha-\beta} = \frac{0.0448}{0.0093} = 4.82$$

  - Returns are in %: $\bar{\sigma} = 2.20\%$ a day (34.7% a year), against a sample standard deviation of 1.40% a day (22.1% a year)

Step 2: Today's Volatility and the Long-Run Level (2/2)

- Why the gap is not a contradiction: the denominator $1 - \alpha - \beta = 0.0093$ is small and imprecise
  - Holding ω fixed, $p = 0.9867$ gives $\bar{\sigma}^2 = 3.37$ and $p = 0.9947$ gives 8.47
  - the profile CI reaches $p = 1$, where $\bar{\sigma}^2$ is infinite
  - Use $\hat{\sigma}_{T+1}$ and the short-horizon forecast path
  - Treat the long-run level as weakly identified
- The forecast path returns geometrically to $\bar{\sigma}^2$, at the rate $\hat{\alpha} + \hat{\beta}$

$$\hat{\sigma}^2_{T+h} = \bar{\sigma}^2 + (\hat{\alpha} + \hat{\beta})^{h-1}(\hat{\sigma}^2_{T+1} - \bar{\sigma}^2), \qquad h = 21\colon\ 4.82 + 0.9907^{20}\,(3.11 - 4.82) = 3.40$$

  - So $\hat{\sigma}_{T+21} = 1.84\%$: after one month the forecast has covered only 17% of the distance to the long-run level

Step 3: Four VaR 1% Forecasts (1/2)

- Unconditional, on the last 500 days: one quantile for the whole window
  - **HS** (historical simulation): the loss $L = -r$ exceeded on 1% of the last 500 days (empirical quantile)
  - **Normal**: $\mathrm{VaR} = -(\hat{\mu} + \hat{\sigma} z_{\mathbf{0.01}})$
  - $\hat{\mu}$, $\hat{\sigma}$: sample mean and standard deviation; $z_{\mathbf{0.01}} = -2.326$: the 1% quantile of the standard Normal distribution
- Conditional, GARCH-t on the last 1000 days
  - the quantile scales with today's volatility forecast $\hat{\sigma}_t$

$$\textbf{GARCH-}t\text{:}\ \ \mathrm{VaR}_t = -\left(\hat{\mu} + \hat{\sigma}_t\, t_\nu^{-\mathbf{1}}(0.01)\sqrt{(\nu-2)/\nu}\right), \qquad \textbf{FHS}\text{:}\ \ \mathrm{VaR}_t = -\left(\hat{\mu} + \hat{\sigma}_t\, \hat{q}_{\mathbf{0.01}}(z)\right)$$

  - $t_\nu^{-\mathbf{1}}(0.01)$: the 1% quantile of a Student-t with ν degrees of freedom
  - $\sqrt{(\nu-2)/\nu}$ rescales it to variance 1
  - **FHS** (filtered historical simulation): $\hat{q}_{\mathbf{0.01}}(z)$, the empirical 1% quantile of the standardised residuals $z_t = (r_t - \hat{\mu})/\hat{\sigma}_t$, replaces the Student-$t$ quantile McNeil & Frey (2000)

Step 3: Four VaR 1% Forecasts (2/2)

- Real-time design
  - Forecast for day t uses data up to $t-1$ only
  - GARCH re-estimated every 21 trading days
  - $\hat{\sigma}_t$ updated daily
- Evaluation: 3 Jan 2005 – 18 Sep 2026, $T = 5\,432$ days
  - A breach (hit) on day t: the loss $-r_t$ exceeds VaR_t
  - Expected breaches at 1%: $0.01\,T \approx 54$
- Unconditional models react only as the window rolls
  - conditional ones react through $\hat{\sigma}_t$, within days

Step 4: the Three Backtests

- Kupiec coverage test Kupiec (1995): is the breach rate equal to $\alpha = 1\%$?
  - LR_{uc} as in the Chapter 8 review, rejected at 5% when $LR_{uc} > 3.84$
- Christoffersen independence test Christoffersen (1998): is a breach today more likely after a breach yesterday?

$$LR_{ind} = -2 \ln \frac{(1-\hat{\pi})^{n_{00}+n_{10}}\, \hat{\pi}^{\,n_{01}+n_{11}}}{(1-\hat{\pi}_{01})^{n_{00}}\, \hat{\pi}_{01}^{n_{01}}\, (1-\hat{\pi}_{11})^{n_{10}}\, \hat{\pi}_{11}^{n_{11}}} \;\sim\; \chi^2_1$$

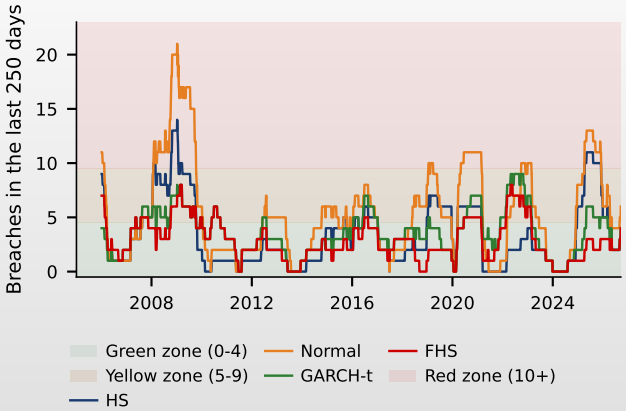
  - n_{ij}: number of days in state j after a day in state i (1 = breach, 0 = no breach)
  - $\hat{\pi}_{01} = n_{01}/(n_{00}+n_{01})$, $\hat{\pi}_{11} = n_{11}/(n_{10}+n_{11})$: breach rates after a quiet day and after a breach
  - $\hat{\pi} = (n_{01}+n_{11})/(n_{00}+n_{01}+n_{10}+n_{11})$: the rate over the $T-1$ transitions, within $1/(T-1)$ of x/T
- Joint test: $LR_{cc} = LR_{uc} + LR_{ind} \sim \chi^2_2$, with 5% critical value 5.99

Step 4: Coverage and Independence Tests

Model	Breaches	Rate	p_{uc}	p_{ind}	p_{cc}	Red zone
HS	79	1.45%	0.002	<0.001	<0.001	5.5%
Normal	122	2.25%	<0.001	<0.001	<0.001	18.6%
GARCH-t	80	1.47%	0.001	0.864	0.005	0.0%
FHS	67	1.23%	0.095	0.267	0.134	0.0%

- Red zone: share of 250-day windows with 10 or more breaches (Basel traffic light)
- Only FHS is rejected by none of the three tests at 5%
- HS and Normal also fail independence
  - their breaches cluster
- GARCH-t: independence is not rejected, but there are too many breaches
  - the Student-t tail is not heavy enough at 1%

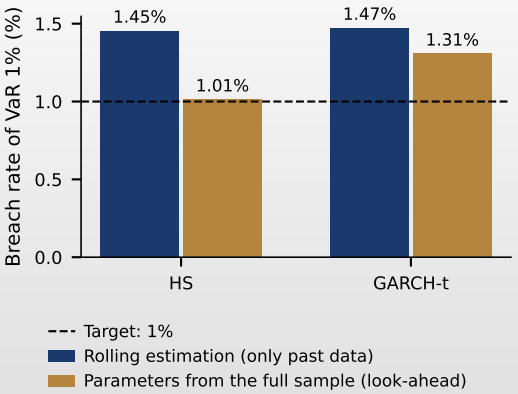
Step 4: Breaches in the Last 250 Days



- Number of breaches in a moving 250-day window, with the Basel green, yellow and red zones
- Normal and HS enter the red zone in episodes, when their breaches cluster
  - shares: last column of the Step 4 table
- The two GARCH-filtered models never enter it

MFM_ch19_case_study

Look-Ahead Bias in the Case Study



- Full-sample HS quantile: 1.01% breaches, by construction
- GARCH-t with full-sample parameters: 1.31% of days
  - with rolling estimation 1.47%
- Both look better than their real-time versions
  - the gap is the value of information nobody had
- Only the rolling result is evidence about a forecasting model

MFM_ch19_case_study

Step 5: Three Tools for a Formal Comparison (1/2)

- **Dynamic Quantile (DQ)** test Engle & Manganelli (2004): are the breaches predictable?
  - Regress $I_t - 0.01$ ($I_t = 1$ if the loss exceeds VaR_t, else 0) on a constant, $I_{t-1}, \ldots, I_{t-4}$ and VaR_t
  - DQ, the Wald statistic that all six coefficients are zero, is χ^2_6 under a correct model
  - It detects breaches predictable from the past or from the level of VaR, which the Christoffersen test (one lag) can miss
- **FZ0 score** Patton, Ziegel & Chen (2019): one number per day that scores a (VaR, ES) forecast pair

$$S(v, e; y) = -\frac{1\{y \le v\}(v - y)}{\alpha e} + \frac{v}{e} + \ln(-e) - 1$$

  - lower is better
  - $\alpha = 2.5\%$: the ES level; y: the realised return
  - $v = -\mathrm{VaR}$, $e = -\mathrm{ES}$: the forecasts written as return quantiles ($e \le v < 0$)
  - Strictly consistent for (VaR, ES) Fissler & Ziegel (2016): on average the true pair has the lowest score

Step 5: Three Tools for a Formal Comparison (2/2)

- **Pairwise Diebold–Mariano tests** Diebold & Mariano (1995): t_{DM} as in the Chapter 14 review
  - d_t: the FZ0 difference of the two models on day t
  - $t_{DM} \approx N(0,1)$ under equal accuracy
- **Model confidence set** at 10% Hansen, Lunde & Nason (2011)
  - it contains the best model with probability 90%
  - a model stays in it while $p_{MCS} \geq 0.10$
- **Comparative ES backtest** Nolde & Ziegel (2017), against HS as the standard model
  - red if the model is significantly worse
  - green if significantly better
  - yellow otherwise

Step 5: Which Forecast Pair Does FZ0 Score?

- A second forecast pair from the same four models: VaR 2.5% and ES 2.5%
  - at the Basel ES level $\alpha = 2.5\%$
  - Not the VaR 1% of Steps 3–4: FZ0 is consistent only when v and e are the quantile and the tail mean at the same level α
- Each model's ES 2.5% is the mean loss beyond its own VaR 2.5%
  - from the same loss distribution it uses for VaR in Step 3
  - **HS**: the estimator of Chapter 7; of the 500 losses, the 12 largest plus half of the 13th, divided by 12.5
  - **Normal**: $\mathrm{ES} = -\hat{\mu} + \hat{\sigma}\,\varphi(z_{\mathbf{0.025}})/0.025 = -\hat{\mu} + 2.338\,\hat{\sigma}$, with φ the standard Normal density
  - **GARCH**-t: $-\hat{\mu}$ plus $\hat{\sigma}_t$ times the upper-tail mean of the Student-t rescaled to variance 1
  - **FHS**: the negative of the sum of the conditional mean and $\hat{\sigma}_t$ times the mean of the 25 smallest of the 1,000 standardised residuals z_t (the same estimator)

Step 5: The ES 2.5% Forecasts in Numbers

Model	Average VaR 1% (Steps 3–4)	Average ES 2.5% (Step 5)
HS	3.66%	3.99%
Normal	2.90%	2.92%
GARCH-t	2.95%	3.08%
FHS	3.17%	3.41%

- Averages of the daily forecasts over the $T = 5\,432$ evaluation days, losses in % of the position
- Normal model: ES 2.5% / VaR 1% $= 2.338/2.326 \approx 1.005$
  - the ratio in the table (2.92/2.90)
  - an ES 1% would be about 15% higher
- The gap between the columns grows with the weight of the tail the model allows
  - about 9% for HS and 7.5% for FHS (empirical tails)
  - 4% for GARCH-t
  - 0.5% for the Normal

Methods: From Question to Chapter

If your question is about	Start from
Return distribution, tails, predictability	Chapters 1–2: moments, robust tests, variance ratios
Expected returns, factors, portfolios	Chapters 3–4: Fama–MacBeth, shrinkage, out-of-sample Sharpe tests
Volatility and correlation	Chapters 5, 6, 9: GARCH, DCC, copulas, log-HAR
Risk measures and capital	Chapters 7–8: FHS, EVT, backtests, FZ0
Liquidity, options, continuous time	Chapters 10–12: Amihud, Black–Scholes, VIX, variance risk premium
Forecasting with machine learning (ML), deep learning or text	Chapters 13–15: purged cross-validation (CV), foundation models, sentiment
Crypto, bubbles, systemic risk	Chapters 16–18: stablecoins, GSADF, CoVaR, networks

- Starting points: the Part C question of every seminar (table in the Appendix)
  - Your own question is welcome if it passes the five questions of this chapter

Identification: from Correlation to a Claim (1/2)

- Event questions need a counterfactual
  - e.g. "did spot ETFs, the FTSE Russell upgrade or a bank tax change X?"
  - A control market not hit by the event: difference-in-differences on the risk measure Y (e.g. volatility or VaR 1%)

$$\hat{\delta} = \left(\bar{Y}^{\text{after}}_{\text{treated}} - \bar{Y}^{\text{before}}_{\text{treated}}\right) - \left(\bar{Y}^{\text{after}}_{\text{control}} - \bar{Y}^{\text{before}}_{\text{control}}\right)$$

  - Equivalently: δ in $Y_{it} = a + b\,D_i + c\,P_t + \delta\,D_i P_t + \varepsilon_{it}$, with $D_i = 1$ for the treated market, $P_t = 1$ after the event and HAC standard errors
  - Identifying assumption: parallel trends (without the event Y would have moved alike in both markets), checked on the pre-event period

Identification: from Correlation to a Claim (2/2)

- A synthetic control built from several untreated markets Abadie, Diamond & Hainmueller (2010)
- Placebo event dates: the event moved one year earlier or later
- Unknown timing: a sup-Wald break test Andrews (1993)
  - report the estimated break date, not an imposed one
- Forecasting questions: out of sample only, with a test period declared in the plan
- One event is $n = 1$
  - say what the design can and cannot identify

Power and Robustness without p-Hacking (1/2)

- **Power**: before the main test, compute the minimum detectable effect (MDE)
  - MDE: the smallest effect the design detects with 80% power at 5%
  - For an approximately Normal estimate with standard error SE, a two-sided 5% test and 80% power

$$\mathrm{MDE} = (z_{0.975} + z_{0.80})\,\mathrm{SE} = 2.80\,\mathrm{SE}, \qquad T_{\min} = \left(\frac{2.80\,\sigma}{\delta}\right)^2 \text{ when } \mathrm{SE} = \sigma/\sqrt{T}$$

  - z_p: the p-quantile of the standard Normal distribution; δ: the effect you look for; σ: the standard deviation of one observation
  - Deliverable: a simulated power curve of the main test at your T
  - Short crypto, ETF or post-event samples make many questions underpowered (Kupiec at $T = 250$: MDE 3.65%)

Power and Robustness without p-Hacking (2/2)

- Specification curve
  - estimate every defensible combination of window, lags, α and sample start
  - Plot all estimates sorted, with their intervals Simonsohn, Simmons & Nelson (2020)
  - Report the spread across specifications, not the best cell
- Correct the family
  - Holm for any false discovery
  - Benjamini–Hochberg for their share
  - or White's bootstrap reality check for data snooping White (2000)
  - Significance thresholds must rise with the number of tests tried Harvey & Liu (2020), Harvey (2017)

Forecast Evaluation for Risk Projects

- Volatility: a robust loss (MSE or QLIKE) against a conditionally unbiased proxy (evaluation error)
  - Many models: the model confidence set Hansen, Lunde & Nason (2011), not the lowest average loss
- VaR: coverage, independence and the DQ test Engle & Manganelli (2004)
  - with a benchmark, conditional predictive ability tests Giacomini & White (2006)
  - State the power of every backtest at your T
- ES: score the pair (VaR, ES) with FZ0
  - comparative backtests with three zones Nolde & Ziegel (2017)
  - dynamic (VaR, ES) models Patton, Ziegel & Chen (2019)
- The integrated case study (Steps 1–5) is the template

The Report and the Presentation (1/2)

The report: an 8–12-page paper

- **1. Introduction**: question, hypothesis, contribution
- **2. Literature**: 10–15 clickable references, at least 3 from the last five years
  - including Franke, Härdle & Hafner (2019), McNeil, Frey & Embrechts (2015)
- **3. Data and replication**: sources, period, cleaning rules
  - the replicated table cell by cell
- **4. Design**: plan, identification, power, benchmark, loss, tests
- **5. Results** with their uncertainty
- **6. Robustness**: specification curve with the correction
- **7. Conclusion**: the answer in one sentence, and its limits

The Report and the Presentation (2/2)

The presentation (about 12 minutes): the slides, in order

- Question (1 slide)
- Data and one chart that shows the problem (2)
- Method and benchmark (2)
- Main result with its uncertainty (2–3)
- Robustness, limits, answer (2)
- One message per slide, readable axes, legend below the chart
- Every member presents and can answer questions on the code
  - then the oral defence, without AI

Typical AI Errors in Financial Modelling (1/2)

- **Look-ahead**: a rolling statistic or a scaler that includes day t in the forecast for day t
  - Check: change the last observation; the forecast for that day must not move
  - shift by one day (`.shift(1)`)
- **Sign of VaR**: the 1% quantile of returns reported as VaR
  - or the 99% quantile of returns (a gain) used as the loss
  - Check: $\mathrm{VaR}_\alpha = -q_\alpha(r)$, with $q_\alpha(r)$ the α-quantile of returns, is positive whenever $q_\alpha(r) < 0$
  - a one-day VaR 1% of an index is a few percent
- **Annualisation**: volatility multiplied by 252 instead of $\sqrt{252}$
  - or 252 days used for crypto that trades every day
  - Check: over h periods, means scale with h, volatilities with $\sqrt{h}$ (i.i.d. returns)
  - use the real number of trading days (365 for crypto)

Typical AI Errors in Financial Modelling (2/2)

- **Random cross-validation (CV) on time series**
  - Shuffled K-fold trains on the future (pitfall of design, Chapter 13)
  - Check: is any test date earlier than a training date? Use walk-forward or purged CV with an embargo López de Prado (2018)
- **Invented references**, wrong degrees of freedom
  - library defaults you did not choose
  - numbers stated without running code
  - Check: search every DOI on Crossref and find the claim in the paper, as in the AI-critique exercise of every seminar (Part C)

Key Results: Feedback, Beliefs and Forks

- **Feedback**: from stage 1 to stage 4 the IQR falls by 47.2%
  - the interdecile range falls by 68.2%
  - peer review narrows the dispersion but does not remove it
- **Beliefs**: the teams' predicted standard deviation of estimates is 71.7% below the realised one
- **Forks**: model choice and sampling frequency move the estimates most
  - A relative-change model gives a median annual trend of $-23\,000\%$ on daily data, -200% monthly and -4.56% yearly
  - Log differences add up exactly across frequencies and a linear trend is linear in the data, so neither has this frequency wedge
  - For a positive measure, Jensen's inequality biases the mean relative change upwards; the negative values in the paper need another mechanism (toy example below)

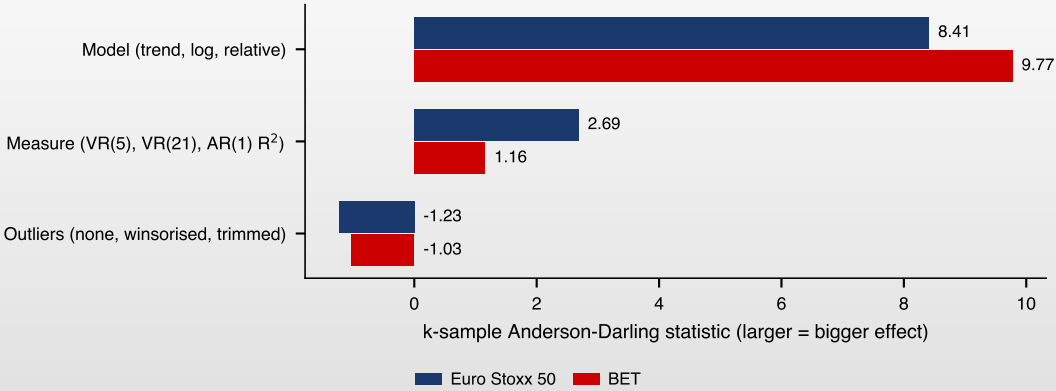
Why Relative Changes Explode: a Toy Example

- A measure that returns to its start: $m = (1,\ 0.01,\ 1)$
  - Relative changes $m_t/m_{t-1} - 1$: -99%, then $+9\,900\%$; their mean is $+4\,900.5\%$
  - Log changes: $\ln 0.01$, then $\ln 100$; their mean is exactly 0, the right answer for a measure with no trend
  - For a positive measure Jensen's inequality pushes the mean relative change up (Appendix): the course multiverse explodes upwards
- A positive measure can never fall by more than 100%
  - a trend of $-23\,000\%$ needs a measure that changes sign near zero ($m_{t-1} = 0.001$, $m_t = -0.2$ gives $-20\,100\%$)
  - or daily changes summed rather than compounded (250 changes of -1% add up to -250%)
- Lesson for your plan: inspect denominators near zero
  - pre-specify the transformation (log differences for positive measures, or a linear trend)

Which Fork Moves the Estimates? The Statistic

- k-sample Anderson–Darling statistic of the 27 estimates, grouped by the option at each fork
  - It compares the distributions of the k groups; larger means more separated
  - Standardised, so it is centred at 0 when the groups are equal: a negative value (outliers below) means less separation than chance
- A descriptive ranking, not a test
  - The 27 estimates share the same data

Which Fork Moves the Estimates?



MFM_ch19_nonstandard_errors

Which Fork Moves the Estimates? Results

- The statistic for each fork, Euro Stoxx 50 and BET
  - Model: 8.41 and 9.77
  - Measure: 2.69 and 1.16
  - Outliers: −1.23 and −1.03
- Same ranking as the paper: model choice moves the estimates most
  - on a large euro-area index and on the BET, a frontier-market index in 2002–2018
- For your project: fix the model of the yearly change in the plan (log differences or a linear trend)
  - report the whole grid with the fork that drives it

Inference by Chapter: Further Formulas

- **Chapter 3**: Fama–MacBeth standard errors ignore that the betas are estimated
  - Shanken's covariance, with $\hat{\gamma}$ the risk premia, $\hat{\Sigma}_f$ the factor covariance and $\hat{V}_{FM}$ the Fama–MacBeth (FM) covariance Shanken (1992)

$$\hat{V}_{\mathbf{sh}} = (1+c)\big(\hat{V}_{FM} - \hat{\Sigma}_f/T\big) + \hat{\Sigma}_f/T, \qquad c = \hat{\gamma}^\top \hat{\Sigma}_f^{-\mathbf{1}} \hat{\gamma}$$

  - valid for a correctly specified model; with an intercept, $\hat{\Sigma}_f$ is bordered by zeros in the row and column of the intercept
  - the empirical reference of Chapter 3: Kan–Robotti–Shanken errors, robust to misspecification, with Newey–West lags and a sensitivity to their number
- **Chapter 5**: QMLE (quasi-maximum likelihood) of GARCH Bollerslev & Wooldridge (1992)
  - A: the expected Hessian (sign changed); B: the variance of the score

$$\sqrt{T}(\hat{\theta} - \theta) \Rightarrow N\big(0, A^{-\mathbf{1}} B A^{-\mathbf{1}}\big)$$

  - in the block of the variance parameters, $A = B$ when $\kappa_z = 3$: Normal innovations, but not only
  - otherwise Hessian-only standard errors ($A^{-\mathbf{1}}$) are wrong; the factor $\sqrt{(\kappa_z - 1)/2}$ applies to the Gaussian QMLE only, not to the Student-t estimator
- **Chapter 7**: the empirical quantile, for i.i.d. returns with density $f(q_\alpha) > 0$
  - in the tail f is small, so historical VaR is noisy

$$\sqrt{T}(\hat{q}_\alpha - q_\alpha) \Rightarrow N\big(0,\ \alpha(1-\alpha)/f(q_\alpha)^{\mathbf{2}}\big)$$

Back

Project Ideas: Part C of Every Seminar (1/2)

Ch.	Open question (link to the seminar)
0	Did spot ETFs change Bitcoin's correlation with equities?
1	Did BET-TR (BET total return) autocorrelation fall after the FTSE reclassification?
2	Did the BVB become more efficient after the 2020 FTSE upgrade?
3	Do factor premia exist on the Bucharest Stock Exchange?
4	Does an optimised portfolio of BVB blue chips, in RON, beat 1/N out of sample?
5	Is the Romanian market's volatility driven by the US?
6	Did Romanian and euro-area equities become more integrated after 2020?
7	How much capital does a BVB portfolio need compared with an S&P 500 portfolio?

- Each seminar gives the data, a starting method and a reference analysis discussed in class

Back

Project Ideas: Part C of Every Seminar (2/2)

Ch.	Open question (link to the seminar)
8	Which risk model would you certify for a Romanian bank's trading book?
9	Is Bitcoin's volatility more predictable than the S&P 500's?
10	Does illiquidity carry a premium on the BVB?
11	Which continuous-time model fits BET and Bitcoin?
12	Is the variance risk premium a tradable signal?
13	Does a foundation model beat boosting?
14	Can a zero-shot foundation model replace GARCH for the BET?
15	Does headline sentiment pay after costs?
16	Is Bitcoin an alternative asset, a risk asset or a hedge?
17	Is there an AI bubble?
18	How exposed are Romanian banks to euro-area banking stress?

- Chapter 20 also suits a forecasting project
  - realised volatility with signatures, VOLARE data

Back

Turning an Idea into a Project

- Replicate first: one published table or figure from a paper behind the idea (step 1)
- Narrow it: one market, one period, one number to forecast or test
  - Example: Seminar 8 becomes "Which of HS, FHS and GARCH-EVT gives a correct VaR 1% for a 50/50 portfolio of the BET and EUR (EUR/RON exchange rate) in 2015–2026?"
- Extend it: a second market, a new model, a longer or more recent sample
  - Example: add a foundation model (Chapter 14) as a volatility filter inside FHS
- Keep the reference analysis as your benchmark and try to improve on it honestly

Back

Data Conventions for the Project

- Use the course data where possible
  - daily series of indices, ETFs, BVB stocks, crypto and FX (Foreign eXchange), plus intraday bars
  - Public sources: BNR reference rates, FRED, Kenneth French Data Library, DefiLlama; the BVB archive for Romanian indices
- Apply the course conventions
  - Adjusted prices for ETFs and stocks, closing values for indices, FX and crypto
  - Remove holiday rows where the vendor repeated the previous price, and weekend quotes where the market is closed
  - Annualise with the real frequency
  - Several assets: align prices on common trading days first, then compute returns
- Look for erroneous prices (bad ticks) before modelling
  - jumps of $\pm 50\%$ in adjusted prices are usually data errors

Back

Checklist for Evaluating Empirical Results

Before you look at the results

- Fix in the plan the hypothesis, the benchmark and the loss
  - the estimation window and the test period
  - the number of models, with their correction
  - the identification and the power

When you report

- Out-of-sample results only for claims about forecasting
- Every headline number with a standard error, CI or p-value
- Costs for trading strategies
- Subperiods and a second market for robustness
- What did **not** work, and why

Back

Where Projects Usually Lose Points

- A question that changes after the results are known
- In-sample fit presented as forecasting power
- No benchmark, or a weak one chosen to be beaten
- Numbers without uncertainty
- Many tests without correction
- Charts without units, legends inside the data, unexplained acronyms
- Code that does not run, or numbers in the report that the code does not produce
- AI output kept without checks
  - undeclared use, unchecked code, references that do not exist

Back

Nonstandard Errors: Data and Stages

- **Data**: all trades in the Euro Stoxx 50 index futures, 2002–2018
  - 720 million records
  - each side flagged as client or own account (Sec. I.A)
- **Six hypotheses**, each "no annual trend" in
  - market efficiency, realised bid–ask spread, client share of volume
  - realised spread on client orders, share of market orders among client orders
  - gross trading revenue of clients
- Each team chooses its own measure of each quantity
- **Four stages** in 2021 (Sec. I.A)
  - independent analysis (paper and code), feedback from two peer evaluators
  - the five best papers shared
  - final estimates without the code constraint
- A pre-analysis plan was filed publicly before the experiment started

Back

Nonstandard Errors: Quality and Inference

- **Quality**
  - Team quality: first principal component of five proxies
  - reproducibility score 0–100 by the certification agency Cascad
  - demeaned peer rating (Sec. I.B)
  - Quantile regressions of the estimates at the 10th, 25th, 50th, 75th and 90th percentiles
  - a shrinking IQR means less dispersion (Sec. I.C)
- **Inference**
  - Each estimate is tested against the median estimate, with a Bonferroni correction at a family level of 0.5% (Sec. I.C)
  - Multiverse analysis: the teams report their choice at each fork of Table V
  - The k-sample Anderson–Darling statistic ranks the forks by how much they move the estimates (Sec. II.C)

Back

Nonstandard Errors, Replication Details

- **Forks from Table V**
  - Outliers: none; the annual measure winsorised or trimmed at its 2.5 and 97.5 percentiles
  - Measure: variance ratio day to week, variance ratio day to month, R^2 of an autoregressive AR(1) model
  - Model: linear trend, log difference, relative difference
  - Daily data allow the day-to-week and day-to-month comparisons ($VR(5)$, $VR(21)$); the intraday measure frequencies of Table V need intraday data
- **Why relative differences explode**: $m_t = |VR - 1|$ or R^2 is close to zero in some years (paper: Fig. 7)
  - No trend, stationary m_t: the linear slope and the mean log change are centred at zero
  - but for independent m_t, $\mathbb{E}(m_t/m_{t-1}) = \mathbb{E}m \cdot \mathbb{E}(1/m) \geq 1$ (Jensen), so the mean relative change is positive
- **Anderson–Darling statistic**: the 27 estimates are strongly correlated
  - so tabulated critical values and p-values do not apply

Back

The Discovery Loop, Prompts (1/2)

- **Week 1, the question**: sharpen a project idea into H_0/H_1 with a replication target
  - "Here is our project idea from Part C of Seminar N. Give three sharper, testable versions, the published table each would replicate and the series from the course data it needs."
- **Weeks 1–2, literature and target**: the number to replicate is one you have seen in the PDF
  - "List the most cited and the most recent papers on this question with DOI; for each, the headline estimate, its table and its sample."
  - Resolve every DOI on Crossref
  - Search with Semantic Scholar, Elicit or Google Scholar; open the table yourselves
- **Week 2, hypotheses and pre-registration**, recorded in a dated commit before the first result
  - "For our H0 and H1, list competing hypotheses and the implication that separates each; then draft a pre-registration: sample, filters, windows, tests, correction."

Back

The Discovery Loop, Prompts (2/2)

- **Weeks 2–3, code and replication**: the replicated table becomes a test in the repository
  - "Write a function that reproduces every cell of the published table from the course data and an assertion with the tolerance we accept."
- **Weeks 4–5, extension and robustness**: run the whole pre-registered grid
  - "Here are all the cells we ran. Which conclusions survive a Holm correction, and which ones would a referee call data snooping?"
- **Before the defence, critique**: the assistant asks, the team answers without AI, as at the defence itself
  - "Act as the examiner of our oral defence: ask the ten hardest questions on identification, inference and data. Do not answer them."
- **Report**: the distribution of results over the grid, not the best cell
  - prompts in `AI_USE.md`
  - caught errors in `AI_ERRORS.md`

Back

What the Researcher Must Check

- **The replication target exists**
  - number, table and sample taken from the paper, not from the assistant's summary
  - the report cites the page and the table, and a second member finds the same number there
- **The plan came first**
  - in the commit history, the pre-registration of the extension precedes the first extension results file
- **Forking paths** (more specifications tried than reported)
  - every run writes its configuration and result to a log
  - the report states their number and corrects the family
- **Reproducibility** (no numbers entered by hand)
  - a fresh clone and one command regenerate every number to the reported digit
- **Understanding**: each member can derive the key formula and defend every choice
  - the test is an internal mock defence, without AI, one week before the real one

Back

Acronyms used in this material (1/3)

- **ACF**: AutoCorrelation Function
- **ADF**: Augmented Dickey–Fuller (test)
- **AI**: Artificial Intelligence
- **AR**: AutoRegressive model
- **ASE**: Academia de Studii Economice din București — *the Bucharest University of Economic Studies*
- **BET**: Bucharest Exchange Trading (index)
- **BET-TR**: BET Total Return (index)
- **BH**: Benjamini–Hochberg (procedure)
- **BNR**: Banca Națională a României — *the National Bank of Romania*
- **BSADF**: Backward Sup Augmented Dickey–Fuller (statistic)
- **BV**: Bipower Variation
- **BVB**: Bursa de Valori București — *the Bucharest Stock Exchange*
- **CAGR**: Compound Annual Growth Rate
- **CAPM**: Capital Asset Pricing Model
- **CI**: Confidence Interval
- **CoVaR**: Conditional Value-at-Risk (of the system, given an institution)
- **COVID-19**: COronaVIrus Disease 2019
- **CV**: Cross-Validation
- **DCC**: Dynamic Conditional Correlation
- **DeFi**: Decentralised Finance
- **DM**: Diebold–Mariano (test)
- **DOI**: Digital Object Identifier
- **DQ**: Dynamic Quantile (test)
- **EDGE**: Efficient Discrete Generalized Estimator (of the bid-ask spread)
- **EEM**: iShares MSCI Emerging Markets ETF
- **EFA**: iShares MSCI EAFE ETF (developed markets outside the US and Canada)

Acronyms used in this material (2/3)

- **ES**: Expected Shortfall
- **ETF**: Exchange-Traded Fund
- **EUR**: euro, international currency code
- **EVT**: Extreme Value Theory
- **FDR**: False Discovery Rate
- **FHS**: Filtered Historical Simulation
- **FM**: Fama–MacBeth (regressions)
- **FRED**: Federal Reserve Economic Data
- **FTSE**: Financial Times Stock Exchange (FTSE Russell, index provider)
- **FWER**: Family-Wise Error Rate
- **FX**: Foreign eXchange
- **FZ0**: Fissler–Ziegel loss, zero-homogeneous version
- **GARCH**: Generalised AutoRegressive Conditional Heteroskedasticity
- **GARCH-EVT**: GARCH filter combined with Extreme Value Theory for the tail
- **GBM**: Geometric Brownian Motion
- **GJR**: Glosten–Jagannathan–Runkle (asymmetric GARCH)
- **GLD**: SPDR Gold Shares
- **GMV**: Global Minimum Variance (portfolio)
- **GRS**: Gibbons–Ross–Shanken (test)
- **GSADF**: Generalised Sup Augmented Dickey–Fuller (test)
- **HAC**: Heteroskedasticity and Autocorrelation Consistent (standard errors)
- **HAR**: Heterogeneous AutoRegressive (model)
- **HS**: Historical Simulation
- **HYG**: iShares iBoxx High Yield Corporate Bond ETF
- **IEF**: iShares 7–10 Year Treasury Bond ETF
- **IGARCH**: Integrated GARCH
- **IQR**: InterQuartile Range

Acronyms used in this material (3/3)

- **JB**: Jarque–Bera (test)
- **LLM**: Large Language Model
- **LQD**: iShares iBoxx Investment Grade Corporate Bond ETF
- **LR**: Likelihood Ratio
- **LRMES**: Long-Run Marginal Expected Shortfall
- **LSTM**: Long Short-Term Memory (network)
- **MCS**: Model Confidence Set
- **MDE**: Minimum Detectable Effect
- **ML**: Machine Learning
- **MSE**: Mean Squared Error
- **PDF**: Portable Document Format
- **QLIKE**: Quasi-LIKElihood (loss function)
- **QMLE**: Quasi-Maximum Likelihood Estimator
- **RON**: Romanian leu, international currency code
- **RV**: Realised Variance
- **SADF**: Sup Augmented Dickey–Fuller (test)
- **SE**: Standard Error
- **SML**: Security Market Line
- **SPY**: SPDR S&P 500 ETF Trust
- **SR**: Sharpe Ratio
- **SRISK**: Systemic RISK (capital shortfall measure)
- **TLT**: iShares 20+ Year Treasury Bond ETF
- **US**: United States
- **USD**: US dollar, international currency code
- **UTC**: Coordinated Universal Time
- **VaR**: Value-at-Risk
- **VIX**: CBOE Volatility IndeX
- **VOLARE**: VOLatility Archive for Realized Estimates (database)
- **VR**: Variance Ratio
- **VRP**: Variance Risk Premium

References (1/6)

- Abadie, A., Diamond, A., Hainmueller, J. (2010). Synthetic control methods for comparative case studies: estimating the effect of California's tobacco control program. *Journal of the American Statistical Association*, 105(490), 493–505.
- Andrews, D.W.K. (1993). Tests for parameter instability and structural change with unknown change point. *Econometrica*, 61(4), 821–856.
- Andrews, D.W.K. (1999). Estimation when a parameter is on a boundary. *Econometrica*, 67(6), 1341–1383.
- Athreya, K.B. (1987). Bootstrap of the mean in the infinite variance case. *Annals of Statistics*, 15(2), 724–731.
- Bailey, D.H., Borwein, J., López de Prado, M., Zhu, Q.J. (2017). The probability of backtest overfitting. *Journal of Computational Finance*, 20(4), 39–69.
- Bailey, D.H., López de Prado, M. (2014). The deflated Sharpe ratio: correcting for selection bias, backtest overfitting and non-normality. *Journal of Portfolio Management*, 40(5), 94–107.
- Barndorff-Nielsen, O.E., Shephard, N. (2006). Econometrics of testing for jumps in financial economics using bipower variation. *Journal of Financial Econometrics*, 4(1), 1–30.
- Benjamini, Y., Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1), 289–300.
- Black, F., Scholes, M. (1973). The pricing of options and corporate liabilities. *Journal of Political Economy*, 81(3), 637–654.
- Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3), 307–327.

References (2/6)

- Bollerslev, T., Wooldridge, J.M. (1992). Quasi-maximum likelihood estimation and inference in dynamic models with time-varying covariances. *Econometric Reviews*, 11(2), 143–172.
- Brownlees, C., Engle, R.F. (2017). SRISK: a conditional capital shortfall measure of systemic risk. *Review of Financial Studies*, 30(1), 48–79.
- Campbell, J.Y., Serfaty-de Medeiros, K., Viceira, L.M. (2010). Global currency hedging. *Journal of Finance*, 65(1), 87–121.
- Chow, K.V., Denning, K.C. (1993). A simple multiple variance ratio test. *Journal of Econometrics*, 58(3), 385–401.
- Christoffersen, P.F. (1998). Evaluating interval forecasts. *International Economic Review*, 39(4), 841–862.
- Cont, R. (2001). Empirical properties of asset returns: stylized facts and statistical issues. *Quantitative Finance*, 1(2), 223–236.
- Corsi, F. (2009). A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics*, 7(2), 174–196.
- DeMiguel, V., Garlappi, L., Uppal, R. (2009). Optimal versus naive diversification: how inefficient is the 1/N portfolio strategy?. *Review of Financial Studies*, 22(5), 1915–1953.
- Diebold, F.X., Mariano, R.S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3), 253–263.
- Engle, R. (2002). Dynamic conditional correlation: a simple class of multivariate generalized autoregressive conditional heteroskedasticity models. *Journal of Business & Economic Statistics*, 20(3), 339–350.
- Engle, R.F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica*, 50(4), 987–1007.

References (3/6)

- Engle, R.F., Manganelli, S. (2004). CAViaR: conditional autoregressive value at risk by regression quantiles. *Journal of Business & Economic Statistics*, 22(4), 367–381.
- Fama, E.F. (1970). Efficient capital markets: a review of theory and empirical work. *Journal of Finance*, 25(2), 383–417.
- Fissler, T., Ziegel, J.F. (2016). Higher order elicitability and Osband's principle. *Annals of Statistics*, 44(4), 1680–1707.
- Forbes, K.J., Rigobon, R. (2002). No contagion, only interdependence: measuring stock market comovements. *Journal of Finance*, 57(5), 2223–2261.
- Franke, J., Härdle, W.K., Hafner, C.M. (2019). Statistics of Financial Markets: An Introduction, 5th ed. Springer.
- Giacomini, R., White, H. (2006). Tests of conditional predictive ability. *Econometrica*, 74(6), 1545–1578.
- Gibbons, M.R., Ross, S.A., Shanken, J. (1989). A test of the efficiency of a given portfolio. *Econometrica*, 57(5), 1121–1152.
- Hansen, P.R., Lunde, A., Nason, J.M. (2011). The model confidence set. *Econometrica*, 79(2), 453–497.
- Härdle, W.K., Chen, C.Y.-H., Overbeck, L. (Eds.) (2017). Applied Quantitative Finance, 3rd ed. Springer.
- Harvey, C.R. (2017). Presidential address: the scientific outlook in financial economics. *Journal of Finance*, 72(4), 1399–1440.

References (4/6)

- Harvey, C.R., Liu, Y. (2020). False (and missed) discoveries in financial economics. *Journal of Finance*, 75(5), 2503–2553.
- Harvey, C.R., Liu, Y., Zhu, H. (2016). . . . and the cross-section of expected returns. *Review of Financial Studies*, 29(1), 5–68.
- Hill, B.M. (1975). A simple general approach to inference about the tail of a distribution. *Annals of Statistics*, 3(5), 1163–1174.
- Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2), 65–70.
- Hou, K., Xue, C., Zhang, L. (2020). Replicating anomalies. *Review of Financial Studies*, 33(5), 2019–2133.
- Jensen, T.I., Kelly, B., Pedersen, L.H. (2023). Is there a replication crisis in finance?. *Journal of Finance*, 78(5), 2465–2518.
- Kupiec, P.H. (1995). Techniques for verifying the accuracy of risk measurement models. *Journal of Derivatives*, 3(2), 73–84.
- Ledoit, O., Wolf, M. (2004). Honey, I shrunk the sample covariance matrix. *Journal of Portfolio Management*, 30(4), 110–119.
- Ledoit, O., Wolf, M. (2008). Robust performance hypothesis testing with the Sharpe ratio. *Journal of Empirical Finance*, 15(5), 850–859.
- Lo, A.W. (2002). The statistics of Sharpe ratios. *Financial Analysts Journal*, 58(4), 36–52.
- Lo, A.W., MacKinlay, A.C. (1988). Stock market prices do not follow random walks: evidence from a simple specification test. *Review of Financial Studies*, 1(1), 41–66.

References (5/6)

- Lobato, I., Nankervis, J.C., Savin, N.E. (2001). Testing for autocorrelation using a modified Box–Pierce Q test. *International Economic Review*, 42(1), 187–205.
- López de Prado, M. (2018). Advances in Financial Machine Learning. Wiley.
- Markowitz, H. (1952). Portfolio selection. *Journal of Finance*, 7(1), 77–91.
- McNeil, A.J., Frey, R. (2000). Estimation of tail-related risk measures for heteroscedastic financial time series: an extreme value approach. *Journal of Empirical Finance*, 7(3–4), 271–300.
- McNeil, A.J., Frey, R., Embrechts, P. (2015). Quantitative Risk Management: Concepts, Techniques and Tools, revised ed. Princeton University Press.
- Menkveld, A.J., Dreber, A., Holzmeister, F., Huber, J., Johannesson, M., Kirchler, M., et al. (2024). Nonstandard errors. *Journal of Finance*, 79(3), 2339–2390.
- Nelson, D.B. (1990). Stationarity and persistence in the GARCH(1,1) model. *Econometric Theory*, 6(3), 318–334.
- Newey, W.K., West, K.D. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55(3), 703–708.
- Nolde, N., Ziegel, J.F. (2017). Elicitability and backtesting: perspectives for banking regulation. *Annals of Applied Statistics*, 11(4), 1833–1874.
- Patton, A.J. (2011). Volatility forecast comparison using imperfect volatility proxies. *Journal of Econometrics*, 160(1), 246–256.

References (6/6)

- Patton, A.J., Ziegel, J.F., Chen, R. (2019). Dynamic semiparametric models for expected shortfall (and value-at-risk). *Journal of Econometrics*, 211(2), 388–413.
- Pele, D.T., Lazar, E., Dufour, A. (2017). Information entropy and measures of market risk. *Entropy*, 19(5), 226.
- Pele, D.T., Mazurencu-Marinescu-Pele, M. (2026). Finite-sample precision limits for expected shortfall forecast comparisons. *Mathematics*, 14(13), 2316.
- Pele, D.T., Wesselhöfft, N., Härdle, W.K., Kolossiatis, M., Yatracos, Y.G. (2023). Are cryptos becoming alternative assets?. *European Journal of Finance*, 29(10), 1064–1105.
- Phillips, P.C.B., Shi, S., Yu, J. (2015). Testing for multiple bubbles: historical episodes of exuberance and collapse in the S&P 500. *International Economic Review*, 56(4), 1043–1078.
- Shanken, J. (1992). On the estimation of beta-pricing models. *Review of Financial Studies*, 5(1), 1–33.
- Simonsohn, U., Simmons, J.P., Nelson, L.D. (2020). Specification curve analysis. *Nature Human Behaviour*, 4(11), 1208–1214.
- Stambaugh, R.F. (1999). Predictive regressions. *Journal of Financial Economics*, 54(3), 375–421.
- Tsay, R.S. (2010). Analysis of Financial Time Series, 3rd ed. Wiley.
- White, H. (2000). A reality check for data snooping. *Econometrica*, 68(5), 1097–1126.
- Zhang, L., Mykland, P.A., Aït-Sahalia, Y. (2005). A tale of two time scales: determining integrated volatility with noisy high-frequency data. *Journal of the American Statistical Association*, 100(472), 1394–1411.