








Modelling Financial Markets

Seminar 19: Review and Project Workshop — Hands-on



Daniel Traian PELE
Drd. Antoaneta Amza, antoaneta.amza@csie.ase.ro

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

Acronyms used in this material (1/2)

- **AI**: Artificial Intelligence
- **AIC**: Akaike Information Criterion
- **AR**: AutoRegressive (model); in event studies: Abnormal Return
- **BET**: Bucharest Exchange Trading (index)
- **BH**: Benjamini–Hochberg (procedure)
- **BIC**: Bayesian Information Criterion
- **BVB**: Bursa de Valori București — *the Bucharest Stock Exchange*
- **CI**: Confidence Interval
- **COVID-19**: COronaVIrus Disease 2019
- **DM**: Diebold–Mariano (test)
- **DQ**: Dynamic Quantile (test)
- **ES**: Expected Shortfall
- **ETF**: Exchange-Traded Fund
- **FDR**: False Discovery Rate
- **FHS**: Filtered Historical Simulation
- **FTSE**: Financial Times Stock Exchange (FTSE Russell, index provider)
- **FWER**: Family-Wise Error Rate
- **FZ0**: Fissler–Ziegel loss, zero-homogeneous version

Acronyms used in this material (2/2)

- **GARCH**: Generalised AutoRegressive Conditional Heteroskedasticity
- **GEO**: Government Emergency Ordinance
- **GJR**: Glosten–Jagannathan–Runkle (asymmetric GARCH)
- **GJR-GARCH**: Glosten–Jagannathan–Runkle GARCH
- **HAC**: Heteroskedasticity and Autocorrelation Consistent (standard errors)
- **HS**: Historical Simulation
- **IGARCH**: Integrated GARCH
- **IQR**: InterQuartile Range
- **LR**: Likelihood Ratio (test)
- **MCS**: Model Confidence Set
- **OLS**: Ordinary Least Squares
- **QML**: Quasi-Maximum Likelihood
- **QMLE**: Quasi-Maximum Likelihood Estimator
- **SE**: Standard Error
- **STOXX50E**: Euro Stoxx 50 index (symbol of the data file)
- **US**: United States
- **VR**: Variance Ratio
- **WIG20**: Warszawski Indeks Giełdowy 20 — *the Warsaw Stock Exchange index of the 20 largest companies*

Does the Risk Forecast Pass a Backtest?

- Question of the seminar
  - **would your project survive the checks reviewed in Chapter 19?**
  - from raw index closing prices to a backtested Value at Risk (VaR) 1% with valid inference, and from an idea to a one-page proposal
- Deliverables
  - the worked A1–A2 checks and the independent forecast A6 with a Generalised Autoregressive Conditional Heteroskedasticity (GARCH) model
  - one chart and one table for B1–B4; B5–B8 are optional extensions
  - the corrected code written by artificial intelligence (AI) in C1
  - one interpretation sentence per exercise
- Notebook for today: open the seminar notebook in Colab
  - the notebook sections carry the same labels as the exercises (A1–A12, B1–B8, C1–C3)
- Seminar: Prof. Daniel Traian Pele, danpele@ase.ro; Antoaneta Amza, PhD candidate, antoaneta.amza@csie.ase.ro

Which Observations Can We Compare?

- BET: Bucharest Exchange Trading index, the benchmark of Bursa de Valori București (BVB)
- WIG20: the Warsaw Stock Exchange index of 20 major companies
- SPY: an Exchange-Traded Fund (ETF) tracking the US (United States) S&P 500 (Standard & Poor's) index
- USD: the US dollar
- Each market keeps its own trading calendar
- Bitcoin trades every day, including days when the stock exchanges are closed
- Returns are log price changes in percent
  - From 100 to 110, the simple return is 10% and the log return about 9.53%
  - The loss is the negative of the log return
- For a joint analysis, align the prices on common dates first, then compute the returns
  - Otherwise two returns bearing the same date can cover different intervals

Data Used

Series	Price used	Calendar	Sample	Used in
BET	official close	Romanian trading days	6 Jan 2000 – 18 Sep 2026, 6 670 returns	A5, B1–B5
Bitcoin	close in USD	every day	18 Sep 2014 – 18 Sep 2026, 4 384 returns	B6–B8
SPY (S&P 500 ETF)	adjusted close	US trading days	2015 – 2026	B8
WIG20	close	Polish trading days	2014 – 2026	worked proposal
Euro Stoxx 50	close	local trading days	2002 – 2018 and later	C3

- Daily prices of the course data, one series per asset (field used: close / adjusted close); duplicate records and weekend-dated rows of the indices removed
- Two series on different calendars (B8, C2): align the *prices* on common trading days first, then compute returns

Setup: Run and Verify

```
bet = 100 * log_returns('bet')
btc = 100 * log_returns('btc')
print(len(bet), bet.index[0].date(),
      bet.index[-1].date())
print(round(bet.std(), 2), len(btc))
```

- The notebook section "Data" defines `log_returns(name)`: daily log returns on the calendar of each series
- Expected output
  - 6670 2000-01-06 2026-09-18
  - 1.4 4384
- If your count or first date differs, stop and check the file, the field and the cleaning rules

Tools from Earlier Chapters, Task by Task

- Almost every tool of today was introduced in earlier chapters
  - Three new ideas: the power of a backtest, an interval at a boundary, the look-ahead trap
- A2: the Holm correction, which controls the Family-Wise Error Rate (FWER), and the Benjamini–Hochberg (BH) correction, which controls the False Discovery Rate (FDR) (Chapter 3) Holm (1979), Benjamini & Hochberg (1995)
- B1: Hill tail index, robust portmanteau Q^*, moving-block bootstrap (Chapter 1) Hill (1975), Lobato, Nankervis & Savin (2001), Künsch (1989)
- B2: GARCH(1,1) with Student-t innovations, robust standard errors, the boundary $\alpha + \beta = 1$ (Chapter 5) Bollerslev & Wooldridge (1992)
- B3, B4 (Chapter 8) Kupiec (1995), Christoffersen (1998), Engle & Manganelli (2004)
  - Historical Simulation (HS), Normal, GARCH-t and Filtered Historical Simulation (FHS) VaR; Kupiec, Christoffersen and Dynamic Quantile (DQ) tests
  - The Fissler–Ziegel (FZ0) score, Diebold–Mariano (DM) tests, the Model Confidence Set (MCS), the comparative Expected Shortfall (ES) backtest

Moments, Autocorrelation and the Ljung–Box Test (B1)

- Sample moments of $r_1, \ldots, r_T$, with $x_t = r_t - \bar{r}$ and s the standard deviation

$$\widehat{\text{skew}} = \frac{1}{T}\sum_{t=1}^{T}\frac{x_t^3}{s^3}, \qquad \hat{\kappa} = \frac{1}{T}\sum_{t=1}^{T}\frac{x_t^4}{s^4} - 3$$

  - skewness < 0: a longer left tail; excess kurtosis $\hat{\kappa} > 0$: heavier tails than the Normal distribution (for which it is 0)
  - extreme days: under the Normal distribution, $T \cdot 2\Phi(-4)$ days with $|x_t| > 4s$ are expected; Φ: the standard Normal distribution function

- Sample autocorrelation at lag k and the **Ljung–Box** test of no autocorrelation up to lag m

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

  - reject at 5% if $Q(m) > \chi^2_{m;0.95}$ (18.31 for $m = 10$); on r_t^2 it detects volatility clustering
  - $Q(m)$ assumes i.i.d. returns; under volatility clustering it rejects too often, hence the robust $Q^*(m)$ (the slide "The Robust Q^* Test and the Block Bootstrap")

The Tail Index and the Hill Estimator (B1)

- Power-law (Pareto-type) tail of the losses L with **tail index** $\alpha > 0$

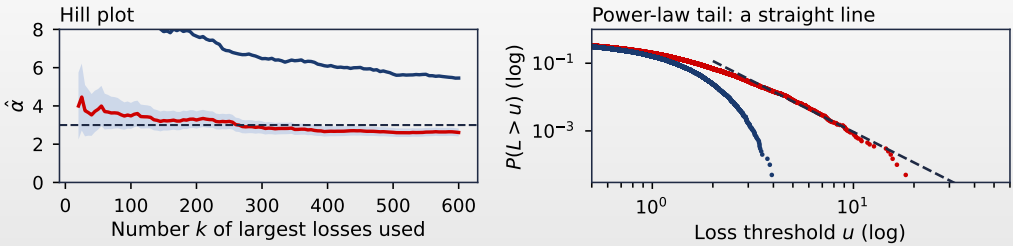
$$P(L > u) \approx c\, u^{-\alpha} \quad \text{for large } u$$

  - $c > 0$: a scale constant; smaller α: heavier tail; $\mathbb{E}|L|^p < \infty$ only if $p < \alpha$ (variance: $\alpha > 2$; kurtosis: $\alpha > 4$)
- **Hill estimator** Hill (1975) on the k largest losses $L_{(\mathbf{1})} \geq L_{(\mathbf{2})} \geq \cdots \geq L_{(k+\mathbf{1})}$

$$\hat{\alpha} = \Big(\frac{1}{k}\sum_{i=\mathbf{1}}^{k} \ln L_{(i)} - \ln L_{(k+\mathbf{1})}\Big)^{-\mathbf{1}}, \qquad \mathrm{CI}_{\mathbf{95\%}} = \hat{\alpha}\Big(1 \pm \frac{1.96}{\sqrt{k}}\Big)$$

  - the average log-excess over the threshold $L_{(k+\mathbf{1})}$; the confidence interval (CI) assumes i.i.d. losses (B1 compares it with a block bootstrap)
  - k trades bias (large k reaches into the centre) against variance (small k): look for a flat region of the Hill plot

A Hill Plot on Simulated Losses



- $T = 20\,000$ simulated returns: Student-$t(3)$, true $\alpha = 3$, and Normal (no power-law tail)
  - left: for the Student-t, $\hat{\alpha}$ is 3.48 at $k = 100$ and 2.67 at $k = 400$; for the Normal it keeps drifting with k
  - right: on log scales a power-law tail is a straight line of slope $-\alpha$; the Normal tail bends down

The Robust Q^* Test and the Block Bootstrap (B1)

- Robust portmanteau Lobato, Nankervis & Savin (2001)
  - $Q^*(m) = T \sum_{k=1}^{m} \hat{\rho}_k^2 / \hat{\tau}_k$, $\hat{\tau}_k = \overline{x_t^2 x_{t-k}^2} / \hat{\gamma}_0^2$, $x_t = r_t - \bar{r}$, $\hat{\gamma}_0 = \overline{x_t^2}$
  - $\hat{\rho}_k$: sample autocorrelation at lag k; bar: average over t
  - The χ_m^2 calibration needs finite fourth moments
- Moving-block bootstrap Künsch (1989)
  - Resample blocks of 20 consecutive days (999 resamples): the dependence within blocks is kept

GARCH(1,1)-t (B2)

- GARCH(1,1) Bollerslev (1986): today's variance is a weighted sum of a constant, yesterday's squared shock and yesterday's variance

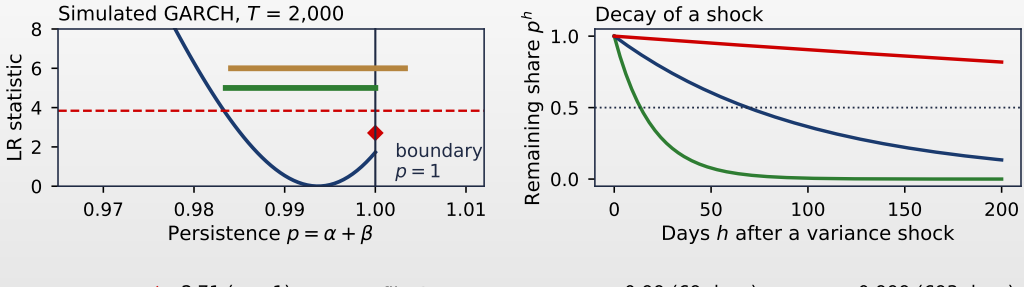
$$r_t = \mu + \varepsilon_t, \qquad \varepsilon_t = \sigma_t z_t, \qquad \sigma_t^2 = \omega + \alpha\,\varepsilon_{t-1}^2 + \beta\,\sigma_{t-1}^2$$

- Notation
  - μ: mean return; σ_t: conditional volatility, known at $t-1$
  - z_t: i.i.d. innovation with mean 0, variance 1: a Student-t with ν degrees of freedom rescaled by $\sqrt{(\nu-2)/\nu}$
  - $\omega > 0$; $\alpha \geq 0$: reaction to news; $\beta \geq 0$: memory; small ν: heavy tails
- Persistence $p = \alpha + \beta$
  - A variance shock decays by the factor p each day; $p < 1$: covariance stationarity

An Interval at a Boundary (B2)

- A Wald interval $\hat{p} \pm 1.96\,\mathrm{se}(\hat{p})$ can leave the parameter space
  - GARCH persistence $p = \alpha + \beta \leq 1$, but 0.99 ± 0.02 gives [0.97; 1.01]
  - Near the boundary $\hat{p}$ is not Normal Andrews (1999), and the interval contains values outside the imposed region $p \leq 1$
- **Profile likelihood**: $\ell_P(p)$ = maximum of the log-likelihood over the other parameters, p fixed
  - 95% interval: all p with $2(\ell_{\max} - \ell_P(p)) \leq 3.84$; $\ell_{\max}$: the maximised log-likelihood
  - It stays inside $p \leq 1$
  - At $p = 1$ the statistic is $\frac{1}{2}\chi^2_0 + \frac{1}{2}\chi^2_1$ (χ^2_0: a point mass at 0), critical value 2.71 (Chapter 5)
- Half-life of a shock: $h = \ln 0.5 / \ln p$
  - 14 days at $p = 0.95$, 69 at $p = 0.99$, 693 at $p = 0.999$, infinite at $p = 1$

Wald and Profile Intervals; the Half-Life of a Shock

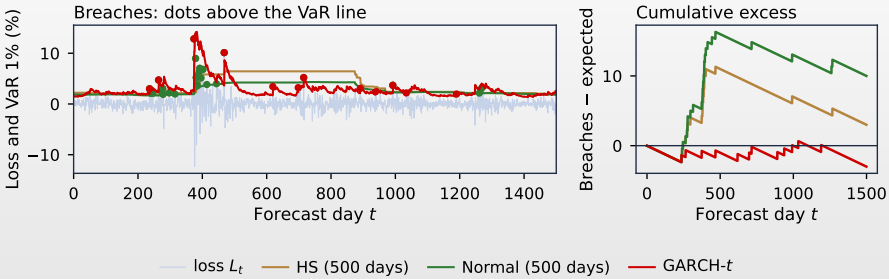


- Left: simulated GARCH(1,1), true $p = 0.995$, $T = 2\,000$; estimate $\hat{p} = 0.994$
  - Wald $[0.984, 1.003]$ crosses the boundary; profile $[0.984, 1]$: $2(\ell_{\max} - \ell_P(1)) = 1.73 < 2.71$, so $p = 1$ is not rejected
- Right: the share p^h of a variance shock left after h days; half-life $\ln 0.5 / \ln p$
  - near $p = 1$ a small change in p moves the half-life by months: report it as an interval

Four VaR Forecasts

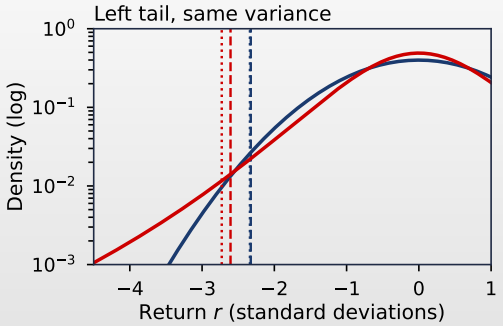
- One-day VaR 1% for day t, from data up to $t-1$
  - **HS**: the empirical 1% quantile of the losses of the last 500 days
  - **Normal**: $\mathrm{VaR} = -(\hat{\mu} + \hat{\sigma} z_{0.01})$, $\hat{\mu}$, $\hat{\sigma}$ from the last 500 days, $z_{0.01} = -2.326$
- **GARCH-t**: $\mathrm{VaR}_t = -(\hat{\mu} + \hat{\sigma}_t \, \tilde{t}_{\hat{\nu}}^{-1}(0.01))$
  - $\tilde{t}_{\nu}^{-1}(0.01) = t_{\nu}^{-1}(0.01)\sqrt{(\nu-2)/\nu}$: the 1% quantile of the standardised Student-t
- **FHS** (filtered historical simulation) McNeil & Frey (2000)
  - The same formula with $\hat{q}_{0.01}(z)$ in place of the Student-t quantile
  - $\hat{q}_{0.01}(z)$: the empirical 1% quantile of the standardised residuals $\hat{z}_t = (r_t - \hat{\mu})/\hat{\sigma}_t$

Rolling VaR Forecasts and Their Breaches



- A simulated GARCH-t(5) path, 1 500 forecast days, VaR 1% from data up to $t-1$; 15 breaches expected
  - Question for the class: which of the three models in the chart (HS, Normal, GARCH-t) will have the most breaches?
  - Answer: Normal, 25 against 15 expected; GARCH-t, the true model, has 12
  - HS has 18 breaches; HS and Normal react late and their breaches cluster after the volatility burst
  - right: cumulative breaches minus the expected 0.01 t; a correct model wanders around zero

VaR and ES under the Normal and the Student-t Distribution



- Normal returns $N(\mu, \sigma^2)$, tail probability α

$$\mathrm{VaR}_\alpha = -(\mu + \sigma z_\alpha), \qquad \mathrm{ES}_\alpha = -\mu + \sigma \frac{\varphi(z_\alpha)}{\alpha}$$

  - $z_\alpha = \Phi^{-1}(\alpha)$; φ: standard Normal density
- Student-t innovations with ν degrees of freedom, scaled to variance 1
  - quantile $\tilde{t}_\nu^{-1}(\alpha)$ of the slide "Four VaR Forecasts" in place of z_α; ES has its own closed form (A8)
- Chart (left tail, log scale, variance 1)
  - Normal distribution: VaR 1% = 2.33, ES 2.5% = 2.34, almost equal
  - Student-t(5): compute the figures in A8

Independence and DQ Tests

- Christoffersen independence test Christoffersen (1998): is a breach more likely after a breach?

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

  - n_{ij}: number of days in state j after a day in state i (1 = breach); $\hat{\pi}_{01} = n_{01}/(n_{00}+n_{01})$, $\hat{\pi}_{11} = n_{11}/(n_{10}+n_{11})$
  - $\hat{\pi} = (n_{01}+n_{11})/(n_{00}+n_{01}+n_{10}+n_{11})$: the common breach rate (numerator); the denominator lets the rate depend on yesterday; large LR_{ind}: clustered breaches

- Dynamic quantile (DQ) test Engle & Manganelli (2004): can today's breach be predicted?

$$I_t - \alpha = \delta_0 + \sum_{k=1}^{4} \delta_k \left(I_{t-k} - \alpha\right) + \delta_5 \,\mathrm{VaR}_t + u_t, \qquad DQ = \frac{\hat{\delta}^\top X^\top X\, \hat{\delta}}{\alpha(1-\alpha)} \;\sim\; \chi^2_6$$

  - I_t: breach indicator; X: the matrix of the 6 regressors; $\hat{\delta}$: the Ordinary Least Squares (OLS) coefficients; H_0: all $\delta = 0$
  - large DQ: breaches are predictable from past breaches or from the level of VaR

Scores and Model Comparison

- FZ0 score Patton, Ziegel & Chen (2019): scores a (VaR, ES) pair on day t; lower is better

$$\mathrm{FZ0}(v, e; y) = -\frac{1\{y \le v\}(v - y)}{\alpha e} + \frac{v}{e} + \ln(-e) - 1$$

  - $\alpha = 2.5\%$, $v = -\mathrm{VaR}$, $e = -\mathrm{ES}$, y: the realised return; $1\{\cdot\}$: 1 if true, else 0

- Two models: Diebold–Mariano statistic $t = \bar{d}/\sqrt{\hat{\Omega}/T}$ Diebold & Mariano (1995)
  - d_t: the FZ0 difference on day t, $\bar{d}$: its mean; $\hat{\Omega}$: its Newey–West long-run variance with $L = \lfloor 4(T/100)^{2/9} \rfloor$ lags; T: the number of matched forecast days
  - FZ0 is computed on percentage returns: compare models within the same asset and units
  - $t \approx N(0, 1)$ under equal accuracy

- Many models: the 10% model confidence set Hansen, Lunde & Nason (2011)
  - Keeps every model with $p_{MCS} \ge 0.10$; it contains the best model with probability 90%

Which Probability Is Called p Here?

- In the GARCH discussion, p is the persistence
  - the sum of the two dynamic coefficients, $\alpha + \beta$
  - a value near 1: a shock fades slowly
- In the Kupiec test and in A1, p is the nominal breach probability
  - set to 1%
  - it corresponds to the tail probability α of the VaR, ES, DQ and FZ0 formulas and of Chapter 19
- The observed breach fraction estimates the true breach probability π
  - H_0: the true probability equals the nominal one
- Do not mix the two meanings: the GARCH persistence does not enter the Kupiec statistic
  - likewise, α also denotes the ARCH (AutoRegressive Conditional Heteroskedasticity) coefficient of GARCH and the Hill tail index: the formula tells which

Size, Power and the Kupiec Test (A1)

- **Size** and **power**
  - Size: probability of rejecting a correct model; power: probability of rejecting a wrong one
  - A backtest is useful only if its power is high for the errors that matter
- With independent breaches, their number X in T days is binomial with the true rate π

$$X \sim \mathrm{Bin}(T, \pi), \qquad P(X = x) = \binom{T}{x} \pi^x (1-\pi)^{T-x}$$

- Kupiec coverage test Kupiec (1995): nominal rate $p = 1\%$ against observed rate $\hat{\pi} = x/T$, with a Likelihood Ratio (LR) statistic

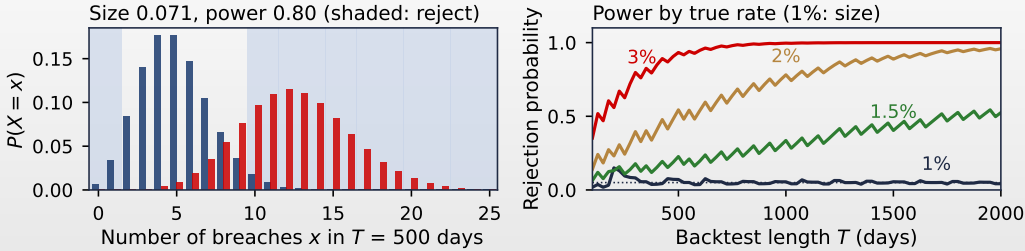
$$LR_{uc}(x) = -2 \ln \frac{(1-p)^{T-x} p^x}{(1-\hat{\pi})^{T-x} \hat{\pi}^x} \sim \chi^2_1 \text{ under } H_0 \colon \pi = p$$

  - x: observed breaches; $LR_{uc} = 0$ when $\hat{\pi} = p$, larger as $\hat{\pi}$ moves away from p; convention $0 \ln 0 = 0$
  - Reject at 5% when $LR_{uc}(x) > 3.84$, the 95% quantile of χ^2_1

Exact Size and Power (A1)

- Exact size and power from the binomial distribution
  - Rejection region $R =$ the counts x with $LR_{uc}(x) > 3.84$
  - Exact size $= \sum_{x \in R} P(X = x \mid \pi = 1\%)$; exact power $= \sum_{x \in R} P(X = x \mid \pi)$ at the true π
  - X is discrete, so the exact size is not exactly 5%
- Example: $T = 100$ days (1 breach expected)
  - $LR_{uc}(0) = -2 \cdot 100 \ln 0.99 = 2.01$, $LR_{uc}(3) = 2.63$, $LR_{uc}(4) = 5.18$: $R = \{x \geq 4\}$
  - Zero breaches never reject
  - Size $P(X \geq 4 \mid 1\%) = 0.018$; power 0.14 if the true rate is 2%, 0.74 if it is 5%
- A non-rejection with low power is "no evidence", not "no problem"
  - State the power before the test (A1; details: Chapter 19)

Size and Power of the Kupiec Test



- Left: breaches in $T = 500$ days under the nominal 1% and under a true 2.5%; shaded counts reject ($x \leq 1$ or $x \geq 10$)
  - size 0.071 (not exactly 5%: X is discrete); power 0.80 against 2.5%
  - right: power grows with T and with the distance of the true rate from 1%; at 1.5% even 2 000 days give only about 50%

Many Tests at Once: Bonferroni, Holm and BH (A2)

- m tests, p-values sorted $p_{(1)} \leq \cdots \leq p_{(m)}$; V: false rejections, R: all rejections

$$\mathrm{FWER} = P(V \geq 1), \qquad \mathrm{FDR} = \mathbb{E}\Big[\frac{V}{\max(R,1)}\Big]$$

  - FWER (family-wise error rate): probability of at least one false discovery; FDR (false discovery rate): expected share of false ones among the discoveries

- **Bonferroni**: reject $H_{(k)}$ if $p_{(k)} \leq 0.05/m$; controls FWER, conservative
- **Step-down correction** Holm (1979): FWER $\leq 5\%$ under any dependence, never fewer rejections than Bonferroni

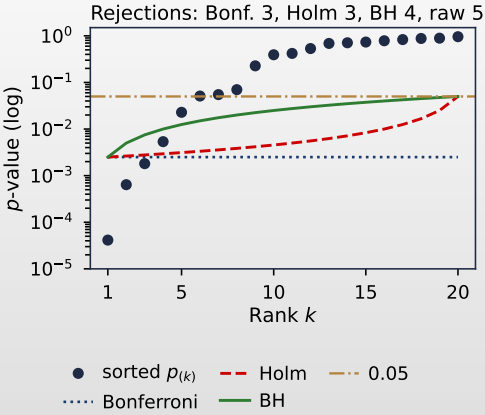
$$\text{reject } H_{(k)} \text{ while } p_{(k)} \leq \frac{0.05}{m-k+1},\ k = 1, 2, \ldots; \text{ stop at the first failure}$$

- **Step-up procedure** Benjamini & Hochberg (1995): FDR $\leq 5\%$ for independent or positively dependent p-values

$$k^* = \max\Big\{k : p_{(k)} \leq \frac{0.05\,k}{m}\Big\}; \quad \text{reject } H_{(1)}, \ldots, H_{(k^*)}$$

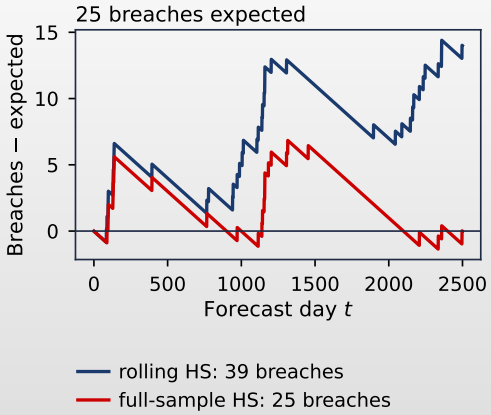
  - BH (Benjamini–Hochberg) rejects at least as many hypotheses as Holm

The Three Thresholds on Simulated p-Values



- 20 simulated tests, 6 with a true effect
  - dots: sorted $p_{(k)}$; a point below a line is rejected by that rule (Holm and BH as on the previous slide)
- Rejections
  - unadjusted 5%: 5; Bonferroni: 3; Holm: 3; BH: 4
  - Holm's line rises to 0.05 at the last rank; BH's line rises linearly, so it admits more discoveries
- Which to use
  - screening many candidates: BH (a few false leads are acceptable)
  - committing capital to one result: FWER control (Holm)

Look-Ahead Bias (B4)



- **Look-ahead bias**: a forecast for day t that uses information from after day $t-1$
  - e.g. parameters or a quantile estimated on the full sample
  - the forecast looks better than any forecast that could have been made in real time
- Chart: simulated GARCH-t losses, 2 500 forecast days, 25 breaches expected
  - rolling HS (500 days): 39 breaches; full-sample HS: 25, exactly 1% by construction
- Test for your code
  - every forecast for day t must be computable from a data file truncated at $t-1$ (B4)

Robust Errors and GARCH Forecasts

- Quasi-Maximum Likelihood Estimation (QMLE), also called QML Bollerslev & Wooldridge (1992) (A3): Gaussian likelihood, any innovation distribution with $\mathbb{E}z_t^4 < \infty$

$$\widehat{\mathrm{Var}}(\hat{\theta}) = \frac{1}{T} A^{-1} B A^{-1}, \qquad A = -\mathbb{E}\Big[\frac{\partial^2 \ell_t}{\partial\theta\,\partial\theta^\top}\Big], \quad B = \mathbb{E}\big[s_t s_t^\top\big]$$

  - ℓ_t: log-likelihood of day t; $s_t = \partial\ell_t/\partial\theta$: the score; with Normal innovations $A = B$ and the sandwich reduces to A^{-1}/T
- Delta method (A4): the standard error of a smooth function $g(\hat{\rho})$
  - $\mathrm{se}(g(\hat{\rho})) \approx |g'(\hat{\rho})|\,\mathrm{se}(\hat{\rho})$; unreliable when g' explodes, as for the half-life near $p = 1$
- GARCH(1,1) variance forecasts (A5, A6): geometric return to the long-run level

$$\sigma^2_{T+h} = \bar{\sigma}^2 + (\alpha+\beta)^{h-1}\big(\sigma^2_{T+1} - \bar{\sigma}^2\big), \qquad \bar{\sigma}^2 = \frac{\omega}{1-\alpha-\beta}$$

  - the h-day variance is the sum $\sum_{j=1}^{h} \sigma^2_{T+j}$, not $h\,\sigma^2_{T+1}$ (the square-root-of-time rule)

Exercise Map

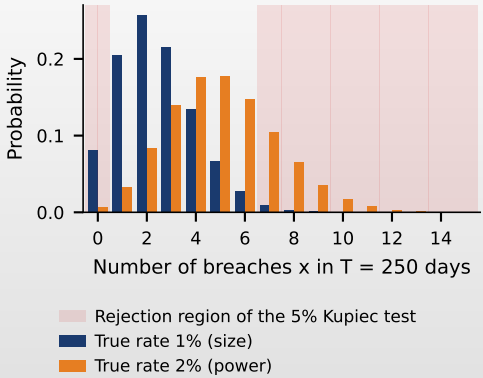
Exercise	Question	Type	Model
A1	how often does one year catch a bad VaR model?	Solved	–
A2	which strategies survive multiple testing?	Solved	–
A6	GARCH forecast with supplied parameters	Proposed, independent	GARCH prerequisites
B1	how heavy are the BET tails?	Solved	–
B2	which interval for GARCH persistence?	Solved	–
B3	which VaR 1% model would we have trusted?	Solved	–
B4	how much does look-ahead flatter a backtest?	Solved	–
C1	what is wrong in an AI-written backtest?	Proposed	A1, B3, B4
C2	your team's proposal (homework)	Proposed	worked proposal
Extension packs	A3–A5, A7–A12, B5–B8, C3 (optional, after the lecture)	Solved and Proposed (marked on each task)	named on each task

- **[Solved]**: full solution in the slides and in the notebook, a model to copy; **[Proposed]**: you solve it, the solution is discussed in class

A1 [Solved]: The Power of the Kupiec Test — Task

- **Question**: How often does one year of data catch a VaR 1% model whose true breach rate is 2%?
- **Data and assumptions**
  - A VaR 1% model is backtested over T days; the number of breaches is $X \sim \mathrm{Bin}(T, \pi)$; $H_0 : \pi = 0.01$.
  - Kupiec statistic $LR_{uc}(x)$ with nominal rate $p = 0.01$ (formula: the slide "Size, Power and the Kupiec Test"); reject at 5% if $LR_{uc} > 3.84$ Kupiec (1995).
- **Tasks**
  - (a) For $T = 250$, compute $LR_{uc}(x)$ for $x = 0, 1, 2, \ldots$ and find the rejection region R; repeat for $T = 1\,000$.
  - (b) Compute the exact size $P(X \in R \mid \pi = 1\%)$ for both T.
  - (c) Compute the exact power $P(X \in R \mid \pi = 2\%)$ for both T.
- **Report**
  - The two rejection regions, the exact sizes and the exact powers.
- **Interpretation**
  - How often does one year of data catch a model whose true breach rate is twice the nominal one?
  - Why does Basel add the traffic-light multiplier (green, yellow and red zones; Chapter 8)?

A1 [Solved]: Solution (1/2)



- (a) Step 1 ($T = 250$, expected 2.5 breaches): $LR_{uc}(0) = -2 \cdot 250 \ln 0.99 = 5.03 > 3.84$, so $x = 0$ is rejected; $LR_{uc}(1) = 1.18$.
- Step 2: $LR_{uc}(6) = 3.56 < 3.84$ and $LR_{uc}(7) = 5.50 > 3.84$, so $R = \{0\} \cup \{x \geq 7\}$.
- Step 3 ($T = 1\,000$, expected 10 breaches): the same scan gives $R = \{x \leq 4\} \cup \{x \geq 17\}$.
- (b) Step 4: at $T = 250$, size $= P(X = 0) + P(X \geq 7) = 0.0811 + 0.0137 = 0.095$ under $\pi = 1\%$; at $T = 1\,000$ the size is 0.055.
- Reading: X is discrete, so the exact size is not 5%; with few expected breaches it can be almost twice the nominal level.

MFM_ch19_seminar

A1 [Solved]: Solution (2/2)

- (c) Step 5: at $T = 250$, power $= P(X = 0) + P(X \geq 7) = 0.0064 + 0.2363 = 0.243$ under $\pi = 2\%$.
- Step 6: at $T = 1\,000$, $P(X \leq 4) + P(X \geq 17) = 0.782$ under $\pi = 2\%$.

	$T = 250$	$T = 1\,000$
Rejection region	$x = 0$ or $x \geq 7$	$x \leq 4$ or $x \geq 17$
Exact size ($\pi = 1\%$)	0.095	0.055
Power ($\pi = 2\%$)	0.243	0.782

- Interpretation: one year of data rejects a model with twice the breach rate only about one time in four; four years are needed for a power near 80%; that is why Basel adds the traffic-light multiplier instead of relying on the test alone.

A2 [Solved]: Holm and Benjamini–Hochberg — Task

- **Question**: Which of ten trading strategies survive a correction for multiple testing?
- **Data and assumptions**
  - Ten trading strategies, each tested against a zero excess return; sorted p-values $p_{(\mathbf{1})}, \ldots, p_{(\mathbf{10})}$: 0.001, 0.004, 0.012, 0.019, 0.028, 0.041, 0.09, 0.20, 0.46, 0.73; $m = 10$; target level 5%.
  - FWER, FDR and the Bonferroni, Holm and BH rules: the slide "Many Tests at Once".
- **Tasks**
  - (a) Count the rejections without correction ($p \leq 0.05$) and with Bonferroni ($p \leq 0.05/m$).
  - (b) Apply Holm at FWER 5% Holm (1979): compare $p_{(i)}$ with $0.05/(m - i + 1)$ for $i = 1, 2, \ldots$ and stop at the first $p_{(i)}$ above its threshold.
  - (c) Apply Benjamini–Hochberg at FDR 5% Benjamini & Hochberg (1995): find the largest k with $p_{(k)} \leq 0.05k/m$ and reject $H_{(\mathbf{1})}, \ldots, H_{(k)}$.
- **Report**
  - A table with $p_{(i)}$ and both thresholds, and the number of rejections of each procedure.
- **Interpretation**
  - What does each procedure promise?
  - Which one would you use to screen trading strategies?

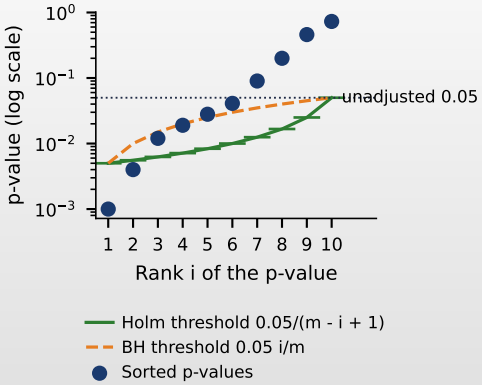
Back Exercise list

A2 [Solved]: Solution (1/2)

i	1	2	3	4	5	6	7	8	9	10
$p_{(i)}$	0.001	0.004	0.012	0.019	0.028	0.041	0.09	0.20	0.46	0.73
Holm	0.0050	0.0056	0.0063	0.0071	0.0083	0.0100	0.0125	0.0167	0.0250	0.0500
BH	0.005	0.010	0.015	0.020	0.025	0.030	0.035	0.040	0.045	0.050

- (a) Unadjusted: 6 p-values below 5%; Bonferroni ($p \leq 0.005$): 2 rejections.
- (b) Holm, step down: $0.001 \leq 0.0050$ and $0.004 \leq 0.0056$ pass; $0.012 > 0.0063$ stops the procedure: 2 rejections.

A2 [Solved]: Solution (2/2)



- (c) BH, step up from the largest p: $0.041 > 0.030$, $0.028 > 0.025$, but $0.019 \leq 0.020$, so $k = 4$: 4 rejections (also $0.012 \leq 0.015$ is not needed, since BH rejects all $H_{(i)}$ with $i \leq k$).
- Summary: none 6, Bonferroni 2, Holm 2, BH 4.
- Holm promises: probability of any false discovery $\leq 5\%$ (FWER), uniformly more powerful than Bonferroni.
- BH promises: expected share of false discoveries among the declared ones $\leq 5\%$ (FDR) when the p-values are independent or positively regression dependent on the true nulls Benjamini & Yekutieli (2001); it rejects more.
- Interpretation: for screening many candidate strategies BH is the usual choice (a few false leads are acceptable); before committing capital to one strategy, FWER control (Holm) is the safer promise.

MFM_ch19_seminar

A6 [Proposed]: GARCH Forecast with Given Parameters — Task

Solution discussed in class

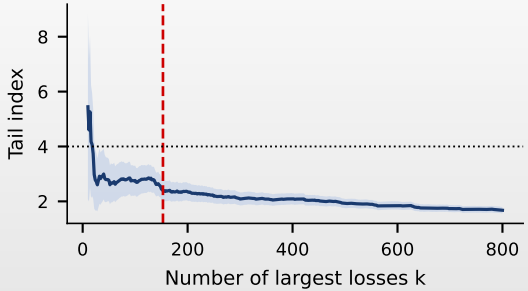
- **Question**: What happens to the GARCH forecast when tomorrow's volatility is above its long-run level
- **Data and assumptions**
  - A GARCH(1,1) for daily returns in % with $\omega = 0.02$, $\alpha = 0.10$, $\beta = 0.88$; tomorrow's variance $\sigma^2_{T+1} = 4$ ($\sigma_{T+1} = 2\%$).
  - Forecast formula and geometric sum: the slide "Robust Errors and GARCH Forecasts"; 252 trading days a year.
- **Tasks**
  - (a) Compute the persistence, the long-run variance $\bar{\sigma}^2$ and the annualised long-run volatility.
  - (b) Compute the half-life of a shock.
  - (c) Forecast the variance on day 5, σ^2_{T+5}.
  - (d) Compute the variance of the 5-day return and the 5-day volatility, and compare them with $5\sigma^2_{T+1} = 20$ and $\sqrt{5}\,\sigma_{T+1}$.
- **Report**
  - Every number of (a)–(d) with the substituted formula.
- **Interpretation**
  - Is the 5-day volatility above or below $\sqrt{5}\,\sigma_{T+1}$?
  - Why?

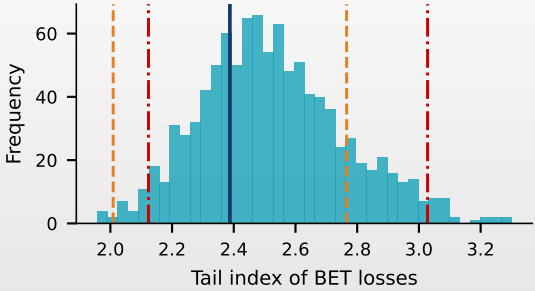
B1 [Solved]: Stylised Facts of BET with Robust Inference — Task

- **Questions**
  - How heavy are the tails of the BET?
  - Does its autocorrelation survive once volatility clustering is accounted for?
- **Data and assumptions**
  - BET official daily closes (course data), 6 Jan 2000 – 18 Sep 2026: $T = 6\,670$ log returns r_t in %; losses $L_t = -r_t$; moving-block bootstrap, 999 resamples of 20-day blocks; all tests at 5%.
- **Tasks**
  - (a) Compute the standard deviation, the skewness and the excess kurtosis of r_t.
  - (b) Estimate the Hill tail index of losses with $k = 5\%$ of the loss days, with the asymptotic 95% CI of the slide "The Tail Index and the Hill Estimator" and the 95% percentile CI of the block bootstrap.
  - (c) Compute the Ljung–Box $Q(10)$ on r_t and on r_t^2 and the robust $Q^*(10)$ of the slide "The Robust Q^* Test and the Block Bootstrap"; compare each with $\chi^2_{10;0.95} = 18.31$.
  - (d) Count the days with $|r_t - \bar{r}| > 4s$ and compare with the number expected under the Normal distribution (the slide "Moments, Autocorrelation and the Ljung–Box Test").
- **Report**
  - The statistics of (a), $\hat{\alpha}$ with both intervals and a Hill plot, the three portmanteau statistics with p-values and decisions, the count of (d).
- **Interpretation**
  - Why is the block interval for the tail index wider than the asymptotic one?
  - Does the autocorrelation of r_t survive the robust test?
  - How far can you trust its p-value, given the tail index?

Back Exercise list

B1 [Solved]: Solution (1/3)





Asymptotic 95% band
Hill estimate $\hat{\alpha}(k)$
Chosen k = 153 (5% of loss days)
$\alpha = 4$: finite fourth moment above
Block-bootstrap estimates (999)
Estimate
Asymptotic 95% CI
Block-bootstrap 95% CI

MFM_ch19_seminar

B1 [Solved]: Solution (2/3)

- (a) $s = 1.40\%$, skewness -0.65, excess kurtosis 11.1: fat, left-skewed tails.
- (b) Step 1: sort the losses; $k = 153$ (5% of the loss days); $\hat{\alpha} = 2.39$.
- Step 2 (asymptotic): $2.39(1 \pm 1.96/\sqrt{153}) = [2.01;\ 2.77]$, standard error 0.193.
- Step 3 (block bootstrap): rebuild 999 series from random 20-day blocks, re-estimate $\hat{\alpha}$ with the same k, take the 2.5% and 97.5% percentiles: [2.12; 3.03], standard error 0.233.
- Reading: both intervals lie below 4, supporting an infinite fourth moment under the assumed regularly varying tail model, and the sample kurtosis (11.1) is not reported as an estimate (A11).

MFM_ch19_seminar

B1 [Solved]: Solution (3/3)

- (c) $Q(10)$ on r_t: 108.0 (p <0.001) > 18.31: reject no autocorrelation; robust $Q^*(10) = 24.4$ (p 0.007): still reject, but 4.4 times smaller.
- $Q(10)$ on r_t^2: 2 440 (p <0.001): strong volatility clustering, read descriptively, since if the fourth moment is infinite r_t^2 has no finite variance; first-order autocorrelation of r_t: 0.109.
- (d) 51 days beyond ± 4 standard deviations, against $6\,670 \cdot 2\Phi(-4) = 0.42$ expected under the Normal distribution.
- Interpretation 1: extremes cluster in crises; blocks keep the clusters, so the block interval accounts for the dependence that the asymptotic formula ignores.
- Interpretation 2: the robust test still rejects, but with $\hat{\alpha} < 4$ its χ^2 calibration is not guaranteed: sample autocorrelations of heavy-tailed GARCH-type returns converge more slowly and to non-Normal limits Davis & Mikosch (1998); read it as evidence of a small autocorrelation; stale prices of less liquid stocks (Chapters 2, 10) are a plausible explanation that needs its own test.

MFM_ch19_seminar

B2 [Solved]: GARCH(1,1)-t for BET: an Interval at the Boundary — Task

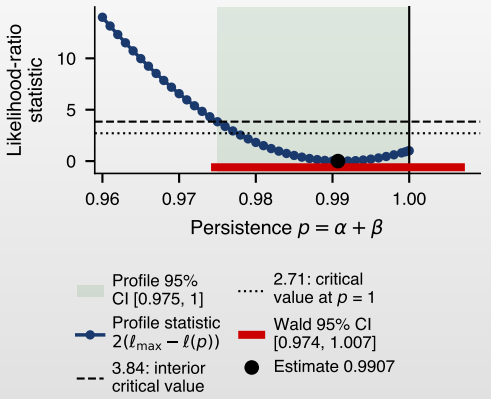
- **Question**: Which interval for GARCH persistence can be reported when the estimate is close to 1?
- **Data and assumptions**
  - BET daily log returns in % (B1); constant mean; the model defined in the slide "GARCH(1,1)-t"; robust standard errors Bollerslev & Wooldridge (1992); level 95%.
  - Persistence $p = \alpha + \beta$: a finite unconditional variance needs $p < 1$; strict stationarity only needs $\mathbb{E}\ln(\alpha z_t^2 + \beta) < 0$ Nelson (1990), with $z_t = \varepsilon_t/\sigma_t$.
  - $\hat{p}$ lies near the boundary $p = 1$ of the imposed parameter space $p \leq 1$ Andrews (1999).
- **Tasks**
  - (a) Estimate the model by maximum likelihood; report $\hat{\alpha}$, $\hat{\beta}$, $\hat{\nu}$ with robust standard errors and $\hat{p}$.
  - (b) Compute $\mathrm{se}(\hat{p})$ from the robust covariance of $(\hat{\alpha}, \hat{\beta})$ and the Wald CI $\hat{p} \pm 1.96\,\mathrm{se}(\hat{p})$.
  - (c) Profile likelihood: for each p in $0.960, 0.961, \ldots, 0.999, 0.9995, 0.9999, 1$ maximise the log-likelihood over μ, ω, ν and $s = \alpha/p$ ($\alpha = sp$, $\beta = (1-s)p$); keep $p < 1$ if $2(\ell_{\max} - \ell(p)) \leq 3.84$, and $p = 1$ if the statistic is ≤ 2.71 (the slide "An Interval at a Boundary").
  - (d) Transform the estimate and both intervals into half-lives (the slide "An Interval at a Boundary").
  - (e) Compare the standardised Student-t quantile $\tilde{t}_{\hat{\nu}}^{-1}(0.01)$ (the slide "Four VaR Forecasts") with $z_{0.01}$.
- **Report**
  - Estimates with standard errors; both CIs for p and for the half-life; the profile curve; the two quantiles.
- **Interpretation**
  - Which interval can you report?
  - Why is the other one uninformative under this restriction?
  - What does $\hat{\nu} \approx 5$ mean for VaR?

Back Exercise list

B2 [Solved]: Solution (1/2)

- (a) $\hat{\alpha} = 0.192$ (se 0.025), $\hat{\beta} = 0.799$ (se 0.024), $\hat{\nu} = 5.01$ (se 0.31); $\hat{\rho} = 0.9907$.
- (b) With the robust covariance, $\mathrm{se}(\hat{\rho}) = 0.0084$ (the negative covariance of $\hat{\alpha}$ and $\hat{\beta}$ makes it smaller than either standard error).
- Wald CI: $0.9907 \pm 1.96 \cdot 0.0084 = [0.974;\ 1.007]$: the upper end is above 1, outside the parameter space $\rho \leq 1$.
- Why it is unusable under the imposed restriction: at the boundary the distribution of $\hat{\rho}$ is not Normal Andrews (1999), and the interval contains values $\rho > 1$, for which the variance explodes and the half-life is not defined.
- (d) Half-life $\ln 0.5 / \ln 0.9907 = 74$ days; from the Wald interval: 26 days at the lower end and no value at the upper end.

B2 [Solved]: Solution (2/2)



- (c) Profile statistic $2(\ell_{max} - \ell(p))$: 4.32 at $p = 0.974$ (excluded), 3.83 at $p = 0.975$ (included), 0.99 at $p = 0.9999$ (included); at $p = 1$: $1.01 < 2.71$, boundary p-value 0.16 (included).
- Profile CI [0.975; 1]: the likelihood excludes no value from 0.975 up to and including Integrated GARCH (IGARCH), $p = 1$.
- (d) Half-life interval: at least $\ln 0.5 / \ln 0.975 = 27$ days, with no finite upper bound, since $p = 1$ is in the interval.
- (e) $t_{\hat{\nu}}^{-1}(0.01)\sqrt{(\hat{\nu}-2)/\hat{\nu}} = -2.606$ against $z_{0.01} = -2.326$: for the same σ_t the Student-t VaR 1% is 1.12 times the Normal one (zero mean).
- Interpretation 1: report the profile interval; shocks last at least a month, but the data cannot tell a month from several years, and IGARCH ($p = 1$) is not rejected.
- Interpretation 2: $\hat{\nu} \approx 5$ means heavy tails even conditionally on σ_t; a Normal GARCH VaR 1% would be too low.

MFM_ch19_seminar

B3 [Solved]: Backtesting VaR 1% and ES 2.5% for BET, 2005–2026 — Task

- **Question**: Which of four VaR 1% models for the BET would we have trusted in real time, 2005–2026?
- **Data and assumptions**
  - BET daily losses $L_t = -r_t$, evaluation 3 Jan 2005 – 18 Sep 2026, $T = 5\,432$ days; information through $t-1$ only.
  - Four one-day models McNeil & Frey (2000): HS (500 days), Normal (500 days), GARCH-t (1000-day window, re-estimated every 21 days), FHS (GARCH-t filter), as in the slide "Four VaR Forecasts"; VaR 1% and the pair (VaR 2.5%, ES 2.5%).
- **Tasks**
  - (a) Count the breaches $L_t > \mathrm{VaR}_t$; compute the rate, the Kupiec p_{uc} and the Christoffersen p_{ind} Kupiec (1995), Christoffersen (1998).
  - (b) Run the DQ test of the slide "Independence and DQ Tests" (4 lagged hits and VaR_t) Engle & Manganelli (2004).
  - (c) Compute the mean FZ0 score of (VaR 2.5%, ES 2.5%), Diebold–Mariano tests (Newey–West, 9 lags) and the 10% model confidence set (the slide "Scores and Model Comparison").
  - (d) Comparative ES backtest against HS Nolde & Ziegel (2017): with t the DM statistic of FZ0(model) − FZ0(HS), red if $t > 1.645$, green if $t < -1.645$, yellow otherwise.
- **Report**
  - Coverage table (a)–(b), cumulative-breach chart, scores, MCS p-values, ES zones and mean VaR.
- **Interpretation**
  - Which model meets all the backtesting criteria of this exercise?
  - Which one would a trader prefer?
  - Do the formal comparisons confirm the backtests?

Back Exercise list

B3: Empirical Quantiles and Tail Means

- HS: the most recent 500 daily losses
  - quantile: linear interpolation between adjacent order statistics
  - empirical ES: the estimator of Chapter 7; $500 \times 2.5\% = 12.5$, so the 12 largest losses plus half of the 13th, divided by 12.5
- FHS: the 1 000 standardised residuals of the GARCH estimation window
  - tail mean: the same estimator; $1\,000 \times 2.5\% = 25$, so the mean of the 25 smallest residuals
  - transform it with the forecast conditional mean and volatility, then reverse the sign to express a loss
- When $n\alpha$ is not an integer, the boundary observation gets the fractional weight $n\alpha - \lfloor n\alpha \rfloor$
- GARCH parameters and the FHS residual pool are refitted every 21 trading days
- Conditional volatility updates daily, using only past information

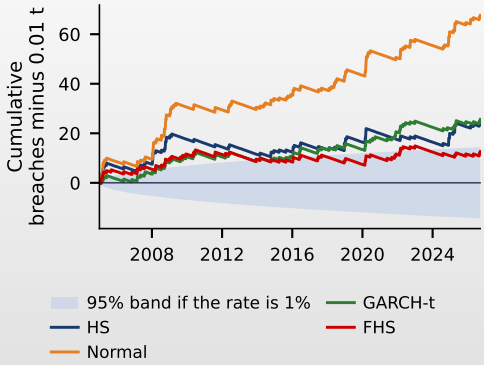
B3 [Solved]: Solution (1/3)

- (a)–(b) Steps: $\mathrm{Hit}_t = 1\{L_t > \mathrm{VaR}_t\}$; expected breaches $0.01 \cdot 5\,432 \approx 54$; tests at 5%.

Model	x	Rate (%)	p_{uc}	p_{ind}	p_{DQ}
HS	79	1.45	0.002	<0.001	<0.001
Normal	122	2.25	<0.001	<0.001	<0.001
GARCH-t	80	1.47	0.001	0.864	0.007
FHS	67	1.23	0.095	0.267	0.129

- HS and Normal fail coverage, independence and DQ: they react too slowly and their breaches cluster.
- For GARCH-t independence is not rejected, but it has too many breaches, so it fails coverage and DQ; only FHS is rejected by none of the three tests.

B3 [Solved]: Solution (2/3)



- Chart: cumulative breaches minus the expected $0.01\,t$; a line above the light-blue band has too many breaches for a 1% model
- At the end only FHS lies inside the band (counts: table of the previous slide)
- The Normal line jumps in 2008, 2020 and 2022: its breaches arrive in clusters, the pattern that the independence and DQ tests detect

MFM_ch19_seminar

B3 [Solved]: Solution (3/3)

- (c)–(d) Lower FZ0 is better; DM t vs HS < -1.645 means the model beats HS.

Model	FZ0	p_{MCS}	DM t vs HS	ES zone	Mean VaR (%)
HS	1.422	0.00	–	standard	3.66
Normal	1.495	0.00	2.57	red	2.90
GARCH-t	1.175	0.18	−5.82	green	2.95
FHS	1.157	1.00	−6.01	green	3.17

- No statistically detectable FZ0 difference between FHS and GARCH-t (DM $t = 1.39$); together they form the model confidence set {GARCH-t, FHS}; Normal is red, GARCH-t and FHS are green.
- Interpretation 1: FHS is the only model that meets all the stated criteria (coverage, independence, DQ, green ES zone), which summarises the evaluation criteria used in this exercise (not the regulatory checklist of the Basel backtesting standard); a trader wants low capital, and the Normal model has the lowest average VaR (2.90% vs FHS 3.17%) while failing every test.
- Interpretation 2: the formal comparisons confirm the backtests and add one thing: GARCH-t fails coverage, but its FZ0 score for (VaR, ES) is not significantly different from that of FHS.

MFM_ch19_seminar

B4 [Solved]: Look-Ahead Bias — Task

- **Question**: How much better does a backtest look when it uses information from the future?
- **Data and assumptions**
  - BET losses of B3: evaluation 3 Jan 2005 – 18 Sep 2026, $T = 5\,432$ days, $0.01T \approx 54$ expected breaches; rolling HS and GARCH-t results from B3.
  - The look-ahead versions use data that were not available at $t-1$; Kupiec test at 5%.
- **Tasks**
  - (a) HS with look-ahead: set every day $\mathrm{VaR}_{1\%} = q_{0.99}(L)$, the empirical quantile of all losses of the evaluation period; count the breaches and compute the rate.
  - (b) GARCH-t with look-ahead: estimate once on the full sample 6 Jan 2000 – 18 Sep 2026, filter $\hat{\sigma}_t$ day by day, set $\mathrm{VaR}_t = -(\hat{\mu} + \hat{\sigma}_t\, t_{\hat{\nu}}^{-1}(0.01))$ (the slide "Four VaR Forecasts"); count the breaches and run the Kupiec test.
  - (c) Compare both with the rolling versions of B3 in one table.
- **Report**
  - A table with breaches, rate and p_{uc} for the four versions.
- **Interpretation**
  - Why does the look-ahead version look better?
  - Could anyone have used it in 2008?

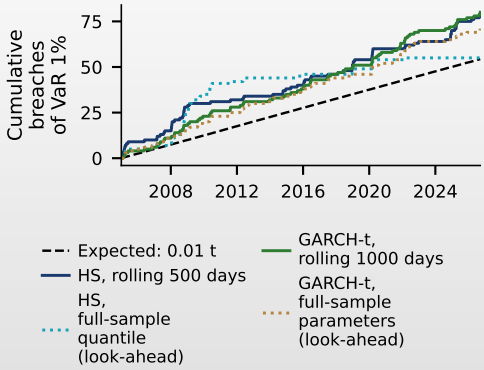
Back Exercise list

B4 [Solved]: Solution (1/2)

- (a) Full-sample HS: $\mathrm{VaR}_{1\%} = 4.24\%$ every day; 55 breaches, rate 1.01%: about 1% by construction, because 1% of the losses of a sample exceed its own 0.99 quantile.
- (b) Full-sample GARCH-t: 71 breaches, rate 1.31%, p_{uc} 0.030: still rejected at 5%, but much closer to 1% than the rolling model.

Model	Parameters	x	Rate (%)	p_{uc}
HS	full sample (look-ahead)	55	1.01	0.93
HS	rolling, 500 days	79	1.45	0.002
GARCH-t	full sample (look-ahead)	71	1.31	0.030
GARCH-t	rolling, 1000 days	80	1.47	0.001

B4 [Solved]: Solution (2/2)



- (c) Reading: both look-ahead versions have fewer breaches than their rolling counterparts and look better calibrated.
- Why: the full sample includes 2008, 2020 and every later crash, so the quantile and the parameters have already seen the losses they are tested on.
- Interpretation: no one could have used these numbers in 2008, since they depend on losses from 2008 to 2026; only the rolling result is evidence about a forecasting model.
- Check for your own code: the truncation test of the slide "Look-Ahead Bias".

MFM_ch19_seminar

C1 [Proposed]: Critique an AI-Written Backtest — the AI Answer

Solution discussed in class

- **Student prompt**: "bet holds the daily BET closes. Write a backtest of historical-simulation VaR 1% with a 500-day window and a Kupiec test, and give me a reference for the result."
- **Answer of the AI assistant**

```
import numpy as np
from scipy.stats import chi2
r = np.log(bet).diff().dropna()
var = -r.rolling(500).quantile(0.01)
hit = (-r > var)[var.notna()]
x, n, p = hit.sum(), hit.size, 0.01
pi = x / n
lr = -2 * (x*np.log(p) + (n-x)*np.log(1-p) - x*np.log(pi) - (n-x)*np.log(1-pi))
print(x, n, chi2.sf(lr, df=2))
```

- "The breach rate is close to 1% and the Kupiec p-value is above 5%: historical simulation is unbiased for BET, as shown by Kupiec and Christoffersen (2004), *Backtesting historical simulation VaR*, Journal of Risk Model Validation, 8(2), 45–67."

C1 [Proposed]: Critique an AI-Written Backtest — Task

Solution discussed in class

- The answer on the previous slide contains at least **three** errors that often appear in AI-written project code; also assess its conclusion and check how the code behaves when there are zero breaches
  - Model: B3 and B4 (backtest and the look-ahead trap), A1 (Kupiec test)
- Your tasks
  - (a) Find each error: which line or sentence, and why it is wrong
  - (b) Say how you would detect it without knowing the answer: a timing test, the number of restrictions, a DOI (Digital Object Identifier) search
  - (c) Give the corrected code and a correct reference
  - (d) Log each error as in `AI_ERRORS.md`: what the AI answered, what was wrong, how you found out, the correct version
- **Report**
  - The list of errors (line, reason, detection test), the corrected code, the correct references and the `AI_ERRORS.md` entries.
- **Interpretation**
  - Which claim about BET can the corrected backtest support?

A Worked Proposal (1/2)

- **Question**: did the tail risk of the BET change after the index provider FTSE (Financial Times Stock Exchange, FTSE Russell, index provider) Russell reclassified the BVB as a Secondary Emerging market (21 September 2020)?
- **Hypothesis**: H_0: same Hill tail index of losses and same rolling HS and FHS VaR 1% breach rates in the six years before and after
- **Data**: BET official closes, 22 Sep 2014 – 18 Sep 2026; control: WIG20 (Poland); duplicate records (rows identical to the previous one) and weekend-dated rows removed; BET and WIG20 prices aligned on common trading days first, then returns
- **Identification**: difference-in-differences of Hill indices (BET minus WIG20); placebo dates $\pm$ one year
  - Identifying assumptions: without the event the two tail indices would have changed alike; no spillover of the reclassification to Poland; no concurrent shock specific to one of the two markets
  - Complementary scale diagnostic: a sup-Wald break test with unknown date Andrews (1993) on $\ln|r_t|$ (it detects shifts in volatility, not in the tail index)

A Worked Proposal (2/2)

- **Inference**: moving-block bootstrap (999 resamples, 20-day blocks)
  - Within each period the same blocks of days for BET and WIG20, so the dependence between the two markets is kept; independent draws in the two disjoint periods
- **Power**: minimum detectable difference-in-differences at 80% power and 5% level, $(z_{0.975} + z_{0.80})\,\mathrm{se} = 2.80\,\mathrm{se}$, with z_p the standard Normal p-quantile and se the standard error of the difference-in-differences (the standard 80%-power formula; C2(e))
- **Specification grid**: $k = 2.5\%$, 5%, 10% of the loss days $\times$ blocks of 10, 20, 40 days: 9 tests of the difference-in-differences, Holm correction (A2)
- **Risks**: the pre-period contains the COVID-19 crash and the BET fall of 19 December 2018, after the announcement of the bank-asset tax (Government Emergency Ordinance, GEO 114/2018); one event, so no causal claim without the control

C2 [Proposed]: Your Team's Proposal — Task (1/2)

Solution discussed in class

- **Questions**
  - What will your team study?
  - Can the design answer the question?
- **Data and assumptions**
  - One market or asset of your choice: a series from the course data (the list is in the course repository) or a public source named by its official owner; daily data unless the question needs another frequency.
  - The proposal has the sections (a)–(g) below; the worked proposal of the previous slides is a partial model; ideas: Part C of Seminars 0–18.
- **Tasks**
  - (a) Write the question in one sentence, with H_0, H_1 and the comparison model the study must beat.
  - (b) Name the published paper closest to your question and the table or figure you compare your results with, with DOI and table number.
  - (c) Specify the data: series, owner, period, frequency and cleaning rules.
  - (d) Describe the identification: control market, placebo dates, break test with unknown date, or out-of-sample design.
  - (e) State the main test and its loss function, and compute the smallest effect your design detects with 80% power at your T (as in A1).
  - (f) List the specification grid (windows, lags, α levels, samples), count the tests and choose the multiplicity correction (A2, A10).

Back Exercise list

C2 [Proposed]: Your Team's Proposal — Task (2/2)

Solution discussed in class

- **Tasks (continued)**
  - (g) List the main risks: data problems, too few events, overfitting.
- **Report**
  - The proposal on one page, with 10–15 references, at least 3 from the last five years, each checked to exist (DOI or link), within one week.
  - Model: the worked proposal
- **Checklist for the proposal**
  - Is the question answerable with data you already have?
  - Is the comparison model strong enough?
  - Is the loss function fixed in advance?
  - Does every forecast use only information available at $t - 1$?

Team Project: Proposal and Repository

- C2: the one-page proposal, within one week
- README: question, data, how to run
- The pre-analysis plan as a commit, before the test period is opened
- `AI_USE.md` and `AI_ERRORS.md`: every AI use and every caught error, as in C1
- Extension packs are optional; further study: Statistics of Financial Markets
- Assessment criteria
  - a precise question and an analysis plan fixed before the test period
  - correct inference: HAC (Heteroskedasticity and Autocorrelation Consistent) standard errors or bootstrap errors, multiple testing controlled
  - risk forecasts validated by out-of-sample backtesting
  - a reproducible repository: the code runs from the raw data to the tables

Extension Packs: What Each One Needs

Pack	Exercises	Prerequisites
1. Volatility inference	A3 [Solved], A4, A5 [Solved], B5	B2
2. Risk measures and backtests	A7 [Solved], A8, A9, B6	A1, B3
3. Tails and multiple testing	A10, A11 [Solved], B7, A12	A2, B1
4. Data traps	B8	B4
5. Nonstandard errors	C3	A2, the lecture case study

- Do the exercises of a pack in the order shown; each one names its model, and the notebook sections carry the same labels
- Useful for the team project: C3 is a possible project start

A3 [Solved]: The QMLE Sandwich for GARCH(1,1) — Task

- **Question**: How much do Hessian-only standard errors understate GARCH uncertainty when the innovations have heavy tails?
- **Data and assumptions**
  - Model: $r_t = \sigma_t z_t$, $\sigma_t^2 = \omega + \alpha r_{t-1}^2 + \beta \sigma_{t-1}^2$, $\theta = (\omega, \alpha, \beta)^\top$; z_t i.i.d. with $\mathbb{E}z_t = 0$, $\mathbb{E}z_t^2 = 1$, $\kappa = \mathbb{E}z_t^4 < \infty$.
  - Estimation by the Gaussian quasi-likelihood $\ell_t(\theta) = -\frac{1}{2}(\ln \sigma_t^2 + r_t^2/\sigma_t^2)$; notation $g_t = \sigma_t^{-2}\, \partial \sigma_t^2 / \partial \theta$, known at $t-1$ Bollerslev & Wooldridge (1992).
  - For (c): the GARCH(1,1)-t fit to BET in B2 gives $\hat{\nu} = 5.011$; for Student-t innovations $\kappa = 3(\nu-2)/(\nu-4)$ when $\nu > 4$.
- **Tasks**
  - (a) Differentiate ℓ_t with respect to θ, write the score $s_t = \partial \ell_t / \partial \theta$ in terms of z_t and g_t, and show that $\mathbb{E}[s_t \mid \mathcal{F}_{t-1}] = 0$.
  - (b) Compute $A = -\mathbb{E}[\partial s_t / \partial \theta^\top]$ and $B = \mathbb{E}[s_t s_t^\top]$ using $\mathbb{E}z_t^2 = 1$ and $\mathbb{E}z_t^4 = \kappa$, and show that the sandwich equals $A^{-1} B A^{-1} = \frac{\kappa - 1}{2} A^{-1}$.
  - (c) Substitute $\hat{\nu} = 5.011$: compute κ, the correction factor $(\kappa - 1)/2$ for variances and $\sqrt{(\kappa-1)/2}$ for standard errors.
- **Report**
  - The expressions for s_t, A and B; the values of κ and of both correction factors.
- **Interpretation**
  - When are the Hessian-only standard errors (square roots of the diagonal of A^{-1}/T) correct?
  - By how much do they inflate the t-statistics for BET?
  - What happens when $\nu \leq 4$?

Back Exercise list

A3 [Solved]: Solution (1/2)

- (a) Step 1: $\partial \ell_t / \partial \sigma_t^2 = -\frac{1}{2}\left(\frac{1}{\sigma_t^2} - \frac{r_t^2}{\sigma_t^4}\right) = \frac{1}{2\sigma_t^2}(z_t^2 - 1)$.
- Step 2 (chain rule): $s_t = \frac{1}{2}(z_t^2 - 1)\, g_t$; g_t is known at $t-1$ and $\mathbb{E}[z_t^2 - 1] = 0$, so $\mathbb{E}[s_t \mid \mathcal{F}_{t-1}] = 0$: the score is a martingale difference.
- (b) Step 3: $\partial z_t^2 / \partial \theta = -z_t^2 g_t$, so $\frac{\partial s_t}{\partial \theta^\top} = \frac{1}{2}(z_t^2 - 1)\frac{\partial g_t}{\partial \theta^\top} - \frac{1}{2} z_t^2\, g_t g_t^\top$.
- Step 4: the first term has conditional mean 0 and $\mathbb{E} z_t^2 = 1$, hence $A = \frac{1}{2}\mathbb{E}[g_t g_t^\top]$.
- Step 5: $\mathbb{E}(z_t^2 - 1)^2 = \kappa - 2 + 1 = \kappa - 1$, hence $B = \frac{1}{4}(\kappa - 1)\mathbb{E}[g_t g_t^\top] = \frac{\kappa - 1}{2} A$.

A3 [Solved]: Solution (2/2)

- Step 6: $A^{-1}BA^{-1} = \frac{\kappa-1}{2}A^{-1}AA^{-1} = \frac{\kappa-1}{2}A^{-1}$; it equals the Hessian-only matrix A^{-1} only if $\kappa = 3$: the kurtosis of the Normal distribution, but also of other distributions; the equality concerns the block of the variance parameters in the Gaussian QMLE.
- (c) Step 7: $\kappa = 3(\hat{\nu}-2)/(\hat{\nu}-4) = 3 \cdot 3.011/1.011 = 8.93$.
- Step 8: variances $\times\,(\kappa-1)/2 = 3.97$; standard errors $\times\sqrt{3.97} = 1.99$.
- Interpretation: with BET-like tails the Hessian-only standard errors are about half the correct ones, so their t-statistics are about twice too large; report the sandwich Bollerslev & Wooldridge (1992); the factor applies to the Gaussian QMLE, not to the Hessian of the Student-t estimator of the case study.
- With $\nu \leq 4$, $\kappa = \infty$, B does not exist and the QMLE is no longer $\sqrt{T}$-Normal: the usual intervals lose their justification.

A4 [Proposed]: The Delta Method for the Half-Life — Task

Solution discussed in class

- **Question**: Can a delta-method interval describe the uncertainty of a volatility half-life?
- **Data and assumptions**
  - Half-life of a volatility shock $h(p) = \ln 0.5 / \ln p$, where $p = \alpha + \beta$ is the persistence of a GARCH(1,1).
  - BET estimates of B2 with robust covariance: $\hat{p} = 0.9907$, $\mathrm{se}(\hat{p}) = 0.0084$; assume for this delta-method calculation that $\hat{p}$ is asymptotically Normal (A3); assess this approximation near the boundary; level 95%.
- **Tasks**
  - (a) Derive $h'(p)$ and write the delta-method standard error $\mathrm{se}(\hat{h}) = |h'(\hat{p})| \, \mathrm{se}(\hat{p})$.
  - (b) Compute $\hat{h}$, $|h'(\hat{p})|$, $\mathrm{se}(\hat{h})$ and the Wald interval $\hat{h} \pm 1.96 \, \mathrm{se}(\hat{h})$.
  - (c) Using $\ln p \approx -(1-p)$, show that $|h'(p)| \approx \ln 2/(1-p)^2$ as $p \to 1$; evaluate the exact and the approximate slope at $p = 0.99$, 0.995 and 0.999.
- **Report**
  - $\hat{h}$ with its standard error and Wald interval; a table of $|h'(p)|$, exact and approximate.
- **Interpretation**
  - Why is a Normal interval for the half-life meaningless here?
  - What should you report instead?
- **Model**: A3

Back Exercise list

A5 [Solved]: GARCH(1,1) Forecast for BET — Task

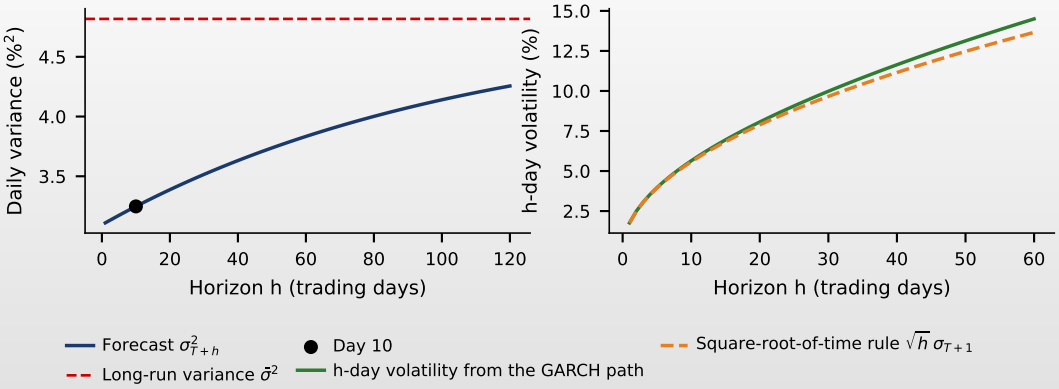
- **Questions**
  - How much risk does the BET carry over the next 10 days according to GARCH?
  - Does the square-root-of-time rule get it right?
- **Data and assumptions**
  - BET daily log returns in % (official closes, course data), 6 Jan 2000 – 18 Sep 2026; GARCH(1,1)-t on the full sample: $\omega = 0.0448$, $\alpha = 0.1916$, $\beta = 0.7991$.
  - Tomorrow's variance, the one-step forecast $\sigma^2_{T+1} = \omega + \alpha\varepsilon^2_T + \beta\sigma^2_T = 3.11$ ($\sigma_{T+1} = 1.76\%$); forecasts σ^2_{T+h} and long-run variance $\bar{\sigma}^2$ as in the slide "Robust Errors and GARCH Forecasts"; 252 trading days a year.
- **Tasks**
  - (a) Compute the persistence $\alpha + \beta$, the long-run variance $\bar{\sigma}^2$, the daily volatility $\bar{\sigma}$ and the annualised volatility $\sqrt{252\,\bar{\sigma}^2}$.
  - (b) Compute the half-life of a shock (the slide "An Interval at a Boundary").
  - (c) Forecast the variance on day 10, σ^2_{T+10}.
  - (d) Compute the variance of the 10-day return $\sum_{h=1}^{10} \sigma^2_{T+h}$ with the closed form of the geometric sum, the 10-day volatility, and compare it with the square-root-of-time rule $\sqrt{10}\,\sigma_{T+1}$.
- **Report**
  - Every number of (a)–(d), each with the substituted formula.
- **Interpretation**
  - Why does the forecast move towards the long-run level?
  - When does the square-root-of-time rule understate or overstate the 10-day risk?

Back Exercise list

A5 [Solved]: Solution (1/3)

- (a) Step 1: $\alpha + \beta = 0.9907$, so $1 - \alpha - \beta = 0.0093$.
- Step 2: $\bar{\sigma}^2 = 0.0448/0.0093 = 4.82$, i.e. $\bar{\sigma} = 2.20\%$ a day.
- Step 3: $\sqrt{252 \cdot 4.82} = \sqrt{1\,214.6} = 34.9\%$ a year.
- (b) Step 4: $h_{1/2} = \ln 0.5 / \ln 0.9907 = 74.2$ trading days, about 3.5 months of 21 trading days.
- Reading: a shock to variance loses half of its effect in about 74 trading days, so BET volatility is very persistent.
- Reading: $\bar{\sigma}^2$ is imprecise because the denominator $1 - \alpha - \beta = 0.0093$ is small; the profile interval of B2 reaches $p = 1$, where $\bar{\sigma}^2$ is infinite.

A5 [Solved]: Solution (2/3)



MFM_ch19_seminar

A5 [Solved]: Solution (3/3)

- (c) Step 5: gap to the long run $\sigma^2_{T+1} - \bar{\sigma}^2 = 3.11 - 4.82 = -1.71$; $(\alpha+\beta)^9 = 0.919$.
- Step 6: $\sigma^2_{T+10} = 4.82 + 0.919 \cdot (-1.71) = 3.25$: today is below the long run, so the forecast rises.
- (d) Step 7: $\sum_{h=1}^{10} \sigma^2_{T+h} = 10\bar{\sigma}^2 + (\sigma^2_{T+1} - \bar{\sigma}^2)\frac{1-(\alpha+\beta)^{10}}{1-\alpha-\beta} = 10 \cdot 4.82 + (-1.71) \cdot 9.592 = 31.80$.
- Step 8: 10-day volatility $\sqrt{31.80} = 5.64\%$, against $\sqrt{10} \cdot 1.76 = 5.57\%$ by the square-root-of-time rule.
- Interpretation: variance reverts to $\bar{\sigma}^2$ at rate $\alpha+\beta$; the square-root-of-time rule understates the 10-day risk when volatility is below its long-run level (as now) and overstates it after a crisis spike.

MFM_ch19_seminar

Is the Asymmetric Response Worth an Extra Parameter?

- Glosten–Jagannathan–Runkle (GJR) GARCH adds a response specific to negative shocks
  - its extra coefficient is zero under the symmetric GARCH model
  - both models use the same data
- Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) balance the fitted likelihood against the parameter count
  - lower values are preferred, comparing models with the same dependent variable and sample
  - for one extra parameter, AIC adds a penalty of 2 and BIC one of $\ln T$
  - with a large sample, the same likelihood improvement can satisfy AIC but not BIC
- The LR test asks whether the extra response is supported by the data
- The information criteria choose between specifications
- For risk forecasting, neither replaces the out-of-sample VaR backtest

B5 [Proposed]: Asymmetry: GJR-GARCH-t for BET — Task

Solution discussed in class

- **Question**: Do negative shocks raise BET volatility more than positive ones?
- **Data and assumptions**
  - BET returns as in B2; the GJR GARCH(1,1)-t (previous slide): $\sigma_t^2 = \omega + (\alpha + \gamma\, 1\{\varepsilon_{t-1} < 0\})\varepsilon_{t-1}^2 + \beta\sigma_{t-1}^2$; the GARCH(1,1)-$t$ of B2 is the restricted model $\gamma = 0$.
  - $T = 6\,670$; $\mathrm{AIC} = -2\ell + 2k$, $\mathrm{BIC} = -2\ell + k \ln T$ (ℓ: maximised log-likelihood, k = number of parameters; lower is better); level 5%.
- **Tasks**
  - (a) Estimate the GJR model and report $\hat{\gamma}$, its robust standard error and $t = \hat{\gamma}/\mathrm{se}(\hat{\gamma})$.
  - (b) Compute $LR = 2(\ell_{\mathrm{GJR}} - \ell_{\mathrm{GARCH}})$ and its p-value from $\chi^2(1)$.
  - (c) Compute AIC and BIC of both models and show that $\Delta\mathrm{AIC} = 2 - LR$ and $\Delta\mathrm{BIC} = \ln T - LR$ (GJR minus GARCH).
  - (d) Compute $(\hat{\alpha} + \hat{\gamma})/\hat{\alpha}$, the impact of a negative shock relative to a positive one.
- **Report**
  - $\hat{\gamma}$ with standard error and t; LR and p; both criteria for both models; the ratio of (d).
- **Interpretation**
  - Do the tests and the criteria agree?
  - Which model would you use for VaR?
  - How would you decide?
- **Model**: B2

Back Exercise list

A7 [Solved]: Normal VaR 1% and ES 2.5% — Task

- **Question**: How large are VaR 1% and ES 2.5% of a BET position under the Normal distribution?
- **Data and assumptions**
  - A position of 1 million RON (Romanian leu) in the BET; one-day log return $r \sim N(\mu, \sigma^2)$ with $\mu = 0.05\%$, $\sigma = 1.4\%$; loss $L = -r$; amounts in RON use the linear approximation $10^6 \cdot L/100$ (exact loss: $10^6\,[1 - e^{r/100}]$).
  - Definitions and Normal formulas of VaR and ES: the slide "VaR and ES under the Normal and the Student-t Distribution".
- **Tasks**
  - (a) Find $z_{0.01}$ and compute $\mathrm{VaR}_{1\%}$ in % and in RON.
  - (b) Find $z_{0.025}$ and $\varphi(z_{0.025})$ and compute $\mathrm{ES}_{2.5\%}$ in % and in RON.
  - (c) Compute the ratio $\mathrm{ES}_{2.5\%}/\mathrm{VaR}_{1\%}$.
- **Report**
  - Both measures in % and in RON, and their ratio.
- **Interpretation**
  - Why did the Basel Committee replace VaR 1% by ES 2.5%, if the two numbers are almost equal under the Normal distribution?

Back Exercise list

A7 [Solved]: Solution

- (a) Step 1: $z_{0.01} = -2.326$, so $\sigma z_{0.01} = 1.4 \cdot (-2.326) = -3.257$.
- Step 2: $\mathrm{VaR}_{1\%} = -(0.05 - 3.257) = 3.207\%$, i.e. $3.207\% \cdot 10^6 = 32\,069$ RON.
- (b) Step 3: $z_{0.025} = -1.960$, $\varphi(z_{0.025}) = 0.0584$, $\varphi(z_{0.025})/0.025 = 2.338$.
- Step 4: $\mathrm{ES}_{2.5\%} = -0.05 + 1.4 \cdot 2.338 = -0.05 + 3.273 = 3.223\%$, i.e. 32 229 RON.
- (c) Step 5: $\mathrm{ES}_{2.5\%}/\mathrm{VaR}_{1\%} = 3.223/3.207 = 1.005$: under the Normal distribution the two numbers are almost equal.
- Interpretation: the switch is not about the level under the Normal distribution; ES looks inside the tail and is subadditive, while VaR only marks the entrance of the tail (Chapter 7); with heavy tails ES rises more (A8).

A8 [Proposed]: Student-t VaR 1% and ES 2.5% — Task

Solution discussed in class

- **Question**: How do the two measures change when the tails are heavy but the variance is the same?
- **Data and assumptions**
  - Same position, $\mu = 0.05\%$ and $\sigma = 1.4\%$ as in A7, but $r = \mu + s\,T_\nu$, T_ν Student-t with $\nu = 5$, scale $s = \sigma\sqrt{(\nu-2)/\nu}$ so that the variance is σ^2.
  - Formulas: $\mathrm{VaR}_\alpha = -(\mu + s\,q_\alpha)$ with $q_\alpha = t_\nu^{-1}(\alpha)$; $\mathrm{ES}_\alpha = -\mu + s\,\frac{f_\nu(q_\alpha)}{\alpha}\,\frac{\nu+q_\alpha^2}{\nu-1}$, with f_ν the Student-t density.
- **Tasks**
  - (a) Compute the scale s.
  - (b) Find $t_5^{-1}(0.01)$ and compute $\mathrm{VaR}_{1\%}$ in % and in RON.
  - (c) Find $q = t_5^{-1}(0.025)$, compute the factors $f_5(q)/0.025$ and $(5+q^2)/4$, then $\mathrm{ES}_{2.5\%}$ in % and in RON.
  - (d) Compute the ratios Student-t/Normal for VaR and for ES (Normal values from A7), and the ratio ES/VaR.
- **Report**
  - Both measures in % and RON, and the three ratios.
- **Interpretation**
  - Which number changes more when the tails become heavy, VaR or ES?
  - What does this mean for capital?
- **Model**: A7

Back Exercise list

A9 [Proposed]: The Power of the Christoffersen Test — Task

Solution discussed in class

- **Question**: Can one year of breaches reveal that they cluster?
- **Data and assumptions**
  - Breaches I_t follow a two-state Markov chain with the correct unconditional rate $\pi = 1\%$ but clustered: $\pi_{11} = P(I_t = 1 \mid I_{t-1} = 1) = 0.20$ (independence would mean $\pi_{11} = 1\%$).
  - Christoffersen independence test $LR_{ind} \sim \chi^2(1)$, rejection at 5% if $LR_{ind} > 3.84$ Christoffersen (1998); a path without any breach counts as a non-rejection.
- **Tasks**
  - (a) Derive $\pi_{01} = P(I_t = 1 \mid I_{t-1} = 0)$ from the stationarity condition $\pi = \pi_{01}/(1 - \pi_{11} + \pi_{01})$.
  - (b) Simulate 20 000 paths of $T = 250$ days (first day a breach with probability π), compute LR_{ind} on each path and estimate the power as the share of rejections, with its Monte Carlo standard error.
  - (c) Repeat (b) for $T = 1\,000$ days with 5 000 paths.
  - (d) Compute the expected number of consecutive breach pairs $(T-1)\pi\pi_{11}$ for both T.
- **Report**
  - π_{01}, both powers with the Monte Carlo standard error, the expected numbers of pairs.
- **Interpretation**
  - Why is independence almost untestable with one year of data?
- **Model**: A1

Back Exercise list

B6 [Proposed]: Backtesting VaR and ES for Bitcoin, 2019–2026 — Task

Solution discussed in class

- **Question**: Does the ranking of B3 hold for Bitcoin?
- **Data and assumptions**
  - Bitcoin daily losses (close in USD, course data), evaluation 1 Jan 2019 – 18 Sep 2026, $T = 2\,818$ days, every calendar day; information through $t - 1$ only.
  - The four models, windows and tests of B3; Newey–West variance with 8 lags; level 5% (10% for the model confidence set).
- **Tasks**
  - (a) Count the breaches of VaR 1% and compute the rate, p_{uc} and p_{ind} for each model.
  - (b) Run the DQ test with 4 lagged hits and VaR_t.
  - (c) Compute the FZ0 scores, the Diebold–Mariano tests and the model confidence set at 10%.
  - (d) Run the comparative ES backtest against HS.
  - (e) Plot the daily losses and the HS and FHS VaR 1% from 1 January 2022.
- **Report**
  - The two tables of B3 for Bitcoin and the chart of (e).
- **Interpretation**
  - Is the ranking of the models the same as for BET?
  - Can the tests separate the models at all?
- **Model**: B3

Back Exercise list

A10 [Proposed]: Hurdles for One Hundred Factors — Task

Solution discussed in class

- **Question**: How high must the $|t|$ hurdle be when one hundred factors are tested?
- **Data and assumptions**
  - 100 candidate factors, each tested with a two-sided t-statistic that is approximately $N(0,1)$ under H_0; target level 5%.
  - Bonferroni and BH as in A2; under arbitrary dependence the Benjamini–Yekutieli procedure replaces $0.05k/m$ by $0.05k/[m\,c(m)]$, $c(m) = \sum_{i=1}^{m} 1/i$ Benjamini & Yekutieli (2001).
- **Tasks**
  - (a) Compute the Bonferroni per-test level at FWER 5% and the hurdle $|t| > z_{1-0.05/(2m)}$.
  - (b) Compute the BH hurdle for $|t|$ when exactly 10 factors are declared and when only 1 is declared.
  - (c) Compute $c(100)$ and the Benjamini–Yekutieli hurdle when 10 factors are declared.
  - (d) Compute the hurdles with $|t| > 3$ proposed by Harvey, Liu & Zhu (2016).
- **Report**
  - The per-test levels and the four $|t|$ hurdles.
- **Interpretation**
  - Why does the BH hurdle depend on how many factors are declared?
  - Is BH still valid when factors are positively correlated?
- **Model**: A2

Back Exercise list

A11 [Solved]: Why Kurtosis Fails with a Tail Index below 4 — Task

- **Question**: What does the sample kurtosis estimate when the tail index is below 4?
- **Data and assumptions**
  - $|X|$ has a Pareto-type tail: $P(|X| > x) = Cx^{-\alpha}$ for $x \geq x_0 > 0$, $C > 0$; α is the tail index.
  - Simulation design: Student-t with 3 degrees of freedom (tail index 3, finite variance), sample sizes $T = 10^3$, 10^4, 10^5, 200 samples of each size.
- **Tasks**
  - (a) Use $\mathbb{E}|X|^4 = \int_0^\infty 4x^3 P(|X| > x)\, dx$ to show that $\mathbb{E}|X|^4 = \infty$ when $\alpha \leq 4$.
  - (b) Simulate the samples, compute the sample excess kurtosis of each, and report its median and 90% quantile for each T.
  - (c) Explain what a bootstrap interval for the kurtosis estimates in this case Athreya (1987).
- **Report**
  - The proof of (a) and a table of the median and 90% quantile for the three sample sizes.
- **Interpretation**
  - What should you report instead of the kurtosis when the tail index is below 4?

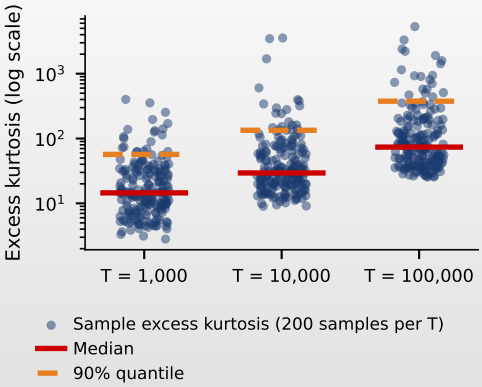
Back Exercise list

A11 [Solved]: Solution (1/2)

- (a) Step 1: for $x \geq x_0$ the integrand is $4x^3 \cdot Cx^{-\alpha} = 4Cx^{3-\alpha}$, and it is non-negative everywhere.
- Step 2: $\mathbb{E}|X|^4 \geq 4C \int_{x_0}^{\infty} x^{3-\alpha}\, dx$, and this integral diverges when $3 - \alpha \geq -1$, i.e. $\alpha \leq 4$.
- Step 3: for $2 < \alpha \leq 4$ the variance is finite but $\mathbb{E}X^4 = \infty$, so the population kurtosis $\mathbb{E}X^4/(\mathbb{E}X^2)^2$ is infinite and the sample kurtosis has no finite limit.
- (b) Steps: draw 200 samples of each size, compute the sample excess kurtosis, take the median and the 90% quantile across the 200 values.

T	10^3	10^4	10^5
Median excess kurtosis	14.5	29.4	73.7
90% quantile	56.9	134.0	376.4

A11 [Solved]: Solution (2/2)



- Reading: the median grows with T and the 90% quantile grows faster: the statistic does not settle down, it grows without limit.
- (c) There is no population value for an interval to cover, and the bootstrap of such moments is inconsistent Athreya (1987): a bootstrap interval for the kurtosis has no coverage guarantee.
- Interpretation: when the tail index may be below 4, report a tail index with its interval (Hill with block bootstrap, B1) and not the kurtosis; a large sample kurtosis says only that the tails are heavy, not how heavy.

MFM_ch19_seminar

B7 [Proposed]: Stylised Facts of Bitcoin — Task

Solution discussed in class

- **Question**: Do the stylised facts of B1 hold for Bitcoin?
- **Data and assumptions**
  - Bitcoin in USD (daily close, course data), every calendar day, 18 Sep 2014 – 18 Sep 2026: $T = 4\,384$ log returns in %.
  - Same settings as B1: $k = 5\%$ of the loss days, 999 block-bootstrap resamples with 20-day blocks, 10 lags, level 5%; BET results from B1.
- **Tasks**
  - (a) Compute the standard deviation, skewness and excess kurtosis.
  - (b) Estimate the Hill tail index with the asymptotic and the block-bootstrap 95% CI.
  - (c) Compute $Q(10)$ on r_t and r_t^2 and the robust $Q^*(10)$, and decide at 5%.
  - (d) Count the days beyond ± 4 standard deviations and compare with the Normal distribution.
  - (e) Test $H_0 : \alpha_{\mathrm{BTC}} = \alpha_{\mathrm{BET}}$ with $z = (\hat{\alpha}_{\mathrm{BTC}} - \hat{\alpha}_{\mathrm{BET}}) / \sqrt{\mathrm{se}_{\mathrm{BTC}}^2 + \mathrm{se}_{\mathrm{BET}}^2}$ using the block-bootstrap standard errors and assuming independent estimates (the test is detailed in A12).
- **Report**
  - A table Bitcoin vs BET with (a)–(d); the test of (e) with z, p and decision.
- **Interpretation**
  - Are Bitcoin's tails heavier than the BET's, or is Bitcoin only more volatile?
  - Does the linear autocorrelation of Bitcoin survive the robust test?
- **Model**: B1

Back Exercise list

A12 [Proposed]: Comparing Two Tail Indices — Task

Solution discussed in class

- **Question**: Are the tails of Bitcoin heavier than those of the BET?
- **Data and assumptions**
  - Hill estimator $\hat{\alpha}$ of the slide "The Tail Index and the Hill Estimator", $k = 5\%$ of the loss days; $\mathrm{se}(\hat{\alpha}) \approx \hat{\alpha}/\sqrt{k}$ for i.i.d. data Hill (1975).
  - Supplied estimates from B1 and B7 (no prior run needed): BET $\hat{\alpha} = 2.39$ ($k = 153$), Bitcoin $\hat{\alpha} = 2.74$ ($k = 104$); block-bootstrap standard errors (999 resamples, 20-day blocks): BET 0.233, Bitcoin 0.227; simplifying assumption: the two estimates are independent (the samples overlap in time, so common shocks would add $-2\,\mathrm{Cov}(\hat{\alpha}_{\mathrm{BTC}}, \hat{\alpha}_{\mathrm{BET}})$ to the variance of the difference); level 5%.
- **Tasks**
  - (a) Compute the i.i.d. standard errors of both estimates.
  - (b) Test $H_0 : \alpha_{\mathrm{BTC}} = \alpha_{\mathrm{BET}}$ with $z = (\hat{\alpha}_{\mathrm{BTC}} - \hat{\alpha}_{\mathrm{BET}})/\sqrt{\mathrm{se}^2_{\mathrm{BTC}} + \mathrm{se}^2_{\mathrm{BET}}}$, two-sided, with the i.i.d. standard errors.
  - (c) Repeat the test with the block-bootstrap standard errors.
- **Report**
  - All standard errors, both z-statistics, p-values and decisions.
- **Interpretation**
  - Why can the i.i.d. formula misstate the uncertainty for returns?
  - Does the conclusion depend on it?
- **Model**: A11, B1

Back Exercise list

B8 [Proposed]: Aligning the Trading Calendars of Bitcoin and SPY — Task

Solution discussed in class

- **Question**: What does aligning returns instead of prices do to Bitcoin–SPY statistics?
- **Data and assumptions**
  - Bitcoin (daily close in USD, every calendar day) and SPY, the S&P 500 ETF (close adjusted for dividends and splits, weekdays only), from the course data; daily log returns in %.
  - Two periods: 1 January 2015 – 31 December 2019 (1 258 common days) and 1 January 2020 – 18 September 2026 (1 687 common days).
- **Tasks**
  - (a) Wrong alignment: compute the log returns of each asset on its own calendar, then keep the common days.
  - (b) Right alignment: keep the common days of the prices, then compute the log returns (a Monday return then covers Friday to Monday for both assets).
  - (c) For each period and each alignment compute the Bitcoin–SPY correlation, the standard deviation of Bitcoin returns on all days and on Mondays, and the cumulative Bitcoin return in %, $100\,[\exp(\sum_t r_t/100) - 1]$ (with r_t in %).
  - (d) Plot the 250-day rolling correlation from 2016 for both alignments.
- **Report**
  - A table (statistics × periods × alignments) and the chart of (d).
- **Interpretation**
  - Which statistic is most damaged by the wrong alignment?
  - Would the VaR of a long-only Bitcoin–SPY portfolio, from a variance–covariance model built on it, be too low or too high?
- **Model**: B4

Back Exercise list

C3: Reproducing the Standard-Error Convention

- Data: the close fields of the files BET.csv and STOXX50E.INDX.csv (STOXX50E: the symbol of the Euro Stoxx 50 index) in the course market data, initially for 2002–2018
- The paper uses simple conventional standard errors; first reproduce that convention
- Trend: multiply the ordinary OLS slope standard error (SE) by $100/\bar{m}$ ($\bar{m}$: the mean of the annual measure m_t of C3), treating the denominator as fixed
- Log and relative changes: the sample standard deviation of the annual changes divided by the square root of their count
- Keep only changes between consecutive observed calendar years
- These SEs ignore the dependence induced by differencing and the uncertainty of the trend denominator
  - they are the paper's convention, not a claim of universally valid inference
- Separate sensitivity analysis: HAC or block-bootstrap errors; identify the decisions that change
- Keep this table separate from the one that reproduces the paper's convention.

C3 [Proposed]: Nonstandard Errors: A Mini-Multiverse — Task (1/2)

Solution discussed in class

- **Question**: How much do the analysis choices alone move a market-efficiency trend?
- **Data and assumptions**
  - Context: in Menkveld et al. (2024) 164 teams tested the same hypotheses on the same data; the interquartile range (IQR) of their yearly market-efficiency trend estimates (the nonstandard error) was 6.7 percentage points, against a median SE of 2.5% (Table I, Panel C).
  - Data: daily closes of the Euro Stoxx 50 index and of the BET (course data), 2002–2018, then up to the last full year; H_0: "market efficiency has not changed over time" (Sec. I.A).
- **Tasks**
  - (a) Before any run, commit a plan with the forks of Table V that daily data allow: outliers (none; annual m_t winsorised or trimmed at its 2.5 and 97.5 percentiles, trimmed years left missing); measure, within each year (Variance Ratio, VR: $|VR(q) - 1|$, $VR(q) = \mathrm{Var}(\sum_{i=0}^{q-1} r_{t-i})/(q\mathrm{Var}(r_t))$, overlapping sums with the Lo–MacKinlay corrections of Chapter 2, $q = 5, 21$; R^2 of an autoregressive AR(1) model with intercept); model (linear trend; log difference; relative difference).
  - (b) For each index and each of the 27 paths (3 measures $\times$ 3 outlier rules $\times$ 3 models), compute the annual measure m_t and its average yearly % change with an SE: $100\, b/\bar{m}$ from the regression of the measure on a linear trend, the mean of $100\,\Delta \ln m_t$, or the mean of $100\,(m_t/m_{t-1} - 1)$. Nulls: for a trendless stationary m_t the slope and mean log change are centred at zero, the mean relative change is not.
  - (c) Report the median, the IQR (the nonstandard error of your multiverse) and the 10–90 percentile range of the 27 estimates; compare the IQR with $1.35\times$ the median SE.

Back Exercise list

C3 [Proposed]: Nonstandard Errors: A Mini-Multiverse — Task (2/2)

Solution discussed in class

- **Tasks (continued)**
  - (a, continued) The plan also fixes: the rule for zero or missing m_t; the Holm family (the 18 paths with a linear trend or a log difference, per index and period); the last complete year at your data cutoff.
  - (d) Rank the forks by the k-sample Anderson–Darling statistic (`scipy.stats.anderson_ksamp`) of the estimates grouped by the alternative at each fork (Sec. II.C).
  - (e) Test the trend against zero at 5% for the 18 paths centred at zero under H_0, with the Holm correction over these 18 tests and standard errors that respect the dependence (HAC, block bootstrap); report how many paths reject with each sign (positive trend = declining efficiency).
  - (e, continued) The 9 mean relative changes are not centred at zero under H_0: report them descriptively, without a test against zero.
- **Report**
  - A table of the 27 estimates with their SEs for each index; the median, IQR and 10–90 range; the Anderson–Darling ranking of the forks; the number of Holm rejections by sign; a box plot of the estimates grouped by the most influential fork.
  - Model: A2 (Holm and BH over a family of tests); the lecture case study
- **Interpretation**
  - Is the IQR over 27 paths an estimate of the nonstandard error of 164 human teams? Use footnote 32 of the paper in your answer.
  - If the sign of the trend changes with the model fork, what should your project report?
  - Which rule would you have written into the pre-analysis plan?

References (1/3)

- 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.
- Benjamini, Y., Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B*, 57(1), 289–300.
- Benjamini, Y., Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, 29(4), 1165–1188.
- Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3), 307–327.
- 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.
- Christoffersen, P.F. (1998). Evaluating interval forecasts. *International Economic Review*, 39(4), 841–862.
- Davis, R.A., Mikosch, T. (1998). The sample autocorrelations of heavy-tailed processes with applications to ARCH. *Annals of Statistics*, 26(5), 2049–2080.
- Diebold, F.X., Mariano, R.S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3), 253–263.
- Engle, R.F., Manganelli, S. (2004). CAViaR: conditional autoregressive value at risk by regression quantiles. *Journal of Business & Economic Statistics*, 22(4), 367–381.

References (2/3)

- Franke, J., Härdle, W.K., Hafner, C.M. (2019). Statistics of Financial Markets: An Introduction, 5th ed. Springer.
- Hansen, P.R., Lunde, A., Nason, J.M. (2011). The model confidence set. *Econometrica*, 79(2), 453–497.
- 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.
- Künsch, H.R. (1989). The jackknife and the bootstrap for general stationary observations. *Annals of Statistics*, 17(3), 1217–1241.
- Kupiec, P.H. (1995). Techniques for verifying the accuracy of risk measurement models. *Journal of Derivatives*, 3(2), 73–84.
- 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.
- 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.
- 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.

References (3/3)

- Nelson, D.B. (1990). Stationarity and persistence in the GARCH(1,1) model. *Econometric Theory*, 6(3), 318–334.
- Nolde, N., Ziegel, J.F. (2017). Elicitability and backtesting: perspectives for banking regulation. *Annals of Applied Statistics*, 11(4), 1833–1874.
- 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., Wesselhöfft, N., Härdle, W.K., Kolossiatis, M., Yatracos, Y.G. (2023). Are cryptos becoming alternative assets?. *European Journal of Finance*, 29(10), 1064–1105.