

Modelling Financial Markets

Seminar 8: Backtesting and Evaluating Risk Forecasts — 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

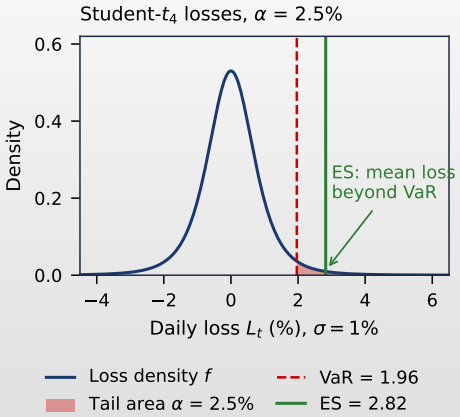
Which Risk Forecast Is Supported by the Evidence?

- Question of the seminar
  - **which forecasting procedure would you provisionally choose for Value-at-Risk (VaR) 1% and Expected Shortfall (ES) 2.5%, and how strong is the evidence?**
- Deliverables
  - your reconstruction of the solved A calculations and answers to the independent checks
  - one reproduced table and figure per assigned B exercise, with your own interpretation
  - the corrected C1 code
  - one interpretation sentence per exercise
- Notebook for today: open the seminar notebook in Colab
- B3 and B4 use the residual bootstrap and conformal prediction, developed in Lecture 8: solve them after the lecture; B5 is homework
- Seminar: Prof. Daniel Traian Pele, danpele@ase.ro; Antoaneta Amza, PhD candidate, antoaneta.amza@csie.ase.ro

How Are Observations and Forecasts Aligned?

- BET means Bucharest Exchange Trading index; BVB is Bursa de Valori București. BNR is Banca Națională a României; the EUR/RON rate is the number of Romanian lei (RON) per euro (EUR). US assets are quoted in United States dollars (USD).
- UTC means Coordinated Universal Time. Bitcoin trades every day; an equity-market trading day is a different calendar unit. Compare forecasts only on the stated evaluation dates.
- The Probability Integral Transform (PIT) places the observation inside its forecast distribution. If forecasts are correct and continuous, its values should behave like independent and identically distributed (i.i.d.) Uniform draws.
- Losses in code are decimals: a price decline from 100 to 95 gives log loss about 0.0513, or 5.13%. Course charts express the same loss in percent.

VaR and ES of a Loss



- Details: Chapter 8
- **Loss** $L_t = -r_t$; $\mathcal{F}_{t-1}$: what is known at the close of day $t-1$
- **VaR** at tail probability α: the loss exceeded with probability α

$$P\left(L_t > \mathrm{VaR}_t \mid \mathcal{F}_{t-1}\right) = \alpha$$

  - VaR_t is the $(1-\alpha)$-quantile of the forecast loss distribution
- **ES**: the mean loss on the days beyond VaR

$$\mathrm{ES}_t = E\left[L_t \mid L_t > \mathrm{VaR}_t,\ \mathcal{F}_{t-1}\right]$$

  - always $\mathrm{ES}_t \geq \mathrm{VaR}_t$; the gap grows with the tail weight
  - chart: Student-t_4 with $\sigma = 1\%$, $\alpha = 2.5\%$: VaR = 1.96%, ES = 2.82%

Breaches and the Null Hypothesis

- A backtest compares out-of-sample forecasts, issued at $t-1$, with the realised losses
- **Breach** (hit) on day t

$$I_t = 1\{L_t > \mathrm{VaR}_t\} = \begin{cases} 1, & L_t > \mathrm{VaR}_t \\ 0, & \text{otherwise} \end{cases}$$

  - $1\{\cdot\}$: indicator function, 1 when the event in braces occurs
  - $x = \sum_{t=1}^{T} I_t$: number of breaches in T evaluation days; $\hat{\pi} = x/T$: observed breach rate
- **Null hypothesis** of every VaR backtest: a correct forecast at level α

$$H_0: \; P(I_t = 1 \mid \mathcal{F}_{t-1}) = \alpha \quad \text{for every } t$$

  - it implies: the I_t are i.i.d. Bernoulli(α), independent 0/1 variables equal to 1 with probability α; the converse fails when $\mathcal{F}_{t-1}$ holds other information
  - the tests below check only some consequences: the right *number* of breaches (unconditional coverage)
  - and no *memory*: a breach today carries no information about tomorrow (independence)

What Distinguishes the Forecasting Methods?

- Historical Simulation (HS) uses the empirical loss distribution. Filtered Historical Simulation (FHS) first removes changing volatility, then rescales the residual quantile for tomorrow.
- Generalised Autoregressive Conditional Heteroskedasticity (GARCH) models conditional variance. Student-t innovations permit heavier tails than Normal innovations.
- Extreme Value Theory (EVT) models the far tail. GARCH-EVT (a GARCH filter plus EVT for the tail) fits a Generalised Pareto Distribution (GPD) to the filtered losses above a threshold.
- For example, HS retains a crisis observation until it leaves the window. FHS changes its impact as forecast volatility changes, even while the historical window is unchanged.

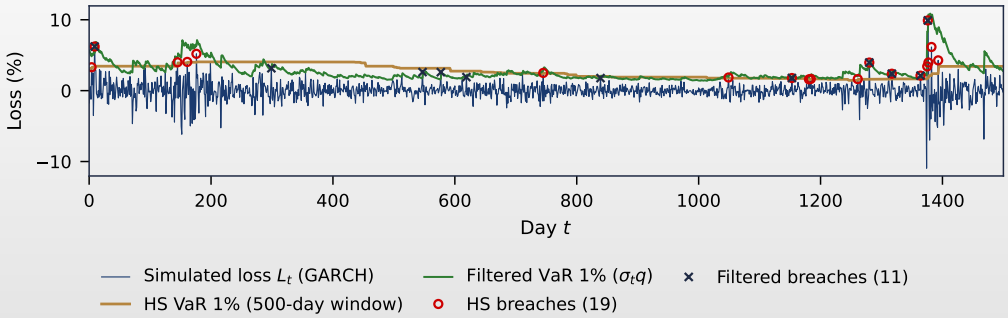
The Six Forecasting Procedures

- Panel of six VaR/ES procedures (Chapter 8); W: rolling window of 1 000 days. Parameter fields and refit calendar
- **Unconditional**: one distribution for the whole window
  - HS: $\mathrm{VaR}_t = \hat{q}_{\mathbf{1}-\alpha}(L_{t-W}, \ldots, L_{t-\mathbf{1}})$, the empirical $(1-\alpha)$-quantile of the window's losses
  - Normal: $\mathrm{VaR}_t = -\hat{\mu} + \hat{\sigma}\, z_{\mathbf{1}-\alpha}$; $\hat{\mu}$, $\hat{\sigma}$: window mean and standard deviation of r; $z_{\mathbf{1}-\alpha} = \Phi^{-\mathbf{1}}(1-\alpha)$
  - Student-t: fit location m, scale s and ν by maximum likelihood in each window; VaR is $-m + s\, t_\nu^{-\mathbf{1}}(1-\alpha)$. Equivalently, for $\nu > 2$, use $\sigma = s\sqrt{\nu/(\nu-2)}$ and the quantile of a unit-variance t_ν, $t_\nu^{-\mathbf{1}}(1-\alpha)\sqrt{(\nu-2)/\nu}$
- **Filtered**: a GARCH(1,1) volatility first, then a distribution for $z_t = (r_t - \mu)/\sigma_t$

$$\sigma_t^{\mathbf{2}} = \omega + a\,(r_{t-\mathbf{1}} - \mu)^{\mathbf{2}} + b\,\sigma_{t-\mathbf{1}}^{\mathbf{2}}, \qquad \mathrm{VaR}_t = -\mu + \sigma_t\, q_{\mathbf{1}-\alpha}(-z)$$

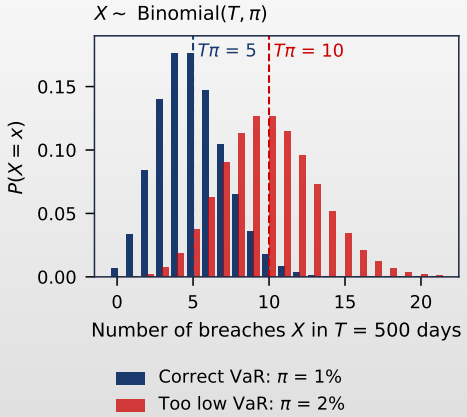
  - $\omega > 0$, $a, b \geq 0$, $a + b < 1$; $q_{\mathbf{1}-\alpha}(-z)$: the $(1-\alpha)$-quantile of the standardised loss $-z_t$
  - GARCH-t: z_t unit-variance Student-t_ν; FHS: the empirical quantile of the fitted $\hat{z}_t$
  - GARCH-EVT: a GPD tail for $-\hat{z}_t$ above a threshold u: $q = u + \frac{\beta}{\xi}\left[\left(\frac{n}{N_u}\alpha\right)^{-\xi} - 1\right]$
  - ξ: tail shape, β: scale, N_u: number of the n residuals above u

Breaches of an Unconditional and a Filtered VaR



- Simulated GARCH(1,1) losses with Student-t_5 shocks, 1 500 days, VaR 1%: 15 breaches expected
  - HS (500-day window) moves in steps and lags the volatility: 19 breaches, bunched in the calm-to-storm transitions
  - the filtered VaR $\sigma_t q$ rises with σ_t: 11 breaches, spread over the sample

The Number of Breaches Is Binomial



- Under H_0 the count $X = \sum_t I_t$ of T i.i.d. Bernoulli(π) breaches is Binomial(T, π)

$$P(X = x) = \binom{T}{x} \pi^x (1-\pi)^{T-x}, \quad x = 0, \ldots, T$$

  - $\binom{T}{x} = \frac{T!}{x!(T-x)!}$: the number of ways to place x breaches among T days
  - mean $E[X] = T\pi$, variance $\mathrm{Var}(X) = T\pi(1-\pi)$
- **Exact** one-sided p-value for too many breaches

$$P(X \geq x) = \sum_{k=x}^{T} \binom{T}{k} \alpha^k (1-\alpha)^{T-k}$$

  - the probability, under $\pi = \alpha$, of a count at least as large as the observed one
- Chart, $T = 500$: means 5 and 10 breaches, yet the two distributions overlap
  - a count near 7 is compatible with both $\pi = 1\%$ and $\pi = 2\%$

Which Hypothesis Does Each Backtest Examine?

- Likelihood Ratio (LR) compares restricted and unrestricted likelihoods. Maximum Likelihood Estimation (MLE) chooses the unrestricted parameter estimates. Kupiec's Proportion of Failures (POF) test checks the overall breach rate.
- Conditional Coverage (CC) combines the correct breach rate with independence. Dynamic Quantile (DQ) regression asks whether known information predicts breaches; Ordinary Least Squares (OLS) estimates its coefficients.
- A model may produce ten breaches in a thousand days and still fail because all ten cluster in a crisis. A correct count and unpredictable timing are separate requirements.

The Kupiec Test

- Unconditional coverage Kupiec (1995): is the observed breach rate compatible with α?
  - $H_0 : \pi = \alpha$, with π the true breach probability; $H_1 : \pi \neq \alpha$
- Binomial log-likelihood of the count if the breach probability were π

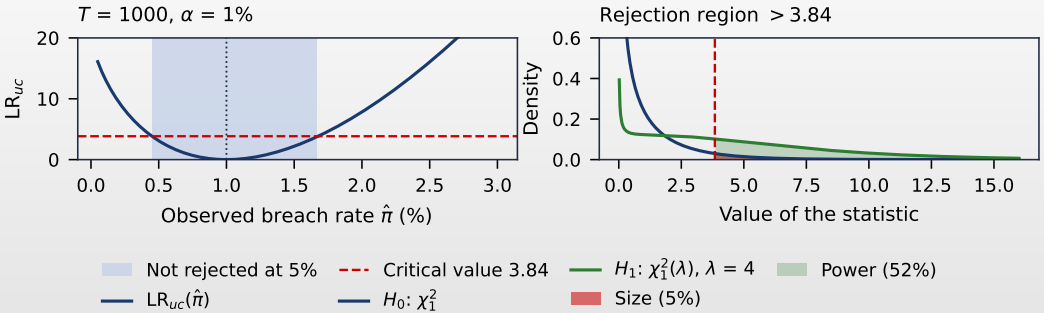
$$\ell(\pi) = (T - x)\ln(1 - \pi) + x \ln \pi$$

  - maximised at $\hat{\pi} = x/T$ (unrestricted); $\ell(\alpha)$: the value under H_0 (restricted)
- Likelihood ratio: how much better the observed rate explains the count than α

$$\mathrm{LR}_{uc} = 2\big[\ell(\hat{\pi}) - \ell(\alpha)\big] \xrightarrow{d} \chi^2_1 \text{ under } H_0$$

  - $\mathrm{LR}_{uc} \geq 0$, zero when $\hat{\pi} = \alpha$; χ^2_1: chi-square with 1 degree of freedom (one restriction)
  - reject at 5% if $\mathrm{LR}_{uc} > 3.84$, the 95% quantile of χ^2_1; p-value $= P(\chi^2_1 > \mathrm{LR}_{uc})$
  - the LR principle of Chapter 5 (GJR against GARCH), here for a Binomial count
- Example: $T = 250$, $\alpha = 1\%$: 2.5 breaches expected
  - $x = 2$ gives $\mathrm{LR}_{uc} = 0.11$: no evidence against the model
  - with so few expected breaches, the exact Binomial p-value is safer than the χ^2_1 approximation

Size and Power of a Backtest



- **Size**: $P(\text{reject} \mid H_0)$, nominally 5%; **power**: $P(\text{reject} \mid \pi \neq \alpha)$
  - left: LR_{uc} against $\hat{\pi}$ for $T = 1\,000$; not rejected for $\hat{\pi}$ between 0.46% and 1.67%
  - right: under H_1, $\mathrm{LR}_{uc} \approx \chi^2_1(\lambda)$, noncentral, with $\lambda = T(\pi - \alpha)^2/[\alpha(1-\alpha)]$; local power $P(\chi^2_1(\lambda) > 3.84)$
  - $\lambda = 4$ gives a power of 52%; with a discrete x the exact size and power come from Binomial sums

Independence: the Christoffersen Test

- Christoffersen (1998): the hits follow a two-state Markov chain (0 = no breach, 1 = breach)

$$\Pi = \begin{pmatrix} 1-\pi_{01} & \pi_{01} \\ 1-\pi_{11} & \pi_{11} \end{pmatrix}, \quad \pi_{01} = P(I_t = 1 \mid I_{t-1} = 0), \quad \pi_{11} = P(I_t = 1 \mid I_{t-1} = 1)$$

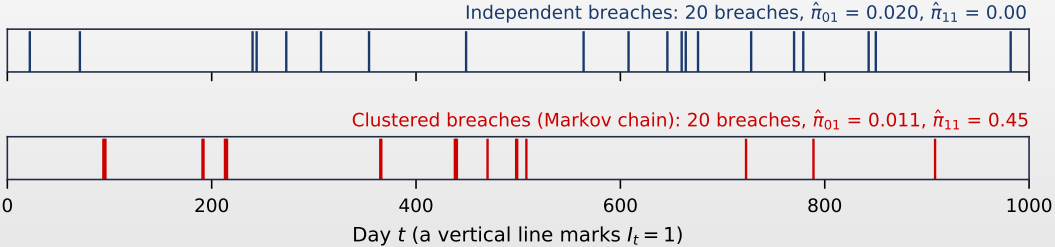
  - n_{ij}: number of days in state j after a day in state i; estimates $\hat{\pi}_{01} = \frac{n_{01}}{n_{00}+n_{01}}$, $\hat{\pi}_{11} = \frac{n_{11}}{n_{10}+n_{11}}$, $\hat{\pi} = \frac{n_{01}+n_{11}}{n}$, $n = \sum n_{ij}$

- Log-likelihood of the transitions, unrestricted

$$\ln L(\pi_{01}, \pi_{11}) = n_{00} \ln(1-\pi_{01}) + n_{01} \ln \pi_{01} + n_{10} \ln(1-\pi_{11}) + n_{11} \ln \pi_{11}$$

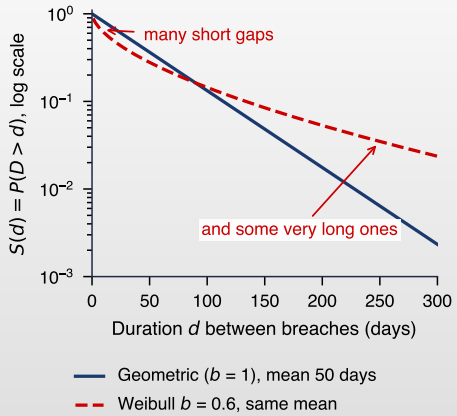
- Two likelihood ratio tests
  - $\mathrm{LR}_{ind} = 2[\ln L(\hat{\pi}_{01}, \hat{\pi}_{11}) - \ln L(\hat{\pi}, \hat{\pi})]$, $H_0 : \pi_{01} = \pi_{11}$ (no memory), χ^2_1; large when a breach makes the next one more likely
  - $\mathrm{LR}_{cc} = \mathrm{LR}_{uc} + \mathrm{LR}_{ind}$, $H_0 : \pi_{01} = \pi_{11} = \alpha$ (right rate, no memory), two restrictions, χ^2_2, 5% critical value 5.99
- Example: 12 days $0, 0, 1, 1, 0, 0, 0, 1, 0, 0, 0, 0$
  - the 11 pairs (I_{t-1}, I_t) give $n_{00} = 6$, $n_{01} = 2$, $n_{10} = 2$, $n_{11} = 1$

Independent and Clustered Breaches



- Two simulated sequences of 1 000 days, each with 20 breaches: the same count, so the same LR_{uc}
  - top: i.i.d. Bernoulli(2%); no breach follows another, $\hat{\pi}_{\mathbf{11}} = 0$
  - bottom: Markov chain with $\pi_{\mathbf{01}} = 0.012$, $\pi_{\mathbf{11}} = 0.40$; $\hat{\pi}_{\mathbf{11}} = 0.45$ against $\hat{\pi}_{\mathbf{01}} = 0.011$: breaches arrive in bursts
  - only LR_{ind} and LR_{cc} can tell the two apart

Durations between Breaches



- Duration test Christoffersen & Pelletier (2004): D_i, the number of days from one breach to the next
  - under H_0 geometric, memoryless, mean $1/\alpha$: $P(D = d) = \alpha(1-\alpha)^{d-1}$, $S(d) = P(D > d) = (1-\alpha)^d$
- Alternative: Weibull durations with $a > 0$ (the inverse scale) and shape $b > 0$

$$S(d) = e^{-(ad)^b}, \qquad h(d) = a^b\, b\, d^{\,b-1}$$

  - $h(d)$: hazard rate, a rate rather than a probability; $h(d)\Delta$ approximates the breach probability in a short interval Δ after d days without one
  - $b = 1$: constant hazard, memoryless; $b < 1$: hazard falls with d, clusters then long silences
- $\mathrm{LR}_{dur} = 2[\ln L(\hat{a}, \hat{b}) - \ln L(\tilde{a}, 1)] \approx \chi^2_1$, $H_0 : b = 1$
  - the first and last spells are censored: only $S(d)$ enters for them

The Dynamic Quantile Test

- Dynamic quantile (DQ) test Engle & Manganelli (2004): can anything known yesterday predict today's breach?
  - $\mathrm{Hit}_t = I_t - \alpha$: the centred breach, mean 0 and unpredictable under H_0
- Regression of the centred breach on information dated $t-1$

$$\mathrm{Hit}_t = \beta_0 + \sum_{k=1}^{4} \beta_k \,\mathrm{Hit}_{t-k} + \beta_5 \,\mathrm{VaR}_t + u_t$$

  - β_0: a wrong breach rate; $\beta_1, \ldots, \beta_4$: memory; β_5: breaches depending on the VaR level
  - H_0: $\beta_0 = \cdots = \beta_5 = 0$
- Wald statistic, with X the matrix of the six regressors and $\hat{\beta}$ the OLS estimates

$$\mathrm{DQ} = \frac{\hat{\beta}^\top X^\top X \,\hat{\beta}}{\alpha(1-\alpha)} \xrightarrow{d} \chi^2_6$$

  - $\alpha(1-\alpha)$: the variance of Hit_t under H_0; large DQ = predictable breaches
  - with one year of data the size of these tests is far from 5% and their power is low: B1 reads them with simulated sizes

How Does Backtesting Affect Regulatory Capital?

- The Basel Committee on Banking Supervision (BCBS) sets the framework discussed here. Fundamental Review of the Trading Book (FRTB) is the market-risk reform behind the newer capital multipliers.
- Keep the rule's window, tail probability and multiplier together. A breach count from a longer evaluation sample cannot be inserted directly into the 250-day traffic light.

The Basel Traffic Light

- Basel traffic light BCBS (1996): count the exceptions X of VaR 1% in the last 250 days

$$F(x) = P(X \leq x) = \sum_{k=0}^{x} \binom{250}{k} 0.01^k \, 0.99^{250-k}$$

  - correct model: $X \sim \text{Binomial}(250,\, 0.01)$; $F(x)$: the probability of at most x exceptions
- Zones defined by $F(x)$
  - green while $F(x) < 95\%$; red from $F(x) \geq 99.99\%$; yellow in between
  - yellow and red raise the capital multiplier: 1996, 3 plus a factor; FRTB BCBS (2019), from 1.5 to 2
- A zone is a lookup rule for a supervisor, not a 5% test
  - it fixes the probability that a correct model lands in red, not the power against a wrong one

Backtesting ES: Z_1 and Z_2

- Acerbi–Szekely Acerbi & Szekely (2014): breach losses in units of the ES forecast, with breaches of VaR at the same level α (2.5%)

$$Z_1 = 1 - \frac{1}{N_T}\sum_{t=1}^{T}\frac{L_t I_t}{\mathrm{ES}_t}, \qquad Z_2 = 1 - \frac{1}{T\alpha}\sum_{t=1}^{T}\frac{L_t I_t}{\mathrm{ES}_t}$$

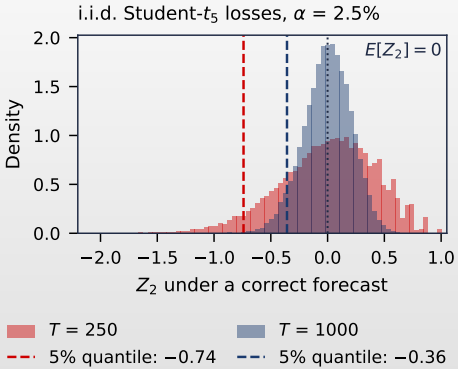
  - $N_T = \sum_t I_t$: the observed number of breaches; $T\alpha$: the expected number
  - Z_1: are the breach losses, on average, larger than ES? (size only)
  - Z_2: too many *or* too large tail losses (number and size together)

- Why $E[Z_2] = 0$ for a correct forecast (proof in A4)

$$E\left[L_t I_t \mid \mathcal{F}_{t-1}\right] = P(L_t > \mathrm{VaR}_t \mid \mathcal{F}_{t-1})\, E[L_t \mid L_t > \mathrm{VaR}_t, \mathcal{F}_{t-1}] = \alpha\,\mathrm{ES}_t$$

  - $Z_2 < 0$: ES underestimated; the distribution of Z_2 under H_0 has no closed form and is simulated (B3)
  - Example: $T\alpha = 6.25$, seven breaches, each with $L_t/\mathrm{ES}_t = 1.1$: $Z_2 = 1 - 7.7/6.25 = -0.23$

The Distribution of Z_2 under a Correct Forecast



- Simulation: i.i.d. Student-t_5 losses with the true VaR and ES at 2.5%, 20 000 samples
  - both histograms are centred at $E[Z_2] = 0$
  - the standard deviation shrinks roughly like $1/\sqrt{T\alpha}$: about 6 expected breaches for $T = 250$, 25 for $T = 1\,000$
- One-sided test: reject when Z_2 is below the 5% quantile of its null distribution
  - here -0.74 for $T = 250$ and -0.36 for $T = 1\,000$
  - the left tail is long: a single large loss can push Z_2 far below 0

How Do We Rank Forecasts and Quantify Uncertainty?

- The Fissler–Ziegel loss with zero-homogeneous score differences (FZ0) scores the VaR–ES pair: $\mathrm{FZ0}(cv, ce; cL) = \mathrm{FZ0}(v, e; L) + \ln c$. A smaller average score is better; differences of scores are not differences of monetary losses.
- Diebold–Mariano (DM) tests equal average scores; Giacomini–White (GW) tests conditional predictive ability. A Model Confidence Set (MCS) retains the models whose performance cannot be distinguished by its procedure.
- Heteroskedasticity and autocorrelation consistent (HAC) standard errors account for dependent score differences. In an autoregressive AR(1) process, a large difference today tends to persist tomorrow; the Autocorrelation Function (ACF) measures this dependence by lag.

Consistent Scores for VaR and ES

- **Consistent score** Gneiting (2011): its expected value is smallest at the true VaR (and ES)
  - a tail forecast is not compared by a squared error: its expectation is minimised by the mean, not by a quantile
- VaR: the pinball (quantile) loss, v: the VaR forecast

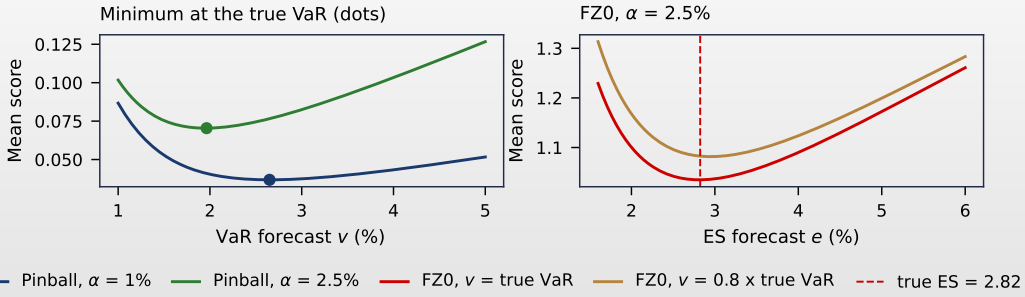
$$S(v, L) = (1\{L > v\} - \alpha)(L - v)$$

  - a breach costs $(1 - \alpha)(L - v)$, a non-breach $\alpha(v - L)$: both too low and too high VaR are penalised
  - e.g. $\alpha = 1\%$, $v = 3$: a loss $L = 5$ scores $0.99 \times 2 = 1.98$, a gain $L = -1$ scores $0.01 \times 4 = 0.04$
- (VaR, ES) jointly: FZ0 Patton, Ziegel & Chen (2019), with $v = \mathrm{VaR}$ and $e = \mathrm{ES} > 0$

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

  - ES alone is not elicitable: no score is minimised by ES by itself, but the pair (VaR, ES) is Fissler & Ziegel (2016)
  - for both scores a lower mean over the evaluation days indicates a better forecast

Mean Scores as Functions of the Forecast



- Simulated Student-t_4 losses with unit variance (400 000 days)
  - left: the mean pinball loss is lowest exactly at the true VaR, 2.65 for $\alpha = 1\%$ and 1.96 for $\alpha = 2.5\%$
  - right: with v at the true VaR, the mean FZ0 is lowest at the true ES $= 2.82$; a too low v raises the whole curve

Comparing Two Forecasts: the Diebold–Mariano Test

- Diebold & Mariano (1995) (Chapter 5): is the mean score difference of models A and B different from 0?

$$d_t = S_{A,t} - S_{B,t}, \qquad \bar{d} = \frac{1}{T}\sum_{t=1}^{T} d_t, \qquad \mathrm{DM} = \frac{\bar{d}}{\sqrt{\widehat{\mathrm{LRV}}/T}} \xrightarrow{d} N(0,1)$$

  - $S_{A,t}$: score of A on day t; H_0: $E[d_t] = 0$ (equal expected scores); $\mathrm{DM} > 1.96$: B significantly better at 5%

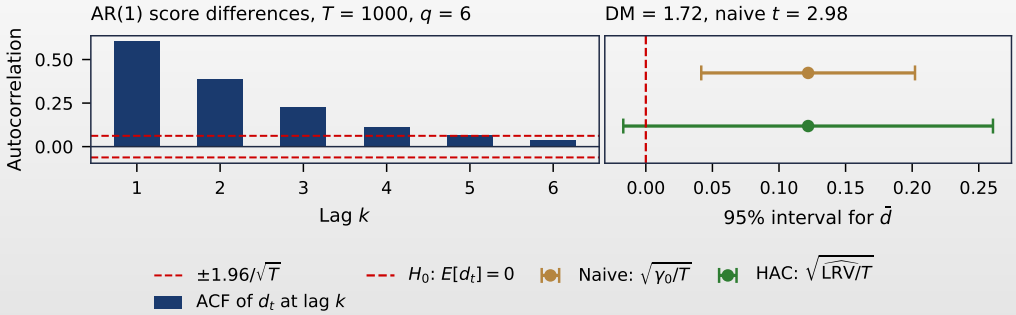
- d_t is autocorrelated, so the long-run variance (LRV; Newey–West Newey & West (1987)) replaces the variance γ_0

$$\widehat{\mathrm{LRV}} = \gamma_0 + 2\sum_{k=1}^{q}\Big(1 - \frac{k}{q+1}\Big)\gamma_k, \qquad \gamma_k = \frac{1}{T}\sum_{t=k+1}^{T}(d_t - \bar{d})(d_{t-k} - \bar{d})$$

  - γ_k: autocovariance of d_t at lag k; weights $1 - k/(q+1)$ keep $\widehat{\mathrm{LRV}} \geq 0$; $q = \lfloor 4(T/100)^{2/9} \rfloor$ lags, $\lfloor\cdot\rfloor$: integer part
  - Example, $q = 1$: $\gamma_0 = 4$, $\gamma_1 = 1$: $\widehat{\mathrm{LRV}} = 4 + 2 \times 0.5 \times 1 = 5$; with $\bar{d} = 0.1$, $T = 2000$: $\mathrm{DM} = 0.1/\sqrt{5/2000} = 2.0$

- Giacomini–White Giacomini & White (2006) (optional in B2): form $Z_t = (d_t, d_{t-1}d_t)^\top$ and test its zero mean: $W = T\bar{Z}^\top \hat{\Omega}^{-1}\bar{Z} \sim \chi^2_2$, where $\hat{\Omega} = T^{-1}\sum_t Z_t Z_t^\top$
  - it asks whether one model is better *conditionally* on what was known at $t-1$

Why the HAC Standard Error Matters



- Simulated score differences $d_t = 0.17 + x_t$, x_t an AR(1) with coefficient 0.6, $T = 1\,000$, $q = 6$
  - left: positive autocorrelations at the first lags, so $\widehat{\mathrm{LRV}}/\gamma_0 = 3.0$
  - right: the naive interval excludes 0 ($t = 2.98$), the HAC interval does not ($\mathrm{DM} = 1.72$): ignoring autocorrelation overstates significance

Tools for B3: Exceedance Residuals and the Bootstrap

- McNeil–Frey McNeil & Frey (2000): on the n breach days, the loss beyond ES in units of σ_t

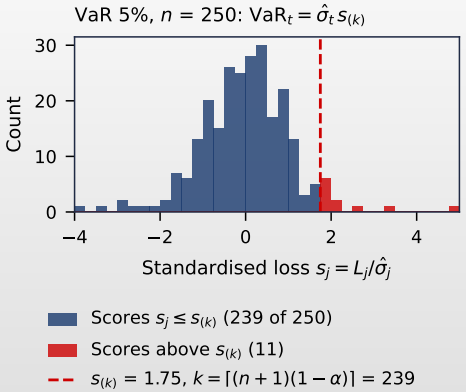
$$e_t = \frac{L_t - \mathrm{ES}_t}{\sigma_t}, \qquad t = \frac{\bar{e}}{s_e/\sqrt{n}}$$

  - correct ES: $E[e_t] = 0$; H_1: $E[e_t] > 0$ (ES too low); $\bar{e}$, s_e: mean and standard deviation of the e_t
- Bootstrap p-value: the null distribution of t from the data, without a Normal or t approximation
  - centre the residuals, $\tilde{e}_t = e_t - \bar{e}$, so that H_0 holds; draw n of them with replacement B times
  - $p = \frac{1}{B}\sum_{b=1}^{B} 1\{t_b^* \geq t\}$, t_b^*: the statistic of draw b
  - with B draws the Monte Carlo standard error of p is $\sqrt{p(1-p)/B}$

What Is Calibrated by Conformal Prediction?

- Split conformal prediction chooses a rank of calibration scores. The guarantee depends on exchangeability: observations can be reordered without changing their joint distribution.
- Independent and identically distributed (i.i.d.) observations satisfy this property. Financial time series need not; a marginal coverage guarantee is not a guarantee for each day.
- Adaptive Conformal Inference (ACI), used in A10, updates the working tail probability after each breach. It is a separate adaptation step, not another name for the fixed split-conformal quantile.

Tools for B4: Split Conformal VaR



- Split conformal prediction Lei et al. (2018): a quantile of past standardised losses, rescaled by today's volatility
  - calibration scores $s_j = L_j/\hat{\sigma}_j$ on the last n days; $\hat{\sigma}_j$: a volatility estimate known at $j-1$
- $s_{(k)}$: the k-th smallest score

$$k = \lceil (n+1)(1-\alpha) \rceil, \qquad \mathrm{VaR}_t = \hat{\sigma}_t \, s_{(k)}$$

  - $\lceil \cdot \rceil$: rounding up; chart: $n = 250$, $\alpha = 5\%$, $k = 239$, 11 scores above $s_{(k)}$
- Guarantee: $P(L_t > \mathrm{VaR}_t) \le \alpha$ if the $n+1$ scores are exchangeable
  - exchangeable: their joint distribution does not change when their order is permuted (e.g. i.i.d.)
  - the guarantee is marginal, an average over all days, not day by day

Data Used and the Forecast Panel

Market	Series	Calendar	Evaluation days
S&P 500	S&P 500 index, daily close	US trading days	03.01.2007–18.09.2026, 4 959
BET	BET index, daily close	BVB trading days	03.01.2007–18.09.2026, 4 937
Bitcoin	Bitcoin in USD, daily close	every day, 24:00 UTC	14.06.2017–18.09.2026, 3 384
EUR/RON	BNR reference rate	Romanian working days	11.06.2009–18.09.2026, 4 352

- Daily closing prices of the course data, one series per market, loaded automatically by the notebook; losses $L_t = -\ln(P_t/P_{t-1})$ in **decimals** ($0.01 = 1\%$), each series on its own calendar, no filling of missing days
- Forecast panel: for each market and model, one row per evaluation day with the loss L_t, VaR 1%, 2.5% and 5%, ES 2.5%, the forecast volatility σ_t, the mean μ_t and the PIT u_t; the six models of Chapter 8 (rolling window of 1 000 days, refit every 20 days, data up to $t-1$ only)
- The first call computes and caches the forecast panel; subsequent calls reuse it

Setup: Run and Verify

```
F, tails = fc('sp500')
g = F['GARCH-t']
print(len(g), g.index[0], g.index[-1])
# loss, VaR 1%, VaR 2.5%, ES 2.5%, scale
print(g.iloc[:3, [-1, 0, 1, 3, 6]])
```

- "Style" defines `fc` and `hits`; "Common helpers" defines `kupiec`, `christoffersen`, `duration_test`, `dq_test`, `fz0`
- Expected output: 4 959 rows, 03.01.2007 to 18.09.2026
- A negative `L` is a gain; VaR and ES are positive loss thresholds
- If your count or first row differs, stop and check the market, the start date and the model name

date	L	VaR1	VaR2.5	ES2.5	scale
2007-01-03	0.001199	0.012031	0.009944	0.012132	0.005298
2007-01-04	−0.001228	0.011894	0.009830	0.011995	0.005241
2007-01-05	0.006103	0.011745	0.009706	0.011844	0.005178

Exercise Map

Exercise	Question	Type	Worked example
A1	is 7 breaches in 250 days evidence against a VaR 1%?	Solved	–
A2	do the BET breaches of 2007–2010 cluster?	Solved	–
A3, A4	traffic lights for VaR counts and for ES (Z_2)	Solved	–
B1	which of six VaR models passes on the S&P 500?	Solved	–
B2	is FHS significantly better?	Solved	–
Checks	new-input Kupiec and ES checks after A4	Proposed	A1, A4
C1	what is wrong in an answer from an artificial intelligence (AI) assistant?	Proposed	A2, A3
B5	would the supervisor have penalised the GARCH-t model on the BET? (after the lecture)	Proposed	A3, B1
B3	is ES underestimated? (after the lecture)	Solved	–
B4	does conformal VaR keep its coverage in crises? (after the lecture)	Solved	–
Extensions	A1 Extended, A5–A10, B6–B9, C2, C3 (optional, after the seminar)	Solved and Proposed	named on each task

- **[Solved]**: full solution in the slides and in the notebook, a worked example to follow; **[Proposed]**: you solve it, the solution is discussed in class
- Part A: derivations on paper; Part B: verification and inference on real forecasts, each ending with an interpretation question; Part C: open questions and the critique of an AI (Artificial Intelligence) answer

A1 [Solved]: Kupiec from a Breach Count — Task

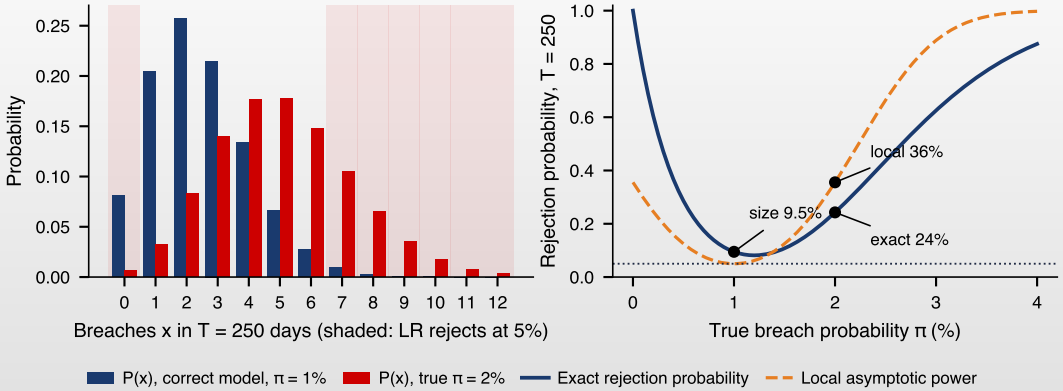
- **Question**: is $x = 7$ breaches in $T = 250$ days evidence against a VaR 1% model?
- **Inputs**: $T = 250$, $x = 7$, $\alpha = 0.01$; the hits are i.i.d. Bernoulli(π); no data file
- **Tasks**
  1. Compute $\hat{\pi} = x/T$ and the log-likelihoods $\ell(\alpha)$ (restricted) and $\ell(\hat{\pi})$ (unrestricted) in "The Kupiec Test".
  2. Compute $\mathrm{LR}_{uc} = 2[\ell(\hat{\pi}) - \ell(\alpha)]$, its χ^2_1 p-value, and the exact one-sided probability $P(X \geq 7)$ under Binomial(250, 0.01).
  3. List the integers x with $\mathrm{LR}_{uc}(x) \leq 3.84$: the acceptance region of the rule "reject if $\mathrm{LR}_{uc} > 3.84$".
  4. Sum Binomial probabilities outside that region to get the actual size ($\pi = 1\%$) and the power against $\pi = 2\%$; compare with the local power $P(\chi^2_1(\lambda) > 3.84)$ in "Size and Power of a Backtest".
- **Report**: $\hat{\pi}$, the two log-likelihoods, LR_{uc} with both p-values, the acceptance region, nominal size, actual size and the two powers
- **Interpretation**: can one year of VaR 1% data detect a model that doubles the breach rate?
- Notebook section: A1

A1 [Solved]: Solution (1/3) — The Test Step by Step

1. $\hat{\pi} = 7/250 = 0.028$; $\ell(\alpha) = 243 \ln 0.99 + 7 \ln 0.01 = -34.678$; $\ell(\hat{\pi}) = 243 \ln 0.972 + 7 \ln 0.028 = -31.930$.
2. $\mathrm{LR}_{uc} = 2\,(-31.930 - (-34.678)) = 5.50$; χ^2_1 $p = 0.019$ (two-sided alternative); one-sided exact $P(X \geq 7) = 0.014$: both reject at 5%, but their alternatives differ.
3. Evaluating $\mathrm{LR}_{uc}(x)$ for $x = 0, 1, 2, \ldots$: $\mathrm{LR}_{uc}(0) = -500 \ln 0.99 = 5.03 > 3.84$, $\mathrm{LR}_{uc}(6) = 3.56$, $\mathrm{LR}_{uc}(7) = 5.50$: the acceptance region is $\{1, \ldots, 6\}$.
4. Three different quantities: **nominal size** 5% (the χ^2_1 critical value), **actual size** of this rule (next slide), **power** against $\pi = 2\%$ (next slide).

⊡ Interpretation: 7 breaches are 2.8 times the 2.5 expected; a correct VaR 1% produces 7 or more breaches in only 0.014 of years

A1 [Solved]: Solution (2/3) — Size and Power



MFM_ch8_sem_partA

A1 [Solved]: Solution (3/3) — Size and Power

- Actual size: $P(X = 0) + P(X \geq 7) = 8.11\% + 1.37\% = 9.5\%$ under $\pi = 1\%$, not 5%
- Exact power at $\pi = 2\%$: $1 - P(1 \leq X \leq 6) = 1 - 75.7\% = 24\%$
- Local: $\lambda = 250 \cdot 0.01^2/0.0099 = 2.53$, $P(\chi^2_1(\lambda) > 3.84) = 36\%$: the asymptotics overstate the power
- Interpretation: with 2.5 expected breaches the lattice of x dominates; a model that doubles the breach rate passes one year of testing about three times in four

MFM_ch8_sem_partA

A2 [Solved]: From a Hit Sequence to Transitions — Task

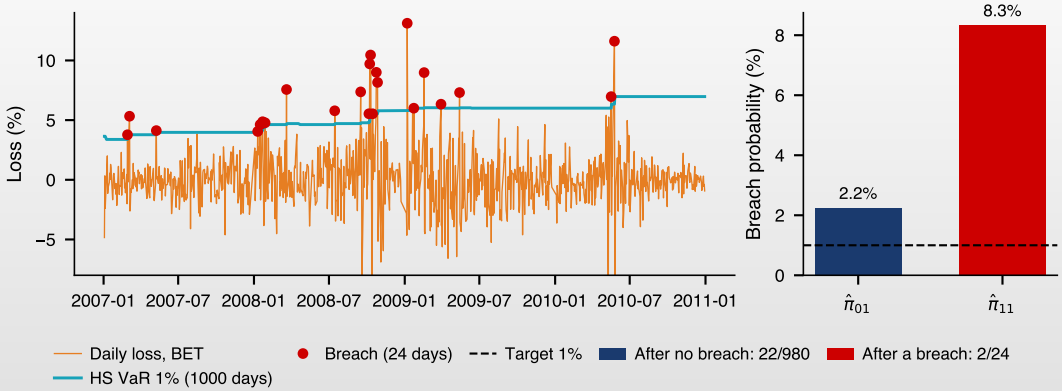
- **Questions**
  - do the breaches of historical simulation on the BET in 2007–2010 cluster?
  - are there simply too many of them?
- **Inputs**
  - A second 12-day sequence (compare its transition counts with the primer): $I_t = 0\ 0\ 1\ 1\ 0\ 0\ 0\ 0\ 1\ 0\ 0\ 0$
  - BET daily closes, HS VaR 1% (window of 1 000 days), 2007–2010: $T = 1005$ days, so $n = T - 1 = 1004$ transitions (the first day is conditioned on)
  - Transition counts: $n_{\mathbf{00}} = 958$, $n_{\mathbf{01}} = 22$, $n_{\mathbf{10}} = 22$, $n_{\mathbf{11}} = 2$
- **Tasks**
  1. Count the 11 transitions of the 12-day sequence into a 2×2 table ($n_{\mathbf{00}}, n_{\mathbf{01}}, n_{\mathbf{10}}, n_{\mathbf{11}}$).
  2. For the BET, compute $\hat{\pi}_{\mathbf{01}}$, $\hat{\pi}_{\mathbf{11}}$ and $\hat{\pi}$ (estimators: "Independence: the Christoffersen Test").
  3. Write the restricted log-likelihood $\ln L(\hat{\pi}, \hat{\pi})$ and the unrestricted $\ln L(\hat{\pi}_{\mathbf{01}}, \hat{\pi}_{\mathbf{11}})$; compute LR_{ind} and its $\chi^2_{\mathbf{1}}$ p-value.
  4. Compute LR_{uc} on the same n transitions ($x = 24$), then $\mathrm{LR}_{cc} = \mathrm{LR}_{uc} + \mathrm{LR}_{ind}$ and its $\chi^2_{\mathbf{2}}$ p-value.
- **Report**: the toy table, three probabilities, LR_{ind}, LR_{uc}, LR_{cc} with p-values
- **Interpretation**: does HS fail because of too many breaches or because breaches cluster?
- Notebook section: A2

A2 [Solved]: Solution (1/3) — Counts and Likelihoods

1. Toy sequence: pairs (I_{t-1}, I_t) give $n_{00} = 6$, $n_{01} = 2$, $n_{10} = 2$, $n_{11} = 1$ (total 11).
2. BET: $\hat{\pi}_{01} = 22/(958 + 22) = 0.0224$; $\hat{\pi}_{11} = 2/(22 + 2) = 0.0833$; $\hat{\pi} = (22 + 2)/1004 = 0.0239$.
3. $\ln L(\pi_{01}, \pi_{11}) = n_{00} \ln(1 - \pi_{01}) + n_{01} \ln \pi_{01} + n_{10} \ln(1 - \pi_{11}) + n_{11} \ln \pi_{11}$; each part is a Bernoulli likelihood, so the MLEs are the ratios of step 2; $\mathrm{LR}_{ind} = 2.32$, $p = 0.128$: independence not rejected.
4. $\mathrm{LR}_{uc} = 14.11$ on the 1004 transitions; $\mathrm{LR}_{cc} = 14.11 + 2.32 = 16.43$, $p = 0.0003$ (χ^2_2: both π_{01} and π_{11} fixed at α): rejected.

- On the same transitions the decomposition $\mathrm{LR}_{cc} = \mathrm{LR}_{uc} + \mathrm{LR}_{ind}$ is exact; with LR_{uc} on all T days it holds only asymptotically

A2 [Solved]: Solution (2/3) — BET Breaches and Transitions



MFM_ch8_sem_partA

A2 [Solved]: Solution (3/3) — Breach Count versus Clustering

- The HS VaR moves in steps and stays flat while losses cluster in late 2008: 24 breaches where 10 are expected
- After a breach the breach probability is almost four times higher (0.0833 versus 0.0224), yet it rests on only 2 consecutive pairs
- Interpretation: HS fails mainly through too many breaches ($\hat{\pi} \approx 2.4\%$); the clustering is visible but four years of data cannot confirm it with a first-order test

MFM_ch8_sem_partA

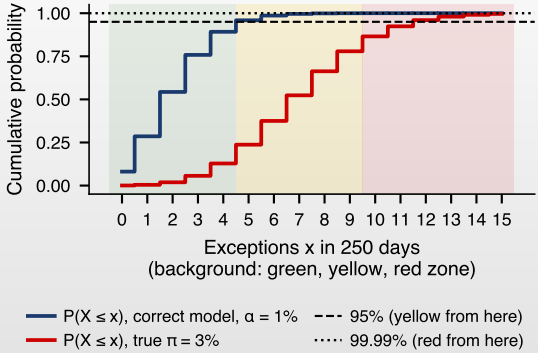
A3 [Solved]: The Traffic Light from the Binomial — Task

- **Questions**: a bank reports 6 exceptions of its VaR 1% in 250 days
  - which zone is it in?
  - what is the capital consequence?
  - how unusual is it?
- **Inputs**
  - Correct model: $X \sim \text{Binomial}(250, 0.01)$; zones defined by $P(X \leq x)$ as in "The Basel Traffic Light" (BCBS (1996))
  - Table: 1996 plus factor (added to 3); FRTB multiplier (BCBS (2019))
- **Tasks**
  1. Compute $P(X \leq x)$ for $x = 4, 5, 9, 10$ and find where the zones start.
  2. For $x = 6$: the zone, the 1996 and FRTB multipliers, $P(X \geq 6)$.
  3. Compute $P(X \leq 4)$ when the true breach probability is 3%.
  4. Compare $P(X \geq 5)$ (a correct model leaves green) with the chance of reaching red.
- **Report**: four probabilities, the zones, the answers to 2–4
- **Interpretation**
  - which error does the traffic light control?
  - which error does it leave large?
- Notebook section: A3

Exceptions	1996 plus factor	FRTB multiplier
0–4	0	1.50
5	0.40	1.70
6	0.50	1.76
7	0.65	1.83
8	0.75	1.88
9	0.85	1.92
≥ 10	1.00	2.00

Back Exercise list

A3 [Solved]: Solution — The Traffic Light from the Binomial



- Zones: green 0–4, yellow 5–9, red 10 or more
  - $P(X \leq 4) = 89.22\%$, $P(X \leq 5) = 95.88\%$
  - $P(X \leq 9) = 99.97\%$, $P(X \leq 10) = 99.99\%$
- $x = 6$: yellow
  - multiplier 1996: $3 + 0.50 = 3.50$; FRTB: 1.76
  - $P(X \geq 6) = 100\% - 95.88\% = 4.12\%$
- True rate 3%: $P(X \leq 4) = 12.8\%$
  - a model that triples the breach rate stays green about one year in eight
- A correct model leaves green: $P(X \geq 5) = 10.78\%$
  - but almost never reaches red: the zones control false red alarms
  - a zone is a lookup rule, not a 5% significance test

MFM_ch8_sem_partA

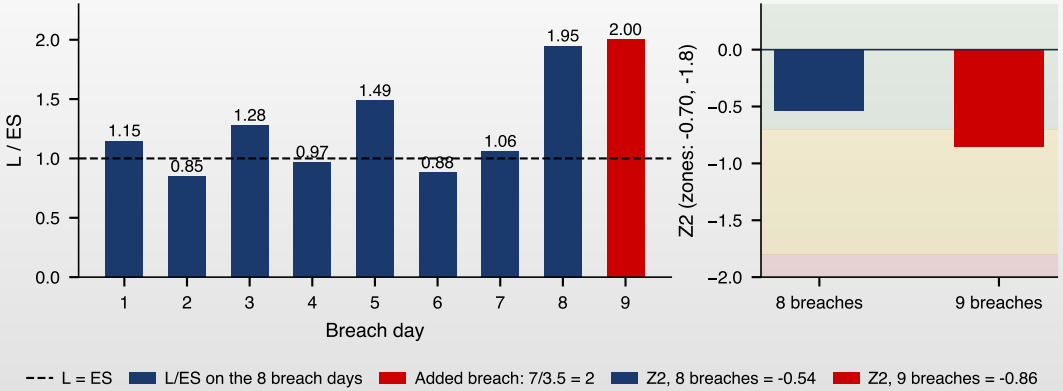
A4 [Solved]: Z_1 and Z_2 on One Year — Task

- **Question**: do eight breach days in one year show that ES 2.5% is underestimated?
- **Inputs**
  - $T = 250$ days, $\alpha = 0.025$, so $T\alpha = 6.25$; the table lists the eight days that are **already identified** as breaches of VaR 2.5%
  - The statistics Z_1 and Z_2 (Acerbi & Szekely (2014)) in "Backtesting ES: Z_1 and Z_2"
- **Tasks**
  1. Complete the column L_t/ES_t and compute each day's contribution $L_t/(\mathrm{ES}_t\, T\alpha)$ to the sum in Z_2.
  2. Compute Z_1 and Z_2; classify Z_2 as green above -0.70 and red below -1.8 (illustrative thresholds; B3 calibrates them for a model).
  3. Replace one non-breach day of the same window by a breach with $L = 7\%$, $\mathrm{ES} = 3.5\%$ (T stays 250) and recompute Z_2.
  4. Show $E[L_t I_t \mid \mathcal{F}_{t-1}] = \alpha\,\mathrm{ES}_t$ for a correct continuous forecast and deduce $E[Z_2] = 0$.
- **Report**: the ratios, Z_1, Z_2 before and after, the zone, the proof
- **Interpretation**: which days drive Z_2?
- Notebook section: A4

Day	L_t (%)	ES_t (%)	L_t/ES_t
1	3.9	3.4	...
2	2.8	3.3	...
3	4.6	3.6	...
4	3.1	3.2	...
5	5.2	3.5	...
6	2.9	3.3	...
7	3.6	3.4	...
8	7.4	3.8	...

Back Exercise list

A4 [Solved]: Solution (1/2) — Z_1 and Z_2 on One Year



MFM_ch8_sem_partA

A4 [Solved]: Solution (2/2) — Z_1 and Z_2 on One Year

- Ratios: $1.147, 0.848, 1.278, 0.969, 1.486, 0.879, 1.059, 1.947$; sum 9.613; contributions = ratio/6.25, from 0.136 (day 2) to 0.312 (day 8)
- $Z_1 = 1 - 9.613/8 = -0.202$: tail losses 20% above ES on average; $Z_2 = 1 - 9.613/6.25 = -0.538 > -0.70$: green
- Added breach, ratio 2: $Z_2 = -0.86 < -0.70$: yellow
- Proof: $E[L_t I_t \mid \mathcal{F}_{t-1}] = P(L_t > \mathrm{VaR}_t \mid \mathcal{F}_{t-1})\, E[L_t \mid L_t > \mathrm{VaR}_t, \mathcal{F}_{t-1}] = \alpha\, \mathrm{ES}_t$; $T\alpha$ is constant, so $E[Z_2] = 0$ (Z_1 divides by the random N_T)
- Interpretation: Z_2 reacts to both the number and the size of breaches; one large loss can change the zone

MFM_ch8_sem_partA

Kupiec and ES: Independent Checks [Proposed]

- **Context**: two separate cases
  - case 1: a VaR 1% has 9 exceptions in 500 days
  - case 2: VaR and ES 2.5% forecasts over 250 days, so $T\alpha = 6.25$, with 6 VaR breaches; the loss/ES ratios are 1.10, 0.90, 1.25, 1.05, 0.80, 1.20
- **Tasks**
  1. Case 1: compute the expected number of exceptions, the Kupiec statistic and the asymptotic decision at 5%.
  2. Case 1: compute the exact one-sided p-value $P(X \geq 9)$ for $X \sim \mathrm{Binomial}(500, 0.01)$ (too many exceptions).
  3. Case 2: compute $Z_2 = 1 - \sum_t (I_t L_t / \mathrm{ES}_t)/(T\alpha)$ and compare it with the 5% reference critical value, -0.70 for $T = 250$ (Acerbi & Szekely (2014)).
  4. For each test, state the null hypothesis and the sign of the deviation.
- **Report**: the expected number, LR_{uc}, both p-values, the sum of the ratios and Z_2

Back Exercise list

B1 [Solved]: Six VaR Models on the S&P 500 — Task

- **Questions**
  - which of the six procedures pass the VaR 1% backtests?
  - which test decides?
- **Inputs**
  - S&P 500 (Standard & Poor's) panel, notebook B1: loss L_t, VaR 1% of six models; $T = 4\,959$; table: ten GARCH-t days, September 2008
  - Durations: first and last spells censored; DQ: constant, 4 lagged hits, VaR_t
- **Tasks**
  1. Fill the column I_t of the table and count the breaches.
  2. For GARCH-t in 2008 ($T = 253$): compute x, LR_{uc}, the χ^2_1 and the exact two-sided Binomial p-values; decide at 5%.
  3. All six models, full sample: x, rate, Kupiec (asymptotic, exact), independence, CC, Weibull b, duration and DQ p-values.
  4. Plot the breach days and the survival curves of the durations.
  5. Read a DQ $p = 0.03$ at $T = 250$ against the simulated size of the tests.
- **Report**: one table row per model (task 3 statistics), two figures
- **Interpretation**: which models pass all tests? Why can a model with the right count still fail?

Date	L_t (%)	VaR_t (%)	I_t
04.09.2008	3.04	2.98	?
05.09.2008	−0.44	3.45	?
08.09.2008	−2.03	3.35	?
09.09.2008	3.47	3.46	?
10.09.2008	−0.61	3.99	?
11.09.2008	−1.37	3.88	?
12.09.2008	−0.21	3.84	?
15.09.2008	4.83	3.72	?
16.09.2008	−1.74	4.68	?
17.09.2008	4.83	4.64	?

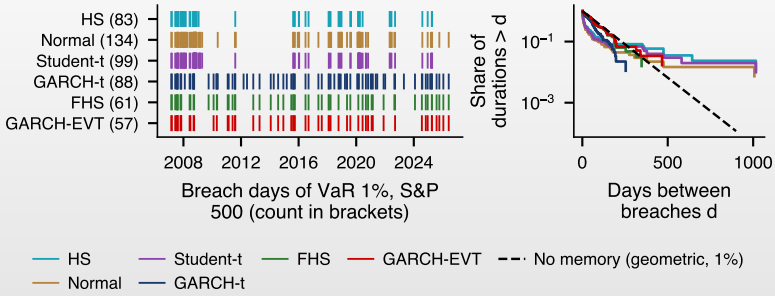
Back Exercise list

B1 [Solved]: Solution (1/4) — One Model Step by Step

Date	L_t	VaR_t	I_t
04.09.2008	3.04	2.98	1
05.09.2008	−0.44	3.45	0
08.09.2008	−2.03	3.35	0
09.09.2008	3.47	3.46	1
10.09.2008	−0.61	3.99	0
11.09.2008	−1.37	3.88	0
12.09.2008	−0.21	3.84	0
15.09.2008	4.83	3.72	1
16.09.2008	−1.74	4.68	0
17.09.2008	4.83	4.64	1

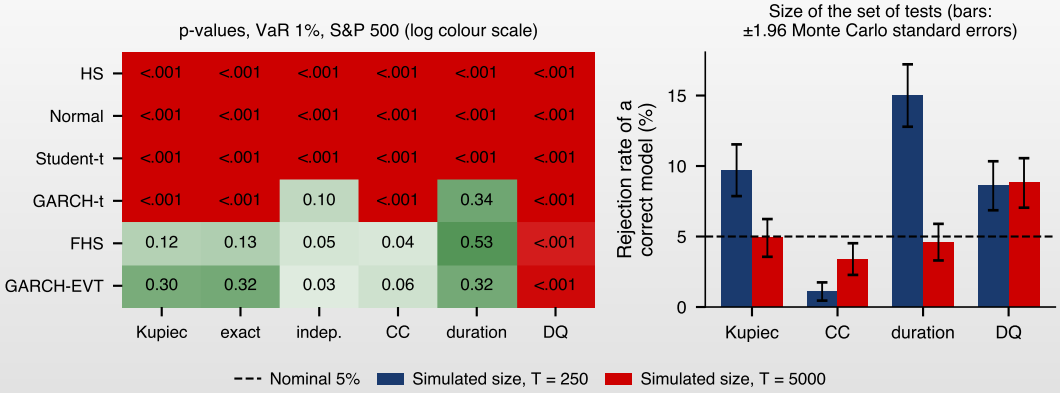
1. $I_t = 1$ when $L_t > \mathrm{VaR}_t$: 4 breaches in 10 days, against 0.1 expected: a burst, while the VaR rises only after each loss.
2. 2008: $x = 8$ in $T = 253$, $\hat{\pi} = 3.16\%$; $\mathrm{LR}_{uc} = 7.60$, $p = 0.0058$; exact two-sided p ($2P(X \geq 8)$) $= 0.0086$: rejected at 5%.
3. Full sample: the same hit series goes into all five tests; the counts give the rates, the transitions give independence, the gaps between breaches give the durations.
4. With 4 959 days the asymptotic and exact Kupiec p-values agree; in one year they differ (0.0058 versus 0.0086).

B1 [Solved]: Solution (2/4) — When the Breaches Happen



- Figures in notebook B1
- Unconditional models (HS, Normal, Student-t): bursts in 2008–2009, 2020 and 2022, then years of silence; Weibull b: HS 0.52, Normal 0.56, Student-t 0.53 (duration $p < 0.001$)
- Filtered models: survival curves close to the geometric line; GARCH-t 0.92 ($p = 0.344$), GARCH-EVT 0.90 ($p = 0.317$)

B1 [Solved]: Solution (3/4) — All Tests and Their Size



MFM_ch8_size_power

B1 [Solved]: Solution (4/4) — All Tests and Their Size

- Only GARCH-EVT passes Kupiec, CC and durations; FHS fails CC narrowly ($p = 0.040$); DQ rejects all six (GARCH-EVT $p < 0.001$): hits are predictable from lagged hits and the VaR level
- Size at $T = 250$: POF 9.7%, CC 1.1%, durations 15.0%, DQ 8.6% (Monte Carlo standard error about 0.9 points): a DQ $p = 0.03$ in one year is weak evidence; simulate the critical value
- Interpretation: a right count (FHS: 1.23%) can still fail on timing; only filtered models with a flexible tail survive most tests

MFM_ch8_size_power

B2 [Solved]: Is FHS Significantly Better? — Task

- **Question**: is FHS significantly better than GARCH-t, HS and GARCH-EVT under the FZ0 score on the S&P 500?
- **Inputs**
  - Daily FZ0 scores of (VaR 2.5%, ES 2.5%) for the six models on the S&P 500, aligned on the same 4 959 days (Patton, Ziegel & Chen (2019))
  - Score differential $d_t = S_{A,t} - S_{\mathrm{FHS},t}$ (a difference of scores, not of losses); Newey–West with $q = \lfloor 4(T/100)^{2/9} \rfloor = 9$ lags, $\lfloor \cdot \rfloor$: integer part (Newey & West (1987))
- **Tasks**
  1. Compute the mean FZ0 of each model.
  2. For A = GARCH-t: $\bar{d}$, the autocovariances $\gamma_0, \ldots, \gamma_9$, LRV with $q = 9$ ("Comparing Two Forecasts: the Diebold–Mariano Test"), $\mathrm{se} = \sqrt{\mathrm{LRV}/T}$, $\mathrm{DM} = \bar{d}/\mathrm{se}$ and its p-value (Diebold & Mariano (1995)).
  3. Repeat for HS and GARCH-EVT; compute the naive t (ordinary standard error) and the ratio LRV/γ_0.
  4. Plot the cumulative d_t, the ACF of d_t and the naive and HAC 95% intervals of $\bar{d}$.
  5. Optional: the Giacomini–White test with $h_{t-1} = (1, d_{t-1})$ (Giacomini & White (2006)).
- **Report**: mean scores, one complete DM calculation, a DM table and one figure
- **Interpretation**
  - which differences are significant?
  - when does ignoring autocorrelation overstate significance?
- Notebook section: B2

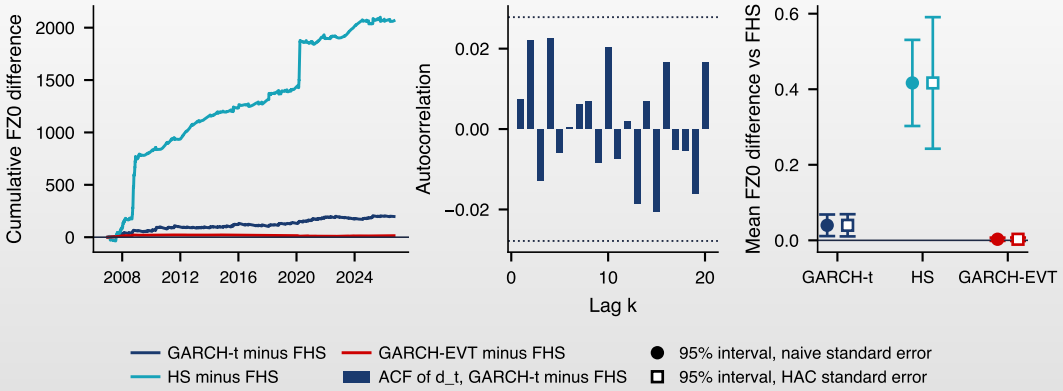
Back Exercise list

B2 [Solved]: Solution (1/3) — One DM Statistic Step by Step

1. Mean FZ0: FHS -3.551, GARCH-EVT -3.548, GARCH-t -3.511, HS -3.134, Student-t -3.036, Normal -2.958.
2. GARCH-t minus FHS: $\bar{d} = 0.0398$; $\gamma_0 = 1.0537$; $\gamma_1, \ldots, \gamma_9$: 0.0077, 0.0231, -0.0135, 0.0237, -0.0062, 0.0003, 0.0064, 0.0073, -0.0088.
3. $2\sum_k (1 - k/10)\gamma_k = 0.0595$, so $\mathrm{LRV} = 1.1132$; $\mathrm{se} = \sqrt{1.1132/4\,959} = 0.0150$ (naive 0.0146).
4. $\mathrm{DM} = 0.0398/0.0150 = 2.66$, $p = 0.008$: GARCH-t is significantly worse at 5%.

A minus FHS	$\bar{d}$	DM	p	naive t	LRV/γ_0
GARCH-t	0.0398	2.66	0.008	2.73	1.06
HS	0.4167	4.69	<0.001	7.16	2.33
GARCH-EVT	0.0029	1.24	0.214	1.32	1.12

B2 [Solved]: Solution (2/3) — Differentials, ACF and Intervals



MFM_ch8_sem_compare

B2 [Solved]: Solution (3/3) — Differentials, ACF and Intervals

- HS loses steadily and in jumps (2008, 2020); its LRV is 2.33 times γ_0, so the naive t (7.16) overstates the HAC DM (4.69); for GARCH-t the ACF is small (lag 1: 0.007) and the two intervals almost coincide
- Provisional ranking: FHS and GARCH-EVT tie ($p = 0.214$); GARCH-t and HS are significantly worse
- GW (optional), GARCH-EVT against FHS: 6.70 ($p = 0.035$); the conditional tie is rejected, but what drives the rejection needs a separate test

MFM_ch8_sem_compare

A Reproducible Test Series for the Code Audit

- Generate the toy returns with `rng = np.random.default_rng(8)`, then `r = 0.01*rng.standard_t(4,400)/np.sqrt(2)`.
- Set `r[320] = -0.08`; construct `P = 100*np.exp(np.r_[0,np.cumsum(r)])`. Thus 400 returns produce 401 prices; return index 320 connects price indices 320 and 321.
- For real data use `load_returns("sp500")` defined in the notebook's data section; it reads closes from 1990 and returns decimal log returns. Keep the same 250-observation window in both implementations.
- Compare both codes on both series. Save the forecast and exception indicator by date; locate the first differing forecast before interpreting test statistics.

C1 [Proposed]: Critique an AI Answer — Task

- **Questions**
  - is the AI answer on the next slide correct?
  - does its conclusion survive once the code is fixed?
- **Worked example to follow**: A2, A3
- **Inputs**
  - The prompt, the answer and its line-numbered code (next slide); `lr_ind(I)` is `christoffersen(I, 0.01)['LR_ind']` from the notebook helpers
  - Test series: 401 prices from 100, daily log returns $0.01 \times t_4/\sqrt{2}$ (seed 8) with $r = -8\%$ at return-array index 320 (0-based); real data: S&P 500 daily closes from 1990
- **Tasks**
  1. Find the three conceptual errors: in the forecast window, in the test distribution, in the regulatory reading.
  2. For each error, say how you would detect it: by a numerical check of the output, by the formulas of Chapter 8, or by a published table (BCBS).
  3. Run the answer's code and your corrected code on the test series and on the S&P 500; report VaR for return index 320 (zero-based, VaR[70] when W=250) and the exception counts.
  4. Give the correct zone, plus factor and multiplier for 7 exceptions (A3).
  5. Optional checks: the evaluation window of the traffic light; $x = 0$ in the Kupiec formula.
- **Report**: the three errors with detection and correction, the corrected code, four exception counts (two codes on two series)
- **Interpretation**: does "the model must be withdrawn" survive?
- Notebook section: C1

Back Exercise list

C1 [Proposed]: Critique an AI Answer — the Answer

- **Prompt**: "I have daily S&P 500 closes in a pandas Series `P`. Write Python for a rolling 250-day historical-simulation VaR 1%, count the exceptions and run the Kupiec and Christoffersen tests. My last 250 days have 7 exceptions: what does that mean?"

```
import numpy as np, scipy.stats as stats
r = np.log(P).diff().dropna().values
W, p = 250, 0.01
VaR = np.array([-np.quantile(r[t-W:t+1], p) for t in range(W, len(r))])
I = (-r[W:] > VaR).astype(int)
T, x = len(I), I.sum(); ph = x / T
LR_uc = -2*((T-x)*np.log(1-p) + x*np.log(p) - (T-x)*np.log(1-ph) - x*np.log(ph))
LR_cc = LR_uc + lr_ind(I)
p_uc, p_cc = stats.chi2.sf(LR_uc, df=1), stats.chi2.sf(LR_cc, df=1)
```

- **Answer**: "Your last 250 days: $x = 7$, $\mathrm{LR}_{uc} = 5.50$, $p = 0.019$: coverage rejected at 5%. Under the Basel traffic light 7 exceptions is the red zone, so the model must be withdrawn."
- AI code usually runs without an error message: the mistakes are in the finance and the statistics, not in the syntax

What the Forecast Panel Stores

- Call `panels, tails = fc(market)`. For each model, `panels[model]` contains dated columns `mu`, `scale`, and `nu` for Student models; GARCH-EVT also stores `u`, `xi`, `beta`, `pu`.
- `tails["HS"]` and `tails["FHS"]` contain 25 ordered losses per forecast day, in that day's loss units. The model parameters are panel columns; they are not stored in this tail array.
- Fit Student and GARCH parameters every 20 observations on the preceding 1,000 observations; use a constant mean and floor estimated degrees of freedom at 2.2. The GPD threshold is the 90th percentile of standardised losses.
- In B7, u is the GPD threshold, not the uniform draw U. The tail sampler is valid here because $U > 0.975 > 0.90$; the same fitted model and tail protocol apply to the C2 portfolio.

B3 [Solved]: Is ES Underestimated? — Task (1/2)

- **Question**: do the ES 2.5% forecasts of GARCH-t, FHS and HS understate the tail of the S&P 500?
- **Inputs**
  - S&P 500 forecast panel: loss L_t, VaR 2.5%, ES 2.5%, forecast volatility σ_t (for HS the window standard deviation), and for HS and FHS the 25 largest window losses of each day; $T = 4\,959$
  - One day step by step, GARCH-t, 12.03.2020: $L = 9.99\%$, $\mathrm{VaR} = 9.40\%$, $\mathrm{ES} = 13.43\%$, $\sigma = 4.81\%$
- **Tasks**
  1. For that day compute the exceedance residual $e_t = (L_t - \mathrm{ES}_t)/\sigma_t$ and the contribution $L_t/(\mathrm{ES}_t\, T\alpha)$ to the sum in Z_2.
  2. Run Algorithm 1 (next slide) for the three models: n, $\bar{e}$, t, bootstrap p and Student-t p.
  3. Run Algorithm 2 for the three models on the full sample and on the last 250 days: Z_2, p, simulated 5% critical value.
  4. Plot the two null distributions of Z_2 (GARCH-t) and the bootstrap distribution of t.
- **Report**: one table (model, n, $\bar{e}$, t, two p, Z_2, p, critical value, both samples) and one figure
- **Interpretation**: what does a rejection say about the joint forecasting procedure?
- Notebook section: B3

Back Exercise list

B3 [Solved]: Two Algorithms — Task (2/2)

Algorithm 1: McNeil–Frey residual bootstrap

1. Keep the breach days $L_t > \mathrm{VaR}_t$; $e_t = (L_t - \mathrm{ES}_t)/\sigma_t$, n residuals.
2. $t = \bar{e}/(s_e/\sqrt{n})$ ($\bar{e}$, s_e: mean and standard deviation of the residuals), H_1: $E[e_t] > 0$.
3. Centre: $\tilde{e}_t = e_t - \bar{e}$ (imposes H_0).
4. 5000 times: draw n values of $\tilde{e}$ with replacement (seed 0), compute t^*.
5. p = share of $t^* \geq t$; assumes independent residuals.

Algorithm 2: Z_2 under the model's own forecasts

1. Hold every VaR_t, ES_t fixed; Z_2 from the data.
2. 5000 times (seed 11), each day: $U_t \sim U(0,1)$; if $U_t > 1-\alpha$, $L_t^* = F_t^{-1}(U_t)$, else no breach.
3. F_t^{-1}: the forecast quantile function of L_t; GARCH-t $-\mu_t + \sigma_t t_\nu^{-1}(U)\sqrt{(\nu-2)/\nu}$ (t_ν^{-1}: Student-t quantile, rescaled to variance 1); HS, FHS: the stored tail losses.
4. $Z_2^* = 1 - \sum_t L_t^* 1\{U_t > 1-\alpha\}/(\mathrm{ES}_t T\alpha)$.
5. p = share $Z_2^* \leq Z_2$; critical value: 5% quantile of Z_2^*.
6. Assumes independent U_t and fixed forecast distributions: estimation risk and GARCH refitting are ignored.

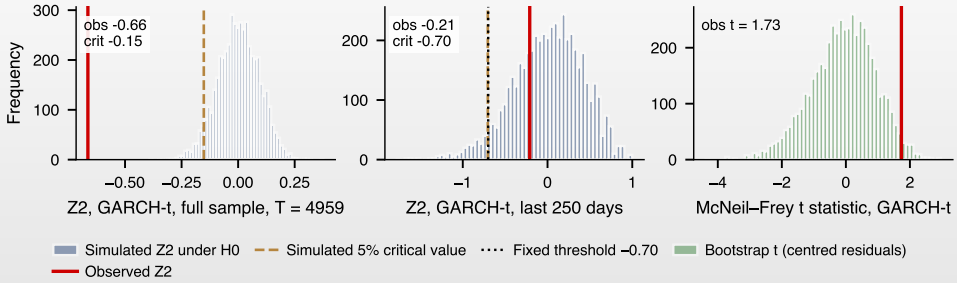
B3 [Solved]: Solution (1/2) — Step by Step and in Full

1. 12.03.2020: $e_t = (9.99 - 13.43)/4.81 = -0.71$ (the loss was below the forecast ES); contribution $0.744/123.98 = 0.0060$, out of 198 breach days.

Model	n	$\bar{e}$ / t	p_{boot} / p_t	Z_2 ($T = 4\,959$)	p / crit.	Z_2, p (250 d)
GARCH-t	198	0.100 / 1.73	0.025 / 0.042	-0.664	<0.001 / -0.152	-0.210, 0.289
FHS	128	0.104 / 1.34	0.072 / 0.091	-0.079	0.191 / -0.150	0.142, 0.602
HS	156	0.363 / 2.88	<0.001 / 0.002	-0.405	<0.001 / -0.154	0.723, 0.975

- GARCH-t: $t = 1.73$, bootstrap $p = 0.025$ against Student-t $p = 0.042$: skewed residuals make the t approximation poor
- Full sample: GARCH-t ($Z_2 = -0.664$) and HS rejected, FHS passes; last 250 days: no model is rejected, GARCH-t has only 6 breaches
- FHS: $p = 0.191$, critical value -0.150 with 5000 draws (seed 11); with 2000 draws, $p = 0.209$: same decision

B3 [Solved]: Solution (2/2) — Null Distributions and Limits



- The null distribution of Z_2 shrinks like $1/\sqrt{T\alpha}$: 5% critical value -0.152 on 4 959 days, -0.705 on 250 days (close to the reference -0.70)
- Interpretation: the GARCH-t rejection is evidence against the whole procedure (volatility filter, t tail, estimation); FHS, with the same filter, passes, which points to the tail but does not prove it; 5000 draws give a Monte Carlo standard error of about 0.3 points at $p = 5\%$

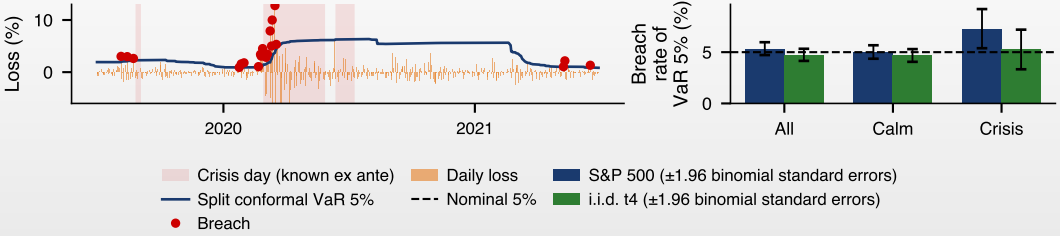
MFM_ch8_sem_es

B4 [Solved]: Conformal VaR in Calm and Crisis Periods — Task

- **Question**: does split conformal VaR keep its coverage on crisis days of the S&P 500?
- **Inputs**
  - Timeline for forecast day t (all windows end at $t-1$): scale $\hat{\sigma}_t$ = standard deviation of $r_{t-\mathbf{250}}, \ldots, r_{t-\mathbf{1}}$; calibration scores $s_{t-\mathbf{250}}, \ldots, s_{t-\mathbf{1}}$, $s_j = L_j/\hat{\sigma}_j$; crisis flag: standard deviation of $r_{t-\mathbf{20}}, \ldots, r_{t-\mathbf{1}}$ above the 90th percentile of its values over the previous 1000 days
  - Split conformal VaR in "Tools for B4: Split Conformal VaR" with $n = 250$; evaluation 03.01.2007–18.09.2026 (4 959 days); benchmark: i.i.d. Student-$t_{\mathbf{4}}$ returns $\times 1\%$, seed 5, same length
- **Tasks**
  1. Build VaR 5% and VaR 1% and the breach series on the S&P 500.
  2. Fill the grid (data source $\times$ level $\times$ all/calm/crisis) with rates and counts; add the Kupiec p (all days) and the Christoffersen independence p.
  3. Repeat on the i.i.d. benchmark series: its returns are exchangeable, but the scores standardised on a rolling window are only approximately exchangeable.
  4. Plot a VaR path with crisis shading and the rates with their Wald binomial 95% intervals.
- **Report**: the grid, two p-values per series, one figure
- **Interpretation**
  - is the marginal guarantee met on real data?
  - what changes between real and exchangeable data: coverage or independence?
- Notebook section: B4

Back Exercise list

B4 [Solved]: Solution — Conformal VaR in Calm and Crisis Periods



- S&P 500, VaR 5%: all 265/4 959 = 5.34% (Kupiec $p = 0.272$), calm 213/4 246 = 5.02%, crisis 52/713 = 7.29%; independence rejected ($p < 0.001$); VaR 1%: 1.29% overall ($p = 0.049$), 2.10% in crises
- i.i.d. benchmark: 4.74%, calm 4.68%, crisis 5.26% (27/513), independence not rejected ($p = 0.965$); the binomial intervals ignore dependence, so the real-data ones are too narrow
- Interpretation: the average is close to 5%, but volatility clustering breaks exchangeability, so breaches cluster and crisis coverage fails

MFM_ch8_sem_conformal

B5 [Proposed]: Traffic Light for GARCH-t on the BET — Task

- **Question**: would the supervisor have penalised the GARCH-t model (VaR 1%) on the BET since 2007?
- **Worked example to follow**: A3, B1
- **Inputs**
  - GARCH-t VaR 1% and VaR 2.5% forecasts on the BET; evaluation 03.01.2007–18.09.2026 ($T = 4\,937$); the first complete 250-day window ends on 21.12.2007, so there are 4 688 windows; zones and multipliers: A3
- **Tasks**
  1. Count the VaR 1% exceptions in the last 250 days; give the zone, $P(X \leq x)$ and the FRTB multiplier.
  2. For every window, the count and zone; report the share of windows in each zone and the worst window with its end date.
  3. Count the VaR 2.5% exceptions in the last 250 days and compare with the FRTB desk limit of 30 exceptions.
  4. Compute the full-sample breach rate and Kupiec p-value.
- **Report**: a table of counts, zones, zone shares and test p-values, plus the worst-window date; replicate the FRTB desk comparison linked below
- **Interpretation**: is time in yellow evidence against the model? (overlapping windows: shares are descriptive, not a test)
- Public replication benchmark: FRTB desk comparison
- Notebook section: B5

Back Exercise list

Exit Questions and Team Project

Exit questions

- Which model would you provisionally choose for the S&P 500?
- On what evidence?
- What is the main limitation of that evidence?

Assessed team project (one)

- C2: a provisional VaR/ES recommendation for a Romanian trading book (extension pack 6), as a Quantlet folder in the team repository
- Assessment criteria: data and timing correct (no look-ahead), minimum test set, evidence table, recommendation with its limitation
- Report actual AI use in `AI_USE.md` and every error you found in `AI_ERRORS.md`: request, answer, error, how found, correction; no minimum number of errors
- Choose one assessed project: C2 or C3. The other extension exercises are optional.

Extension Packs: What Each One Needs

Pack	Exercises	Prerequisites
1. Test theory	A1 Extended [Solved], A5, A6	A1, A2
2. Scores and ES theory	A7 [Solved], A8, A9	A4, B2
3. Conformal VaR	A10 [Solved], B6	B4
4. More markets	B7, B8	B2, B3
5. Estimation risk	B9	B1; Chapter 8, slide "Backtests Under Estimation Risk"
6. Projects	C2 or C3 (choose one)	B1–B3, B5, B7; C3 also needs R

- Each exercise has its own notebook section (C3 is solved in R; its notebook section gives only the setup); do them in order within a pack
- C2 and C3 are possible starts for the team project

A1 Extended [Solved]: Why LR_{uc} Is χ^2_1 — Task

- **Question**: where do the χ^2_1 limit and the noncentrality λ of A1 come from?
- **Inputs**: i.i.d. Bernoulli(π) hits, $\ell(\pi) = x \ln \pi + (T - x) \ln(1 - \pi)$, $\mathrm{LR}_{uc} = 2[\ell(\hat{\pi}) - \ell(\alpha)]$; local alternatives $\pi_T = \alpha + h/\sqrt{T}$ with h fixed
- **Tasks**
  1. Expand $\ell(\alpha)$ to second order around $\hat{\pi}$ and show $\mathrm{LR}_{uc} = T(\hat{\pi} - \alpha)^2/(\alpha(1 - \alpha)) + o_p(1)$ under H_0.
  2. Use the CLT (central limit theorem: $\sqrt{T}(\hat{\pi} - \alpha)$ is approximately Normal) to conclude $\mathrm{LR}_{uc} \xrightarrow{d} \chi^2_1$.
  3. Under π_T, derive the limit $\chi^2_1(\lambda)$ and give λ.
- **Report**: the three steps
- **Interpretation**: why is the local approximation poor at $\alpha T = 2.5$?
- Notebook section: A1

A1 Extended [Solved]: Solution — Why LR_{uc} Is χ^2_1

1. $\ell'(\hat{\pi}) = 0$ at the MLE; $\ell(\alpha) = \ell(\hat{\pi}) + \frac{1}{2}\ell''(\bar{\pi})(\alpha - \hat{\pi})^2$ with $\bar{\pi}$ between α and $\hat{\pi}$, and $-\ell''(\bar{\pi})/T = \frac{\hat{\pi}}{\bar{\pi}^2} + \frac{1-\hat{\pi}}{(1-\bar{\pi})^2} \xrightarrow{p} \frac{1}{\alpha(1-\alpha)}$.
2. So $\mathrm{LR}_{uc} = Z_T^2 + o_p(1)$, $Z_T = \sqrt{T}(\hat{\pi} - \alpha)/\sqrt{\alpha(1-\alpha)} \xrightarrow{d} N(0,1)$ by the CLT: χ^2_1, critical value 3.84.
3. Under π_T: $\sqrt{T}(\hat{\pi} - \alpha) = \sqrt{T}(\hat{\pi} - \pi_T) + h$, so $Z_T \xrightarrow{d} N(h/\sqrt{\alpha(1-\alpha)}, 1)$ and $\mathrm{LR}_{uc} \xrightarrow{d} \chi^2_1(\lambda)$, $\lambda = h^2/(\alpha(1-\alpha)) = T(\pi - \alpha)^2/(\alpha(1-\alpha))$.

⊡ Interpretation: the expansion needs many breaches; with 2.5 expected, x takes few integer values, so its distribution is strongly discrete, hence the gap between 36% (local) and 24% (exact) in A1

How Do We Search for the Required Sample Size?

- Keep three probabilities separate: the forecast tail level α, the true breach rate π, and the test's nominal significance level. The threshold 3.84 concerns the last of these.
- For each candidate T, enumerate $x = 0, \ldots, T$, compute the likelihood-ratio statistic and retain the rejection counts. At endpoints use the limiting convention $0 \log 0 = 0$, not a missing value.
- Use the same rejection counts twice: sum their probabilities with the null breach rate to obtain size, and with the alternative rate to obtain power. Do not change the rejection rule between the two sums.
- The count boundary changes in integer steps. After finding the first sample size that reaches the target power, continue through the requested range; the first crossing need not be a permanent crossing.

A5 [Proposed]: How Many Days to Detect a Doubled Rate? — Task

- **Question**: how many days are needed to detect, with 80% power, a VaR 2.5% whose true breach rate is 5%?
- **Worked example to follow**: A1
- **Inputs**: $\alpha = 0.025$, $\pi = 0.05$, the rule "reject if $\mathrm{LR}_{uc} > 3.84$"; the FRTB desk limit of 30 exceptions of VaR 2.5% is a separate regulatory rule
- **Tasks**
  1. At $T = 250$: λ, the local power, the exact acceptance region, the exact power and the actual size.
  2. Find λ^* with $P(\chi^2_1(\lambda^*) > 3.84) = 0.80$ by root finding and solve for the asymptotic T.
  3. For $T = 100, 101, \ldots, 700$ (step 1): acceptance set $\{x : \mathrm{LR}_{uc}(x) \leq 3.84\}$, power $= 1 - P(X \in \text{set} \mid \pi)$; plot power against T and find the first T with power $\geq 80\%$.
  4. Check every integer T up to 1,500; explain any saw-tooth pattern in the exact power curve.
  5. Compute $P(X > 30)$ for $X \sim \text{Binomial}(250, 0.05)$.
- **Report**: λ, powers, size, λ^*, the two values of T, $P(X > 30)$, one figure
- **Interpretation**: what does the desk limit screen for?
- Notebook section: A5

Back Exercise list

A6 [Proposed]: LR versus DQ at $x = 7$ — Task

- **Question**: why do LR_{uc} and the intercept-only DQ statistic give different numbers for the same breaches?
- **Worked example to follow**: A2
- **Inputs**
  - Hit: $T \times 1$ vector of $I_t - \alpha$; X: $T \times k$ regressors known at $t-1$, full rank; $\mathrm{Hit} = X\beta + u$; $Q = \operatorname{plim} X'X/T$ (plim: probability limit as $T \to \infty$) positive definite (Engle & Manganelli (2004))
  - Example design with one lag: rows $X_t = (1, \mathrm{Hit}_{t-\mathbf{1}})$, so $k = 2$
- **Tasks**
  1. Show $E[\mathrm{Hit}_t \mid \mathcal{F}_{t-\mathbf{1}}] = 0$ and $\mathrm{Var}(\mathrm{Hit}_t \mid \mathcal{F}_{t-\mathbf{1}}) = \alpha(1-\alpha)$ under $H_{\mathbf{0}}$.
  2. With $X_t = 1$ show $\mathrm{DQ} = T(\hat{\pi} - \alpha)^{\mathbf{2}}/(\alpha(1-\alpha))$; compute it for $T = 250$, $x = 7$ and compare with $\mathrm{LR}_{uc} = 5.50$ ($p = 0.019$).
  3. Compute both statistics for $x = 0, \ldots, 15$ and their acceptance sets at 3.84.
  4. Advanced: derive $\mathrm{DQ} \xrightarrow{d} \chi^{\mathbf{2}}_k$ from a martingale-difference CLT for $T^{-\mathbf{1}/\mathbf{2}} X'\mathrm{Hit}$.
- **Report**: the two statistics and p-values, the acceptance sets, the derivation
- **Interpretation**: do the two tests disagree in their decision or only in their numbers?
- Notebook section: A6

Back Exercise list

A7 [Solved]: Consistency of the Pinball and FZ0 Losses — Task

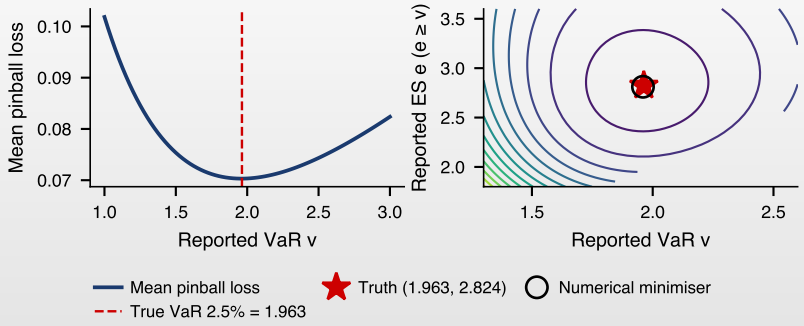
- **Questions**
  - why does the truth minimise the expected pinball and FZ0 scores?
  - does a simulation confirm it?
- **Inputs**
  - L with $E|L| < \infty$, density $f > 0$ near VaR_α, $\mathrm{ES}_\alpha > 0$; forecasts $e \geq v > 0$; $\alpha = 2.5\%$
  - Simulation: $L = \sqrt{(4-2)/4}\, Z$, $Z \sim t_4$, 10^6 draws, seed 5; truth $\mathrm{VaR} = 1.963$, $\mathrm{ES} = 2.824$ (unit-variance t_4 formulas)
- **Tasks**
  1. Differentiate $E[(1\{L > v\} - \alpha)(L - v)]$ in v, find the zero and show it is a minimum.
  2. For $\mathrm{FZ0}(v, e; L) = \frac{\mathbf{1}\{L>v\}(L-v)}{\alpha e} + \frac{v}{e} + \ln e - 1$ compute ∂_v and ∂_e of the expectation and show that the true pair is the unique minimiser.
  3. Minimise the average pinball and FZ0 losses on the simulated draws and plot the score curve and the FZ0 contours.
- **Report**: derivatives, the argument for the minimum, numerical minimisers, one figure
- **Interpretation**: why does consistency make ranking by average score meaningful?
- Notebook section: A7

Back Exercise list

A7 [Solved]: Solution (1/2) — Derivations

1. $\frac{d}{dv} E[(1\{L > v\} - \alpha)(L - v)] = F(v) - (1 - \alpha)$: negative below and positive above VaR_α, zero at $F(v) = 1 - \alpha$; second derivative $f(v) > 0$: a unique minimum.
2. $\partial_v E[\mathrm{FZ0}] = \frac{1}{e}\left(1 - \frac{P(L>v)}{\alpha}\right)$: same sign pattern, zero only at $v = \mathrm{VaR}_\alpha$, for every e.
3. At that v, $\partial_e E[\mathrm{FZ0}] = \frac{1}{e^2}\left(e - v - E[(L - v)1\{L > v\}]/\alpha\right) = \frac{1}{e^2}(e - \mathrm{ES}_\alpha)$: negative for $e < \mathrm{ES}_\alpha$, positive above: the profiled objective in e has its unique minimum at ES_α.
4. Identification functions: $V_1 = \alpha - 1\{L > v\}$, $V_2 = e - v - \frac{1}{\alpha}1\{L > v\}(L - v)$; their means vanish only at the true pair (Fissler & Ziegel (2016)).

A7 [Solved]: Solution (2/2) — Numerical Check



- Minimisers: pinball $v = 1.960$; FZ0 $(v, e) = (1.960,\ 2.812)$; truth $(1.963,\ 2.824)$: equal up to simulation error
- Interpretation: a forecaster who knows F minimises the expected score by reporting the truth, so a lower average score is evidence of a better forecast

MFM_ch8_sem_partA

A8 [Proposed]: ES Alone Is Not Elicitable — Task

- **Question**: why can no score of ES alone reward honest forecasts?
- **Worked example to follow**: A7
- **Inputs**
  - $T(F)$ is elicitable if $T(F) = \arg\min_x E_F[S(x, L)]$ with a unique minimiser (Gneiting (2011)); δ_c: the point mass at c (loss c for sure)
  - ES for any distribution, atoms included: $\mathrm{ES}_\alpha = \frac{1}{\alpha}\int_{1-\alpha}^{1} q_u\, du$, the mean of the upper-$\alpha$ tail quantiles of the loss (an atom may be split)
- **Tasks**
  1. Show that if T is elicitable, every level set $\{F : T(F) = t\}$ is closed under mixtures $(1-\lambda)F_0 + \lambda F_1$.
  2. $F_0 = \delta_1$; F_1: $P(L = 2) = \alpha/2$, $P(L = 0) = 1 - \alpha/2$: compute ES_α of F_0, F_1 and $\frac{1}{2}F_0 + \frac{1}{2}F_1$ from the tail-quantile picture.
  3. Explain, conceptually, why this does not contradict the joint elicitability of (VaR, ES).
- **Report**: the argument, three ES values, one sentence
- **Interpretation**: what does this imply for ranking ES forecasts?
- Notebook section: A8

Back Exercise list

A9 [Proposed]: Moments of the Cumulative Violation — Task

- **Questions**
  - what are the null moments of H_t?
  - is the Normal critical value adequate at $T = 250$?
- **Worked example to follow**: A4
- **Inputs**: PITs u_t i.i.d. $U(0,1)$ under a correct forecast; $H_t = \frac{1}{\alpha}(u_t - (1-\alpha))1\{u_t > 1-\alpha\}$ (Du & Escanciano (2017)): zero on 97.5% of days, rising linearly to 1 in the tail; $\alpha = 2.5\%$
- **Tasks**
  1. Plot $H(u)$ for $u \in (0.9, 1)$.
  2. Compute $E[H]$ and $E[H^2]$ by integrating over $(1-\alpha, 1)$; deduce $\mathrm{Var}(H) = \alpha(1/3 - \alpha/4)$.
  3. Check both moments on $4 \cdot 10^6$ uniform draws (seed 6).
  4. Compute the one-sided Normal 5% cut-off for H at $T = 250$; simulate 20 000 samples of 250 PITs (seed 12) and compare with their 95th percentile.
- **Report**: the moments, the check, both cut-offs and the actual size of the Normal cut-off
- **Interpretation**: why does H_t carry more information than I_t?
- Notebook section: A9

Back Exercise list

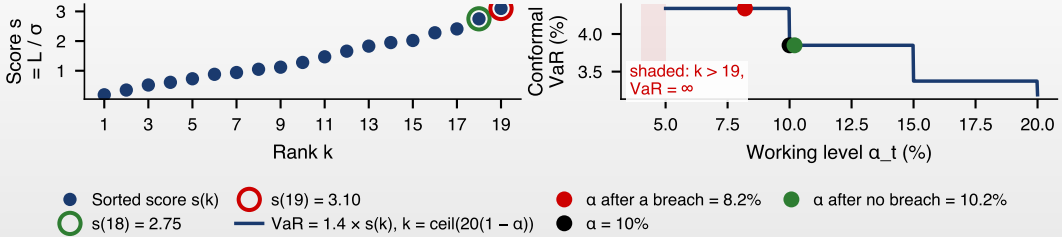
A10 [Solved]: Conformal Quantile and One ACI Step — Task

- **Question**: how does split conformal VaR 10% react to one breach under ACI?
- **Inputs**
  - $n = 19$ calibration scores (sorted, table); tomorrow's $\hat{\sigma} = 1.4\%$; target $\alpha = 0.10$ (fixed); ACI step $\gamma = 0.02$, $\alpha_1 = 0.10$ (Gibbs & Candès (2021))
  - For this one-step illustration the scores and $\hat{\sigma}$ are frozen; if $k > n$ the quantile is $+\infty$
- **Tasks**
  1. Compute $k = \lceil (n+1)(1-\alpha) \rceil$ ($\lceil \cdot \rceil$: round up), $s_{(k)}$ (the k-th smallest score) and $\mathrm{VaR} = \hat{\sigma} s_{(k)}$.
  2. Compute $\alpha_2 = \alpha_1 + \gamma(\alpha - \mathrm{err}_1)$ after a breach and after no breach ($\mathrm{err}_1 = 1$ if day 1 is a breach, else 0).
  3. Recompute k and VaR for both values of α_2; below which α_t does the VaR become infinite?
- **Report**: k, three VaR values, the infinite-VaR threshold
- **Interpretation**: how smoothly does ACI move the VaR?
- Notebook section: A10

k	$s_{(k)}$	k	$s_{(k)}$	k	$s_{(k)}$
1	0.19	8	1.05	15	2.02
2	0.35	9	1.12	16	2.28
3	0.52	10	1.28	17	2.41
4	0.61	11	1.47	18	2.75
5	0.73	12	1.66	19	3.10
6	0.88	13	1.83		
7	0.94	14	1.95		

Back Exercise list

A10 [Solved]: Solution — Conformal Quantile and One ACI Step



- $k = \lceil 20 \cdot 0.9 \rceil = 18$; $s_{(18)} = 2.75$; VaR $= 1.4 \cdot 2.75 = 3.85\%$
- After a breach: $\alpha_2 = 0.10 + 0.02(0.10 - 1) = 0.082$, $k = \lceil 18.36 \rceil = 19$, VaR $= 1.4 \cdot 3.10 = 4.34\%$
- No breach: $\alpha_2 = 0.102$, $k = \lceil 17.96 \rceil = 18$: the VaR does not move; for $\alpha_t < 1/(n+1) = 5\%$, $k = 20 > n$: VaR $= \infty$
- Interpretation: with small n ACI moves in jumps of one order statistic; small changes of α_t are invisible until they cross a grid point $j/(n+1)$

MFM_ch8_sem_conformal

B6 [Proposed]: Conformal VaR on the BET and Bitcoin — Task

- **Question**: does ACI repair crisis coverage on the BET and on Bitcoin, or only the long-run average?
- **Worked example to follow**: B4
- **Inputs**
  - BET daily closes from 03.01.2007 and Bitcoin daily closes (24:00 UTC) from 14.06.2017, to 18.09.2026; VaR 5%: split conformal and ACI on both markets, GARCH-t on the BET only (GARCH forecasts on Bitcoin are withheld, Chapter 8); scores, windows and crisis flag as in B4
  - ACI: $\alpha_1 = 0.05$, $\gamma = 0.005$, started on the first evaluation day; an infinite VaR counts as no breach and is reported separately
- **Tasks**
  1. For each market and method, report breach counts and rates on all, calm and crisis days.
  2. Report the number of days with an infinite ACI VaR.
  3. Attach the Wald binomial 95% interval of B4 (which ignores dependence) to each crisis rate and say which differences from 5% are clear.
- **Report**: a rate-and-count table, the infinite-VaR count, one figure
- **Interpretation**: an evidence-qualified comparison of the three methods, not a single winner
- Public replication benchmark: coverage in calm and crisis periods
- Notebook section: B6

Back Exercise list

B7 [Proposed]: Z_2 by Simulation on the BET and the S&P 500 — Task

- **Question**: are the ES 2.5% forecasts of GARCH-t and GARCH-EVT calibrated on the BET and on the S&P 500?
- **Worked example to follow**: B3
- **Inputs**
  - Forecast panels of the BET and the S&P 500 (both from 2007); Algorithm 2 of B3 with 5000 draws, seed 11, forecasts held fixed
  - Samplers: GARCH-t as in B3; GARCH-EVT $L^* = -\mu_t + \sigma_t\big(u + \frac{\beta}{\xi}\big[\big(\frac{\mathbf{1}-U}{p_u}\big)^{-\xi} - 1\big]\big)$ with the stored GPD threshold u, shape ξ, scale β and share p_u of standardised losses above u
- **Tasks**
  1. On the full sample, build a table with one row per market and model (BET and S&P 500 $\times$ GARCH-t and GARCH-EVT): Z_2, p, 5% critical value, breach days.
  2. Repeat on the last 250 days.
  3. Plot the four null distributions with the observed Z_2.
- **Report**: the eight-row table and one figure
- **Interpretation**: a provisional comparison with its limitations: why does the verdict change between the full sample and the last year?
- Public replication benchmark: results scorecard
- Notebook section: B7

Back Exercise list

B8 [Proposed]: Comparing Models on the BET and the S&P 500 — Task

- **Question**: can the data separate the six models on the BET and on the S&P 500?
- **Worked example to follow**: B2
- **Inputs**
  - Daily FZ0 scores of the six models on the BET and on the S&P 500; DM with Newey–West lags $\lfloor 4(T/100)^{2/9} \rfloor$
  - MCS (Hansen, Lunde & Nason (2011)): T_{max} statistic (the largest standardised mean loss differential among the models still in the set), circular block bootstrap, blocks of 10 days, 1000 replications, seed 2, 90% set
  - Murphy diagram (Ehm, Gneiting, Jordan & Krüger (2016)), the mean elementary score over a grid of thresholds: BET VaR 2.5%, HS versus GARCH-EVT, 240 thresholds from 1.0% to 7.3%, pointwise HAC bands with the same automatic lag rule as DM
- **Tasks**
  1. Find the lowest mean FZ0 in each market and run exploratory DM tests of every other model against it.
  2. Run the MCS; show the first elimination step (statistic, p-value, model removed) and list the 90% set.
  3. Plot the Murphy difference with its bands and say whether the curves cross.
- **Report**: a DM/MCS table, the elimination steps, one figure
- **Interpretation**
  - which models remain in the MCS for each market?
  - what limits the separation?
- Public replication benchmark: models retained by the MCS
- Notebook section: B8

Back Exercise list

The Simulation Starts from a Fully Specified Model

- The daily log returns are decimals. Parameters (full precision): $\mu = 0.0007491951$, $\omega = 1.2150511 \times 10^{-6}$, $a = 0.10233295$, $b = 0.89211334$, $\nu = 6.14504545$.
- Initialize the generating variance at $\omega/(1 - a - b)$. Draw Student-t innovations and multiply by $\sqrt{(\nu - 2)/\nu}$ for unit variance; discard the first 500 returns.
- Generate 400 paths with NumPy `default_rng(9)`. Each retained path has 1000 estimation returns followed by 5000 evaluation returns. Compare both forecast methods on the same path.
- After re-estimation, initialize its variance filter with the population variance of the first 50 estimation returns (divisor 50). Update the filter daily, but hold the fitted parameters fixed during evaluation.

B9 [Proposed]: Kupiec Size Under Estimation Risk — Task

- **Question**: how often does Kupiec reject a correct GARCH-t model when its parameters are estimated?
- **Worked example to follow**: B1
- **Inputs**
  - Fixed Data-Generating Process (DGP): GARCH(1,1)-t fitted to the S&P 500 1990–2026 ($a = 0.102$, $b = 0.892$, $\nu = 6.1$); 400 paths of $R + 5000$ days after a burn-in of 500, seed 9
  - Fixed scheme: estimate once on the first $R = 1000$ days (ν floored at 2.2), then forecast VaR 1% for $P = 250$, 1000, 5000 days; on the same path also the true-parameter VaR
  - A path with $\hat{a} + \hat{b}$ within 10^{-8} of 1, or a failed fit, has no estimated forecast and is left out
- **Tasks**
  1. Count the paths with an admissible fit; on them, record the Kupiec rejection at 5% for the estimated and the true forecasts, paired on each path.
  2. Also report the true-parameter rate on all paths: does the selection change it?
  3. Give each rate $\hat{p}$ its Monte Carlo standard error $\sqrt{\hat{p}(1-\hat{p})/n}$, n: the paths used.
  4. Compare the standard deviation of the breach rate across paths with $\sqrt{\alpha(1-\alpha)/P}$; plot the rates against P/R.
  5. Going further: bootstrap calibration: simulate under the fitted null, re-estimate on each replicate, use the 95th percentile of LR_{uc} as critical value.
- **Report**: a table (P, two rates with standard errors, two standard deviations) and one figure
- **Interpretation**: how does the distortion grow with P/R (Escanciano & Olmo (2010))?
- Public replication benchmark: fixed-scheme estimation risk
- Notebook section: B9

Back Exercise list

C2 [Proposed]: A Provisional Model for a Romanian Trading Book — Task (1/2)

- **Question**: which VaR/ES model would you provisionally recommend for a portfolio in lei, half in the BET and half in euro?
- **Inputs**
  - BET daily closes and the BNR EUR/RON reference rate; align the **prices** on common trading days, then simple returns, then the portfolio log return $r_t^p = \ln(1 + \frac{1}{2} R_t^{BET} + \frac{1}{2} R_t^{EUR})$ (R: simple returns), daily rebalanced, no costs
  - The six models of Chapter 8 (window of 1 000 days); evaluation from 09.07.2009 (after 2008: the 2008 crisis enters estimation windows but is outside the evaluation sample) to 18.09.2026
- **Tasks**
  1. Build the portfolio and plot its losses with the VaR 1% of HS, FHS and GARCH-EVT.
  2. Run the minimum test set (next slide) for all six models.
- Notebook section: C2

Back Exercise list

C2 [Proposed]: Evidence and Recommendation — Task (2/2)

- **Questions**
  - what does the evidence support?
  - what does it not support?
- **Worked example to follow**: B1–B3, B5, B7
- **Inputs**: minimum test set: frequency (Kupiec, CC), timing (durations or DQ), ES (Z_2 by simulation), comparison (FZ0, DM, 90% MCS), regulation (worst and last 250-day windows, desk limits)
- **Tasks**
  1. Fill the evidence table: model $\times$ rate, p_{cc}, p_{dur}, Z_2 / p, worst window, last window (VaR 1% / VaR 2.5%), FZ0 / MCS p.
  2. For each test report simulated size at $T = 250$ (B1); for the frequency test compute power at breach rate 2α as in A1. Discuss the family-wise false-alarm rate; individual 5% sizes do not imply a 5% rate for the set.
  3. Write one provisional recommendation with a challenger model, and list what FRTB approval would also require (backtests on desk profit and loss, P&L; P&L attribution).
- **Report**: a Quantlet folder and one slide: figure, evidence table, recommendation, main limitation
- **Interpretation**: would the choice survive a year like 2008 or 2020?
- Notebook section: C2

C3: The EGARCH Simulation Model

- Use the Exponential GARCH (EGARCH) model with the parameters of Bayer–Dimitriadis, Eq. (2.1): $Y_t = \sigma_t z_t$, zero conditional mean.
- $\log \sigma_t^2 = -0.0012 - 0.161 z_{t-1} + 0.136(|z_{t-1}| - \mathbb{E}|z|) + 0.978 \log \sigma_{t-1}^2$.
- The `rugarch` distribution `std` is Student-t with 7.39 degrees of freedom, scaled to unit variance. Thus $z = \sqrt{(7.39-2)/7.39}\, t_{7.39}$.
- Follow the authors' simulation functions: `ugarchpath` with `n.start=100` and its default initialisation; retain the original coefficients rather than refit them.
- Set `n.sim=750`, in addition to `n.start=100`; discard the first 500 returned observations and evaluate the remaining $n = 250$. Use seeds $1, \ldots, 1000$ for the 1000 paths requested here.
- For a size experiment, compute VaR and ES from the true conditional distribution at each date. Report the package version and the number of failed estimates; do not count failed fits as non-rejections.

C3 [Proposed]: Backtesting ES with ES Forecasts Only — Task (1/2)

- **Question**: do the regression-based ES backtests of Bayer & Dimitriadis (2022) agree with Z_2 and McNeil–Frey on the Chapter 8 forecasts?
- **Inputs**
  - The available ES 2.5% forecasts: six models on the S&P 500 and the BET; HS, Normal and Student-t on Bitcoin and EUR/RON (18 series; GARCH forecasts on Bitcoin and EUR/RON are withheld); returns convention of the paper: $r_t = -L_t$, $\hat{e}_t = -\widehat{\mathrm{ES}}_t < 0$
  - R (4.x) with the authors' package: `esback::esr_backtest(r, e, alpha = 0.025, version = v)`; $v = 1$: Strict Expected Shortfall Regression (ESR) (two-sided), $v = 3$: Intercept ESR (one-sided H_0: $\gamma_1 \geq 0$, γ_1: intercept of $r_t - \hat{e}_t$, negative when ES is underestimated); misspecification-robust covariance (default)
- **Tasks**
  1. Work one case first: S&P 500, GARCH-t: print the coefficients, their standard errors, the statistic and the p-value of both versions.
  2. Run both versions on all 18 series and draw a market-by-model heat map of the p-values.
  3. Compare the decisions at 5% with Z_2 and McNeil–Frey (Chapter 8, section "Backtesting Expected Shortfall").
- Notebook section: C3

Back Exercise list

C3 [Proposed]: Blocks and Size — Task (2/2)

- **Questions**
  - how do the tests behave on one-year blocks?
  - what is their size at $T = 250$?
- **Worked example to follow**: B3
- **Inputs**: non-overlapping 250-day blocks of the S&P 500 from the first evaluation day (drop the incomplete last block); the EGARCH(1,1)-t process of Eq. 2.1 of the paper, 1000 paths, seeds 1 through 1000
- **Tasks**
  1. Report the share of blocks in which each model is rejected, and the number of failed fits (excluded from the share).
  2. Simulate 1000 paths of 250 days, backtest the true ES 2.5% with the Intercept ESR, two-sided and one-sided, and report the rejection rates with their Monte Carlo standard errors.
  3. Compare the two-sided rate with the 0.04 of Table 1 (robust covariance) and the one-sided rate with the nominal 5% (Table 2 reports it only at $T = 2\,500$).
- **Report**: a Quantlet folder and two slides: the decision heat map with Z_2 and McNeil–Frey, and the block shares with the size check
- **Interpretation**
  - when ESR and Z_2 disagree, which would you trust as a supervisor who receives only ES forecasts?
  - why would you trust it?

Acronyms used in this material (1/2)

- **ACF**: AutoCorrelation Function
- **ACI**: Adaptive Conformal Inference
- **AI**: Artificial Intelligence
- **AR**: AutoRegressive (model)
- **BCBS**: Basel Committee on Banking Supervision
- **BET**: Bucharest Exchange Trading (index)
- **BNR**: Banca Națională a României — *the National Bank of Romania*
- **BVB**: Bursa de Valori București — *the Bucharest Stock Exchange*
- **CLT**: Central Limit Theorem
- **DGP**: Data-Generating Process
- **DM**: Diebold–Mariano (test)
- **DQ**: Dynamic Quantile (test)
- **EGARCH**: Exponential GARCH
- **ES**: Expected Shortfall
- **ESR**: Expected Shortfall Regression (backtest)
- **EVT**: Extreme Value Theory
- **FHS**: Filtered Historical Simulation
- **FRTB**: Fundamental Review of the Trading Book

Acronyms used in this material (2/2)

- **FZ0**: Fissler–Ziegel loss with zero-homogeneous score differences
- **GARCH**: Generalised AutoRegressive Conditional Heteroskedasticity
- **GARCH-EVT**: GARCH filter combined with Extreme Value Theory for the tail
- **GJR**: Glosten–Jagannathan–Runkle (asymmetric GARCH)
- **GPD**: Generalised Pareto Distribution
- **GW**: Giacomini–White (test)
- **HAC**: Heteroskedasticity and Autocorrelation Consistent (standard errors)
- **HS**: Historical Simulation
- **LR**: Likelihood Ratio (test)
- **LRV**: Long-Run Variance
- **MCS**: Model Confidence Set
- **MLE**: Maximum Likelihood Estimation
- **OLS**: Ordinary Least Squares
- **PIT**: Probability Integral Transform
- **POF**: Proportion Of Failures (test)
- **US**: United States
- **UTC**: Coordinated Universal Time

References (1/4)

- Acerbi, C., Szekely, B. (2014). Backtesting expected shortfall. MSCI Inc., December 2014.
- Basel Committee on Banking Supervision (1996). Supervisory framework for the use of "backtesting" in conjunction with the internal models approach to market risk capital requirements. Bank for International Settlements.
- Basel Committee on Banking Supervision (2019). Minimum capital requirements for market risk (rev. February 2019). Bank for International Settlements.
- Bayer, S., Dimitriadis, T. (2022). Regression-based expected shortfall backtesting. *Journal of Financial Econometrics*, 20(3), 437–471.
- Christoffersen, P.F. (1998). Evaluating interval forecasts. *International Economic Review*, 39(4), 841–862.
- Christoffersen, P., Pelletier, D. (2004). Backtesting value-at-risk: a duration-based approach. *Journal of Financial Econometrics*, 2(1), 84–108.

References (2/4)

- Diebold, F.X., Mariano, R.S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3), 253–263.
- Du, Z., Escanciano, J.C. (2017). Backtesting expected shortfall: accounting for tail risk. *Management Science*, 63(4), 940–958.
- Ehm, W., Gneiting, T., Jordan, A., Krüger, F. (2016). Of quantiles and expectiles: consistent scoring functions, Choquet representations and forecast rankings. *Journal of the Royal Statistical Society, Series B*, 78(3), 505–562.
- Engle, R.F., Manganelli, S. (2004). CAViaR: conditional autoregressive value at risk by regression quantiles. *Journal of Business & Economic Statistics*, 22(4), 367–381.
- Escanciano, J.C., Olmo, J. (2010). Backtesting parametric value-at-risk with estimation risk. *Journal of Business & Economic Statistics*, 28(1), 36–51.
- Fissler, T., Ziegel, J.F. (2016). Higher order elicitability and Osband's principle. *Annals of Statistics*, 44(4).

References (3/4)

- 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.
- Gibbs, I., Candès, E. (2021). Adaptive conformal inference under distribution shift. *Advances in Neural Information Processing Systems*, 34; arXiv:2106.00170.
- Gneiting, T. (2011). Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106(494), 746–762.
- Hansen, P.R., Lunde, A., Nason, J.M. (2011). The model confidence set. *Econometrica*, 79(2), 453–497.
- Kupiec, P.H. (1995). Techniques for verifying the accuracy of risk measurement models. *Journal of Derivatives*, 3(2), 73–84.

References (4/4)

- 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.
- Newey, W.K., West, K.D. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55(3), 703–708.
- 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., Mazurencu-Marinescu-Pele, M. (2026). Finite-sample precision limits for expected shortfall forecast comparisons. *Mathematics*, 14(13), 2316.