クオンツ短期トレード予測モデルの正当性・頑健性 — 検査指標・バックテスト手法・注意点の網羅サーベイ

作成日: 2026-07-16　|　**対象:** 短期（クオンツ）トレードの予測モデル（ニューラルネット等）の正当性・頑健性の検査

本サーベイの目的: 過去データで良好なバックテスト結果が得られた予測モデルについて、それが「本物の予測力」なのか「時系列へのカーブフィッティング／既知リスクファクターへの便乗／偶然の産物」なのかを見分けるための、検査指標・統計的検定・バックテスト設計・確率的リスク評価・実務上の注意点を、一次情報（査読論文・定評ある教科書・公式ドキュメント）に基づいて網羅的に整理したもの。

構成（6本柱）:
1. 過学習とカーブフィッティングの検出（§2–§3）
2. バックテスト過剰最適化と多重検定の補正（§4–§5）
3. ファクター中立なアルファの正当性検証（§6）
4. リスク指標とリターン分布の確率的推定（§7–§9）
5. バイアス・取引コスト・レジューム頑健性（§10–§12）
6. 金融機械学習の実務フレームワーク（§13）

注意（免責）: 本書は生成時点で取得・照合した公開情報に基づく技術サーベイであり、投資助言ではない。各主張には本文内引用を付し、末尾に統合参考文献（123件）を掲げた。数式は可読性のためプレーンテキスト／コード表記で示している。引用URL・書誌情報はリサーチ時点で実在を確認したものだが、リンク切れや版の更新が生じうる。

エグゼクティブサマリー

本稿は、短期クオンツ予測モデルが示す良好なバックテスト成績が、再現可能なスキル（アルファ）なのか、多数の試行から偶然選ばれた「偽の発見（false discovery）」なのかを識別するための、正当性・頑健性の検査体系を横断的に整理する。中核の問いは終始一貫して「その数字は、試した試行数・多重性・コスト・レジーム変化に耐えるか」であり、各章はこれを異なる角度から締め上げる道具立てを与える（全体像と運/スキルの分離は第1章）。

(1) 過学習・カーブフィッティングの見抜き方：バイアス-バリアンス分解と学習曲線で「データ不足の過学習かモデル力不足の過小適合か」を診断し、正則化でバリアンスを削る。ただし二重降下が示すとおり「訓練誤差ゼロ＝過学習」は現代の過剰パラメータ化モデルには当てはまらず、判定は必ずOOS指標で行う（第2章）。金融時系列ではk-foldがリークで楽観化するため、パージング・エンバーゴでラベル重複と系列相関を塞ぎ、単一経路のウォークフォワードでなくCPCVで性能の分布を得るのが要点である（第3章）。

(2) 多重検定・過剰最適化の補正：真のシャープがゼロでも試行数Nを増やせば期待最大シャープは膨張するため、単一検定のt>2.0は甘すぎ、多重性を織り込んだ実質t>3.0が要る（第1章・第5章）。Deflated Sharpe Ratioは選択バイアスと非正規性を同時に割り引き、PBOは選択プロセスの健全性を、MinBTLは事前の設計制約を担う——三者は相補的で、いずれも試行数Nを正しく数えることに依存する（第4章）。手元に各戦略のp値しかなければBonferroni/Holm/BHY、PnL系列があればReality Check/SPA/StepMで見かけの有意性を検定し直す（第5章）。

(3) ファクター中立なアルファ検証：戦略リターンをFF5・Carhart・q-factorなど既知ファクター群に時系列回帰し、残る切片（アルファ）が有意な正かをスパニング検定で確認する。GRS検定で全アルファ同時ゼロを、Barra因子に中立化した残差のrank ICで真の予測力を分離し、その超過リターンが「既知プレミアムのリパッケージ」でないことを示す（第6章）。

(4) リスク指標とリターン分布の確率的推定：シャープ比はIID下でも大きな標準誤差を持ち、系列相関で年率値が最大65%膨らむため、SE・信頼区間・MinTRLを添え、ほぼ全指標が操作可能な点はMPPMで検算する（第7章）。リターンは重い裾を持つので、正規VaRでなくGARCH+EVT/GPDで条件付けたES、経路リスクは最大ドローダウン分布・CDaRで評価する（第8章）。そして点推定に頼らず、時系列依存を保つブロックブートストラップ・順列検定で指標の分布そのものを推定する（第9章）。

(5) バイアス・コスト・レジーム頑健性：生存者・先読み・データスヌーピング等の「7つの大罪」とAHMの7項目プロトコルで、成績の水増し源を体系的に点検する（第10章）。グロスのシグナルが統計的に本物でも、凹なマーケットインパクト・キャパシティ制約・公開後のアルファ減衰（平均で約6割）を織り込むとネットでは消えうる（第11章）。加えて市場は非定常であり、構造変化検定・レジームスイッチング・分数次差分・ウォークフォワードで頑健性を測り、真のOOSはライブ運用のみと心得る（第12章）。

(6) 金融ML実務手法：López de Prado の「MLファンドが失敗する10の理由」を、認識論・データ処理・分類問題・評価の4層のゲートとして用いる。ドルバー・分数次差分・トリプルバリア・メタラベリング・一意性重み付け・パージCV・CPCV・DSRを上流から下流へ相補的に積み上げ、各層の既知の罠を項目単位で潰す（第13章）。

要するに本体系は、単一指標・点推定・素朴な年率化・試行数の非報告という四つの落とし穴を排し、運と偽の発見を取り除いて、実運用で持ちこたえるスキルだけを選別するための一貫した検査手順を与える。以降の各章は、この手順を具体的な数式・閾値・実務チェックリストへ落とし込む。

1. 序論：予測モデル検証の全体像と「偽の発見」問題

短期クオンツ予測モデルの検証で最初に立ちはだかるのは、観測された良好な成績が「運(luck)」の産物なのか、それとも再現可能な「スキル(alpha)」なのかという識別問題である。バックテストのシャープレシオ(Sharpe ratio)やファクターの t 比が高く見えても、それが多数の試行の中からたまたま選ばれた見かけ上の勝者であれば、実運用(out-of-sample)では消えてしまう。本セクションは、この「偽の発見(false discovery)」問題の全体像を概観し、本稿が扱う各手法が「何をどうチェックできるか」を位置づける。

運(luck)とスキル(alpha)を分離する2つの道具

Harvey, Liu & Zhu (2016) は、運とスキルを分離する手段を「アウトオブサンプル検定(out-of-sample test)」と「多重検定の統計枠組み(multiple testing framework)」の2つに整理する。前者について同論文は、実行可能な場合にはアウトオブサンプル検定が偽ファクターを排除する最もクリーンな方法だとしつつ、金融では真の OOS 検定に未来のデータが必要で年単位の時間を要し、ホールドアウト法もデータが既に全て観測済みである以上は真の OOS ではない、という致命的弱点を指摘する。OOS 検定の代表例が McLean & Pontiff (2016) のアノマリー減衰研究である。リアルタイムで即座に判断したい実務家にとってこの時間的制約は決定的であり、そのため本稿は、いま手元にある試行結果だけで判断できる多重検定フレームワークを主軸に据える。

「偽の発見」の定義：FWER と FDR

偽の発見の管理には2つの誤り指標がある。Harvey, Liu & Zhu (2016) は、少なくとも1つの偽発見が生じる確率である FWER(family-wise error rate)と、全発見のうち偽発見が占める期待割合である FDR(false discovery rate)を区別する。前者を厳しく抑えるのが Bonferroni・Holm 補正、後者を制御するのが Benjamini-Hochberg-Yekutieli(BHY)補正であり、どの誤りを許容するかで越えるべき有意閾値が変わる。単一戦略の是非を問うなら FWER、多数の戦略群から採否を選別するなら FDR、という使い分けになる。

有意基準の引き上げ：t>2.0 では足りない

Harvey, Liu & Zhu (2016) は少なくとも 316 個の公表ファクターを収集し、「これほど広範なデータマイニングの下では、新規ファクターに通常の有意基準(例: t 比>2.0)を用いることは経済的にも統計的にも意味をなさない」と述べ、推奨ハードルを t 比>3.0 に引き上げた(理論ベースのファクターはやや低くてよいが、それでも 2.0 では不十分)。この閾値は時間とともに上昇しており、Bonferroni 調整 t 比は 1965 年の 1.96 から 2012 年には 3.78(p 値 0.02%)、2032 年には 4.00(p 値 0.01%)に達すると予測される。FDR を制御する BHY implied t 比は 2010 年以降 3.39(p 値 0.07%)で安定し、有意水準を α_d=5% に緩めても閾値は 2.78〜2.81 と、結論として「多重性を考慮した 5% 有意の最低 t 比は約 2.8(p 値 0.5%)」となる (Harvey, Liu & Zhu, 2016)。参考として Fama-MacBeth(1973)の市場ベータの t 比は 2.57 で 3.0 を下回り、物理学のヒッグス粒子発見は t 比>5(p<0.0001%)を要求する。

公表結果の過半は偽発見か：再現性危機(replication crisis)

この基準を過去の実証研究に当てはめると事態は深刻である。Harvey, Liu & Zhu (2016) は 296 個の公表有意ファクターのうち偽発見と判定される数を Bonferroni で 158 個、Holm で 142 個、BHY(1%)で 132 個、BHY(5%)で 80 個と推定し、医学の Ioannidis (2005)「Why Most Published Research Findings Are False」を引きつつ、金融のファクターの多くも偽発見である可能性が高いと結論づける。独立した再現研究もこれを裏づける。Hou, Xue & Zhang (2020) はアノマリー 452 変数を統一手続きで再現し、時価総額で市場の 3.2% ながら銘柄数の 60% を占めるマイクロキャップを NYSE ブレークポイントと時価総額加重で抑えると、452 中 65% が単一検定ハードル |t|≥1.96 すら越えられず、多重検定ハードル 2.78(5% 有意)を課すと失敗率は 82% に上昇すると報告した(元研究の多くがマイクロキャップを等加重・最小二乗で過大加重していたことが主因)。McLean & Pontiff (2016) は 97 個の特性(characteristics)を対象に、アウトオブサンプル(公表前だが標本期間外)で予測力が平均 26%、公表後には平均 58% 減衰すると報告した(26% はデータマイニング/統計的バイアス由来の減衰の上限、公表後との差の約 32% が情報を得た投資家の価格圧力による分に相当する)。同研究の初期草稿(82 特性)では、統計的バイアスによる OOS 減衰は平均約 10%(統計的にゼロと非有意)、公表後の減衰は平均約 35% とし、両者から公表効果の下限を約 25% と見積もっていた。

バックテスト過学習(backtest overfitting)と試行回数 N

これらの「偽の発見」を機械的に生み出す元凶がバックテスト過学習である。Bailey, Borwein, López de Prado & Zhu (2014) は、真の OOS シャープが 0 でも、N 個の独立試行から得られる最大シャープの期待値が

```
E[max_N] ≈ (1−γ)·Z⁻¹(1−1/N) + γ·Z⁻¹(1−1/(Ne)),  γ≈0.5772(オイラー・マスケローニ定数)
```

で近似でき、上界は √(2·ln N) だと示した。例えば N=10 通りの設定を試すだけで、全戦略の真の OOS シャープが 0 でも、インサンプル(IS)で最大シャープ 1.57 の戦略が見つかると期待される。裏返せば、スプリアスな高シャープを避けるのに必要な最小バックテスト長(MinBTL)は

```
MinBTL(年) ≈ 2·ln[N] / E[max_N]²  （簡便な上界）
```

で、データが 5 年しかなければ 45 通り以上、2 年なら 7 通りの試行で、真の OOS シャープ 0 のまま年率 IS シャープ 1 の戦略をほぼ確実に生成してしまう(MinBTL は過剰適合回避の必要条件であって十分条件ではない) (Bailey, Borwein, López de Prado & Zhu, 2014)。したがって、試行した設定数 N を報告しないバックテストは過剰適合リスクを評価できず、実務家がまず問うべきは「幾つの設定を試したか」である。この選択バイアスと収益の非正規性を同時補正する指標が Deflated Sharpe Ratio(DSR)であり、Bailey & López de Prado (2014) は本質的に Probabilistic Sharpe Ratio(PSR)の棄却閾値を試行の多重性に合わせて調整したもので、観測シャープを、収益の歪度 $\hat{\gamma}_3$・尖度 $\hat{\gamma}_4$、系列長、N 試行下でのシャープ推定量の分散から計算される期待最大シャープだけ割り引くと定式化した。

土台にある p 値(p-value)の誤用

以上の補正は、そもそも p 値の解釈を正すことなしには機能しない。Harvey (2017) は AFA 会長講演で、トップ誌の限られた掲載枠を巡る競争が有意な結果を生む誘因となり、未報告の検定・多重検定の未調整・直接/間接の p-hacking が組み合わさって公表結果の多くが将来持ちこたえないと論じた。核心は、p 値が帰無仮説の真偽を示すのではなく、帰無仮説が真の下でその効果以上を観測する確率 p(D|H0) であって p(H0|D) ではないという点であり、「帰無を反証した」「効果が真である確率」等のありがちな解釈はいずれも誤りである。代替として同論文は最小ベイズ因子(minimum Bayes factor; MBF)MBF=exp(−t²/2) を提案し、例えば t=2.6 なら MBF=0.034、事前オッズを 1:1(五分五分、even odds)とすれば帰無が真の確率は 0.034/(1+0.034)=0.033 と「ベイズ化 p 値(Bayesianized p-value)」を得る。ただし、有意閾値を単純に引き上げるだけではデータマイニングや公表バイアスをかえって増やす副作用がありうる点も同論文は警告している (Harvey, 2017)。実データからの裏づけとして、Chordia, Goyal & Saretto (2020) は 200 万を超えるランダム取引戦略を生成し、多重検定(multiple hypothesis testing; MHT)を制御する t 統計量閾値を時系列回帰で 3.8、クロスセクションの Fama-MacBeth 回帰で 3.4 と推定した(Benjamini-Yekutieli 手続きで偽発見割合(false discovery proportion; FDP)を 5% に制御)。多重検定を考慮しない場合の偽棄却の期待割合は約 45%(45.3%)に上る。

このセクションで何をチェックできるか

本稿の手法群を用いれば、短期予測モデルの検証において次を具体的にチェックできる。第一に、報告された t 比やシャープが、試行回数 N と多重検定を織り込んだ閾値(単一検定の 1.96 ではなく 2.8〜3.8 の水準)を越えているか。第二に、データ長が MinBTL を満たし、E[max_N] や DSR に照らして観測シャープを選択バイアス・非正規性の分だけ割り引いても正のままか。第三に、p 値を p(H0|D) と取り違えず、MBF による事前オッズとの掛け合わせでも結論が頑健か。第四に、対象アノマリーが公表後減衰や再現失敗の傾向を持たないか。これらを通じて運と偽の発見を排し、実運用で持ちこたえるスキルだけを選別するのが以降の各セクションの目的である。

2. 機械学習モデルの過学習検出（一般論）

過学習 (overfitting) とは、モデルが訓練データの信号だけでなくノイズや偶然の変動まで再現してしまい、未知データ (out-of-sample; OOS) での汎化性能が劣化する状態を指す。短期クオンツ予測モデルは、シグナル対ノイズ比 (S/N) が極端に低く、かつ多数の特徴量・ハイパーパラメータ・戦略候補を探索するため、過学習が最も起きやすい応用領域である。本セクションでは、過学習を「検出」するための一般理論と診断ツールを、(1) 基礎理論（バイアス-バリアンス分解）、(2) 診断ツール（学習曲線・複雑度上界）、(3) 抑制手法（正則化）、(4) 現代的修正（二重降下）、(5) 金融特有の落とし穴（交差検証のリーク・多重検定）の順に整理する。

バイアス-バリアンス分解：過学習の基礎理論

過小適合と過学習を切り分ける出発点は、予測モデルの期待二乗誤差の分解である (Geman, Bienenstock & Doursat (1992))。

$$E[(y - \hat{f}(x))^2] = \mathrm{Bias}[\hat{f}]^2 + \mathrm{Var}[\hat{f}] + \sigma^2$$

高バイアス (high bias) は学習アルゴリズムの誤った仮定により特徴と目的変数の関連を捉え損ねる状態（＝過小適合, underfitting）、高バリアンス (high variance) は訓練データの微小変動やノイズにまで敏感に反応する状態（＝過学習）、σ^2 は問題固有で除去不能な既約誤差 (irreducible noise) である。モデルを複雑にするほどバイアスは下がるがバリアンスは上がるというトレードオフが生じる (Geman, Bienenstock & Doursat (1992))。実務的含意は、OOS 誤差が大きいときに「モデルが単純すぎる（バイアス）」のか「複雑すぎる（バリアンス）」のかを診断し、前者ならモデル拡張、後者なら正則化・データ追加という逆方向の処方を選び分けることにある。

学習曲線・検証曲線による過学習診断

バイアスとバリアンスのどちらが支配的かは、訓練スコアと検証スコアの水準とギャップのパターンから機械的に診断できる (scikit-learn (2024))。判定則は次のとおり：訓練スコアも検証スコアも低い→過小適合、訓練スコアが高く検証スコアが低い→過学習、両方高い→良好、訓練スコアが低く検証スコアが高いのは通常起こりえない (scikit-learn (2024))。この診断には二種の曲線を使う。validation_curve は単一ハイパーパラメータを横軸に訓練/検証スコアを描き最適値（過学習に転じる直前）を探す。learning_curve は訓練サンプル数を横軸にとり、高バリアンスモデル（例：小データの SVM）は当初「訓練スコア≫検証スコア」だがサンプル増でギャップが縮小しデータ追加が有効、高バイアスモデル（例：ナイーブベイズ）は両者が低い値へ収束しデータ追加は無効、と示す (scikit-learn (2024))。短期予測モデルの検証では、まず learning_curve でギャップの縮小傾向を見て「データ追加で改善する過学習か、モデル自体が力不足の過小適合か」を切り分けるのが第一手になる。

汎化誤差の理論的上界：Rademacher 複雑度と VC 次元

過学習リスクは、モデルが属する関数クラス F の複雑度で汎化ギャップを上界づけることで事前に評価できる (Mohri et al. (2018))。経験 Rademacher 複雑度は、関数クラスがランダム符号とどれだけ相関できるかを測る。

```
Rad_S(F) = (1/m)·E_σ[ sup_{f∈F} Σ_i σ_i f(z_i) ]    (σ_i は ±1 を等確率でとる)
sup_{f}( L_P(f) − L_S(f) ) ≤ 2·Rad_S(F) + O( √(ln(1/δ)/m) )    (確率 1−δ 以上)
```

Massart の補題より有限集合（サイズ N）で `Rad ≤ √(2·logN/m)`、VC 次元 d の族では `Rad ≤ O(√(d/m))` となる (Mohri et al. (2018))。VC 次元境界はデータ非依存で 0/1 分類の最悪ケース向け、Rademacher はデータ分布依存で回帰・多クラス・カーネル法・深層学習にも使え、より緊密な境界を与える。いずれも標本数 m が小さい／複雑度が高いほど境界が緩み、過学習リスクが増すことを定量化する (Mohri et al. (2018))。実務では、特徴量数やモデル容量を増やす前に「手元の標本数 m でその複雑度が汎化ギャップを許容範囲に収めるか」を桁感覚で見積もる根拠として使う。

正則化による過学習抑制：dropout・early stopping・ノイズ注入

過学習が診断されたら、バリアンスを下げる正則化 (regularization) を導入する。第一に dropout は、訓練時に各ユニットを保持確率 p で無作為に脱落させてニューロン同士の過度な共適応 (co-adaptation) を防ぐ手法で、訓練は指数個の間引きネットワーク (thinned network) からのサンプリングに相当し、テスト時は重みを p でスケールした単一ネットでその平均を安価に近似する。複数ドメインで他の正則化を上回る過学習削減を達成したと報告された (Srivastava et al. (2014))。第二に早期終了 (early stopping) は、パラメータを 0 付近の初期値から始め、検証標本の誤差が増加に転じた時点（通常、訓練誤差が最小化される前）で最適化を止める手法で、特定条件下では L2 罰則（weight decay、重みを 0 へ縮小）と等価になり、より低い計算コストの代替になる (Gu, Kelly & Xiu (2020))。第三に、入力へのノイズ注入 (data augmentation の一種) は、二乗和誤差の場合にネットワーク写像の一階微分のみを含む半正定値 (positive semi-definite) な一般化 Tikhonov 正則化子に帰着することが示されており、ノイズ注入・データ拡張が過学習抑制に効く理論的根拠を与える (Bishop (1995))。実務では、L1 が接続を疎に・L2 が重みを 0 へ縮小する役割分担のもと、これらを併用してバリアンスを段階的に削るのが定石である (Gu, Kelly & Xiu (2020))。

二重降下と「複雑さの美徳」：古典的直観の限界

「訓練誤差ゼロ＝過学習」という素朴な判定は、現代の過剰パラメータ化モデルには当てはまらない。古典的な U 字型バイアス-バリアンス曲線は、モデル容量（パラメータ数 N）が標本数 n に達する補間閾値 (interpolation threshold) でテストリスクがピークを打ち、そこを超えて容量をさらに増やすとテストリスクが再び低下する二重降下 (double descent) へ拡張される (Belkin et al. (2019))。機構は、より大きな関数クラスの中からより小さいノルム（＝より滑らかで「単純」）な補間解を選べる暗黙のバイアスであり、ランダムフーリエ特徴・ニューラルネット・ランダムフォレストで普遍的に観測される (Belkin et al. (2019))。二重降下はモデルサイズだけでなく訓練エポック数・訓練標本数の各軸で現れ、有効モデル複雑度 (effective model complexity; EMC) が訓練標本数に近い臨界領域ではテスト誤差がピークをつけ、「訓練

標本数を 4 倍に増やしてもテスト性能がむしろ悪化する」反直観的な領域が生じる。含意として、より大きなモデル/長い訓練/多いデータが常に汎化を改善するとは限らず、early stopping が最悪のピークを回避しうる (Nakkiran et al. (2019))。金融でも同型の現象が理論的に示されている。パラメータ数 P が観測数 T を上回る高複雑度領域 (P>T) では、真のデータ生成過程 (DGP) を近似するにすぎない高次元線形モデルでも（比較的広い仮定の下で）複雑度の増加が期待 OOS 予測精度とポートフォリオ性能を単調に (strictly increasing) 高めうる。この結果は最小 $\ell 2$ ノルム補間解を与えるリッジレス回帰 (ridgeless regression; 無限小の縮小) の時点で既に成り立ち、最適なリッジ縮小はそれをさらに改善する——いわゆる「複雑さの美徳 (virtue of complexity)」である (Kelly, Malamud & Zhou (2024))。過学習検出の実務への含意は、訓練データを完全に補間する（訓練誤差ゼロの）モデルでも汎化しうるため、「訓練誤差がゼロだから棄却」ではなく OOS 指標そのもので判定すべきだという点にある。

金融時系列特有の落とし穴：交差検証のリークと purge/embargo/CPCV

標準的な k-fold 交差検証 (cross-validation) は観測が i.i.d. と仮定するが、金融データは強い時間依存を持つため、そのまま適用すると情報リーク (leakage) が生じ過学習した楽観的な評価になる (López de Prado (2018))。テスト標本のラベルが訓練標本の特徴/ラベルと時間的に重なると訓練↔テスト間で情報が漏れるため、三つの補正が必要になる。Purging＝テスト集合のラベルと時間的に重複する訓練観測を除去。Embargo＝purge で取り切れないテスト直後のごく短い一定期間（実装では全観測に対する割合として指定する）の訓練観測をさらに除外。Combinatorial Purged CV (CPCV)＝訓練/テスト分割を組合せ的に多数生成し、単一パスでなく複数のバックテスト経路（例：1 つの履歴系列から 5 経路）と性能分布を得て頑健な統計的推論を可能にする (López de Prado (2018))。短期戦略の検証では、まず purge/embargo でリークを塞いだうえで CPCV により性能の分布を得て、単一の楽観的な数字ではなく分布の下側で戦略を評価するのが要点である。

多重検定とバックテスト過学習：DSR・PBO・Diebold-Mariano

多数の戦略・モデルを試すと、真の予測力が皆無でも最良の Sharpe 比 (Sharpe ratio; SR) は偶然で膨張する（勝者の呪い, winner's curse）。N 個の独立試行下での期待最大 SR は、真の SR がゼロでも N とともに増大する (Bailey & López de Prado (2014))。

```
E[max{SR_n}] ≈ E[SR] + √V[SR]·( (1−γ)·Z⁻¹[1−1/N] + γ·Z⁻¹[1−(1/N)e⁻¹] )    (γ ≈ 0.5772)
DSR = Z[ (SR − SR_0)·√(T−1) / √(1 − γ̂₃·SR + ((γ̂₄−1)/4)·SR²) ]
```

Deflated Sharpe Ratio (DSR) は、閾値 SR_0 を上式の期待最大 SR にとり、歪度 $\gamma\hat{3}$・尖度 $\gamma\hat{4}$・標本長 T・試行の SR 分散 V[SR]・独立試行数 N の 5 変数で補正する（DSR<0.95 なら 95% 信頼水準に届かず棄却）(Bailey & López de Prado (2014))。ホールドアウトや k-fold CV は試行数を制御しないため単独ではバックテスト過学習を防げず、20 回適用すれば 95% 水準の偽陽性は例外でなく想定内 (expected) になる。金融系列はメモリ効果を持ち、過学習した戦略は将来に損失最大化へ反転しうるため、最も欠けている決定的情報は試行回数 (number of trials) であり、これを非パラメトリックに評価するのが Probability of Backtest Overfitting (PBO) である (Bailey & López de Prado (2014))。モデル間の統計的優劣の検定にも同じ多重比較の補正が要る。予測精度差の Diebold-Mariano 検定では、12 モデル比較を Bonferroni 補正すると片側臨界値が 2.64 まで上がり、ニューラルネットの線形モデルに対する優位が有意

から「限界的に有意 (marginally significant)」へ後退する (Gu, Kelly & Xiu (2020))。多数モデルを比較するほど選択バイアスが入るため、優劣主張には多重検定補正が必須である。

実証的教訓と前提・限界

低 S/N 環境が過学習に与える帰結は実証的にも顕著である。1957–2016 の米国株・約 30,000 銘柄・94 特性（マクロ変数との交差項等で全 920 特徴量に拡張）を用い、正則化なしで全 920 特徴量を投入した OLS は月次 OOS R^2=−3.46%（定数ゼロ予測にすら劣る過学習の典型）となる一方、正則化した非線形モデルでも最良は NN3 の 0.40% にとどまり、4〜5 層に増やしても改善しない（深さの恩恵は限定的）(Gu, Kelly & Xiu (2020))。市場効率がリターン変動を予測不能なニュースに支配させる低 S/N 環境ゆえに、罰則付き正則化が不可欠だという結論である (Gu, Kelly & Xiu (2020))。

前提と限界として次を共有しておく。第一に、バイアス-バリアンス分解や VC/Rademacher 上界は関数クラスと損失に関する仮定の上に立ち、二重降下が示すとおり過剰パラメータ化域では古典的な単調トレードオフの直観が崩れる (Belkin et al. (2019); Nakkiran et al. (2019))。第二に、学習曲線・検証曲線は i.i.d. を暗黙に仮定するため、金融時系列では purge/embargo/CPCV を併用しない限りリークで楽観化する (López de Prado (2018))。第三に、DSR・PBO・多重検定補正はいずれも独立試行数 N を正しく数えられることに依存し、N の過少報告は指標を容易に甘くする (Bailey & López de Prado (2014))。実務チェックリストとしては、(1) learning_curve でバイアス/バリアンスを診断、(2) dropout・early stopping・L1/L2・ノイズ注入でバリアンスを削る、(3) purge/embargo/CPCV でリークを塞いだ OOS 分布を取る、(4) DSR<0.95・PBO・Bonferroni 補正済み Diebold-Mariano で試行数依存の過学習と見かけの優劣を棄却する、の四段構えを組み合わせるのが、短期クオンツ予測モデルにおける過学習検出の標準運用である。

3. 時系列クロスバリデーション（Walk-Forward / Purged K-Fold / CPCV / Embargo）

短期クオンツ予測モデルの検証では、「どのデータで訓練し、どのデータで評価するか」という分割設計そのものが結果の妥当性を決める。ランダムに分割する標準 k-fold クロスバリデーション (cross-validation) は独立同分布 (i.i.d.) データを前提とするが、金融時系列はこの前提を満たさない。本セクションでは、時間順序とラベルの重なりを尊重する検証法——Walk-Forward、Purged K-Fold、Embargo、およびそれらを組合せる CPCV——を、定義・数式・実務上の使い方・限界の順に扱う。これらは「バックテスト結果が特定の歴史的偶然や情報リーク (information leakage) に依存していないか」を具体的にチェックするための道具立てである。

標準 k-fold が金融時系列で破綻する3機構

標準 k-fold が金融データでリーク (leakage) を起こす機構は3つある。第一に、ラベルが重複する時間ウィンドウ（例：triple-barrier 法）から作られるため観測が i.i.d. でない。検定ラベル Y_j が情報集合 Φ_j に依存し、同じ Φ_j に依存する訓練ラベルが訓練集合に残ると情報が漏れる。第二に、財務特徴量は ARMA 過程のように系列相関 (serial correlation) するため、検定期間の直後に位置する訓練観測もリーク源になる。第三に、k-fold は時間順序を無視してシャッフルするため、未来のデータで学習して過去を予測する look-ahead が生じる。結果として k-fold は、リーク（look-ahead バイアスとラベル重複）により性能推定を過度に楽観化し、バックテスト成績を著しく過大評価する (López de Prado (2018), AFML 第7章)。したがって金融では、時間順序・ラベル重複・系列相関の3点に対処した分割が必須になる。

Walk-Forward 分析（anchored / rolling）

Walk-Forward 分析は、検定集合を常に訓練集合より未来に置く前進検証 (forward chaining) である。scikit-learn の TimeSeriesSplit は既定で拡大窓 (expanding window) を採り、「k 番目の分割では最初の k フォールドを訓練集合、(k+1) 番目のフォールドを検定集合とし、後続の訓練集合は先行するものの上位集合になる」 (scikit-learn docs)。これが訓練起点を固定する anchored walk-forward である。訓練サイズに上限 (max_train_size) を課すと固定長のスライド窓 (sliding / rolling window) になる。さらに gap パラメータ（「各訓練集合の末尾から検定集合の手前までに除外する観測数」）で訓練と検定の間に分離帯を作れ、これは後述の embargo に相当する分離を与える (scikit-learn docs)。手法の起源は Robert Pardo が 1992 年の『Design, Testing and Optimization of Trading Systems』で提示した walk-forward 分析にあり、第2版『The Evaluation and Optimization of Trading Strategies』(2008) で拡張された (Wikipedia (Walk-forward optimization))。

```
分割 k：訓練 = フォールド 1..k、　検定 = フォールド (k+1)
（拡大窓＝anchored、max_train_size 上限＝rolling、gap＝embargo 的分離）
```

実務では、パラメータ再最適化を一定間隔で回しながら常に「学習済みの過去 → 未評価の未来」の順で性能を積み上げるため、時間順序を自然に守れる利点がある。ただし本質的な限界は、Walk-Forward が検証で

きるのが市場史の単一経路 (a single path) に過ぎない点である。別のレジーム・ストレス条件・シナリオでの頑健性を評価できず、良好な結果が特定の歴史的偶然に依存していないかを確かめられないため、容易に過剰最適化 (overfitting) する (López de Prado (2018), AFML 第12章)。

Purging（パージング）

Purged K-Fold は、k-fold にパージングとエンバーゴを加えてリークを塞いだものである。パージングは「訓練集合から、検定集合のラベルと時間的に重なる (overlap) すべての観測を除去する」操作と定義される (skfolio docs)。形式的には、検定ラベル Y_j が区間 [t0, t1] の情報に依存するとき、ラベル区間がこの [t0, t1] と重なる訓練観測をすべて取り除く。実装上も「訓練集合は検定ラベル区間と重なる観測をパージされる」とされ、根拠は AFML 第7章にある (mlfinlab; López de Prado (2018), AFML (notes))。これによりラベル重複による第一のリーク機構を直接遮断できる。

Embargo（エンバーゴ）

エンバーゴは、系列相関への対処としてパージングに追加する除去操作である。定義は「検定集合の観測の直後に続く訓練観測を、財務特徴量が ARMA 過程のような系列相関を持つ系列を含むことが多いため、訓練集合から除去する」ことである (skfolio docs)。実装では除去幅を割合パラメータ（mlfinlab の pct_embargo＝「エンバーゴのサイズを決める割合」）で指定する (mlfinlab)。具体例として、資産配分の応用研究ではパージング・エンバーゴともに 21 営業日（約1営業月）を設定している (arXiv:2407.17645 (2024))。パージングがラベル重複を、エンバーゴが系列相関を担うことで、標準 k-fold の3機構のうち残る2つを塞ぐ。両者を課した k-fold が Purged K-Fold であり、これが CPCV の構成ブロックになる。

CPCV（Combinatorial Purged Cross-Validation）

CPCV は、Purged K-Fold を組合せ的に拡張し、単一経路の限界を克服する。データを N 群に分割し、そのうち k 群を検定に使う全組合せを列挙する。訓練/検定の分割 (split) 数は C(N, k) 通りであり、各群が複数の分割で検定側に現れることで、複数の純粋アウトオブサンプル経路 (backtest paths) が生成される。経路数の公式は次式で与えられる (RiskLab AI)。

```
φ[N,k] = (k/N)·C(N, N−k) = ∏_{i=1}^{k−1}(N−i) / (k−1)!
```

k=2 が「スイートスポット」で $\phi[N,2] = N-1$ となる (RiskLab AI)。この式から、例えば N=6, k=2 では C(6,2)=15 分割・5 経路が得られる。実際の応用では 45 の訓練組合せ・36 のバックテスト経路、パージ+エンバーゴ各 21 日、平均訓練≈3.5 年・平均検定≈2 年という構成が用いられている (arXiv:2407.17645 (2024))。Walk-Forward が単一経路しか作れないのに対し、CPCV は多数の代替歴史経路の分布を作る点が本質的な違いである。

CPCV の統計的な効き目は、経路数 ϕ を増やすことで平均パフォーマンス推定量の分散を下げる点にある (RiskLab AI)。

```
σ²(μ̄_i) = φ⁻¹·σ²_i·(1 + (φ−1)·ρ̄_i)
```

ϕ が大きいほど（経路が多いほど）Walk-Forward のような単一経路法より分散が大幅に減り、Sharpe 比を単一点でなく複数経路の分布として評価できる。ただし経路間相関 $\bar{\rho}$ が高いと削減効果は限定的になる (RiskLab AI)。実務的には、CPCV は「Sharpe 比の点推定」を「Sharpe 比の分布」に置き換え、悪い経路の裾（ストレス時の劣化やドローダウン）まで含めて戦略の頑健性を点検できる。

過剰最適化の観点からの評価と手法選択

なぜ Walk-Forward の単一経路では不十分で CPCV の経路分布が要るのかは、選択バイアス (selection bias) の定量から裏づけられる。真の Sharpe 比がゼロでも独立試行数 N を増やせば試行中の最大 Sharpe 比の期待値は上昇し、その大きさは概ね $\sqrt{(2 \cdot \ln N)}$ のオーダーで上に抑えられる（極値分布の近似では、真 SR=0 でも N=10 で期待最大 SR≈1.57、7 個の二値パラメータで $N=2^7=128$ なら期待最大 Sharpe>2.6）。加えてメモリ効果 (memory) 下では「IS を最適化するほど OOS が悪化する」負の関係が生じうる (Bailey, Borwein, López de Prado & Zhu (2014))。単一の Sharpe 比推定は Lo (2002) の漸近分布 $\hat{SR} \to^d N(SR, (1+SR^2/2)/T)$（T は観測数）が示すとおり推定ノイズが大きく、点評価は信頼できない (Bailey, Borwein, López de Prado & Zhu (2014))。単純ホールドアウト (hold-out) も、公開データでは設計に既に使われている可能性・研究者が OOS 期間の市場挙動を知り無意識に混入する可能性・小標本（週次戦略なら 20 年未満）では使えない点・単一 hold-out の分散が大きく別の hold-out で結論が反転しうる点から金融では不適切である (Bailey, Borwein, López de Prado & Zhu (2015))。これらの弱点を、検定集合を対称に総当たりで入れ替えて過剰最適化確率 (Probability of Backtest Overfitting; PBO) を測る CSCV が補い、相対ランク $\bar{\omega}_c$ から logit $\lambda_c = \ln(\bar{\omega}_c/(1-\bar{\omega}_c))$ を作り、$\lambda_c<0$ の確率（IS で最良の構成が OOS で中央値を下回る確率）を PBO とする。PBO が高いほど過剰最適化を示し、0.5 は IS の選抜が OOS に対して無情報であることに対応する (Bailey, Borwein, López de Prado & Zhu (2015))。手法選択の実証指針として、Heston 確率ボラティリティ・Merton ジャンプ拡散・レジームスイッチ等の合成統制環境で K-Fold・Purged K-Fold・CPCV・Walk-Forward を比較した研究は、CPCV が過剰最適化緩和で顕著に優越（より低い PBO・より高い Deflated Sharpe Ratio 検定統計量）する一方、Walk-Forward は偽発見防止で最も劣り時間的変動が大きく定常性が弱いと報告している (Arian, Norouzi & Seco (2024))。

前提・限界と注意点

第一に、パージングとエンバーゴはいずれも各観測のラベル生成区間（予測ホライズン）を正確に把握できることを前提とする。ラベル区間を過少に見積もれば重複除去が不十分になり、リークが残る。第二に、CPCV は分割数が C(N, k) と組合せ的に増えるため、群数 N を大きくすると計算コストが急増する。第三に、分散削減効果は経路間相関 $\bar{\rho}$ に依存し、$\bar{\rho}$ が高い（経路が似通う）ほど ϕ を増やす見返りは小さくなる (RiskLab AI)。第四に、これらの CV 法はいずれも「試した独立試行数 N を数えて最小バックテスト長 (Minimum Backtest Length; MinBTL) 以下に抑える」規律と併用しなければ過剰最適化を防げない。データが5年しかなければ独立な構成を 45 個以下に、2年なら僅か7構成に抑えないと、真の OOS Sharpe=0 の無スキル戦略がほぼ確実に生成される。試行数 N を報告しないバックテストは、そもそも過剰最適化リスクを評価不能にする (Bailey, Borwein, López de Prado & Zhu (2014))。

実務上のチェックリストとしては、(1) ラベルの予測ホライズンを明示し、パージ+エンバーゴでラベル重複と系列相関を塞ぐ、(2) 単一経路の Walk-Forward で頑健性を結論づけず、CPCV で複数のアウトオブサン

プル経路の分布（特に悪い経路の裾）を確認する、(3) 得られた経路群に対し PBO を評価し（値が高いほど過剰最適化を疑う）、同時に MinBTL の観点から試行数 N が手元のデータ長に見合うかを検証設計の段階で先に決める、の三点を組合せる。パージ/エンバーゴがリークを、CPCV が単一経路の脆さを、PBO/MinBTL が選択バイアスを担当するため、これらは代替ではなく相補的に用いるのが正しい運用である。

4. バックテスト過剰最適化の定量化（Deflated Sharpe Ratio / PBO / 最小バックテスト長）

短期クオンツ戦略の検証では、「良好なバックテスト結果」そのものが証拠にならない。多数の候補をスクリーニングすれば、投資スキルが皆無でも高い Sharpe 比 (Sharpe ratio) を示す戦略は必ず見つかるからである。したがって「このバックテストは過剰最適化 (overfitting) か否か」は真偽二択ではなく、試した試行数に依存する非ゼロ確率の問題として扱う必要がある。95% 水準の hold-out 検証を約 20 回適用すれば、無効な戦略が 1 つ通過するのは偶然ではなく期待どおりの帰結であり、hold-out や k-fold クロスバリデーション (cross-validation) はこの試行数依存性を無視するため過剰最適化を防げない (Bailey & López de Prado (2014))。本セクションでは、この確率を実際に測るための Deflated Sharpe Ratio (DSR)、Probability of Backtest Overfitting (PBO)、最小バックテスト長 (Minimum Backtest Length; MinBTL) を扱う。

選択バイアスの定量化：期待最大 Sharpe 比

出発点は「勝者の呪い (winner's curse)」の定量化である。真の Sharpe 比がゼロの無スキル戦略でも、独立試行数 N を増やすほど「試行の中の最大 Sharpe 比」の期待値は上昇する。極値理論 (Gumbel 分布) から、標準化 Sharpe 比の期待最大値は次式で近似される (Bailey & López de Prado (2014))。

```
E[max{SR_n}] ≈ E[{SR_n}] + √V[{SR_n}] · ( (1−γ)·Z⁻¹[1 − 1/N] + γ·Z⁻¹[1 − 1/(N·e)] )
```

ここで $\gamma \approx 0.5772156649$ はオイラー・マスケローニ定数 (Euler–Mascheroni constant)、Z^{-1} は標準正規分位関数、V[{SR_n}] は試行間の Sharpe 比推定値の分散、N は独立試行数である。E[{SR_n}]=0, V=1 でも N=10 で約 1.57、N=100 で 2.5 超まで膨張する (Bailey & López de Prado (2014))。実務上の含意は明快で、候補モデルを大量にスクリーニングした事実を無視して最良戦略の Sharpe 比だけを報告するのは、選択バイアス (selection bias) をそのまま利益として計上していることに等しい。

Deflated Sharpe Ratio (DSR)

DSR は、選択バイアス（多重検定）と非正規性 (non-normality) の両方を同時に補正した有意性指標であり、DSR>0.95 のとき「本物の発見」とみなす (Bailey & López de Prado (2014))。定義は Probabilistic Sharpe Ratio (PSR) の基準 Sharpe 比を多重検定補正後の閾値 SR_0 に置き換えたものである。

```
DSR ≡ PSR(SR_0) = Z[ (SR̂ − SR_0)·√(T−1) / √(1 − γ̂_3·SR̂ + (γ̂_4−1)/4·SR̂²) ]
```

Z は標準正規累積分布関数、SR^は選択戦略の（非年率換算）推定 Sharpe 比、T はサンプル長、γ^3 は歪度 (skewness)、γ^4 は尖度 (kurtosis) である。棄却閾値 SR_0 は「N 個の無スキル戦略から得られる期待最大 Sharpe 比」であり、前節の式そのもの、すなわち `SR_0 = √V[{SR_n}]·((1−γ)Z⁻¹[1−1/N] + γZ⁻¹[1−1/(N·e)])` で与えられる。要するに DSR は、標準 Sharpe 比（平均と標準偏差という 2 推定量の

関数）を、非正規性 ($\gamma\hat{}3$, $\gamma\hat{}4$)・系列長 T・試行間分散 V・独立試行数 N という追加 5 変数で「デフレート（割引）」する (Bailey & López de Prado (2014))。分母の歪度・尖度項は、負の歪度や過大な尖度が見かけ上 Sharpe 比を膨らませる効果を補正する。

論文の数値例は実務的な直感を与える。年率 Sharpe 比 2.5 を主張する国債オークション季節性戦略でも、N=100 試行・V=1/2・T=1250（5 年日次）・$\gamma\hat{}3=-3$・$\gamma\hat{}4=10$ を入力すると SR_0≈0.1132（非年率）、DSR≈0.9004<0.95 となり 95% 水準で棄却される。仮に試行が N=46 だけなら DSR=0.9505 で採用ライン超え、リターンが正規（$\gamma\hat{}3=0$, $\gamma\hat{}4=3$）なら N=88 試行まで採用ラインを維持できた——つまり非正規性は許容試行数を大きく削る (Bailey & López de Prado (2014))。

DSR を回す実務手順は、(1) 最終戦略に至るまでの独立試行数 N を数える、(2) リターンの歪度・尖度と系列長を測る、(3) 上式で SR_0 と DSR を計算し 0.95 と比較する、である。試行が独立でなく相関を持つ場合は、平均ペア相関 $\rho\hat{}$から実効独立試行数 $\hat{N} = \hat{\rho} + (1-\hat{\rho})\cdot M$（M は総試行数、$\rho\in(-1/(M-1), 1]$）を求め、名目 M ではなく N^を N に用いる。$\rho\to1$ で N^→1、$\rho\to0$ で N^→M となり、相関を無視すると E[max] を過大評価する（PCA など次元削減で独立試行数を出す代替法もある）(Bailey & López de Prado (2014))。

PSR と最小実績長 (MinTRL)

DSR の土台にある PSR は、単一戦略の Sharpe 比が基準 c を超える確率を返す (Bailey & López de Prado (2012))。

$$PSR(c) = Z\left[(\hat{SR} - c) / SE(\hat{SR}) \right]$$
$$SE(\hat{SR}) = \sqrt{ (1 - \gamma_3\cdot\hat{SR} + (\gamma_4-1)/4\cdot\hat{SR}^2) / (T-1) }$$

この標準誤差は Lo (2002) の Sharpe 比の漸近分布に基づく。Lo (2002) は年率 Sharpe 比の推定量が漸近的に $\hat{SR} \to^d N(SR, (1 + SR^2/(2q))/y)$ に従うことを示し（q は年間観測数、y は推定年数）、歪度・尖度を含めた一般形が上記 SE である。IID 正規（$\gamma\_3=0$, $\gamma\_4=3$）では $\sqrt{((1 + SR^2/2)/(T-1))}$ に帰着する。この式が、短いサンプル（小さい T）と非正規リターンが Sharpe 比推定を上方に膨らませることの数理的根拠であり、PSR・DSR の分母そのものである (Lo (2002))。

有意化に必要な最短実績長が Minimum Track Record Length (MinTRL) である (Bailey & López de Prado (2012))。

$$MinTRL(c) = 1 + (1 - \gamma_3\cdot\hat{SR} + (\gamma_4-1)/4\cdot\hat{SR}^2) \cdot \left(z_{1-\alpha} / (\hat{SR} - c) \right)^2$$

z_{1−α} は信頼水準 1−α の標準正規臨界値（α=0.05 で 95%）。SR^と c が近いほど、また歪度が負・尖度が大きいほど必要観測数が増える。実務では「主張する Sharpe 比を統計的に確立するには最低何本の実績データが要るか」を事前に見積もり、それに満たない track record は結論保留とする使い方ができる (Bailey & López de Prado (2012))。

Probability of Backtest Overfitting (PBO) と CSCV

DSR が単一の選択戦略を評価するのに対し、PBO は「戦略選択プロセス」そのものの健全性をモデル非依存・ノンパラメトリックに測る。定義は、インサンプル (IS) で最良の戦略が、アウトオブサンプル (OOS) で全戦略の中央値を下回る確率である (Bailey, Borwein, López de Prado & Zhu (2017))。

```
PBO = Σ_{n=1}^N Prob[ r̄_n < N/2 | r ∈ Ω*_n ] · Prob[ r ∈ Ω*_n ]
```

r̄n は戦略 n の OOS 順位、Ω*_n は「n が IS で最良」となる事象である。ここでの IS/OOS は戦略選択に用いる観測部分集合であって、回帰の推定期間（モデル較正）ではない点が要点で、これゆえ取引ルールや予測式の知識を要さずに定義できる (Bailey, Borwein, López de Prado & Zhu (2017))。

推定は Combinatorially Symmetric Cross-Validation (CSCV) で行う。各列が 1 試行の PnL 系列である (T×N) 行列 M を作り、行方向に偶数 S 個の等サイズ非重複部分に分割し、S/2 個ずつ取る全組合せ C(S, S/2) を列挙する（例：S=16 なら C(16,8)=12,870 通り。原論文のプレプリント版は「12,780」と誤記）。各組合せで一方を訓練集合、補集合を検定集合とし、両者で N 戦略の順位を計算する。IS 最良戦略 n* の OOS 相対順位 $\omega_c = \bar{r}_{n^*}/(N+1) \in (0,1)$ から logit `λ_c = ln(ω_c/(1−ω_c))` を求め、全組合せにわたる負側の質量を PBO とする (Bailey, Borwein, López de Prado & Zhu (2017))。

```
PBO = φ = ∫_{−∞}^0 f(λ) dλ   （λ_c ≤ 0 の相対頻度）
```

$\lambda_c > 0$ は OOS でも中央値超え（過剰最適化なし）、$\lambda_c \le 0$ は IS 最良が OOS で中央値以下を意味する。$\phi \approx 0$ は過剰最適化なし、$\phi \approx 1$ は高度な過剰最適化で、慣行 (Neyman–Pearson) では PBO>0.05 のモデルを棄却する。IS/OOS を対称に総当たりで入れ替えるため「検定集合をどう選ぶか」という hold-out の恣意性を回避でき、Sharpe 比に限らず任意のパフォーマンス指標に適用できる (Bailey, Borwein, López de Prado & Zhu (2017))。

同じ CSCV 分布から副次的な診断も得られる。IS 最良戦略の (IS 性能, OOS 性能) 対を `R̄_{n*} = α + β·R_{n*} + ε` に回帰すると、傾き β は実務ではたいてい負——過剰最適化はメモリ効果 (memory effect) により将来性能を最小化する。さらに OOS 損失確率 `Prob[R̄_{n*}<0]` も算出できる。$\phi \approx 0$ でも損失確率が高いことはあり得る（過剰最適化以外の理由で戦略が弱い場合）ため、両者を併読する (Bailey, Borwein, López de Prado & Zhu (2017))。

最小バックテスト長 (MinBTL)

MinBTL は、検証を始める前に「N 試行を行うなら、真の OOS Sharpe 比がゼロの無スキル戦略を偶然選ばないために最低何年のバックテストが要るか」を与える設計指標である (Bailey, Borwein, López de Prado & Zhu (2014))。

```
MinBTL ≈ ( [ (1−γ)·Z⁻¹[1−1/N] + γ·Z⁻¹[1−1/(N·e)] ] / E[max_N] )²  <  2·ln[N] / E[max_N]²
```

E[max_N] は許容する（偶然でも達成されうる）IS 期待最大年率 Sharpe 比である。実務的近似 `MinBTL < 2·ln[N]/E[max_N]²` を覚えておけばよい。数値の目安として、5 年のデータでは独立モデル構成を 45 個以

下に抑えないと、IS 年率 Sharpe=1・OOS 期待 Sharpe=0 の戦略が生成されてしまう。わずか 7 構成でも 2 年バックテストなら IS 期待最大 Sharpe=1 に達し、二項独立パラメータで $N=2^7=128$ まで増やせば IS 期待最大 Sharpe>2.6 になる。モデルを複雑化してパラメータを増やすと独立試行 N=2^(パラメータ数) が指数的に膨らみ、過剰最適化が容易になる (Bailey, Borwein, López de Prado & Zhu (2014))。教訓は「選択した構成に至るまでの試行数 N を報告しない限り、過剰最適化リスクは評価不能」であり、MinBTL は過剰最適化回避の必要条件（十分条件ではない）である。

前提・限界と注意点

これらの手法を使ううえで共有すべき前提と限界がある。第一に、DSR・MinBTL・E[max] のいずれも独立試行数 N（相関がある場合は実効試行数 $N^\wedge$）を正しく数えられることに依存する。N は「最終戦略に至るまでに試した全構成の数」であり、これを過少報告すれば指標は容易に甘くなる。したがって検証パイプラインでは、パラメータ探索・特徴量選択・閾値調整を含む全試行を記録し N として渡すことが不可欠である (Bailey & López de Prado (2014); Bailey, Borwein, López de Prado & Zhu (2014))。

第二に、これらは分布の前提と近似の上に立つ。E[max] と MinBTL は極値理論の近似であり、DSR/PSR の標準誤差は Lo (2002) の漸近分布に依拠するため、極端に短いサンプルや強い系列相関のもとでは精度が落ちる (Lo (2002))。第三に、金融時系列はメモリ効果を持つため、最も極端なランダムパターンに過剰適合した戦略は、そのパターンが将来「巻き戻される」ことで、OOS で単に劣後するのでなく系統的に損失を出しうる。これが多くのシステマティック運用が宣伝どおり機能しない一因であり、CSCV の傾き β が負に出る背景でもある (Bailey & López de Prado (2014); Bailey, Borwein, López de Prado & Zhu (2017))。

実務上のチェックリストとしては、(1) 最終戦略に対し独立試行数 N と歪度・尖度・系列長を入力して DSR>0.95 を確認する、(2) PnL 行列全体に CSCV をかけて PBO≤0.05 と OOS 損失確率・劣化傾き β を確認する、(3) 検証設計の段階で MinBTL を計算し、手元のデータ長に対して許容できる試行数の上限を先に決める、の三点を組み合わせる。DSR は単一の勝者を、PBO は選択プロセス全体を、MinBTL は事前の設計制約を担当するため、三者は代替ではなく相補的に用いるのが正しい運用である。

5. データスヌーピングと多重検定補正

短期クオンツ戦略の検証では、「1 つの戦略を 1 回だけ検定する」という古典的統計の前提がほぼ常に破れている。同じ価格系列に対してパラメータ・特徴量・閾値を変えながら何百・何千もの候補を試すからである。本セクションは、この「同一データの多重使用」が生む偽陽性を、(1) 全候補の探索を丸ごと 1 つの帰無仮説に押し込むブートストラップ検定（White の Reality Check、Hansen の SPA、Romano-Wolf のステップワイズ法）と、(2) p 値・t 値・Sharpe 比を試行数で割り引く解析的補正（Bonferroni／Holm／BHY、t>3.0 閾値、ハイカット Sharpe 比）の 2 系統で制御する方法を扱う。要するに「大量スクリーニングの末に見つけた最良戦略の見かけの有意性を、試行数を織り込んで検定し直せるか」を本セクションでチェックできるようにする。

データスヌーピングとは何か（定義と発生機序）

データスヌーピング (data snooping) とは、「あるデータ集合を推論またはモデル選択の目的で 2 回以上使うこと。こうしたデータの再利用が起きると、得られた良好な結果が手法固有の価値ではなく単なる偶然に起因する可能性が常に残る」と定義される (Sullivan, Timmermann & White (1999); White (2000))。その本質は選択バイアス (selection bias) であり、Jensen & Bennington (1970) が指摘したとおり、十分な計算時間があれば乱数表の上でさえ「機能する」機械的トレーディングルールを必ず見つけられる。重要なのは、これが単一研究者の意図的な行為に限らないことである。トレーディングルールの母集団全体に対する生存者バイアス (survivorship bias) としても発生し、文献で生き残って報告されたルールだけを見れば、探索空間全体の失敗が視界から消えている (Sullivan, Timmermann & White (1999))。

問題の実体を測った代表例が STW である。Sullivan, Timmermann & White (1999) は Brock-Lakonishok-LeBaron (1992) の 26 ルールを拡張し、7,846 個のトレーディングルールの母集団を構成して DJIA の 100 年（1897–1996）に適用した。平均リターンと Sharpe 比の両方で評価すると、BLL の 90 年サンプル（1897–1986）では一部ルールがデータスヌーピング調整後も残ったが、追加の 10 年アウトオブサンプル (out-of-sample; OOS, 1987–1996) では「最良ルールがベンチマークを上回らない確率が約 12%」となり、従来の有意水準ではテクニカルルールの経済的価値の証拠は乏しかった。S&P500 先物（1984 年以降 13 年）でも上回る証拠はなく、市場効率性の向上と解釈された (Sullivan, Timmermann & White (1999))。この「調整前は有意、調整後は消える」という落差こそ、多重検定補正が捉える対象そのものである。

White の Reality Check（BRC）：全モデル探索を 1 つの帰無仮説にする

White (2000) の Reality Check（Bootstrap Reality Check; BRC）は、「M 個の候補モデルの中で最良のものでもベンチマークを上回らない」という帰無仮説をブートストラップで検定する。性能指標を ϕ_k（k=1,…,M、ベンチマーク比）とすると帰無仮説は次の複合仮説になる (White (2000))。

```
H0: max_{k=1,…,M} φ_k ≤ 0
```

検定統計量は正規化サンプル平均の最大値である。

```
V̄_n = max_{k=1,…,M} √n · f̄_k ,   f̄_k = (1/n) Σ_t f_{k,t}
```

トレーディングでは 1 期損益を `f_{k,t} = ln(1 + y_t · s_{k,t−1})`（s ∈ {1, 0, −1} はロング／中立／ショートのシグナル）で測り、ベンチマーク比で評価する。p 値は Politis & Romano (1994) の定常ブートストラップ (stationary bootstrap) で計算する。すなわちリサンプル系列から `V̄*_n(j) = max_k √n(f̄*_k(j) − f̄_k), j=1,…,B` を多数生成し、White (2000) の Corollary 2.4 により $\bar{V}_n$ と $\bar{V}_n$ の分布が漸近的に等価であることを用いて、$\bar{V}_n$ を経験分布 $\bar{V}_n$ の分位点と比較して p 値を得る (White (2000); Hsu, Hsu & Kuan (Step-SPA))。この枠組みの利点は、検定全体の family-wise 誤りを制御しつつ M 個すべてのモデルの探索を統計量に織り込む点であり、各 p 値を独立に扱う Bonferroni と違って M が大きくても機能する (White (2000))。実務では、パラメータ格子上の全戦略の PnL 系列をそのまま $\bar{V}_n$・V^*_n に流し込めば、「最良戦略の超過リターンが、試した戦略数を考慮しても偶然で説明できないか」を単一の p 値で判定できる。

White RC の保守性と Hansen の SPA 検定

Reality Check には検出力上の弱点がある。White (2000) は複合仮説に対し最小有利配置 (Least Favorable Configuration; LFC) として全モデルの平均 μ=0 を選び、`√n·d̄ →ᵈ N(0, Ω)` から帰無分布 `max{N(0, Ω)}` を得る。ところが Hansen (2005) は、いくつかの $\mu_i<0$ かつ少なくとも 1 つ $\mu_i=0$ のとき、真の極限分布は $\mu_j<0$ の行・列を削除した部分行列 Ω_0 による `max{N(0, Ω_0)}` であり、平均ゼロのモデルだけに依存して劣悪モデルには依存しないことを示した。したがって「非常に劣悪なモデルを候補に含めると RC の経験的 p 値が人為的に膨らみ」、検出力が失われる (Hansen (2005); Hsu, Hsu & Kuan (Step-SPA))。パラメータ格子には必然的に無価値な戦略が大量に混ざるため、これは短期戦略スクリーニングで実際に起きる問題である。

Hansen (2005) の Superior Predictive Ability（SPA）検定は、この LFC を回避して検出力を回復する。統計量自体は RC とほぼ同一で `SPA_n = max( max_{k=1,…,m} √n·d̄_k, 0 )` だが、novel feature は帰無分布の再センタリングにある。$\hat{\Omega}$の対角要素から $\sigma_k^2 = \omega_{kk}$を取り、閾値と選別を次で定義する (Hansen (2005))。

```
A_{n,k} = −ω̂_k · √(2 log log n)
μ̂_k = d̄_k · 1( √n·d̄_k ≤ A_{n,k} )
```

$\mu_k=0$ のとき $\hat{\mu}_k=0$ a.s.、$\mu_k<0$ のとき確率 1 で `√n·d̄_k ≤ A_{n,k}` となり $\hat{\mu}_k \to_p \mu_k$ となる。ブートストラップ分布 `√n(d̄ − μ)` に `√n·μ̂` を加え直すことで、劣悪モデルを帰無分布から実質的に排除し、`max{N(0, Ω_0), 0}` へのより良い近似を得る。結果として SPA の p 値は RC の p 値より小さくなる＝検出力が高い。Hansen はさらにスケール差の影響を除くため studentized 統計量 `√n·d̄_k / ω̂_k` の使用を推奨する (Hansen (2005))。

Romano-Wolf のステップワイズ法（StepM）

White の BRC は最良モデル 1 つしか棄却できないが、実務では「ベンチマークを上回る戦略群」をできるだけ多く同定したい。Romano & Wolf (2005) のステップワイズ法（StepM／Step-RC）は、Family-

Wise Error（FWE）＝「少なくとも 1 つの戦略を誤って優れていると判定する確率」を漸近的に制御しつつ、単一ステップの Reality Check より多くの偽帰無仮説を棄却する。手続きは反復的である。まず各戦略のパラメータ $\theta_s \le 0$（s がベンチマークを上回らない）を帰無仮説とし、$\sqrt{n}\cdot\bar{d}_k$ を降順に並べる。修正 BRC を第 1 ステップとして、統計量がブートストラップ臨界値を超えた戦略を棄却する。棄却した戦略を候補から除去し、残りデータで再ブートストラップして次の臨界値を求める——これを棄却が起きなくなるまで繰り返す (Romano & Wolf (2005))。統計量間の同時依存構造をブートストラップが暗黙に捉えるため、相関した戦略群を過度に罰しない。有限標本では studentized 統計量 $z_{T,s} = w_{T,s} / \hat{o}_{T,s}$ の使用が推奨される。ただし Step-RC の臨界値は依然 LFC で決まるため保守性が残り、この点は Hansen の SPA 系の再センタリングが補う (Romano & Wolf (2005))。

解析的な多重検定補正：FWER と FDR

ブートストラップ検定が使えない場面（各戦略の p 値や t 値だけが手元にある場合）では、解析的な多重検定補正を使う。まず 2 つの誤り率を区別する。Harvey & Liu (2015) の定義では、M 個の仮説のうち R 個を棄却し、そのうち N_r 個が偽陽性（no-skill を誤って有益と判定）だとすると、Family-Wise Error Rate と False Discovery Rate は次のとおりである。

```
FWER = Pr(N_r ≥ 1)
FDP  = N_r / R  (R>0, それ以外は 0),   FDR = E[FDP]
```

FWER は「1 つの誤りも許さない」設計、FDR は「誤りの割合を制御する」設計であり、FDR 制御は FWER 制御より緩い。FWER を制御する Bonferroni や Holm は試行数に関わらず偽陽性を全排除しようとするため、偽陽性数が試行数に比例して増えることを許す FDR 制御（BHY）より厳しくなる。宇宙開発のような致命的用途には FWER が適切だが、資産運用者は偽発見が試行数とともに増えることを受け入れてよい、というのが基本的な使い分けである (Harvey & Liu (2015))。

FWER を制御する 2 手法は次のように定義される。p 値を昇順に $p_{(1)} \le \cdots \le p_{(M)}$ と並べ、Bonferroni は各 p 値を M 倍、Holm は順序統計量で逐次調整する (Harvey & Liu (2015); Holm (1979))。

```
p_(i)^Bonferroni = min[ M · p_(i), 1 ]
p_(i)^Holm       = min[ max_{j≤i} { (M − j + 1) · p_(j) }, 1 ]
```

常に $p_{(i)}^{Holm} \le p_{(i)}^{Bonferroni}$ なので Holm は一様に Bonferroni より強力（緩い）である。数値例（M=6、順序 p 値 0.005, 0.009, 0.0128, 0.0135, 0.045, 0.06）では、Bonferroni 調整後は (0.03, 0.054, 0.0768, 0.081, 0.270, 0.36) で 5% 有意は第 1 のみ、Holm は第 1・第 2 の 2 つが有意となる。両者とも FWER を事前指定水準以下に保証する (Harvey & Liu (2015))。

FDR を制御するのが Benjamini-Hochberg-Yekutieli（BHY）法である。最大 p 値から降順に逐次構成し、依存構造に対応する正規化定数 c(M) を用いる (Harvey & Liu (2015); Benjamini & Yekutieli (2001))。

```
p_(i)^BHY = p_(M)                                                  (i = M)
p_(i)^BHY = min[ p_(i+1)^BHY, (M · c(M) / i) · p_(i) ]              (i ≤ M−1)
c(M) = Σ_{j=1}^M (1/j)
```

Benjamini & Hochberg (1995) は c(M)≡1（p 値が独立または正の依存のとき有効）、Benjamini & Yekutieli (2001) は c(M)=Σ1/j（任意の依存構造で有効）を採用する。M=6 では c(M)=2.45。同じ数値例で BHY 調整後は (0.0496, 0.0496, 0.0496, 0.0496, 0.06, 0.06) となり最初の 4 つが 5% 有意——Holm より 2 つ、Bonferroni より多い発見になる。c(M)>1 は閾値を厳しくして発見を減らす方向に働き、依存構造を許す代償として保守化する (Harvey & Liu (2015))。実務では、戦略群の相関が弱く見逃しコストが高いなら FDR（BHY）、1 件の偽採用が致命的なら FWER（Holm）を選ぶ。

発見の閾値：t>3.0 とハイカット Sharpe 比

これらの補正を実際のファクター発見に当てはめると、単一検定の慣用 t>2.0 では甘すぎることが分かる。Harvey, Liu & Zhu (2016) は 250 本の掲載論文＋63 本のワーキングペーパー＝313 論文、316 ファクターをカタログ化した。ファクター発見率は 1980–91 が約 1 本／年、1991–2003 が約 5 本／年、直近 9 年は約 18 本／年（過去 9 年で 164 ファクター）と加速している。α_w=5%（Holm, FWER）と α_d=1%（BHY, FDR）を置くと、Bonferroni 基準 t は 1.96 から 2012 年に 3.78、2032 年に 4.00 へ上昇し（単一検定 p 値は 3.78 で 0.02%、4.00 で 0.01%）、BHY 基準 t は 2010 年以降 3.39（p=0.07%）で安定する。総合すると「多重性を考慮した 5% 有意の最小 t は約 2.8（単一検定 p≈0.5% 相当）」で、新ファクターは t>2.0 ではなく実質 t>3.0 を要する。参考に Fama & MacBeth (1973) の市場 β は t=2.57 で当時の 2.0 基準を超えていた (Harvey, Liu & Zhu (2016))。

さらに「未観測の試行」＝publication bias を補正すると閾値は上がる。Harvey, Liu & Zhu (2016) は全試行ファクターの 71% が未観測と推定し、Bonferroni／Holm 基準を 4.01／3.96、BHY 基準を 1% 有意で 3.39→3.68、5% 有意で 2.78→3.18 へ引き上げた。頑健性チェック（過大計上を疑って 316→113 ファクターに減らしても Holm は 3.29、真の母集団が標本の 2 倍と仮定しても Bonferroni は 3.78→3.95・Holm は 3.64→3.83）を通じても「t>3.0」という一般的メッセージは不変である (Harvey, Liu & Zhu (2016))。

同じ考え方を Sharpe 比に落とし込んだのがハイカット Sharpe 比 (haircut Sharpe ratio) である。Harvey & Liu (2015) は、多重検定調整後 p 値 p^M を単一検定 Sharpe 比に逆変換して割引後 Sharpe を求める。独立試行なら `p^M = 1 − (1 − p^S)^N`（p^S は単一検定 p 値、N は試行数）で膨らんだ p 値を得たうえで、`p^M = Pr(|r| > HSR·√T)` を解いてハイカット後の HSR を得る（t と Sharpe は `SR = t-ratio / √T` で結ばれる）。例として T=240（月次 20 年）・年率 SR=0.75・p^S=0.0008 でも、N=200 なら p^M=0.15 まで悪化し、調整後年率 SR は 0.32、ハイカットは約 60%=(0.75−0.32)/0.75 に達する。主張の核心は「割引は非線形で、高 Sharpe 比は緩く、限界的な Sharpe 比を重く罰する」ことにあり、慣用の『Sharpe 比を一律 50% 割引』ルールは不適切である——年率 Sharpe<0.4 ではハイカットはほぼ常に 50% 超、Sharpe>1.0 では最大でも 25% にとどまる (Harvey & Liu (2015))。実例（Exhibit 1）では、E/P（年率 SR 0.43, t=2.99）は N=50 で 50.0%・N=100 で 61.6%、MOM（SR 0.67, t=4.70）は 19.2%・23.0%、BAB（SR 0.78, t=7.29）は 7.9%・9.3% と、元の Sharpe 比と試行数に

応じてハイカットが大きく変わる (Harvey & Liu (2015))。実務では「最終戦略に至るまでの試行数 N」を数え、報告 Sharpe をハイカット後の値で読み替えるのが正しい。

なお、選択バイアス・非正規性・標本長を 1 つの指標へ統合する相補的手法として Bailey & López de Prado (2014) の Deflated Sharpe Ratio (DSR) がある。極値理論による独立試行の期待最大 Sharpe 比を閾値に用いる点で本セクションの発想と同根であり（詳細は該当セクション参照）、Harvey & Liu の閾値アプローチと相補的に使える (Bailey & López de Prado (2014))。

前提・限界と注意点

第一に、これらの補正はいずれも「試行数 N を正しく数えられる」ことに依存する。ハイカット Sharpe・DSR・ブートストラップ検定のどれも、パラメータ探索・特徴量選択・閾値調整を含む全試行を N として渡さない限り甘い結論になる。したがって検証パイプラインでは、最終戦略に至るまでの全候補の PnL 系列と試行回数を記録しておくことが前提となる (Harvey & Liu (2015))。

第二に、依存構造をめぐる論争がある。Harvey, Liu & Zhu (2016) は、検定統計量に一定の依存があると「どの調整手続きも過度に厳格になりうる。これらは主に第 I 種過誤 (Type I error) を防ぐ設計のため、一定の相関構造下では第 I 種過誤を抑え込みすぎ (over-control)、高い第 II 種過誤率 (Type II error rate、真の発見の見逃し) をもたらす」と指摘する。極端な例として全検定が同一（相関=1.0）なら調整は不要で、FWER は単一検定の第 I 種過誤に等しく、通常の単一仮説検定で十分になる。同一リスクを代理するファクターや、価格を分母に持つロング・ショートポートフォリオのリターンは相関するため、これは短期戦略でも現実的な問題である。HLZ は Pearson 相関を通じた依存を明示的に取り込む構造モデルを提案しており、Romano & Wolf (2005) の StepM と Hansen (2005) の SPA は統計量の同時依存構造をブートストラップで捉えることで、この検出力損失を軽減する (Harvey, Liu & Zhu (2016))。実務的には、戦略群の相関が強いほど解析的な FWER 補正は保守的になりすぎるため、ブートストラップ系（RC／SPA／StepM）を優先するのが妥当である。

第三に、ホールドアウトや OOS では多重検定を制御できない。Bailey & López de Prado (2014) が示すとおり、ホールドアウト法は「単一試行が行われたかのように一般性を評価し、試行が増えるほど偽陽性が増える点を無視」する——95% 信頼水準でホールドアウトを 20 回適用すれば偽陽性は例外でなく想定される帰結である。Bailey, Borwein, López de Prado & Zhu (2014) はさらに、データマイニングにより理論上わずか 7 試行で、2 年バックテストの in-sample の期待最大 Sharpe 比が 1.0 に達しながら期待 OOS Sharpe=0 となるスプリアスな戦略を作れることを示した（同じ枠組みでは 5 年データなら 45 試行が同じ状況を生む）。したがって「OOS で確認したから安全」は成り立たず、OOS 検証は本セクションのブートストラップ検定・多重検定補正と併用してはじめて意味を持つ (Bailey & López de Prado (2014); Bailey, Borwein, López de Prado & Zhu (2014); Harvey & Liu (2015))。最後に、理論由来のファクターは純粋な経験的探索より低いハードルでよい（ただし 2.0 は低すぎる）こと、これらの検定は無条件検定に限られることも、HLZ 自身が挙げる適用限界である (Harvey, Liu & Zhu (2016))。

6. ファクターモデルによるアルファの正当性（Barra / Fama-French / スパニング検定）

短期予測モデルが生み出す超過リターンが「新規の予測力（アルファ, alpha）」なのか、それとも「既知のリスクファクターへのエクスポージャ（beta）を言い換えただけ」なのかを切り分けることが、本セクションの主題である。ファクターモデルは、この切り分けを回帰の切片（intercept）として定量化するための標準装置であり、実務では「戦略リターンを既知ファクター群に時系列回帰し、残った切片が統計的に有意な正であるか」を検証の出発点にする。

CAPM から マルチファクターへ：Jensen アルファの限界

出発点は CAPM 回帰 `R_it − R_ft = α_i + β_i(R_Mt − R_ft) + ε_it` で、切片 α が Jensen のアルファである。しかし市場ベータのみでは小型株・バリュー株の持続的超過リターンを説明できず、これがマルチファクター化の動機となった（Fama & French (1993)）。分散化ポートフォリオのリターン説明力は、標本内で CAPM が平均約 70% にとどまるのに対し、3 ファクターモデルは 90% 超に達する（Wikipedia, Fama–French three-factor model）。すなわち Jensen アルファは「CAPM ベータの誤指定（misspecification）を吸い込んだ残差」である可能性があり、より広いファクター集合で条件付けし直す必要がある。

Fama-French 3ファクター / 5ファクターモデル

3 ファクターモデルは `R_it − R_ft = a_i + b_i(R_Mt−R_ft) + s_i·SMB_t + h_i·HML_t + e_it`。SMB（size）と HML（value）は、全銘柄を NYSE 時価総額中央値で Small/Big に 2 分割し、B/M を NYSE 基準のボトム 30%・中間 40%・トップ 30% に 3 分割した 2×3=6 ポートフォリオから構成する。SMB＝小型 3 ポートフォリオ平均−大型 3 ポートフォリオ平均、HML＝(Small Value+Big Value)/2 −(Small Growth+Big Growth)/2 であり、負のブック値企業は除外する（Fama & French (1993); Kenneth R. French Data Library）。切片 a_i がモデル調整後アルファで、有意に正なら既知因子で説明できない超過リターンを示す。

5 ファクターモデル（2015）は RMW（収益性, Robust Minus Weak）と CMA（投資, Conservative Minus Aggressive）を追加し、`… + r_i·RMW_t + c_i·CMA_t + e_it` となる。1963 年 7 月〜2013 年 12 月（606 ヶ月）でクロスセクション分散の 71〜94% を説明するが、後述の Gibbons-Ross-Shanken 検定では「全ポートフォリオのアルファ同時ゼロ」を棄却する。すなわち個別アルファの絶対値は小さいものの、統計的にはゼロと言い切れない未説明リターンが残る（Fama & French (2015)）。

ファクター・スパニング検定（冗長性検定）

スパニング（redundancy）検定は、ある因子（またはあなたの戦略リターン）をその他の因子群に時系列回帰し、切片（アルファ）が 0 と有意に異なるかを見る手法である。切片が非有意ならその因子は冗長（既存因子で張られる空間に含まれる）と判定し、有意なら既存モデルに説明力を追加する「真のアルファ源」と判定する。Fama-French (2015) では HML を市場・SMB・RMW・CMA に回帰すると切片が非有意に

なり、value エクスポージャが主に RMW と CMA に吸収されて冗長化することが示された（Fama & French (2015)）。実務上の使い方は明快で、予測モデルの日次/週次リターン系列を左辺、既知ファクター群を右辺に置いて回帰し、切片 α とその t 値を見る。α が有意に正で残れば「既知因子の重ね合わせでは再現できない予測力」を主張でき、非有意なら戦略は本質的に既知プレミアムのリパッケージに過ぎない。

Carhart 4ファクター（モメンタム）と q-factor モデル

Carhart (1997) は FF3 にモメンタム因子 UMD（過去 12 ヶ月リターンの勝者ロング・敗者ショート、1 ヶ月ラグのゼロコストポートフォリオ）を加え、`EXR_t = α^c + β_mkt·EXMKT_t + β_SMB·SMB_t + β_HML·HML_t + β_UMD·UMD_t + ε_t` とした。サバイバーシップバイアスのない標本で、ファンドのパフォーマンス持続性（"hot hands"）がほぼ 4 因子と経費で説明され、運用スキルの存在を支持しないと結論した。切片 α^c が有意に正のときのみ予測能力ありと解釈される（Carhart (1997)）。短期モメンタム系シグナルを検証するなら、UMD を右辺に必ず含めないと「モメンタムの再発見」をアルファと誤認する。

Hou, Xue & Zhang (2015) の q-factor モデルは市場・size・投資（I/A）・収益性（ROE）の 4 因子で、投資 CAPM に基づき「収益性一定なら投資が多い企業ほど期待リターンが低い／投資一定なら収益性が高い企業ほど期待リターンが高い」を原理とする。1972 年 1 月〜2012 年 12 月の標本で、投資因子（I/A）は月次 0.45%（t=4.95）、ROE 因子は 0.58%（t=4.81）、size 因子は 0.31%（t=2.12）であり、約 80 のアノマリーを検証してその多くを吸収、FF3・Carhart 4 因子と同等以上の性能を示した（Hou, Xue & Zhang (2015)）。ベンチマークモデルを 1 つに固定せず、FF5・Carhart・q-factor で並行してスパニングを取ることで、アルファがどの因子系に対して頑健かを確認できる。

Barra 型構造的リスクモデルによる分解

Barra 型の構造的マルチファクターリスクモデルは、銘柄リターンを `r_it = Σ_j X_ij·f_jt + u_it`（X_ij＝ファクターエクスポージャ、f_jt＝クロスセクション回帰で推定されるファクターリターン、u_it＝特異リターン）に分解し、共分散を `Σ = X·F·X' + Δ`（共通ファクターリスク＋特異分散の対角行列 Δ）で表す。これによりリスク分解・ファクターアトリビューション・alpha vs beta 分解・factor-neutral 構築が可能になる（MSCI Barra USE4 Methodology (Menchero, Orr & Wang, 2011)）。Barra USE4 のスタイル因子は Beta・Residual Volatility・Momentum・Size・Non-linear Size・Book-to-Price（Value）・Earnings Yield・Growth・Leverage・Dividend Yield・Liquidity で、descriptor は平均 0・標準偏差 1 に標準化し、共分散推定に Newey-West 系列相関調整・Eigenfactor Risk Adjustment・Volatility Regime Adjustment・特異リスクの Bayesian shrinkage・最適化バイアス調整を適用する（MSCI Barra USE4 Methodology (Menchero, Orr & Wang, 2011)）。実務では、シグナルを Barra スタイル因子に中立化した「残差リターン」を作り、その残差にアルファが残るかを見ることで、産業・スタイルのベット由来の見かけの収益を除去できる。

GRS 検定：全アルファ同時ゼロ検定とシャープレシオ解釈

複数の検定資産（ポートフォリオ）を一括評価するときは、Gibbons-Ross-Shanken (1989) の F 検定を使う。統計量は `W = [(T−N−K)/N] · [α̂' Σ̂_ε⁻¹ α̂] / [1 + f̄' Ω̂_f⁻¹ f̄]`（T＝期間数、N＝検定資産数、K＝因子数、α̂＝切片ベクトル、Σ̂$_\varepsilon$＝残差共分散、f̄＝因子平均、Ω̂f＝因子共分散）で、帰無仮説 `α_1=…`

`=α_N=0` のもとで F(N, T–N–K) 分布に従う。これは線形アセットプライシングモデルが超過リターンを完全に説明するか（切片が同時ゼロか）を検定する最も広く使われる統計量である（Gibbons, Ross & Shanken (1989)）。幾何学的には GRS はシャープレシオ検定であり、統計量は「(検定資産＋因子) 全体の事後接点ポートフォリオの二乗シャープ比」と「因子のみの二乗シャープ比」の差の単調変換に等しい。両者が等しい（＝検定資産が効率的フロンティアを外側に押し出さない＝全アルファ＝0）ときに効率性が受容される（Agrrawal (2023)）。したがって「あなたのシグナル群を加えると達成可能な最大シャープ比が有意に上がるか」という問いと、アルファ同時ゼロ検定は等価であり、複数シグナルをまとめて価値評価する枠組みになる。

情報レシオとアクティブ運用の基本法則

アルファの経済的な大きさは、Grinold & Kahn の基本法則で予算化できる。情報レシオ `IR = 超過リターン平均 / トラッキングエラー` は `IR∗ = IC × √BR`（IC＝スキル＝予測と実現リターンの相関、BR＝年間の独立予測数）で近似され、制約下では移転係数 TC を掛けて `IR = TC × IC × √BR`、ベンチマークにアクティブ運用を加えた最大二乗シャープ比は `SR_P² = SR_B² + IR² = SR_B² + TC²·(IR∗)²` となる。ここで移転係数 TC を組み込んだ制約下の一般化は Clarke, de Silva & Thorley (2002) による（Grinold & Kahn (2000); Clarke, de Silva & Thorley (2002)）。情報係数（IC）は各期に N 資産のシグナルと将来リターンのクロスセクション相関を取り複数期平均したもので、外れ値に頑健な Spearman 順位 IC（rank IC）が実務標準。実務のファクターでは IC が 0.02〜0.05 程度にとどまり 0.10 超は稀とされるが、breadth を増やせば小さな IC でも有意な IR に変換できる（例: IC=0.05 なら BR=200 で IR≈0.71、BR=10 では ≈0.16）（Grinold & Kahn (2000)）。短期戦略の検証では、シグナルを既知因子に中立化した後の残差リターンに対する rank IC を測ることで「真の予測力」を分離でき、breadth と併せて期待 IR の上限を事前見積もりできる。

前提・限界と多重検定への注意

これらの手法の共通前提は、(1) ファクターリターンが線形かつ標本期間で概ね定常、(2) 残差共分散が推定可能なだけ N に対して T が十分大きい（GRS は T–N–K>0 が必要）、(3) 検定資産・因子が同一のデータ生成過程から観測される、である。最大の落とし穴は多重検定（multiple testing）である。Harvey, Liu & Zhu (2016) は、学術文献で報告されたファクターが 2012 年時点で少なくとも 316 に達したとカタログ化した。仮に全ファクターが無価値でも、単一検定の慣用基準 t>1.96（両側 5%）では偶然だけで 5%×316≈16 個が「有意」に見えてしまう計算になる。彼らは Bonferroni・Holm・Benjamini-Hochberg-Yekutieli（FDR 制御）を多重性・依存構造・出版バイアスを織り込んで適用し、新ファクターの信頼閾値を t 統計量約 3.0 超（将来はさらに上昇）とすべきだと結論した（Harvey, Liu & Zhu (2016)）。したがって短期予測モデルのアルファ正当化では、単一回帰の切片有意性だけでは不十分で、(a) FF5・Carhart・q-factor など複数モデルに対するスパニング検定、(b) 検定した候補数を踏まえた多重検定調整（t 閾値の引き上げ・FDR 制御）、(c) アウトオブサンプルでの再現、(d) Barra 因子中立化後の残差アルファ、を組み合わせて初めて「既知因子の重複でない予測力」を主張できる。逆に言えば、これらのいずれかを欠いたアルファ主張は、データマイニングによる偽発見（false discovery）と区別がつかない。

7. パフォーマンス／リスク指標（実現値）とその落とし穴

バックテストや実運用で得た実現リターン系列から Sharpe 比 (Sharpe ratio) 等の指標を計算し「この戦略は優秀だ」と結論づける前に、その指標が推定誤差 (estimation error)・分布仮定・年率化規約・多重検定 (multiple testing)・操作可能性 (gaming) という落とし穴を抱えていないかを点検する必要がある。実現値の指標は真の値ではなく有限標本からの推定量であり、点推定だけを見るのは平均だけを見て信頼区間を無視するのに等しい。本セクションでは Sharpe 比の統計的性質を起点に、Sortino・Omega・Information Ratio・ドローダウン系・トレードシステム系の各指標について、定義・数式・実務での使い方・前提と限界を整理し、短期クオンツ戦略の検証で「どの指標を何と照合してどう割り引くか」を具体的に示す。

Sharpe 比の推定誤差：IID 仮定下でも無視できない

まず最も基本的な落とし穴は、Sharpe 比そのものが推定量であり大きな標準誤差を伴う点である。Lo (2002) は、リターンが IID (independent and identically distributed) という最も好意的な仮定の下でさえ、Sharpe 比推定量が漸近的に $\sqrt{T}(\hat{SR} - SR) \to^{a} N(0, V_{IID})$、$V_{IID} = 1 + (1/2)SR^2$ に従うことを導いた。ここから標準誤差と 95% 信頼区間は次式になる (Lo (2002))。

$$SE(\hat{SR}) = \sqrt{\left[(1 + \tfrac{1}{2}\cdot SR^2) / T \right]}$$
$$95\%CI = \hat{SR} \pm 1.96\cdot\sqrt{\left[(1 + \tfrac{1}{2}\cdot SR^2) / T \right]}$$

重要なのは、標準誤差が Sharpe 比自身の増加関数である点である。T=60ヶ月では真の SR=1.50 で SE=0.188、SR=3.00 では SE=0.303 に拡大し、高 Sharpe を追求するヘッジファンドほど推定精度が落ちる (Lo (2002))。この式は μ と σ^2 が IID 下でのみ漸近独立で共分散項が消えることを前提とする。実務上は、報告された Sharpe 比に必ずこの SE を添え、信頼区間が 0 を跨ぐか、また二戦略の Sharpe 比の差が SE を超えて有意かを確認する。短期戦略でサンプリング頻度を上げて T を増やせば、IID を壊さない範囲で SE を機械的に縮小できる。

年率化バイアスと系列相関：√12 の落とし穴

月次 Sharpe 比を年率化する際に慣習的な √12 倍を使えるのは IID の時だけである。Lo (2002) は q 期間集約 Sharpe 比を $SR(q) = \eta(q)\cdot SR$、$\eta(q) = q / \sqrt{q + 2\cdot\Sigma_{k=1}^{q-1}(q-k)\rho_k)}$（$\rho_k$ は k 次自己相関）で与えた。全 $\rho_k=0$ の IID では $\eta(q)=\sqrt{q}$ に帰着するが、系列相関があると倍率がずれる。AR(1) 過程では次のとおりである (Lo (2002))。

$$\eta(q) = q \cdot \left[q\cdot(1 + 2\rho/(1-\rho)\cdot(1 - (1-\rho^q)/(q(1-\rho)))) \right]^{(-1/2)}$$

年率化（q=12）の倍率は IID で √12=3.46 だが、月次 1 次自己相関 ρ=+20% では 2.88 に低下し（正の自己相関は Sharpe 比を過大評価させる）、ρ=−20% では 4.17 に上昇する。実データではヘッジファンドの年率 Sharpe 比が系列相関で最大 65% も膨らみ、戦略ランキングが劇的に変わりうると報告されている

(Lo (2002))。短期戦略は約定・スリッページ平滑化・在庫効果で正の自己相関を持ちやすいため、`√頻度` で年率化した Sharpe は疑ってかかり、リターンの自己相関を測って $\eta(q)$ で補正した年率値を用いるべきである。

非正規性の補正：PSR と MinTRL

リターンが非正規なら標準誤差の式自体が変わる。Bailey & López de Prado (2012) は Mertens (2002) に基づき、非正規 IID 下の Sharpe 比標準誤差を歪度 $\gamma\hat{}_3$・尖度 $\gamma\hat{}_4$（正規で 3）を含む形に一般化し、これを用いて Probabilistic Sharpe Ratio (PSR) を定義した。

```
σ̂(ŜR) = √[ (1 − γ̂₃·ŜR + ((γ̂₄−1)/4)·ŜR²) / (T−1) ]
PSR(SR*) = Z[ (ŜR − SR*)·√(T−1) / √(1 − γ̂₃·ŜR + ((γ̂₄−1)/4)·ŜR²) ]
```

Z は標準正規 CDF で、正規時には Lo の `(1 + ½SR²)/(T−1)` に帰着する。負の歪度と正の過剰尖度は分母を増やし PSR を下げる、すなわち「観測 Sharpe が基準 SR* を真に超える確信」を弱める (Bailey & López de Prado (2012))。さらに、所与の基準 Sharpe を統計的有意に超えるのに必要な最小トラックレコード長 (Minimum Track Record Length; MinTRL) は次で与えられる。

```
MinTRL = 1 + [ 1 − γ̂₃·ŜR + ((γ̂₄−1)/4)·ŜR² ]·( Z_α / (ŜR − SR*) )²
```

数値例として、月次 mean=2・std=√12（月次 SR≈0.577）・skew=−0.72・kurt=5.78・基準 SR_ref=1/√12 のとき、負の歪度と高い尖度ほど必要サンプル長が伸びる (Bailey & López de Prado (2012))。IID だが Opdyke (2007) により定常・エルゴード過程まで妥当である。実務では、戦略が主張する Sharpe を確立するのに必要な観測数を MinTRL で先に見積もり、それに満たない track record は結論保留とする。

多重検定への補正：Deflated Sharpe Ratio

単一戦略の PSR で足りないのは、実際には多数の候補をスクリーニングしている点である。Bailey & López de Prado (2014) は、独立試行数 N を増やすと真の Sharpe がゼロでも最大 Sharpe 比の期待値が正になることを極値理論で示した。

```
E[max{ŜR_n}] ≈ μ + σ·[ (1−γ)·Z⁻¹(1−1/N) + γ·Z⁻¹(1−1/(N·e)) ]
```

γ≈0.5772 はオイラー・マスケローニ定数、σ は試行間 Sharpe 比の標準偏差、Z^{-1} は標準正規逆 CDF である。Deflated Sharpe Ratio (DSR) は PSR の棄却しきい値 SR* をこの E[max ŜR] に置き換えたもので、①非正規性（$\gamma\hat{}_3$, $\gamma\hat{}_4$）、②サンプル長 T、③試行間 Sharpe 分散、④独立試行数 N の要素で観測 Sharpe をデフレートする (Bailey & López de Prado (2014))。同論文は、バックテストで最も欠落している情報は『試行回数 N』だと指摘する。実務上は、パラメータ探索・特徴量選択・閾値調整を含む全試行数 N を記録し、最終戦略の Sharpe を DSR で割り引いて評価することが不可欠である。

指標の操作可能性と操作不能指標 MPPM

Sharpe 比を含むほぼ全ての性能指標は、情報を持たない動的・オプション的戦略で意図的に膨らませられる。Goetzmann, Ingersoll, Spiegel & Welch (2007) は、深いアウト・オブ・ザ・マネーのオプションを売って収益を債券に投じる等の『情報なし取引』で、稀な負のペイオフ事象を任意に低頻度化すれば見かけ上ほぼ全指標で勝てることを示した。操作不能な性能指標 (Manipulation-Proof Performance Measure; MPPM) は次のべき乗効用の平均形をとる。

```
θ̂ = (1/((1−ρ)·Δt))·ln( (1/T)·Σ_{t=1}^{T} [ (1+r_t)/(1+r_{ft}) ]^(1−ρ) )
```

ρ はリスク回避係数、Δt は観測間隔、r_{ft} は無リスク金利である。指標は (1) 単一値で順位付け可能、(2) ポートフォリオ規模に非依存、(3) 無情報投資家がベンチマーク逸脱で得できない、(4) 市場均衡と整合、の 4 条件を満たす必要があり、これはスコアが『リターン増加関数・凹・時間分離可能・べき乗形』であることを要求する。ρ=2 で $\Theta \hat{=} 6.6\%$（年率 20.4% 相当）、ρ=3 で 1.2%（14.0% 相当）と例示される (Goetzmann, Ingersoll, Spiegel & Welch (2007))。検証では、Sharpe 等が異常に高い戦略のペイオフ分布に稀な巨大損失（裾の売り持ち）が隠れていないかを疑い、操作耐性の確認には MPPM を併用する。

下方リスク指標：Sortino 比と Omega 比

Sharpe 比が上下変動を対称に罰するのに対し、下方リスクのみを分母に置くのが Sortino 比である。`Sortino 比 = (平均ポートフォリオリターン − MAR) / 下方偏差` で、最低許容リターン (Minimum Acceptable Return; MAR) は絶対リターン・指数・無リスク金利・ゼロ等を投資家が定義する (Kidd / CFA Institute (2012))。よくある誤りは二乗偏差和を『MAR 未満の観測数』で割ることで、正しくは全観測数で割る。また実現した過去リターンのみを使う離散法は下方リスクを著しく過小評価し、日本株 1980–89 年の例では月次下方偏差が離散法で 2.74%、連続分布フィット法で 3.20% だった。S&P500 月次（2001–10）・月次 MAR=0.4167%（年 5%）で下方偏差 3.71% を $\sqrt{12}$ で年率化すると 12.84% になるが『下方偏差は標準偏差と同様には年率化できない』と警告される (Kidd / CFA Institute (2012))。正リターンが多いほど下方偏差は過小評価される点に注意する。

分布仮定を一切置かずに全モーメントを織り込むのが Omega 比である。Keating & Shadwick (2002) はしきい値 r 以上の確率加重ゲインと以下の確率加重ロスの比として定義した。

```
Ω(r) = ∫_r^b [1 − F(x)]dx / ∫_a^r F(x)dx
```

F は累積分布関数、[a,b] はリターンの範囲で、分子は確率加重ゲイン、分母は確率加重ロスである。パラメトリック仮定が不要で歪度・尖度・ファットテールを反映し、しきい値 r を平均リターンに取ると Ω=1 になる (Keating & Shadwick (2002))。Sharpe/Sortino が平均・分散に依存するのに対し、Omega は効用関数を導入せずに分布全体で順位付けできる。短期戦略でリターン分布が非正規なとき、複数のしきい値 r について Omega 曲線を描けば、平均分散指標が見落とす裾の優劣を可視化できる。

ベンチマーク相対指標：Information Ratio

ベンチマーク対比の運用力を測るのが Information Ratio (IR) で、`IR = 平均超過（残差）リターン / トラッキングエラー` である。Goodwin (1998) は IR が t 統計量と `t = IR × √T` の関係にあることを示した。IR=0.5 でも 21 期間（自由度 20）なら t=2.29>1.725 で有意だが、9 観測（自由度 8）では t=1.50 で有意でない。Grinold & Kahn (1995) の経験則では IR 0.50 が『good』、0.75 が『very good』、1.0 が『exceptional』であり、ファンダメンタル法則 (Fundamental Law of Active Management) は `value added ∝ IR²_max`、`IR_max ≈ IC·√BR`（IC=情報係数 (information coefficient)＝スキル、BR=独立ベット数 (breadth)）とする (Goodwin (1998))。年率化には 4 通りの方法があるが、Goodwin (1998) は実データではそれらの間に系統的な大差は生じにくく（方法間で一意の順位は保証されない）、例えば四半期 IR に √4=2 を掛ける方法は年率 IR を四半期値のちょうど 2 倍にすると報告した。むしろ IR はベンチマーク依存が強く、同論文のサンプルではベンチマークを不適切なものに替えると IR が最大 0.53 低下した例がある。Goodwin (1998) はさらに、IR は資産クラス間の配分判断には使えないこと、過去 10 年の IR は将来を保証しないことを警告する。実務では IR を必ず観測数 T とともに報告し、t=IR√T で有意性を確認したうえでベンチマークの妥当性を吟味する。

経路依存のドローダウン指標：MDD・Calmar・Ulcer・テールレシオ

損益 (PnL) 曲線の経路そのものを評価するのがドローダウン系指標である。最大ドローダウン (Maximum Drawdown; MDD) はピークから谷への最大の下落幅で、`MDD(T) = max_{τ}[ max_{t≤τ} X(t) − X(τ) ]`、割合では `MDD = (ピーク − 谷)/ピーク`（ピーク超からの下落率）と表す。ドローダウン期間 (duration) はピークからピークまで（新高値までの時間）で、MDD は上方変動を無視し下方のみに焦点を当てる経路依存指標である (Wikipedia (Drawdown))。これを分母に使う代表指標が Calmar 比で、`Calmar = 直近 36ヶ月の年率複利リターン / 同期間の最大ドローダウン` を月次更新で計算する。Sterling 比を月次更新に改変したもので Sharpe/Sterling より滑らかに変化し、MAR 比が運用開始以降の全データを使うのに対し Calmar は 36ヶ月データを使う。分母が MDD＝1 回の最悪イベントに支配されるため単一の極端事象への感度が高い (Young (1991))。

単一事象への集中を緩和するのが Ulcer 指標 (Ulcer Index; UI) で、直前高値からの二乗パーセントドローダウンの平均の平方根である (Martin & McCann (1989))。

```
UI = √( Σ_{i=1}^n D'_i² / n )    (D'_i = 各期の直前ピークからの%DD、新高値なら 0)
```

二乗により大きなドローダウンを不相応に重く罰し、複数期間にわたる測定でドローダウンの持続時間も反映する（深く回復に時間がかかるほど UI が高い）。標準偏差と異なり上昇は罰しない。Ulcer Performance Index（UPI, Martin 比）= (トータルリターン − 無リスクリターン)/UI は Sharpe 比の分母を UI に置換した指標である (Martin & McCann (1989))。裾の非対称性は、`テールレシオ = 95 パーセンタイル・リターン / |5 パーセンタイル・リターン|` で測れる。1 超なら右裾が厚く、0.25 なら損失が利益の 4 倍悪いことを意味するが、外れ値に敏感でパーセンタイル推定が不安定な点に注意する (PyQuant News)。実務では、観測 MDD・Calmar・UI・テールレシオを併読し、単一の指標が一つの極端事象に支配されていないかを相互に照合する。

トレードシステム指標：期待値・プロフィットファクター・勝率

トレード単位で戦略を診断する指標は相互補完的で、勝率単独は誤解を招く。期待値 (Expectancy) は `（平均利益 × 勝率）−（平均損失 × 負率）` で、1 トレード・1 ドルのリスクあたり平均でいくら得られるかを示し、正か負かが『生死の差』とされる (FXStreet Learning Center / Van Tharp)。関連指標として Profit Factor = 総利益/総損失（1 超で利益システム）、Payoff Ratio = 平均利益/平均損失、Win Rate = 勝ちトレード数×100/総トレード数がある。具体例として、勝率 30%・平均利益 $300・平均損失 $80 のとき期待値 = 0.30×$300 − 0.70×$80 = +$34/トレードとなり、勝率が低くてもペイオフレシオが高ければ高収益になりうる (FXStreet Learning Center / Van Tharp)。したがって勝率のみで戦略を採否せず、期待値・Profit Factor・Payoff Ratio を組で評価する。

前提・限界と実務チェックリスト

以上の指標に共通する前提と限界を整理する。第一に、Sharpe をはじめ実現値の指標はすべて有限標本の推定量であり、SE・信頼区間・MinTRL を添えなければ点推定は意味を持たない (Lo (2002); Bailey & López de Prado (2012))。第二に、年率化は分布仮定に依存する。Sharpe の √頻度倍は系列相関で最大 65% 過大評価され (Lo (2002))、下方偏差は √12 で年率化できない (Kidd / CFA Institute (2012))。第三に、多数の候補を試した事実（試行数 N）を報告しない限り、観測された高 Sharpe が選択バイアスか実力かは区別できず、DSR による割引が必須である (Bailey & López de Prado (2014))。第四に、ほぼ全指標は動的・オプション的戦略で操作可能なため、異常に良い指標はペイオフ分布の裾を疑い、操作耐性には MPPM を用いる (Goetzmann, Ingersoll, Spiegel & Welch (2007))。

短期クオンツ戦略の検証チェックリストとしては、(1) 取引頻度に合ったリターン系列で Sharpe を計算し SE と 95%CI・MinTRL を添える、(2) リターンの自己相関を測り $\eta(q)$ で年率化バイアスを補正する、(3) 歪度・尖度を織り込んだ PSR と試行数 N を入れた DSR で有意性を確認する、(4) 下方リスクは Sortino/Omega、経路リスクは MDD/Calmar/UI/テールレシオ、トレード単位は期待値/Profit Factor/Payoff Ratio と、性質の異なる指標を組で読む、(5) 指標が突出して良いときは操作可能性を疑い MPPM で検算する、の 5 点を組み合わせる。単一指標・点推定・素朴な年率化・試行数の非報告が、実現値指標の四大落とし穴である。

8. リターン分布・テールリスク・ドローダウン分布

短期クオンツ戦略の検証において、平均や Sharpe 比だけを見て「リターン分布は正規 (normal) だろう」と暗黙に仮定すると、極端損失とリスク指標を系統的に過小評価する。資産リターンは正規分布からも無限分散の安定分布 (stable laws) からも外れた重い裾 (heavy tails) を持ち、リスク尺度の選択次第では分散投資を罰したり口座分割で見かけのリスクを減らせたりする欠陥が生じ、損益 (PnL) 曲線のドローダウンは戦略の期待リターンが正であっても増え続ける。本セクションでは、(1) リターン分布の定型化事実 (stylized facts)、(2) 極値理論 (Extreme Value Theory; EVT) と Generalized Pareto Distribution (GPD) による裾指数の推定、(3) コヒーレントリスク尺度 (coherent measures of risk) の公理と VaR の欠陥、(4) Conditional Value-at-Risk (CVaR) / Expected Shortfall (ES)、(5) 最大ドローダウン (maximum drawdown) 分布と Conditional Drawdown-at-Risk (CDaR) を扱う。目的は、検証対象モデルの PnL・リターン系列が「重い裾を持つか」「使っているリスク指標がコヒーレントか」「観測されたドローダウンが統計的に想定範囲内か」を具体的にチェックできるようにすることである。

リターン分布の定型化事実：重い裾・尖度・集約的正規性

出発点は、資産リターンの無条件分布 (unconditional distribution) が持つ普遍的な性質である。第一に、無条件リターン分布はべき乗則 (power-law / Pareto 型) の裾を示し、その裾指数 (tail index)——「有限となる最高絶対モーメントの位数」と定義され、Gauss/指数テールでは +∞、べき乗指数 α の裾では α に等しい——は有限で、大半のデータで 2 より大きく 5 未満に収まる。裾指数が有限かつ 2 超であることは無限分散の安定分布を排除し、5 未満であることは正規分布を排除する (Cont (2001))。

第二に、分布は強い超過尖度 (excess kurtosis; leptokurtosis) と（株式では）負の歪度 (skewness) を示す。尖度は次式で定義され、正規分布で 0、正値が fat tail を示す (Cont (2001))。

```
κ = E[(r − r̄)^4] / σ^4 − 3
```

高頻度になるほど非正規性は際立ち、5 分足リターンの実測では S&P500 先物 $\kappa \approx 15.95$（歪度 −0.4）、US$/DM 先物 $\kappa \approx 74$（歪度 −0.11）、US$/Swiss Franc 先物 $\kappa \approx 60$（歪度 −0.1）に達する。IID を仮定した 95% 信頼区間は歪度 ±0.018・尖度 ±0.036 なので、これらは統計的に極めて有意な非正規性である (Cont (2001))。第三に、リターン分布の形状はスケール依存であり、時間スケール Δt を大きくすると分布は正規分布に近づく（集約的正規性 (aggregational Gaussianity)）。したがって「日次で重い裾でも月次では正規に近い」ことは矛盾ではなく、検証はモデルの取引頻度に合ったスケールで行う必要がある (Cont (2001))。加えて株式・株価指数では大きなドローダウンは観測されるが同規模の急騰は観測されないゲイン/ロス非対称性 (gain/loss asymmetry) が成立する一方、為替では上下対称で成立しない (Cont (2001))。

実務上のチェックはこうである。モデルの取引頻度に対応するリターン系列で尖度と歪度を測り、正規性を前提にした VaR やポジションサイジングが妥当かを判定する。ただし重要な注意として、4 次以上の標本モーメントは数値的に不安定で、リスクの定量指標としては信頼できない。Student-t(自由度 4) を用いた実験でも標本 4 次モーメントは激しく揺れる。したがって尖度は「非正規性の存在を示す定性的シグナル」とし

て使い、テールリスクの定量化は次節の EVT に委ねるべきである (Cont (2001))。なお絶対値リターンの自己相関はべき乗則で緩慢に減衰し（指数 $\beta \in [0.2, 0.4]$、長期記憶の兆候）、一方で線形自己相関は組織化された先物で数分・為替でさらに短時間（$\tau \geq 15$ 分で無視可能）で実質ゼロになる。これはボラティリティのクラスタリングが予測可能でもリターン符号は予測困難という、効率的市場仮説 (efficient market hypothesis) と整合した構図である (Cont (2001))。

極値理論（EVT）と GPD による裾指数の推定

尖度が不安定なら、裾そのものを直接モデル化するのが EVT である。IID 系列では、正規化された最大値 $(M_n - \lambda_n)/\sigma_n$ の極限分布は Gumbel（$\xi=0$、指数テール）・Fréchet（$\xi>0$、べき乗テール）・Weibull（$\xi<0$、有限端点）の 3 型に限られ、統一形は次で書ける (Cont (2001))。

$$H_\xi(x) = \exp\left[-(1 + \xi x)^{-1/\xi} \right]$$

金融日次リターンでは形状パラメータ ξ が 0.2〜0.4 と 0 から有意に離れ、Fréchet 領域＝重い裾に属する。$\alpha(T)=1/\xi$ より $2 < \alpha(T) \leq 5$ となり、前節の裾指数 2〜5 と整合する (Cont (2001))。ただし $\xi>0$ は厳密なべき乗テールを含意しない。任意の緩変動関数 (slowly varying function) $L(x)$ で正則変動テール $F(x) \sim L(x)/x^\alpha$ と両立するため、log-log 回帰による裾指数推定は L の影響で歪み、精度は有効数字 1 桁程度にとどまる (Cont (2001))。

実務では最大値そのものより、閾値超過分布 (peaks over threshold) の方が扱いやすい。Balkema-de Haan / Pickands の定理により、高閾値 u を超える超過 y の条件付き分布は GPD に収束する (McNeil & Frey (2000))。

$$G_{\xi,\beta}(y) = 1 - (1 + \xi y/\beta)^{-1/\xi} \quad (\xi \neq 0)$$

$\xi>0$ は power-law 減衰の重い裾（Pareto・Student-t・Cauchy・Fréchet 等）、$\xi=0$ は指数減衰（正規・指数・対数正規）、$\xi<0$ は有限右端点の短い裾に対応する。裾確率の推定量は、閾値 u を超える観測数 N・全標本 n を用いて次で与えられる (McNeil & Frey (2000))。

$$1 - \hat{F}(x) = (N/n) \cdot (1 + \hat{\xi}(x - u)/\hat{\beta})^{-1/\hat{\xi}}$$

実データ（S&P の GARCH 残差, n=1000, k=100）では損失側 $\xi \fallingdotseq 0.224$（標準誤差 0.122、裾指数 $1/\xi \approx 4.5$）、利得側 $\xi \fallingdotseq -0.096$ と推定され、損失側だけが重い裾を持つ。GARCH の Student-t 近似では $1/\xi$ が自由度に対応する。この損失側 ξ の範囲は Cont の裾指数 2〜5 とも整合する (McNeil & Frey (2000))。検証手順は、モデルの標準化 PnL 残差の損失側に GPD を当て、$\hat{\xi}$ が 0 から有意に離れるか（＝正規性が棄却されるか）と裾指数 $1/\hat{\xi}$ が 2〜5 の重い裾域に入るかを確認することである。

コヒーレントリスク尺度の公理と VaR の欠陥

裾を推定したら、それを 1 つの数値に集約するリスク尺度が必要になる。良いリスク尺度が満たすべき性質は 4 公理として公理化されている。すなわち平行移動不変性 (translation invariance) $\rho(X+\alpha \cdot r)=\rho(X)$

$-\alpha$、劣加法性 (subadditivity) $\rho(X1+X2)\le\rho(X1)+\rho(X2)$、正斉次性 (positive homogeneity) $\rho(\lambda X)=\lambda\rho(X)$（$\lambda\ge0$）、単調性 (monotonicity) $X\le Y \Rightarrow \rho(Y)\le\rho(X)$ の 4 つを全て満たす尺度をコヒーレント (coherent) と呼ぶ (Artzner, Delbaen, Eber & Heath (1999))。とりわけ劣加法性は「合併は追加的リスクを生まない」という分散投資の直観を表し、これを満たさない尺度は (a) 二口座への分割で証拠金を減らせる、(b) 分社化の誘因が生じる、(d) 分権的なリスク集計・資本配分ができない、といった実務的害を招く (Artzner, Delbaen, Eber & Heath (1999))。

広く使われる Value-at-Risk はこの基準を満たさない。$\mathrm{VaR}_\alpha(X)=-\inf\{x \mid P[X\le x\cdot r]>\alpha\}$（分位点の負値）は平行移動不変・正斉次・単調は満たすが劣加法性に失敗し、したがってコヒーレントでない (Artzner, Delbaen, Eber & Heath (1999))。反例は 2 つある。第一にデジタルオプションで、$P\{S_{TU}\}=0.008$ のとき、A・B を 1 枚ずつ売るとそれぞれの 1% VaR は負（受容可能）だが、A+B を売ると 1% VaR が正に転じ $\mathrm{VaR}(A+B) > \mathrm{VaR}(A)+\mathrm{VaR}(B)$ となって劣加法性が破れる。加えて、受容可能な 2A・2B の中点が受容不可の A+B になるため受容集合が非凸になる。第二に信用リスクで、100 万を 1 社債に集中すると 5% VaR は「無リスク」に見えるが、独立デフォルト確率 1% の 100 社に分散すると 2 社以上デフォルトの確率が 0.18 超となり 5% VaR が正に転じる——分散でリスク尺度が悪化するのである (Artzner, Delbaen, Eber & Heath (1999))。ただし全価格が同時正規で超過確率 <0.5 の場合に限り $\mathrm{VaR}_\alpha(X)=-(E[X]+\Phi^{-1}(\alpha)\sigma)$ は劣加法になる。つまり VaR の劣加法性は正規性という強い仮定に依存しており、前 2 節で見た重い裾のもとでは一般に成り立たない (Artzner, Delbaen, Eber & Heath (1999))。

理論的な救済策は表現定理が与える。ρ がコヒーレントであることは、状態空間上の確率測度族（一般化シナリオ）$\mathcal{P}$ が存在して $\rho(X)=\sup\{E_P[-X/r] \mid P\in\mathcal{P}\}$ と書けることと同値であり、シナリオが多いほど尺度は保守的になる (Artzner, Delbaen, Eber & Heath (1999))。この表現から具体的なコヒーレント候補として Tail Conditional Expectation（TailVaR）$\mathrm{TCE}_\alpha(X)=-E[X/r \mid X/r \le -\mathrm{VaR}_\alpha(X)]$ と Worst Conditional Expectation $\mathrm{WCE}_\alpha(X)$ が提案され、$\mathrm{TCE}_\alpha \le \mathrm{WCE}_\alpha$ で WCE はコヒーレント、TCE は VaR より保守的で規制上受容されうる中で最も安価と論じられた (Artzner, Delbaen, Eber & Heath (1999))。

CVaR / Expected Shortfall（ES）

TCE の実装しやすい形が CVaR / ES である。CVaR は α-tail 分布の平均として定義され、損失分布が不連続でもコヒーレントである。一方、素朴な上側平均 $\mathrm{CVaR}^{+}=E[f \mid f>\zeta_\alpha]$（mean shortfall）と下側平均 $\mathrm{CVaR}^{-}=E[f \mid f\ge\zeta_\alpha]$（Artzner の tail VaR に相当）は一般にコヒーレントでない。三者は $\mathrm{CVaR}^{-} \le \mathrm{CVaR} \le \mathrm{CVaR}^{+}$ の関係にあり、VaR 閾値 ζ_α に確率のジャンプ（atom）が無ければ一致する。CVaR は VaR と CVaR^{+} の加重平均 $\phi_\alpha=\lambda_\alpha\cdot\zeta_\alpha+(1-\lambda_\alpha)\cdot\phi_\alpha^{+}$ として atom を「分割」でき、常に CVaR ≥ VaR が成り立つ (Rockafellar & Uryasev (2002))。

CVaR の実務的な強みは最適化しやすさにある。Rockafellar-Uryasev の最小化公式は次のとおりで、$[t]^{+}=\max\{0,t\}$ である (Rockafellar & Uryasev (2002))。

```
F_α(x, ζ) = ζ + (1/(1−α)) · E{ [f(x, y) − ζ]⁺ }
φ_α(x) = min_ζ F_α(x, ζ),   VaR ζ_α(x) = argmin の下端点
```

$F_\alpha(x,\cdot)$ は ζ について凸で、x について CVaR を最小化することは (x,ζ) について F_α を同時最小化することと等価になる。シナリオ標本を使えば CVaR 最小化ポートフォリオ問題は線形計画 (LP) に帰着し、VaR 最小化と違って凸性が保たれ大規模計算が可能で、VaR と CVaR が最適化の副産物として同時に得られる (Rockafellar & Uryasev (2002))。

同じ量は Expected Shortfall として「最悪 $100\alpha\%$ の損失の平均」とも定義でき、標本推定量は「最悪 A% の観測の単純平均」という極めて素朴な形になる (Acerbi & Tasche (2002))。

```
ES^(α)(X) = −(1/α) · { E[X·1_{X ≤ x^(α)}] − x^(α)·(P[X ≤ x^(α)] − α) }
```

ES は単調性・平行移動・正斉次・劣加法性を全て満たしコヒーレントである（凸性は劣加法性＋正斉次性から従う）。ES は TCE とは一般に異なり、TCE は分布が不連続だと α を超える割合まで平均してしまい劣加法性を失いうる。VaR が非劣加法なのは定義中の「最小損失 (minimum loss)」に起因する、というのが Acerbi-Tasche の診断である (Acerbi & Tasche (2002))。なお信頼水準 α の規約は文献間で異なり、Rockafellar-Uryasev は α を信頼水準（0.95/0.99）、Acerbi-Tasche は α を裾側確率（A%）として用いるため、実装時に取り違えないこと。

正規性を仮定して ES を計算すると、重い裾のもとで過小評価が生じる。GPD（$\xi<1$, β）に厳密に従う超過 W では $E[W \mid W>w]=(w+\beta)/(1-\xi)$ となり、ES-to-quantile 比は GPD（$\xi>0$）では分位点 $q\to1$ で $(1-\xi)^{-1}>1$ に収束するのに対し、正規では 1 に収束する (McNeil & Frey (2000))。実測（$\xi=0.224$, $\beta=0.568$）では q=0.99 で GPD 1.42 対 正規 1.15、q=0.995 で 1.39 対 1.12 と、正規は ES を系統的に低く見積もる。バックテストでも、S&P（7414 営業日）0.99 分位の期待違反 74 回に対し、条件付き EVT（GARCH+EVT）は 73 回（p=0.91）と良好、条件付き正規は 104 回（p=0.00）と大幅超過だった。したがって短期戦略のテールリスク検証では、正規 VaR ではなく GARCH でボラティリティを条件付けたうえで残差裾を EVT で推定し ES を出す構成が推奨される (McNeil & Frey (2000))。

最大ドローダウン分布と CDaR

テール損失が単発の事象なら、ドローダウンはその累積である。ドリフト付き Brownian 運動 $X(t)=\sigma W(t)+\mu t$ の最大ドローダウン $D(T)=\sup_{t\in[0,T]}[\sup_{s\in[0,t]}X(s)-X(t)]$（ピークからボトムまでの最大下落）は、無ドリフト $\mu=0$ のとき閉形式を持つ (Magdon-Ismail, Atiya, Pratap & Abu-Mostafa (2004))。

```
E[D̄] |_{μ=0} = √(π/2) · σ√T ≈ 1.2533 · σ√T
```

比較として範囲（最大−最小）の期待値は $E[R]=2\sqrt{(2/\pi)}\cdot\sigma\sqrt{T}\approx1.596\cdot\sigma\sqrt{T}$ なので、期待最大ドローダウンは期待レンジより有意に小さい (Magdon-Ismail, Atiya, Pratap & Abu-Mostafa (2004))。時間 T への依存はドリフト符号で 3 レジームに分かれ、一般形は $E[D]=(2\sigma^2/\mu)\cdot Q_D(\alpha^2)$（$\alpha=\mu\sqrt{T}/(\sqrt{2}\cdot\sigma)$、すなわち $\alpha\propto\mu\sqrt{T}/\sigma$）で、漸近挙動 ($T\to\infty$) は $\mu>0$ で対数成長（$Q_p(x)\to(1/4)\log x + 0.49088$）、$\mu=0$ で平方根成長、$\mu<0$ で線形成長（$Q_n(x)\to x+1/2$、すなわち $E[D]\to|\mu|T$）となる (Magdon-Ismail, Atiya, Pratap & Abu-Mostafa (2004))。実務的含意は重い。正の期待リターンを持つ戦略でも最大ドローダウンは log T で増え続け、有限時間で必ず新記録的ドローダウンが起こりうる。したがって観測された最大ドロ

ーダウンが「異常」か「正常」かは、推定した μ・σ・T をこの式に入れた理論期待値と照合して初めて判定でき、$\sigma\sqrt{T}$ のスケールを大きく超えるドローダウンだけを戦略破綻のシグナルと見なすべきである (Magdon-Ismail, Atiya, Pratap & Abu-Mostafa (2004))。

単一の最大ドローダウンは 1 事象に情報が集中しすぎる。これを緩和するのが CDaR で、非複利累積リターン曲線 w（underwater 曲線）の各時点の絶対ドローダウン $\xi_k = \max_{j\le k} w_j - w_k$ から成る系列に CVaR を適用したものである (Chekhlov, Uryasev & Zabarankin (2005))。

```
Δ_α(w) = CVaR_α( AD(w) ) = 最悪 (1−α)·100% のドローダウンの平均
```

許容水準 α を端点にとると MaxDD(w)=Δ_1(w)、AvDD(w)=Δ_0(w) となり、最大ドローダウンと平均ドローダウンを両極限に含む。CDaR は偏差尺度 (deviation measure) の公理（非負性・定数シフト不感・正斉次性・凸性）を満たし、ドローダウンの大きさだけでなく期間 (duration) も考慮するため、単一事象に集中する最大ドローダウンより情報量が多い。reward-CDD 最適化は区分線形凸となり LP に帰着し、数十万銘柄規模まで解ける (Chekhlov, Uryasev & Zabarankin (2005))。実務のドローダウン規制の例として、50%DD は許容不可、20%DD で口座閉鎖、15%DD で警告、DD からの脱出は 1 年以内、といった閾値運用が挙げられており、CDaR はこうした複数水準の制約を最適化に組み込む枠組みを与える (Chekhlov, Uryasev & Zabarankin (2005))。

前提・限界と実務チェックリスト

これらの手法を使ううえでの前提と限界を整理する。第一に、定型化事実（重い裾・尖度・集約的正規性）はモデルの取引頻度に対応するスケールで測る必要があり、尖度など高次標本モーメントは不安定なので定量指標には使わず、裾は EVT/GPD で直接推定する (Cont (2001))。第二に、GPD の形状 ξ 推定は緩変動関数の影響を受け精度が有効数字 1 桁程度にとどまるため、閾値 u の選択に対する感度を確認すべきである (Cont (2001); McNeil & Frey (2000))。第三に、VaR は正規性という強い仮定下でしか劣加法にならず、重い裾のもとでは分散投資を罰しリスク集計を歪めるため、リスク集計・資本配分にはコヒーレントな ES/CVaR を用いる (Artzner, Delbaen, Eber & Heath (1999); Rockafellar & Uryasev (2002); Acerbi & Tasche (2002))。第四に、ドローダウンの理論期待値は Brownian 運動という理想化に基づくので、実データの自己相関やボラティリティのクラスタリングがあると乖離しうるが、観測ドローダウンの「正常/異常」判定の基準線としては有効である (Magdon-Ismail, Atiya, Pratap & Abu-Mostafa (2004))。

短期戦略の検証チェックリストとしては、(1) 取引頻度に合ったリターン系列で歪度・尖度を測り非正規性の有無を定性判定する、(2) 損失側 PnL 残差に GPD を当て $\hat{\xi}$ が 0 から有意に離れるか・裾指数 $1/\hat{\xi}$ が 2〜5 域かを確認する、(3) リスク集計は正規 VaR ではなく GARCH+EVT で条件付けた ES を用い、0.99 分位でバックテスト違反回数が期待値と整合するかを検定する、(4) 観測最大ドローダウンを $E[D]\approx 1.2533\sigma\sqrt{T}$（$\mu$=0 基準）や CDaR と照合し、規制水準に対する余裕を評価する、の 4 点を組み合わせるのが妥当である。分布の裾・リスク尺度・ドローダウンは互いに独立した診断ではなく、重い裾が ES を膨らませ、それが累積してドローダウンになるという一連の因果でつながっているため、三者を一貫して検証することが重要である。

9. 分布推定：ブートストラップ／モンテカルロ／順列検定による確率的評価

前セクションまでの多重検定補正や Deflated Sharpe Ratio が「試行数を織り込んだ有意閾値」を与えるのに対し、本セクションは検定の土台となる「性能指標のサンプリング分布 (sampling distribution) そのものをどう推定するか」を扱う。点推定された Sharpe 比 (Sharpe ratio) や平均リターンだけでは「その値が偶然の範囲か」は判定できない。標準誤差 (standard error; SE)・信頼区間 (confidence interval; CI)・p 値を得るには、統計量の分布を漸近理論かリサンプリング (resampling) で近似する必要がある。短期クオンツ戦略は系列相関 (serial correlation) と条件付き分散変動 (conditional heteroskedasticity) を持つため、IID (independent and identically distributed) を仮定する素朴な分布推定は系統的に誤る。本セクションで具体的にチェックできるのは、(1) 単一戦略の Sharpe 比が主張どおりの精度で推定できているか、(2) 2 戦略の Sharpe 比の差が有意か、(3) シグナルに真の予測力があるか、(4) 大量スクリーニングの末の最良戦略がベンチマークを超えるか、の 4 点であり、いずれも「時系列依存を保ったまま分布を推定する」ことが鍵になる。

Sharpe 比推定量のバイアスと基準分布（IID 正規）

出発点は推定量そのものの偏りである。素朴な推定量 `ζ̂_basic = μ̂/σ̂` は上方バイアスを持ち、小標本ほど顕著になる。IID 正規の下でバイアスはガンマ関数比で厳密に表せる (Riondato / Two Sigma (2018))。

```
E[ζ̂_basic] = ζ · √((n−1)/2) · Γ((n−2)/2) / Γ((n−1)/2)   ( > ζ )
```

バイアス係数は n=12 で約 1.08、n=40 で約 1.02、n=75 で約 1.01 と標本長 n の増加で 1 に収束する。厳密分散も `Var[ζ̂_basic] = (1+nζ²)(n−1)/(n(n−3)) − E[ζ̂_basic]²` で与えられ、非正規の場合は歪度 γ_1・尖度 γ_2 を含む Bao (2009) 展開 `E[ζ̂_basic] = ζ + (3/(4n))ζ − (1/(2n))γ_1 − (3/(8n))γ_2·ζ + o(1/n)` を使う。Fisher の k 統計量で歪度・尖度を推定した近似不偏推定量も提示されている (Riondato / Two Sigma (2018))。分布の中心（推定量の期待値）が真値からずれている以上、その上に組む CI も補正が必要になる。

漸近的には、IID 正規リターン下で Sharpe 比推定量の標準誤差は次で与えられ、95% 信頼区間は `SR ± 1.96·SE` となる (Lo (2002))。

```
SE(SR) = √((1 + ½SR²)/T)
√T(ŜR − SR) →ᵈ N(0, V_IID),   V_IID = 1 + ½SR²
```

数値の直感として、T=60 で SR=1.50 なら SE=0.188、SR=3.00 なら SE=0.303、T=12・SR=3.00 では 0.677、T=250・SR=0.50 では 0.067 である。SE は SR 一定なら T の増加で縮小し、T 一定でも SR が大きいほど絶対値は増大する。相対誤差 SE/SR は SR→∞ で √(1/(2T)) に収束する (Lo (2002))。これが「短いサンプルと高い Sharpe 比ほど推定が不安定」であることの数理的根拠であり、相対 CI 幅の下限を与える。

非 IID への拡張：HAC ロバスト標準誤差と年率化補正

リターンが系列相関や条件付き分散変動を持つ非 IID の場合、上式は誤った SE を返す。この場合は GMM (generalized method of moments) と Newey–West 型の HAC (heteroskedasticity and autocorrelation consistent) 推定で頑健標準誤差を計算する (Lo (2002))。

```
√T(ŜR − SR) →ᵈ N(0, V_GMM),   V_GMM = (∂g/∂θ)·Σ·(∂g/∂θ)',   SE = √(V_GMM/T)
Σ̂ = Ω̂_0 + Σ_{j=1}^m ω(j,m)(Ω̂_j + Ω̂_j'),   ω(j,m) = 1 − j/(m+1)
```

切断ラグ m は `m/T → 0` を満たす必要があり、実証では m=3, 6 が用いられる。系列相関の有無は Ljung–Box の Q 統計量 `Q_{q−1} = T(T+2)Σρ̂_k²/(T−k) ~ χ²_{q−1}` で診断する (Lo (2002))。定常・エルゴード過程では同様に GMM＋HAC（Andrews (1991) カーネル）で漸近分散を計算できる (Riondato / Two Sigma (2018))。

系列相関は年率化にも直接効く。月次 Sharpe 比を √12 倍して年率化するのは系列相関ゼロ（IID）のときだけ正しく、一般の補正係数 $\eta(q)$ は最初の q−1 個の自己相関の関数である (Lo (2002))。

```
SR(q) = η(q)·SR,   η(q) = q / √(q + 2Σ_{k=1}^{q−1}(q−k)ρ_k)
AR(1): η(q) = √q · [1 + (2ρ/(1−ρ))(1 − (1−ρ^q)/(q(1−ρ)))]^{−1/2}
```

ρ=0 で $\eta(12)$=3.46（=√12）に一致し、正の系列相関は $\eta(q)$ を縮小、負は拡大させる。月次一次自己相関が −20% なら年率係数は月次 Sharpe の 4.17 倍、+20% なら 2.88 倍になる。実証では、あるコンバーチブル／オプション裁定ヘッジファンドの年率 Sharpe は √12·ŜR=4.35 だが、系列相関を織り込むと ŜR(12)=2.99 まで低下する（45% の過大評価）。IID 推定は年率 Sharpe を最大 65% 過大評価し、負の相関ではむしろ過小評価してランキングまで変える (Lo (2002))。Two Sigma 技報も同じ年率化式 `ζ_T = q/√(q+2Σ(q−k)ρ_k)·ζ_t` を採用し、Mertens (2002) を引いて「正規性の誤仮定は漸近分散推定を真値から最大 70% ずらす」と警告する (Riondato / Two Sigma (2018))。実務では、まず Ljung–Box で系列相関を検定し、有意なら √12 年率化と IID 系 CI を捨てて HAC・$\eta(q)$ に切り替えるのが正しい運用になる。

Sharpe 比の信頼区間：漸近 plug-in／CDF 反転／スチューデント化ブートストラップ

Sharpe 比の CI には 3 系統がある (Riondato / Two Sigma (2018))。第一は漸近正規に基づく plug-in 区間で、`√n(ζ̂_basic − ζ) →ᵈ N(0, 1+ζ²/2)` から `C = (ζ̂ − z_{α/2}·ŝe, ζ̂ + z_{α/2}·ŝe)` を作る。第二は CDF を反転して得る厳密区間である。第三が時系列に適した percentile-t（スチューデント化）ブートストラップで、リサンプルごとに統計量をスチューデント化して分位点を取る。

```
T = (ζ̂* − ζ̄)/ŝe*
C = (ζ̄ − T_{α/2}·s̄e,  ζ̄ − T_{1−α/2}·s̄e)
```

非 IID データでは、リサンプル時に時系列依存を壊さないよう circular block bootstrap を用い、必要なら VAR/GARCH による半パラメトリック適合を併用する (Riondato / Two Sigma (2018))。正規性を仮定する漸近 plug-in 区間は前述のとおり最大 70% も分散を誤りうるため、依存や非正規性が疑われる実データではスチューデント化ブートストラップ区間を既定にするのが安全である。

定常／循環ブロックブートストラップ：時系列依存を保つリサンプリング

ブロックブートストラップの核心は「1 点ずつではなく連続ブロック単位でリサンプルすることで自己相関構造を保つ」ことにある。Politis & Romano (1994) の定常ブートストラップ (stationary bootstrap) は、幾何分布に従うランダム長のブロックを連結し、生成される擬似系列自体が定常になるように設計されている。ブロック長 L は幾何分布 Geom(p) に従い、期待ブロック長は b=1/p である。実装は、開始位置をランダムに選んだうえで、各時点で確率 p で新しいランダム開始位置へジャンプ、確率 (1−p) で次点へ継続し、系列端は先頭へ折り返す（wrap-around／circular）。固定長ブロックを使う moving block bootstrap と違い、ブロック境界が確率的なため境界の不連続が回避され、擬似系列が定常になる。整合性の条件は `b→∞ かつ b/n→0`（n は標本長）で、circular／stationary いずれの推定量も O(1/b) のバイアスを持つ (Politis & Romano (1994))。実装上は「平均ブロック長 L を望むなら幾何パラメータを p=1/L と置く」だけでよい (Wicklin / SAS (2021))。

ブロック長は結果を左右する調整パラメータであり、恣意的に決めるべきではない。Politis & White (2004) は、自己共分散和とスペクトル密度からデータ適応的に最適期待ブロック長を自動推定する方法を与えた。最適長は標本長の N^{1/3} に比例する。

```
stationary: b_opt,SB = (2G²/D_SB)^{1/3}·N^{1/3}
circular:   b_opt,CB = (2G²/D_CB)^{1/3}·N^{1/3}
G = Σ|k|R(k),   D_SB = 2g²(0),   D_CB = (4/3)g²(0)
```

D_CB は D_SB より小さく、最適 MSE は O(N^{−2/3}) で減衰する。相関が強いほどブロックは長く選ばれる。なお定常ブートストラップの分散定数 D_SB は当初の Politis & White (2004) では別の表式だったが、Lahiri (1999) の分散計算の誤りに起因する訂正を受けて D_SB = 2g²(0) に改められている (Patton, Politis & White (2009))。実務では、この自動選択でブロック長を決めたうえで前節のスチューデント化ブートストラップ区間や次節の差の検定を回すのが再現性の高い手順になる (Politis & White (2004); Patton, Politis & White (2009))。

2 戦略の Sharpe 比差の検定（Ledoit & Wolf）

「戦略 A は戦略 B より Sharpe 比が有意に高いか」を問うには、差 `Δ̂ = μ̂_i/σ̂_i − μ̂_n/σ̂_n = f(û)` の分布が要る。正規性を仮定する Jobson–Korkie／Memmel 検定は時系列依存や厚い裾の下で妥当性を失うため、Ledoit & Wolf (2008) は時系列に頑健なスチューデント化ブートストラップ信頼区間で差を検定する。標準誤差は HAC で `s(Δ̂) = √(∇f'Ψ̂∇f/T)`、`Ψ̂ = (T/(T−4))Σ_j k(j/S_T)·Ĉ_T(j)` とし、Andrews–Monahan (1992) の prewhitened QS カーネルと自動バンド幅を推奨する。時系列依存を保つため circular block bootstrap を用い、ブロック長 b∈{1,2,4,6,8,10} を K=5000 本の擬似系列でキャリブレーションする。対称スチューデント化ブート CI は `Δ̂ ± z*_{|·|,1−α}·s(Δ̂)` で、z* は `|Δ̂*−Δ̂|/s(Δ̂*)` の分布の (1−α) 分位点である。推論は M=4999 リサンプルで行い、p 値は次で計算する (Ledoit & Wolf (2008))。

```
PV = (#{ |Δ̂*^m| ≥ d } + 1) / (M + 1),   d = |Δ̂| / s(Δ̂)
```

分子・分母の +1 は p=0 を回避する連続性補正である。これにより「2 戦略の Sharpe 比差が、依存と非正規性を織り込んでも偶然で説明できないか」を単一の p 値で判定できる。

モンテカルロ順列検定：シグナルの予測力を帰無で崩す

順列検定 (permutation test) は「何をシャッフルするか」が帰無仮説を規定する点が本質である (Susan Potter (2024))。(a) シグナルのみ時間方向に無作為並べ替えすれば帰無は「シグナルの予測力ゼロ」、(b) 循環並べ替え `signal_perm[t] = signal[(t+offset)%T]` なら帰無は「シグナルのタイミングは無関係」で、価格由来シグナルの自己相関を保存できる、(c) ブロック・リターン順列なら帰無は「リターンはシグナルと独立」である。p 値は観測成績以上を出した順列の割合で測る。

```
p = (# 順列で実績以上 + 1) / (N + 1)
```

+1 は Phipson & Smyth (2010) 補正で p=0 を回避する。推奨は 10,000 順列（最小 p 値は約 1/10001≈0.0001）、有意性を主張する最小標準でも 1,000 順列である (Susan Potter (2024))。価格由来シグナルを素朴に時間シャッフルすると自己相関を壊して帰無分布を歪めるため、循環並べ替え (b) を選ぶべき点が実務上の要諦になる。White (2000) の Reality Check も、全戦略集合を同時に考慮してデータスヌーピング (data snooping) を補正するという意味で同系統の多重検定である (Susan Potter (2024))。

大量試行の最良戦略の検定：Reality Check と SPA

多数の戦略を同一データで試したあとに最良戦略を評価するときは、単一戦略の CI では不足で、探索全体を 1 つの帰無に押し込む必要がある。White (2000) の Reality Check は「最良モデルでもベンチマークを超えない」を帰無とし、正規化サンプル平均の最大値を検定統計量として定常ブートストラップで p 値を得る (White (2000); Hsu & Kuan (2005))。

```
H_0: max_{k=1,…,ℓ} E(f_k) ≤ 0
V̄_ℓ = max_k √n·f̄_k
V̄*_{ℓ,i} = max_k √n(f̄*_{k,i} − f̄_k),   i = 1,…,B
```

B 本のブートストラップ実現 $\bar{V}^{*}$を *Politis–Romano 定常ブートストラップで生成し、観測 $\bar{V}_{\ell}$ をその経験分布の分位点と比べて p 値を出す。全モデル集合を同時に考慮するため多重検定バイアスを補正できる。Hsu–Kuan の適用例では、性能指標に平均リターンと Sharpe 比、ベンチマークにニュートラル・ルール（無ポジション）と日次無リスク金利を置き、39,832 ルール／戦略を検定した (White (2000); Hsu & Kuan (2005))。実装上は、パラメータ格子上の全戦略の超過性能系列をそのまま $\bar{V}$・$\bar{V}$に流し込めばよい。*

Reality Check は「最悪配置 (least favorable configuration; LFC)」由来の保守性を持つ。Hansen (2005) の SPA (superior predictive ability) 検定は、統計量のスチューデント化とサンプル依存の帰無分布でこの保守性を改善し、劣モデル混入への頑健性と検出力を高める。

```
RC:  T_n^RC  = max(√n·d̄_1, …, √n·d̄_m),  帰無分布 N_m(0, Ω̂) を μ=0(LFC) で構成
SPA: T_n^SPA = max( max_k √n·d̄_k/ω̂_k, 0 )   (スチューデント化)
     μ̂_k^c = d̄_k · 1{ √n·d̄_k/ω̂_k ≤ −√(2 log log n) }
```

Hansen の Corollary 1 によれば、無関係な劣モデル（μ_k<0）を追加すると m は増えるが RC の p 値の基礎は変わらず、これを悪用すれば RC の検出力を 0 まで下げられる（操作可能）。SPA はサンプル依存の平均推定でこの劣モデルを帰無から実質的に排除する。局所対立の例（m=2）ではスチューデント化により検出力が約 15%→約 53% と 3 倍超に上がる (Hansen (2005))。したがって、パラメータ格子に無価値な戦略が大量に混じる短期戦略スクリーニングでは RC より SPA を優先すべきである（多重検定補正の詳細は該当セクション参照）。

選抜バイアスへの接続：期待最大 Sharpe 比

以上の分布推定は「多数試行の最良値」を評価する段で選抜バイアス (selection bias) と接続する。真の Sharpe 比がゼロでも、独立試行数 N を増やせば試行中の期待最大 Sharpe 比は正になり、極値統計で近似できる (Bailey & López de Prado (2014))。

```
E[max{ŜR_n}] ≈ E[ŜR] + √V[ŜR]·( (1−γ)·Z⁻¹[1−1/N] + γ·Z⁻¹[1−1/(N·e)] ),  γ≈0.5772
```

E[max] は N の増加とともに単調に増える。これが「候補を増やすほど運だけで良好な候補が現れる」理由であり、hold-out や k-fold 検証は多重試行の偽陽性増加を制御できない（95% 水準を 20 回適用すれば偽陽性 1 件は期待どおり）(Bailey & López de Prado (2014))。この期待最大 Sharpe を棄却閾値に据え、非正規性（歪度・尖度）・標本長・試行間分散・試行数 N の 5 変数で Sharpe をデフレートするのが Deflated Sharpe Ratio (DSR) であり、本セクションの分布推定（非正規性・標本長）と多重試行の分布（期待最大）を 1 つの確率に統合する。DSR の定義と数値例は Deflated Sharpe Ratio のセクションで扱う (Bailey & López de Prado (2014))。

前提・限界と注意点

第一に、ブロックブートストラップはブロック長 b に敏感である。短すぎれば依存構造を壊し、長すぎれば実効リサンプル数が減って推定が不安定になる。circular／stationary 推定量は O(1/b) のバイアスを持ち、整合性には `b→∞ かつ b/n→0` が要る。b は恣意的に決めず Politis & White (2004) の自動選択（b∝N^{1/3}）で定めるべきで、HAC の切断ラグ m も同様に `m/T→0` を満たす必要がある (Politis & Romano (1994); Politis & White (2004); Lo (2002))。

第二に、順列検定はシャッフルする対象が帰無仮説を決めるため、検定したい命題に対応する並べ替えを選ばねばならない。価格由来シグナルの自己相関を保つには循環並べ替えが必須で、素朴な時間シャッフルは帰無分布を歪める。また安定した小さい p 値には 10,000 順列規模が要る (Susan Potter (2024))。

第三に、正規性の誤仮定は分布推定を大きく損なう。IID 正規の SE・plug-in 区間・√12 年率化は、系列相関や厚い裾の下で最大 70% も分散を誤り、年率 Sharpe を最大 65% 過大評価しうる。素朴な Sharpe 推定量自体も小標本で上方バイアスを持つため、ガンマ関数比か k 統計量で補正する (Riondato / Two Sigma (2018); Lo (2002))。

第四に、多数試行の検定には固有の落とし穴がある。Reality Check は劣モデルの混入で検出力を意図的に下げられる（操作可能）ため SPA を優先し、最良戦略の評価では期待最大 Sharpe による選抜バイアス補正を併用する (Hansen (2005); Bailey & López de Prado (2014))。実務のチェックリストは、(1) Ljung–Box で系列相関を診断し、有意なら HAC・$\eta(q)$・ブロックブートストラップに切り替える、(2) 単一戦略はスチューデント化ブートストラップ CI、2 戦略比較は Ledoit–Wolf 検定、シグナル評価は循環順列検定を使う、(3) 大量スクリーニングの最良戦略は SPA と期待最大 Sharpe／DSR で二重に割り引く、の 3 点を組み合わせることに尽きる。

10. バックテストのバイアスと落とし穴の体系

短期クオンツ予測モデルの検証で最終的に評価対象となるのは「バックテストの数字」だが、その数字は無数の設計判断・データ処理・試行選択の合成物であり、各段階に成績を系統的に上振れさせるバイアスが潜む。本セクションは、これらのバイアスと落とし穴を体系化し、「どの落とし穴が、どの局面で、どれだけ成績を水増しするか」を定量的に整理する。個別バイアスの土台にある統計的必然（試行数 N と多重検定）を先に据え、その上で実務家向けの網羅的チェックリストとして Deutsche Bank の「7つの大罪 (Seven Sins)」と Arnott, Harvey & Markowitz (2019) の7項目プロトコルを重ねる。

オーバーフィッティングの定義と「試行数 N の非開示」

バックテストは過去の in-sample (IS) シミュレーションであり、検証では IS 性能と out-of-sample (OOS) 性能を峻別しなければならない。オーバーフィッティング (overfitting) とは機械学習由来の概念で、モデルが一般構造ではなく特定標本のノイズを狙い撃ちする状態を指す (Bailey, Borwein, López de Prado & Zhu (2014))。同論文の中心命題は、選択したバックテスト構成に至るまでに試した試行数 N を報告しない研究者は、オーバーフィッティングのリスクを評価不能にする、というものである。AIC などが 5% 有意水準を通っても、それは単一検定の話にすぎず、報告されない数百の試行を考慮していない (Bailey, Borwein, López de Prado & Zhu (2014))。したがってオーバーフィッティングは真偽二択ではなく、N に依存する非ゼロ確率の量として扱う——これが本セクションの体系全体の基点である。

なぜ N がこれほど効くのかは根本問題に遡る。経済・金融は物理と異なり観測が実質1系列しかなく反復実験ができないため「十分なデータが決して手に入らない」——すなわち真のスキルと偶然を原理的に分離しきれない (Luo et al. / Deutsche Bank (2014))。2007年夏のクオンツ危機で value/momentum が数日のうちに暴落し、正規分布の想定をはるかに超える規模のテール事象となったのは、まさにこの過小評価されたテール構造の表れであった (Luo et al. / Deutsche Bank (2014))。

試行数 N がもたらす選択バイアスの数理

真の OOS Sharpe 比 (Sharpe ratio) がゼロの無スキル戦略でも、独立試行 N を増やすほど「最大 IS Sharpe 比」の期待値は上昇する。標準正規化した Sharpe 比の期待最大値は極値理論から次式で近似され、上界は $\sqrt{(2 \cdot \ln N)}$ で与えられる (Bailey, Borwein, López de Prado & Zhu (2014))。

$$E[\max_N] \approx (1-\gamma)\cdot Z^{-1}\left[1-\frac{1}{N}\right] + \gamma \cdot Z^{-1}\left[1-\frac{1}{N}e^{-1}\right], \quad \gamma \approx 0.5772 \text{（オイラー・マスケローニ定数）}$$

具体的な水増し量は直感を裏切る。全戦略の真の OOS Sharpe が 0 でも、N=10 試行で期待最大 IS Sharpe は 1.57、20 試行で期待的に SR>2.0 の仕様が1つ見つかる。モデル複雑性はこの N を指数的に膨らませ、5 パラメータで $N=2^5=32$、7 個の二値パラメータで $N=2^7=128$ となり期待最大 Sharpe は 2.6 を超える (Bailey, Borwein, López de Prado & Zhu (2014))。裏返した設計指標が最小バックテスト長 (Minimum Backtest Length; MinBTL) で、実務近似は $\mathrm{MinBTL} < 2 \cdot \ln(N) / E[\max_N]^2$。含意は明快で、5 年分のデータしかないなら独立モデル構成を 45 個以下に抑えないと、ほぼ確実に IS Sharpe=1・OOS Sharpe=0 の無スキル戦略を生成してしまう (Bailey, Borwein, López de Prado & Zhu (2014))。

さらに深刻なのは、金融時系列が持つ「メモリ効果／補償効果 (memory / compensation effect＝過密化・平均回帰・構造変化)」の存在下では、オーバーフィットの OOS 期待が単に 0 近傍に留まらず負になることである。補償効果を1つ入れるだけで IS と OOS の Sharpe に強い負の線形関係が生じ、Proposition 3 として「SR_IS が高い戦略ほど SR_OOS は低い」——IS を最適化するほど OOS が悪化する関係が成立する (Bailey, Borwein, López de Prado & Zhu (2014))。これが多くのシステマティック戦略が宣伝どおりに機能しない数理的理由であり、以下の個別バイアスはいずれもこの「試行数を稼ぎながら IS を最適化する」プロセスの具体形にほかならない。

クオンツ投資の「7つの大罪」：バイアスの実務的分類

Luo et al. / Deutsche Bank (2014) は実務で成績を歪める落とし穴を7類型に整理する：(1) 生存者バイアス (survivorship bias)、(2) 先読みバイアス (look-ahead bias)、(3) ストーリーテリング (storytelling)、(4) データマイニング・スヌーピング (data mining and snooping)、(5) シグナル減衰・回転率 (signal decay and turnover)、(6) 外れ値 (outliers)、(7) 非対称ペイオフと空売りコスト (asymmetric payoff and shorting cost)。各類型は具体数値で成績を逆転させうる。

生存者バイアスの実測インパクトは大きい。1986年末の Russell 3000 構成銘柄のうち現在まで生存したのは 3,000 中 500 未満で、生存ユニバースを使うと結論が反転する。Merton の distance-to-default ファクターでは正しい point-in-time (PIT) ユニバースなら高品質（低信用リスク）株が勝つが、生存ユニバースでは最悪信用リスク株が最良品質株の7倍のリターンを生む。低ボラティリティ・アノマリーも、現在構成銘柄バイアス (current constituent bias＝先読みの一種) では高ボラ株が低ボラ株を16倍アウトパフォームして見える (Luo et al. / Deutsche Bank (2014))。対策は非アクティブ企業を含む PIT ユニバースの使用と、欠損値を NA として保持し銘柄を除外しないことである。

先読みバイアスは定量化できる。米企業の四半期決算提出は平均30日・中央値28日を要し、非 PIT データ（報告ラグ無視）は益回りファクターの平均 IC を約60%（PIT 5.11% を 8.12% に）、ROE を約10%水増しする (Luo et al. / Deutsche Bank (2014))。より微妙な Restatement bias（多くのデータソースが最終改訂値のみ保存する問題）や、株式分割の先読み（分割調整済み低位株戦略は年率26%/Sharpe 0.82 に見えるが未調整の正価格では約11%/Sharpe 0.30）も同根であり、理想解は Compustat・Capital IQ 等が提供する PIT データベースである (Luo et al. / Deutsche Bank (2014))。

ストーリーテリングと確証バイアス (confirmatory bias) は統計量に現れない。Langer et al. (1978) のコピー機実験のように人は「理由」が付くと承諾しやすく、パターンを見れば必ず事後の物語を作れるが、説明できることと OOS 性能は無関係である。財務レバレッジは「リスクプレミアで高リターン」とも「行動ファイナンス的に低リターン」とも説得的に説明でき、実データも矛盾する (Luo et al. / Deutsche Bank (2014))。データマイニング／スヌーピングはこれを機械化する：72 の銘柄選択ファクターから各スタイルの最良を選んだ「in-sample モデル」は Sharpe 0.7 を示すが、どれが勝つか事前に知らずに組む正しいローリング OOS モデルは5年間ほぼフラットになる (Luo et al. / Deutsche Bank (2014))。Arnott, Harvey & Markowitz (2019) が引く Chordia, Goyal & Saretto (2017) は 210 万戦略を検証し統計的・経済的閾値を通過したのはわずか 17 戦略であり、ティッカー先頭3文字で作る戦略が50年で年率約6%アルファを生む「alpha-bet」の例は、無意味な組合せでも完璧に見えるバックテストが作れることを示す (Arnott, Harvey & Markowitz (2019))。

外れ値処理は結果を左右する設計選択である。winsorization の水準（5%/1%/0.1%）や正規化手法で結論が変わり、価格モメンタムは先にランキング正規化すると予測力を約35%失う（モメンタムは既にリターン空間にあり正規化不要なため）(Luo et al. / Deutsche Bank (2014))。Arnott, Harvey & Markowitz (2019) は「winsorization 水準はモデル構築前に決めよ、『5% では機能するが 1% では失敗し 5% を採用』は不正な研究プロセスの明白な兆候」と警告する。最後にシグナル減衰・回転率と空売りコストが live と backtest の乖離を生む：日本のリバーサル戦略は無コストで年率12%だが取引コストを30bps 課すとアルファが負に転落し、現実的な空売り制約（Data Explorers の DCBS で示される hard-to-borrow 銘柄）を課すとショート側高アルファ戦略の累積性能は約半分に低下する (Luo et al. / Deutsche Bank (2014))。

多重検定の閾値：t>2.0 は甘すぎる

上記の落とし穴を統計的に締め上げるのが多重検定 (multiple testing) の管理である。Harvey, Liu & Zhu (2016) は 313 本の論文（うち公刊 250 本・ワーキングペーパー 63 本）から 316 ファクターを収集し、数百のファクターが試された今、Fama-MacBeth (1973) 当時の t>2.0 基準は不適切だとして、新ファクターに t 比 > 3.0（両側 p≈0.27% 相当）のハードルを課した。ファクター発見のペースはこの数十年で急激に加速しており（同論文は直近10年の産出率が続くと仮定して 2032 年まで必要 t 閾値を外挿する）、結論として「金融経済学で主張される研究発見の大半は偽である可能性が高い」と述べる (Harvey, Liu & Zhu (2016))。具体的な調整は3手法で、いずれも検定統計量間の一般相関を許容する：Bonferroni（FWER 制御）、Holm（逐次、FWER 制御、Bonferroni を常に支配）、Benjamini-Hochberg-Yekutieli＝BHY（逐次、FDR 制御）。M=10 検定・α=5% の例では、単一検定なら10ファクター全てが有意になるが、有意と残るのは Bonferroni で3個・Holm で4個・BHY で6個にすぎず、許容 p 値は Bonferroni で 0.5%、Holm で 0.60%、BHY で 0.85% まで厳格化される (Harvey, Liu & Zhu (2016))。実務では、試した候補数 M を右辺に入れて閾値 t を引き上げ、単一戦略の是非なら FWER、多数戦略の選別なら FDR を使い分ける。

AHM の7項目バックテスト・プロトコル

Arnott, Harvey & Markowitz (2019) は上記を運用に落とす7カテゴリのチェックリストを提示する：(1) Research Motivation（事後のストーリー作りを警戒し ex ante の経済的基盤を確立）、(2) Multiple Testing & Statistical Methods（試した全モデル・変数を記録。20 個の無作為戦略なら1つは偶然 t≧2.0 を超えるので、複数戦略を試すなら t=2.0 は無意味な基準）、(3) Sample Choice and Data（標本を事前定義し、外れ値・winsorization 水準を事前決定）、(4) Cross-Validation（真の OOS はライブ運用のみ、iterated OOS は OOS でない、取引コストを無視するな）、(5) Model Dynamics（構造変化、過密化＝Heisenberg 不確定性、ライブモデルの弄り回し禁止）、(6) Model Complexity（次元の呪い、正則化と解釈可能な ML）、(7) Research Culture（勝ちより科学の質を報奨し、大半の実験は失敗する前提を置く）(Arnott, Harvey & Markowitz (2019))。追試の帰結を「winner's curse」として定式化し、頑健／効果が大幅縮小（例：マイクロキャップ除外時）／効果消滅の3結果に分類する。実際 Arnott-Beck-Kalesnik (2016) では主要8ファクターの公刊後アルファが年 5.8%→2.4%（約60%喪失）へ減衰し、Hou-Xue-Zhang (2017) はアノマリー超過リターンの大半がマイクロキャップ除外で消えると報告する。決定的な制約

として、金融の高品質データは約55年（<700 月次／約200 四半期）しかなく、ML/深層学習にはそもそも小さすぎる (Arnott, Harvey & Markowitz (2019))。

前提・限界と注意点

これらを使ううえで共有すべき前提がある。第一に、「7つの大罪」は投資銀行の実務家向けリサーチノートであり学術査読を経ていない点に留意が必要である (Luo et al. / Deutsche Bank (2014))。第二に、多重検定の閾値も試行数 N・M を正しく数えられることに依存し、探索・特徴量選択・閾値調整を含む全試行を記録しない限り指標は容易に甘くなる (Harvey, Liu & Zhu (2016))。第三に、いかなる in-sample 検証も真の OOS ではない。Bailey, Borwein, López de Prado & Zhu (2015) は単純 hold-out の5欠陥（公開データでは既に hold-out を IS に使っている可能性、熟練研究者は OOS 期間の変数挙動を既知、小標本では観測1,000 未満に適用不可＝週次戦略なら20年未満に使うべきでない、サンプル末尾の直近喪失、試行数を無視）を挙げ、Arnott, Harvey & Markowitz (2019) も「真の OOS はライブ運用のみ (the only true out of sample is the live trading experience)」と一致する。

したがって実務チェックリストは以下を組み合わせる：(1) 検証設計の段階で手元データ長に対する MinBTL を計算し許容試行数の上限を先に決める、(2) データ処理では PIT ユニバースと報告ラグを厳守し、winsorization・正規化・除外ルールをモデル構築前に固定する、(3) 事前定義ルールでバックテストし、単一データセットで開発したら他国・別期間で検証、validation sample はライブ前まで触れない、(4) 最終判定では試した候補数を踏まえた多重検定調整（t 閾値の引き上げ・FDR 制御）を課し、live 性能を最終検定とする (Bailey, Borwein, López de Prado & Zhu (2015); Luo et al. / Deutsche Bank (2014); Harvey, Liu & Zhu (2016); Arnott, Harvey & Markowitz (2019))。本セクションの分類は、Sharpe 比の数値そのものではなく「その数値がどの落とし穴で膨らんでいるか」を系統的に点検するための地図であり、定量指標（DSR・PBO 等）と併用して初めて偽の発見を排除できる。

11. 取引コスト・マーケットインパクト・キャパシティ・アルファ減衰

バックテスト上のシャープレシオ (Sharpe ratio) が統計的に本物 (前節までの多重検定・過学習の関門を通過) であっても、それが実運用で残るとは限らない。短期クオンツ戦略は回転率 (turnover) が高く、シグナルの予測地平 (horizon) が短いほど1トレードあたりのアルファは小さいため、取引コストとマーケットインパクト (market impact) が期待リターンを食い潰しやすい。本セクションは、(1) 執行コストのモデル化 (最適執行)、(2) マーケットインパクトの関数形 (線形の恒久インパクト vs 凹な一時インパクト)、(3) 実測コスト水準と学術推計の乖離、(4) 戦略キャパシティ (capacity) の推定、(5) 公開後のアルファ減衰 (alpha decay)、を扱う。実務家がチェックすべきは「グロスのシグナルが、現実的なコストモデルとファンド規模のもとでネットで生き残るか」であり、本節はそのための道具立てを与える。

実装ショートフォールと最適執行：Almgren-Chriss 枠組み

執行コストの基準量は「実装ショートフォール (implementation shortfall)」である。Almgren & Chriss (2000) は、意思決定時点の紙上価値 XS_0 と実現収益との差 (Perold の implementation shortfall に一致) を、期待ショートフォール E(x) と分散 V(x) に分解し、リスク回避度 λ (トレーダー効用の曲率) のもとで平均分散目的 `E(x) + λV(x)` を最小化する執行軌道 (trading trajectory) を求める枠組みを与えた。ここで E(x) は恒久インパクト (permanent impact) と一時インパクト (temporary impact) の和、$V(x)=\sigma^2\cdot\Sigma\,\tau\cdot x\_k^2$ は執行途中の価格変動リスクである。速く清算すれば価格変動リスクは下がるがインパクトコストは上がり、ゆっくり清算すればその逆になる、というトレードオフを定式化した点が核心である (Almgren & Chriss (2000))。

インパクトを線形と仮定すると閉形式の二次コストモデルが得られる。Almgren & Chriss (2000) は恒久インパクトを `g(v)=γv` (売却量に比例、執行速度に依らず価格を押し下げる)、一時インパクトを `h(n_k/τ)=ε·sgn(n_k)+(η/τ)·n_k` とした。ε は「ビッドアスクスプレッドの半分＋手数料」という固定費、η は流動性の過渡的コストである。結果として期待コストは

```
E(x) = ½γX² + ε|X| + (η̃/τ)·Σ n_k² ,   η̃ = η − ½γτ
```

となり、$\eta>0$ で厳密凸になる (Almgren & Chriss (2000))。この目的関数の最小化解は双曲線関数 (hyperbolic sinh) で減衰する:

```
x_j = [ sinh(κ(T−t_j)) / sinh(κT) ] · X ,   κ² = λσ²/η̃ ,   κ ≈ √(λσ²/η)
```

トレードの「半減期 (half-life) $\theta=1/\kappa$」は、ボラティリティ σ・一時インパクト η・リスク回避度 λ だけで決まり、外生的な締切 T には依存しない。$T\gg\theta$ なら最小分散型 (前倒しの速い売却)、$T\ll\theta$ なら最小コストの直線 (VWAP) 型に近づき、効率的フロンティアは滑らかな凸曲線となる (Almgren & Chriss (2000))。数値例では S_0=\$50・X=100万株・年率30%ボラ・スプレッド1/8・日次出来高500万株に対し、ε=1/16、「日次出来高の1%を取引するとスプレッド1つ分動く」から η=2.5e-6、「日次出来高の10%売却でスプレッド1つ分動く」から γ=2.5e-7 と較正し、λ=1e-6 で $\kappa\approx$0.6/日、5日清算なら

$\kappa T \approx 3$ の中間領域になる (Almgren & Chriss (2000))。実務では、自戦略の想定執行サイズ・保有期間・銘柄ボラから半減期 θ を計算し、シグナルの予測地平が θ より短ければ「執行が間に合わずインパクトでアルファが溶ける」ことを事前診断できる。ただし本モデルは、バスケットサイズに依らず同一時間スケールで清算し (大口でも同じ速度)、著者自身が大口では超線形の η への切替を勧める点に注意が必要である (Almgren & Chriss (2000))。

マーケットインパクトの関数形：線形の恒久インパクト vs 凹な一時インパクト

コストモデルの心臓部はインパクト関数の関数形である。理論の出発点は Kyle の λ (Kyle's lambda) で、Bouchaud (2010) が整理するように、マーケットメイカーの価格更新は符号付き注文流に線形 `Δp = λ·ε·v` で、インパクトは恒久的 (`p_T = p₀ + λ·Σε_n v_n`) となる。λ は市場の深さ (depth) の逆指標である。Huberman & Stanzl (2004) により、恒久インパクトが線形であることが価格操作 (quasi-arbitrage) を排除する必要十分条件である（一時インパクトはより一般の関数形でよい）ことが示されている。ただし価格がランダムウォークになるにはトレード符号が無相関でなければならないが、現実の注文流の符号は長期記憶を持ち、これが後述の凹なメタオーダー・インパクトの動機になる (Bouchaud (2010))。

実証面では、一時インパクトはむしろ強く凹 (concave) である。Tóth et al. (2011) は「平方根則 (square-root law)」 `Δ(Q) = Y·σ·√(Q/V)` (σ=日次ボラ、V=日次出来高、Y は1オーダーの定数) を提示し、これが資産クラスを超えて驚くほど普遍的だと報告した。より一般に `Δ∝Q^δ` で δ は概ね 0.5〜0.7 (Q/V が数×10^{-4}〜数% の範囲)、具体的には「日次出来高の100分の1を取引すると価格が日次ボラの10分の1動く」という強い増幅が生じる。彼らはこれを、価格近傍で潜在流動性 (latent liquidity) が消失するV字型の限界需給曲線 (＝市場は臨界状態) で説明する (Tóth et al. (2011))。Bouchaud (2024) は、この凹依存 ($\delta \approx 0.5$) が先物・株式・オプション・ビットコインまで、地域や期間に依らずほぼ普遍で、しかも「子注文数 N や総執行時間 T にほぼ依存せず総サイズ Q だけで決まる」ため、Kyle の線形モデル (インパクトは Q に線形＝傾きが Kyle の λ) と根本的に矛盾すると強調する (Bouchaud (2024))。

指数の値そのものは文献で割れる。Almgren, Thum, Hauptmann & Li (2005) はシティグループの約70万件 (フィルタ後29,509件、サイズは日次出来高の0.25%〜数%) の実注文を用い、自由較正で恒久指数 $\alpha = 0.891 \pm 0.10$、一時指数 $\beta = 0.600 \pm 0.038$ を得て、「95%水準で平方根モデル $\beta = 1/2$ は棄却される」として $\beta = 3/5$ を採用した (小口では平方根よりやや小さく、大口ではやや大きいコストになる)。較正済みモデルは

```
恒久 I = γ·σ·(X/V)·(θ/V)^(1/4) ,   一時 J = I/2 + sgn(X)·η·σ·(X/(V·T))^(3/5)
```

で、恒久を線形 ($\alpha = 1$) にしたのは無裁定 (Huberman & Stanzl 2004) かつ執行速度に依存させないためであり、回転率 (turnover) が低い銘柄ほどインパクトが大きい。ただし単一トレードの結果はボラが支配するため、フィットの R^2 は1%未満である (Almgren, Thum, Hauptmann & Li (2005))。実務上の含意は、コストモデルの一時インパクトを「線形」で置くと大口で過大・小口で過小に評価しがちなので、平方根 ($\delta \approx 0.5$) か 3/5 の凹関数を採り、係数は σ・出来高・回転率でスケールさせるべき、という点である。恒久成分だけは (無裁定と符号相関の観点から) 線形で置くのが理論的に整合的である (Bouchaud (2010); Almgren, Thum, Hauptmann & Li (2005))。

実測コスト水準と学術推計の乖離：TAQ は桁で過大評価する

コストの絶対水準は、データ源によって桁が変わる。Frazzini, Israel & Moskowitz (2018) は約1.7兆ドル・約7億件・21の先進国市場・約1万銘柄の実約定 (1998年8月〜2016年6月) を用い、平均マーケットインパクト 9.97 bps、平均実装ショートフォール 11.02 bps (バリュー加重＝大口では 15.14 / 16.06 bps) と測定した。インパクトの約85%は恒久的 (翌日反転は約1.26 bps のみ)、実効スプレッドは約4.3 bps で手数料は小さく、これらは TAQ ベースの学術推計より約1桁小さいと報告している (Frazzini, Israel & Moskowitz (2018))。彼らはインパクトを参加率の凹関数

```
MI = a + b·x + c·sign(x)·√|x|    (x = 日次出来高%)
```

でモデル化する (インパクトをサイズに回した log-log 回帰の傾きは 0.35、R^2=95% で平方根に近い)。このモデルでは日次出来高の2%で約 13.73 bps (TAQ線形は 44.89、TAQ平方根は 29.07 bps)、10%で約 32.34 bps となり、TAQ 系はおよそ倍を出す。アウトオブサンプル検証として、S&P500ファンドで年 4.81 bps (Vanguard VFIAX 実績≈4.72)、Russell 2000ファンドで年 12.36 bps (iShares IWM 実績≈12.87) を再現し、線形 TAQ モデルは6〜12倍過大評価すると示した (Frazzini, Israel & Moskowitz (2018))。実務家への教訓は明快で、コストの前提をどのモデルで置くかで戦略の生存性判定が桁で変わるため、可能なら自社の実約定 (パッシブファンドの追跡誤差でも可) でコストモデルを外部妥当性検証すべき、ということである。

戦略キャパシティの推定

キャパシティ (capacity) とは、期待アルファがコストで完全に打ち消される損益分岐 (break-even) の運用規模である。実コストが従来推計の約10分の1なら、損益分岐規模は約10倍になる。Frazzini, Israel & Moskowitz (2012) は約1兆ドルの実約定 (531万件、9,128銘柄、19市場、1998〜2011) を用い、Fama-French のロングショートで米国＝size $103B・value $83B・momentum $52B、グローバル＝$156B・$190B・$89B の損益分岐規模を得た。短期反転 (short-term reversal) は米国 $9B・グローバル $13B までしか生き残らない。負相関のトレードをネッティングする value+momentum の複合ブックは米国 $100B・グローバル $177B に達し、追跡誤差1%以内で取引コスト最適化した版はグローバルで size $1,807B・value $811B・momentum $122B・STR $17B まで拡大する。これは Korajczyk & Sadka (2004) のロングオンリー・モメンタム $2〜5B より1桁大きい (Frazzini, Israel & Moskowitz (2012))。実務では、自戦略の回転率・シグナル減衰・参加率を上記いずれかのインパクトモデルに入れ、想定AUMでのネット期待リターンがゼロを割る点を損益分岐規模として算出する。回転率の高い短期戦略 (STR がその典型) はキャパシティが小さく、コストモデルの選択に対する結論の感度が特に高いことに留意する。

公開後のアルファ減衰と戦略ライフサイクル

最後に、生き残ったアルファも時間とともに減衰する。McLean & Pontiff (2016) は80研究97個のクロスセクション予測子を再現し、平均インサンプルのロングショートリターンは月 58.2 bps (平均 t 値 3.55) だが、アウトオブサンプル (公開前) で26% (≈15.0 bps)、公開後で58% (≈33.7 bps) 減衰すると報告した。前者はデータマイニング・統計的バイアスの上限、両者の差 58%−26%＝32% は公開情報に基づく裁定

(arbitrage) に帰される。減衰はインサンプル・リターンが大きい予測子ほど、また流動性が高く固有リスクの低い (裁定しやすい) 銘柄ほど強く、純粋なデータマイニングやリスクではなくミスプライシング＋裁定限界と整合的である (McLean & Pontiff (2016))。この裁定は実測できる。公開後、予測子ポートフォリオでは取引量 +18.7%・売買代金 +9.7%・分散 −6.5% となり、ショートとロングの空売り残高スプレッド (インサンプル平均 0.143%) が拡大し、既公開の他予測子との相関 (crowding) が上がる — 学術公開が裁定を誘発しアノマリーを侵食する直接証拠である (McLean & Pontiff (2016))。実務家への含意は二つある。第一に、公開済みシグナルはグロスのバックテスト値を額面で信用せず、公開後減衰 (平均で約6割) を織り込んでネット期待値を割り引くこと。第二に、シグナルの取引量・空売り残高・他ファンドとの相関を継続監視し、crowding の進行 (＝アルファ枯渇の先行指標) をモニタリングすることである。

前提と限界・注意点

本節の道具立てには、実務家が検証設計で意識すべき既知の論争点がある。第一に、一時インパクトの指数は平方根 (β=1/2、Tóth et al. (2011)) か 3/5 (Almgren, Thum, Hauptmann & Li (2005)) かで割れており、フィットのレンジ・切片の有無・資産の混ぜ方で結論が変わる。感度分析として複数の指数でネット損益を再評価するのが安全である。第二に、恒久インパクトは無裁定条件から線形が支持される (Kyle / Huberman-Stanzl) 一方、実データのメタオーダー・インパクトは強く凹で、両者は latent order book の V字流動性で調停される (Bouchaud (2010); Bouchaud (2024)) — 恒久成分は線形、一時成分は凹、と使い分けるのが実務的折衷である。第三に、コストの絶対水準は TAQ ベースの学術推計が実約定より約1桁大きく (Frazzini, Israel & Moskowitz (2018))、キャパシティ・生存性の結論はコストモデル依存性が極めて高い。第四に、Almgren-Chriss の線形二次モデルは大口で非現実的 (超線形 η への切替が必要) であり、較正パラメータ (γ, η, ε) は市場・期間・回転率で変わる非定常量である (Almgren & Chriss (2000))。したがって「グロスのシグナルが本物か」を確認した後は、必ず (a) 凹インパクトを含む複数コストモデル、(b) 想定AUMでのキャパシティ、(c) 公開後減衰の割引、の三点をネット期待リターンに反映させて初めて、短期予測モデルが実運用で生き残るかを判定できる。

12. レジーム変化・非定常性・パラメータ頑健性・実運用検証

短期クオンツ予測モデルは、特徴量とリターンの統計的関係が時間を通じて安定であることを暗黙に仮定する。しかし金融市場のデータ生成過程 (data-generating process) は非定常であり、平均・ボラティリティ・相関・予測係数はレジーム (regime) 間で不連続に変化する。ある局面で較正したモデルが次の局面で機能しなくなるのはこのためである。前段までの Deflated Sharpe Ratio (DSR) や Probability of Backtest Overfitting (PBO) が「過去標本に対する過剰最適化 (overfitting)」を締め上げるのに対し、本セクションは「そもそも過去の関係が将来も続くのか」という非定常性・レジーム変化の軸を扱う。具体的には、(a) 構造変化 (structural break) の検定、(b) レジームスイッチングと効率性の非定常性のモデル化、(c) メモリを保持したまま価格系列を定常化する分数次差分、(d) ウォークフォワードによるパラメータ頑健性の検証、(e) 唯一の真の out-of-sample (OOS) としてのライブ運用検証、を順に述べる。DSR・PBO を通過した戦略でも、レジームが変われば、あるいは公表・過密化でエッジが摩耗すれば、ライブで崩れうる——ここが本セクションの検証対象である。

構造変化の検定：ブレーク時点が既知（Chow）／未知（Bai-Perron・CUSUM）

ブレーク時点が事前に既知の場合、部分標本間で回帰係数が等しいか（＝構造変化の有無）は Chow 検定の F 統計量で調べる (Chow (1960))。

```
F = [ (S_C − (S_1 + S_2)) / k ] / [ (S_1 + S_2) / (N_1 + N_2 − 2k) ]
```

S_C は結合データの残差二乗和 (SSR)、S_1・S_2 は各部分標本の SSR、k は総パラメータ数、N_1・N_2 は各標本サイズで、帰無仮説 H0（両部分標本で係数が等しい＝ブレーク無し）の下で自由度 (k, N_1+N_2−2k) の F 分布に従う。限界は、ブレーク時点を事前に与える必要があり未知位置に不向きなこと、誤差の等分散性・正規独立性を仮定することである (Chow (1960))。

ブレーク時点が未知で複数個ありうる一般の場合には Bai & Perron (1998) を用いる。線形回帰で m 個の未知ブレークを SSR 最小化（効率的な動的計画法アルゴリズム）で同時推定し、検定は supF(k)（ブレーク無し対 k 個）、上限 M までの未知個数に対する double-maximum 検定 UDmax・WDmax、そして逐次 supF(l+1|l)（l 個対 l+1 個）で「特殊から一般へ」とブレーク個数を整合的に決定する。一部パラメータのみがシフトする部分構造変化モデルを許容し、ブレーク割合の推定量は一致性を持つ (Bai & Perron (1998))。

ブレーク時点を指定せず、係数の恒常性 (parameter constancy) を逐次・グラフィカルに監視するのが Brown, Durbin & Evans (1975) の CUSUM 検定である。係数一定という帰無仮説の下でゼロ平均・一定分散・無相関となる「再帰残差 (recursive residuals)」を導入し、そのスケール化累積和を一対の直線状信頼境界とともにプロットして、境界を越えたら係数一定を棄却する。CUSUM of squares 検定は残差二乗の累積和で分散シフトを検出する (Brown, Durbin & Evans (1975))。実務上は、予測モデルの係数（ファクター・ローディングやシグナルへのリターン回帰）をローリング窓で走らせ、これら3手法で「特徴量とリターンの関係がいつ壊れたか（＝パラメータがいつ無効化したか）」を検知するのに使う。ブレークが既知イ

ベント（規制変更等）に紐づくなら Chow、位置も個数も未知なら Bai-Perron、常時監視なら CUSUM と使い分ける。

レジームスイッチングと市場効率性の非定常性

構造変化を「少数の持続的レジーム間の遷移」として陽にモデル化するのがマルコフ・スイッチング (Markov switching) モデルである。リターンの平均・ボラティリティ・自己相関・相互共分散がレジーム間で異なることを許すことで、単一レジームでは説明できない定型化事実——fat tails・不均一分散 (heteroskedasticity)・歪度・時変相関——を再現でき、急激なスイッチ後の新しい挙動は数期間持続する (persistence)。レジーム切替は最適ポートフォリオ選択に大きな帰結をもたらし、非線形なリスク・リターン関係を誘発する (Ang & Timmermann (2012))。含意は、短期シグナルの Sharpe 比はレジーム条件付きであり、現在どのレジームにいるかを推定せずに単一の期待値を当てにするのは危険だということである。

なぜエッジが恒久的でないかの理論的背景を与えるのが Adaptive Markets Hypothesis (AMH) である。効率性と非効率性は共存し時間とともに変化し、選好・各種トレーダーの相対規模・税制・規制環境の変化がリスクプレミアム関係を不安定化させ、リスクプレミアムは時変・非定常となる。安定・定常・予測可能な環境では効率的市場仮説 (EMH) が良い近似だが、動的・確率的な環境では行動的規則性が現れ EMH は崩れる。帰結として、裁定機会と定量戦略のパフォーマンスは景気循環的 (cyclical) に盛衰し、恒久的なエッジではなく有限寿命を持つ (Lo (2004))。したがって検証では「このエッジは永続するか」ではなく「どのレジームで効き、いつ減衰するか」を問うのが正しい。

非定常価格系列の定常化とメモリ保持：分数次差分

推論には定常性が要るが、予測にはメモリ（過去の情報）が要る——この二律背反の最適解が分数次差分 (fractional differentiation) である。整数次差分（リターン化、d=1）は価格を（ほぼ）定常化するがメモリをほぼ消去してしまう。そこで実数 $0<d<1$ の分数次差分を用い、ADF 等の定常性検定（95% 臨界値）を通過する最小の d^*（最小限の差分）を選ぶことで、定常性とメモリを両立させる。実例として e-mini S&P500 は小さな d で既に定常化し（ADF が 95% 臨界値を越える）、原系列との相関を 90% 超保持する。López de Prado は世界の流動的な 87 先物すべてで $d<0.6$（多くは非常に小さい d）で定常化を達成したと報告している (López de Prado (2018))。実務では、価格から特徴量を作る際に単純リターン（d=1、メモリ喪失）ではなく分数次差分を用いることで、モデルが仮定する定常性を満たしつつ予測に効くメモリを残せる。

パラメータ頑健性：ウォークフォワード分析

パラメータの頑健性は、単一の分割ではなくウォークフォワード分析 (Walk-Forward Analysis; WFA) で連結 OOS 成績として評価する。インサンプル (IS) 窓でパラメータを最適化し、直後の未見 OOS 窓で検定し、窓を前進 (rolling) または拡張 (anchored) させて反復し、全 OOS 結果を連結して評価する。戦略が「頑健」と判定されるのは、OOS 窓で一貫して利益を出し、OOS 成績が IS 成績の妥当な割合（walk-forward efficiency）を保つときである (Pardo (2008))。決定的な頑健性の証拠は、パラメータ空間で摂動に安定な広い「高原 (plateau)」——近最適が連なる領域——であり、孤立した鋭いピークはノイズへの過

学習の兆候とみなす。窓サイズと再最適化頻度も WFA で決める (Pardo (2008))。実務上は、選んだパラメータ集合が孤立ピークでないこと（隣接パラメータも近最適であること）と walk-forward efficiency が妥当であることの両方を要求し、鋭いピークしか出ないパラメータは採用しない。

実運用検証：真の OOS はライブのみ、公表後の減衰

いかなる過去データ上の検証も真の OOS ではない。ドリフトの無いランダムウォーク（無スキル戦略）では、IS の Sharpe 比と OOS の Sharpe 比は無相関で、両者の散布図は原点 (0,0) を中心とする対称な雲状となり、IS でのオーバーフィットは OOS 成績に何ら寄与しない (Bailey, Borwein, López de Prado & Zhu (2014))。すなわち、過去に高原が見えたことは将来のレジームでも高原が続く保証にはならず、真の OOS は前方向のライブ運用のみである。

公表・過密化による減衰は実証されている。McLean & Pontiff (2016) はアカデミック文献の 82 特性（working-paper 版。最終 Journal of Finance 版では 97 予測因子）を検証し、統計的バイアスによる OOS 減衰は平均約 10%（ゼロと統計的に有意差なし）だが、統計的バイアスと公表を知った投資家の売買圧力の双方に帰される公表後 (post-publication) の減衰は平均約 35% に達すると報告した。減衰は裁定コストが低く流動性の高いアノマリーで大きく、単なるデータマイニングではなく公表を知った裁定資本の流入 (informed arbitrage) と整合的である。最終版の見出し値では OOS で 26% 低下、公表後 58% 低下、差の 32% が公表起因とされる (McLean & Pontiff (2016))。含意は明快で、ライブ移行時にはエッジの減衰を織り込み、実弾投入前にペーパートレード／フォワードテストで検証し、稼働後は CUSUM や Bai-Perron でレジーム・ブレークを監視し、公開シグナルなら過密化による減衰を前提に置くべきである。

前提・限界と注意点

第一に、Chow 検定はブレーク時点既知・等分散・正規独立誤差を仮定するが、これはまさにレジームスイッチングが対象とする fat tails・不均一分散のリターンに対しては誤設定になりうる (Chow (1960); Ang & Timmermann (2012))。第二に、Bai-Perron・CUSUM はいずれも線形回帰の枠組みであり、恒常性を検定する回帰式の設定に結果が依存する。回帰を誤設定すればブレークの誤検出・見落としが起きる (Bai & Perron (1998); Brown, Durbin & Evans (1975))。第三に、分数次差分の d^* は用いる定常性検定とその臨界値に依存し、系列ごとに再推定を要する普遍定数ではない (López de Prado (2018))。第四に、ウォークフォワードもあくまで過去データ上の手続きであり、履歴内で高原が見えてもその高原がレジーム変化を生き延びる保証はない——真の OOS はライブのみである (Pardo (2008); Bailey, Borwein, López de Prado & Zhu (2014))。第五に、McLean & Pontiff (2016) の公表後 35% 減衰は「公開されたアカデミック・アノマリー」に対する値であり、非公開の独自短期シグナルにそのまま適用できる数値ではない。ただしエッジが有限寿命を持ち景気循環的に減衰するという定性的含意は独自シグナルにも及ぶ (McLean & Pontiff (2016); Lo (2004))。

実務チェックリストは以下を組み合わせる：(1) 価格系列は分数次差分で「定常性を満たす最小 d^*」を選び、メモリを保持したまま特徴量化する、(2) レジームスイッチング・モデルの当てはめと構造変化検定（Chow／Bai-Perron／CUSUM）で「現在どのレジームにいるか」「モデル係数がいつ壊れたか」を把握する、(3) ウォークフォワードで高原の存在と walk-forward efficiency を確認し、孤立ピークを排除する、(4) フォワード／ペーパートレードで減衰を実測しつつ CUSUM で恒常性を常時監視し、ライブ運用を唯一

の真の OOS として最終検定に据える。これらは非定常な市場において「過去に効いた関係が、いつまで・どのレジームで効くのか」を系統的に点検するための道具立てであり、DSR・PBO といった過剰最適化指標と併用して初めて、偽の発見とレジーム依存の脆いエッジの双方を排除できる。

13. 金融機械学習の実務フレームワーク（López de Prado ほか）

短期クオンツ予測モデルの検証で「なぜ機械学習 (machine learning; ML) ファンドの多くが失敗するのか」を、López de Prado (2018) の論文「The 10 Reasons Most Machine Learning Funds Fail」(Journal of Portfolio Management 44(6)) は根性論ではなく再現可能な工程管理の問題として捉え直した。失敗を4カテゴリ・10個の落とし穴 (pitfall) に体系化し、各々に機械的な処方箋 (solution) を対応させる枠組みを示す。同論文の Exhibit 1 は各落とし穴に「カテゴリ」欄を与え、認識論的 (Epistemological, 1–2)・データ処理 (Data processing, 3–4)・分類問題 (Classification, 5–7)・評価 (Evaluation, 8–10) の4群に分類している (López de Prado (2018))。本セクションは、この10項目を検証パイプラインのチェックリストとして使い、「自分の短期戦略が認識論・データ処理・分類問題・評価のどの層で既知の罠を踏んでいるか」を項目単位で点検できるようにすることを目的とする。同論文は自ら「本稿は書籍『Advances in Financial Machine Learning』(Wiley, 2018) に部分的に基づく」と述べており、枠組みの実装詳細はこの書籍にある (López de Prado (2018))。

失敗の分類学：4カテゴリ・10ピットフォールと処方箋（Exhibit 1）

10の落とし穴と処方箋の対応表は次のとおりで、4カテゴリに分かれる：認識論的 (1–2)、データ処理 (3–4)、分類問題 (5–7)、評価 (8–10) (López de Prado (2018))。

```
(1) シシュポスのパラダイム          → メタ戦略パラダイム
(2) バックテストによる研究          → 特徴量重要度分析
(3) 時間（時系列）サンプリング      → ボリューム・クロック（ドル/ボリューム・バー）
(4) 整数階差分                      → 分数階差分
(5) 固定時間ホライズン・ラベリング→ トリプルバリア法
(6) 方向とサイズの同時学習          → メタラベリング
(7) 非IID標本の重み付け             → 一意性重み付け／逐次ブートストラップ
(8) 交差検証リーク                  → パージング＆エンバーゴ
(9) ウォークフォワード・バックテスト → 組合せパージ交差検証（CPCV）
(10) バックテスト過学習             → 合成データでの検証／デフレート・シャープレシオ
```

各行は独立した「ゲート」であり、上流を塞がずに下流だけ厳格化しても漏れる点が要点である (López de Prado (2018))。

認識論の層：メタ戦略パラダイムと「バックテストは研究ツールではない」

第一の罠はシシュポスのパラダイム、すなわち個々のクオンツに単独で戦略を作らせる方式である。裁量PMは互いにサイロ化して働くが、この形式をML案件に持ち込むと破綻する。「50人のPhDを雇い各自6ヶ月で戦略を作れ」という発想では、各自が焦って探索した結果、(1)過学習 (overfitting) バックテストの見栄えのよい偽陽性、または (2)アカデミックな裏付けはあるが低シャープの混雑ファクターに落ち着く。仮に50人中5人が真の発見をしても、その利益は50人分の費用を賄えない。処方箋は組立ライン (assembly line) 方式で、タスクをサブタスクに分割し各クオンツを専門特化させるメタ戦略パラダイムであり、これによって幸運 (lucky strikes) に頼らず予測可能なペースで発見が生まれる (López de Prado (2018))。

第二の罠は、同じデータに ML→バックテスト→調整 を繰り返して見栄えのよい結果が出るまで探す行為で、これはウォークフォワードのアウトオブサンプルであっても偽発見に至る科学的不正である。米国統計学会 (ASA) の倫理指針 (2016) が警告するほど有名な誤りで、「標準的な有意水準（偽陽性率）5%で（偽の）戦略を発見するには、およそ20回の反復で足りる」(López de Prado (2018))。処方箋は特徴量重要度分析 (feature importance) を研究ツールに据えることである。分類器を学習し交差検証で汎化誤差を評価したうえで、どの特徴が予測力に寄与し、その重要度が全期間で安定か・レジーム変化で変わるか・他資産へ転移するかを実験する方が、盲目的なバックテストの反復よりはるかに優れた研究法である (López de Prado (2018))。

データ処理の層：ボリューム・クロックと分数階差分

第三の罠は時間バー（一定時間間隔のサンプリング）である。二つの問題がある。(1)市場は一定時間間隔で情報を処理しない（寄付き後は正午付近より活発）ため、低活動期を過剰サンプル・高活動期を過小サンプルする。(2)時間サンプリングされた系列は系列相関・不均一分散・非正規性など統計的性質が悪い（GARCHはこの不均一分散対処のため一部開発された）。処方箋は取引活動の従属過程 (subordinated process) としてバーを形成すること、すなわちボリューム・クロックである。ドル・バーは事前定義の市場価値が取引されるたびに1観測をサンプルし、分割・併合・自社株買い・増資などのコーポレートアクションに頑健で、一定バーサイズでの1日あたりバー数の変動がティック・バーやボリューム・バーより小さい (López de Prado (2018))。

第四の罠は整数階差分、すなわち価格をリターン化して定常化する慣行が、必要以上に記憶 (memory) を消し去ることである。処方箋は非整数（正の実数）階差 d による分数階差分 (fractional differentiation) で、定常性と記憶を両立させる。系列 {x_t} に対し重み付き和で定義する (López de Prado (2018))。

```
x̃_t = Σ_{k=0}^∞ ω_k · x_{t−k},   ω_0 = 1,   ω_k = −ω_{k−1}·(d−k+1)/k
```

d=0 で原系列、d=1 で標準1階差分（対数リターン）に一致し、d∈[0,1] の中間では ω_0=1 以降の重みはすべて負かつ >−1 になる。E-mini S&P500 対数価格の ADF 検定では、d=0 で −0.3387（臨界値 −2.8623、単位根を棄却できず）、d=0.4 で −3.2733（95%超で棄却）かつ原系列との相関 ≈0.995（記憶をほぼ保持）となる。一方 d=1 では ADF=−46.9114 だが原系列との相関はわずか 0.05 で、標準リターンが系列を過剰差分していることが分かる。世界の最も流動的な87先物の対数価格はすべて d<0.6、大半は d<0.3 で定常化する。記憶なき系列は予測不能に見えるため、この過剰差分が効率的市場仮説 (EMH) がアカデミアで根強い一因かもしれない (López de Prado (2018))。実務では、原系列との相関を保ちつつ ADF を臨界値未満に落とす最小の d を探すことで、「予測可能性を殺さずに ML 入力を定常化できているか」を点検できる。

分類問題の層：トリプルバリア法・メタラベリング・標本の一意性

第五の罠は固定時間ホライズン・ラベリングで、r<−τ→y=−1、|r|≤τ→y=0、r>τ→y=1 と固定閾値 τ を当てる方式がボラティリティを無視する。例えば τ=1E-2 を、夜間の実現ボラ σ=1E-4 と寄付きの σ=1E-2 に同じく当てると、リターンが予測可能で有意でも大半のラベルが0になる。処方箋はトリプルバリア法で、水平2本（利確・損切り、推定実現/インプライドボラの動的関数）と垂直1本（建玉後の経過バー

数 h、活動ベースの満期）を置く。上バーに先に触れたら +1、下バーなら −1、垂直バーならリターンの符号または0とする。区間 [t_{i,0}, t_{i,0}+h] 全体を見る経路依存 (path-dependent) のラベリングである (López de Prado (2018))。

第六の罠は方向 (side) とサイズ (size) を単一モデルで同時学習することである。方向決定はフェアバリューに関するファンダメンタルな判断、サイズ決定はリスク予算・確信度に関するリスク管理判断であり、結合は不要な複雑性を招く。処方箋はメタラベリングで、一次モデルで方向を決め、二次（メタ）モデルは市場ではなく「一次モデルの予測が正しい確率」を学習して偽陽性をフィルタする。多くのMLは高い precision（=TP/(TP+FP)）・低い recall（=TP/(TP+FN)）で保守的すぎ機会を逃すため、両者の調和平均である F1 スコアの最大化が望ましく、それを side/size 分離で達成する。追加の利点は、(1)ホワイトボックス（経済理論に基づくファンダメンタルモデル）の上にMLを載せられる（quantamental）、(2)MLが方向でなくサイズのみを決めるため過学習の影響が限定される、(3)買い専用/売り専用に別々の一次モデル・特徴量を使える、(4)「小さい賭けで高精度・大きい賭けで低精度は破滅を招く」ためサイズ決定にMLを専念させられる、の4点である (López de Prado (2018))。

第七の罠は、時間的に重なり非IIDなラベル（「こぼれた血液サンプル」問題）を等重み扱いすることである。二つのラベルは共通リターン r_{t-1,t}=p_t/p_{t-1}−1 の関数であるとき時刻 t で並行 (concurrent) する。処方箋は並行度から一意性 (uniqueness) を計算し、標本重み付けと逐次ブートストラップで補正することである (López de Prado (2018))。

```
1_{t,i} ∈ {0,1}（ラベル i が r_{t-1,t} をまたぐか）
c_t = Σ_i 1_{t,i}（並行ラベル数）
u_{t,i} = 1_{t,i}/c_t（時刻 t での一意性）
ū_i = (Σ_t u_{t,i})/(Σ_t 1_{t,i})（平均一意性）
w̃_i = |Σ_{t=t_{i,0}}^{t_{i,1}} r_{t-1,t}/c_t|（標本重み、総和1に正規化）
```

逐次ブートストラップは全標本を同時抽出せず逐次に抽出し、各ステップで一意性の高い観測の抽出確率を上げる。モンテカルロ実験で標本の平均一意性が有意に上がり「こぼれた標本」効果が緩和されることが示され、バギングでは max_samples を平均一意性に設定する (López de Prado (2018))。

研究ツールとしての特徴量重要度：MDI／MDA／SFI と代替効果

特徴量重要度分析には代表的な3手法がある。MDI (Mean Decrease Impurity) は木ベース分類器から直接得るインサンプル指標（sklearn の feature_importances）で高速だが過学習リスクがある。MDA (Mean Decrease Accuracy) は任意の分類器に適用でき、特徴の値をランダムに置換 (permute) して精度低下を測るアウトオブサンプル指標で、MDIより頑健だが低速である。SFI (Single Feature Importance) は各特徴を単独でアウトオブサンプル評価し、代替効果を受けないが特徴間の結合予測力を無視する (mlfinlab feature-importance docs)。代替効果 (substitution effect) とは、相関する2つ以上の説明変数が予測力を共有するとき一方が重要・他方が冗長と判定され重要度がロバストに定まらない現象で、MDIとMDAはともにこれを受ける。健全性チェックとして、重要度とPCA固有値の加重ケンドール τ 相関を比較する (mlfinlab feature-importance docs)。実務では、この3手法を突き合わせることで「予測力が特定の特徴に真に宿るのか、相関する特徴群に分散して見かけ上どれも中程度に見えているだけか」を判別できる。

評価の層：交差検証リーク・CPCV・偽戦略定理／DSR

第八の罠は交差検証 (cross-validation; CV) リークである。k-fold が金融で失敗するのは観測が非IIDだからで、系列相関する特徴 X_t≈X_{t+1} と重なるデータ点由来のラベル Y_t≈Y_{t+1} があると、t と t+1 を別集合に置くだけで情報が漏れ、無関係な特徴でも Y_{t+1}=E[Y_{t+1}] を当ててしまう。処方箋はパージング（テストのラベルと期間が重なる訓練観測の除去）とエンバーゴ（テスト直後の訓練観測をさらに除去、幅 h≈0.01T で通常リークを完全に防げる）である (López de Prado (2018))。

第九の罠はウォークフォワード (WF) バックテストで、(1)歴史的1経路のみの検証で容易に過学習する、(2)将来を代表せずデータ順序に結果が依存する（順序を逆にして結果が一貫しないのが過学習の証拠）、(3)初期の意思決定が小さい部分標本に基づきウォームアップで情報を無駄にする、の3欠点をもつ。処方箋は組合せパージ交差検証 (CPCV) で、T観測をシャッフルせず N群に分割し k群をテストに選ぶ全組合せから多数のバックテスト経路を生成する (López de Prado (2018))。

```
分割数 = C(N,k) = ∏_{i=0}^{k−1}(N−i)/k!
経路数 φ[N,k] = (k/N)·C(N,k) = ∏_{i=1}^{k−1}(N−i)/(k−1)!
例: N=6, k=2 → 15分割・5経路
```

各分割にパージとエンバーゴを適用する。CPCVの決定的利点は、単一の（過学習した）シャープレシオ推定でなくシャープレシオの分布を導けることで、悪い経路の裾まで含めて頑健性を点検できる (López de Prado (2018))。

第十の罠はバックテスト過学習であり、その理論的基礎が偽戦略定理 (False Strategy Theorem) である。真のシャープが0のマルチンゲールでも、多重検定下では最大シャープは正になる。I回の独立試行 x_i〜Z（標準正規）の期待最大値は (López de Prado (2018)) で与えられる。

```
E[max{x_i}] ≈ (1−γ)·Z⁻¹[1−1/I] + γ·Z⁻¹[1−1/(I·e)] ≤ √(2·ln I),  γ ≈ 0.5772156649
PSR[SR*] = Z[ (SR̂−SR*)·√(T−1) / √(1 − γ̂_3·SR̂ + (γ̂_4−1)/4·SR̂²) ]
SR* = √(V[SR̂_n])·( (1−γ)·Z⁻¹[1−1/N] + γ·Z⁻¹[1−1/(N·e)] )
```

処方箋はデフレート・シャープレシオ (Deflated Sharpe Ratio; DSR)、すなわち基準 SR *をユーザ定義でなく試行の多重性（試行数 N と試行間分散 V[SR^n]）を反映して推定した PSR[SR]* である。DSRは Bailey & López de Prado (2014) が開発し、選択バイアス・バックテスト過学習・標本長・分布の非正規性（歪度 $\gamma\hat{}3$・尖度 $\gamma\hat{}4$）を同時に補正する。SR* は試行数 N と試行間分散が大きいほど急速に増えるため、試行数 I を記録しないと FWER・FDR・PBO を算定できない (López de Prado (2018))。目安として、年率シャープ0.95を95%信頼で有意とするには約3年分の日次リターンが必要である (Deflated Sharpe ratio (Wikipedia))。

発展：Machine Learning for Asset Managers（2020）

López de Prado (2020) は上記フレームワークを最適化・共分散推定側へ拡張する。主な貢献は、(1)ランダム行列理論（Marcenko-Pastur）で信号と雑音の固有値を分離して共分散・相関行列をデノイズし、さらに市場成分を除くデトーニング (detoning) を導入、(2)情報理論の距離指標・エントロピー・コディペンデンスで線形相関を超えた非線形効果・外れ値を捉える、(3)ONC (Optimal Number of Clusters) で特徴を

クラスタリングしてから MDI/MDA を適用するクラスタ化特徴量重要度 (CFI) により線形・非線形の代替効果（多重共線性）に頑健化する、(4)多重共線性下でロバスト性を欠く p値・統計的有意性を MDI/MDA で置き換える、(5)Markowitzの呪い（共分散逆行列の不安定性）に階層クラスタリング／NCO (Nested Clustered Optimization) で対処、(6)モンテカルロと組合せて偽戦略を検出し運と真のアルファを区別する、である (CFA Institute review (2021))。予測ではラベル（予測目標）の定式化が重要で、実務家は価格点推定より方向を正確に当てるモデルを重視する (CFA Institute review (2021))。

前提・限界と注意点

第一に、この枠組みは各観測のラベル生成区間（予測ホライズン）と、最終戦略に至るまでの独立試行数 N を正確に把握できることに依存する。ラベル区間を過少に見積もればパージングが不十分になりリークが残り、N を過少報告すれば DSR・偽戦略定理の閾値が容易に甘くなる (López de Prado (2018))。第二に、分数階差分・トリプルバリア・一意性重み付けはいずれも追加のハイパーパラメータ（d、バリア幅、h）を持ち込むため、これら自体の探索を試行数 N に数え入れないと過学習対策が自己矛盾する。第三に、偽戦略定理と DSR は極値理論と漸近近似に立つため、極端に短い標本や強い系列相関のもとでは精度が落ちる (Bailey & López de Prado (2014))。第四に、SFI・CFI が代替効果を回避する一方で結合予測力を過小評価しうるように、各処方箋は別の偏りとのトレードオフを持つ。

実務上のチェックリストとしては、(1)入力を時間バーでなくドル・バーで作り、原系列との相関を保つ最小の d で定常化する、(2)ラベルを固定ホライズンでなくボラ依存のトリプルバリアで付け、方向とサイズをメタラベリングで分離する、(3)非IIDを一意性重み付け・逐次ブートストラップで補正し、CVはパージ＋エンバーゴを課す、(4)研究の主軸を反復バックテストでなく特徴量重要度（MDI/MDA/SFI・CFI）に置く、(5)WFの単一経路でなく CPCV でシャープの分布を作り、試行数 N を記録して DSR で選択バイアスをデフレートする、の5点を層ごとに積み上げる。10の処方箋は代替でなく上流から下流へ相補的に用いるのが正しい運用である (López de Prado (2018))。

統合チェックリスト：予測モデルを本番投入する前に確認すべきこと

本チェックリストは本稿の全セクションで論じた検証手法を、短期クオンツ予測モデルを実運用へ移す前の最終点検項目として統合したものである。各項目は「なぜ重要か／合格基準の目安」を併記し、数値目安はいずれも本文で示した範囲に基づく。カテゴリは上流（データ・リーク）から下流（実運用移行）へ相補的に積み上げる設計であり、上流のゲートを塞がずに下流だけ厳格化しても偽の発見は漏れる点に注意する。

A. データ／リーク

- □ ポイントインタイム（PIT）ユニバースを用い、生存者バイアスを排除したか（現在まで生き残った銘柄だけで検証すると結論が反転する。1986 年末 Russell 3000 の生存は 500 銘柄未満で、生存ユニバースでは最悪信用リスク株が最良品質株の 7 倍・高ボラ株が低ボラ株を 16 倍アウトパフォームして見える。合格：非アクティブ企業を含む PIT ユニバースを使用し、欠損は NA 保持で銘柄除外しない）
- □ 報告ラグを織り込み先読みバイアスを排除したか（米企業の四半期決算提出は平均 30 日・中央値 28 日を要する。報告ラグを無視すると益回りファクターの IC が約 60% 水増しされる。合格：決算・改訂値・株式分割の全てで発表時点以降の情報のみを使用する PIT データベースを用いる）
- □ 外れ値処理・正規化・除外ルールをモデル構築前に固定したか（「5% winsorization では機能するが 1% では失敗し 5% を採用」は不正な研究プロセスの明白な兆候。価格モメンタムは先にランク正規化すると予測力を約 35% 失う。合格：winsorization 水準・正規化手法をデータを見る前に事前決定し記録）
- □ 各ラベルの予測ホライズン（ラベル生成区間）を明示したか（区間を過少に見積もるとパージ／エンバーゴが不十分になりリークが残る。合格：全ラベルの生成区間を明示し、パージ設計の入力とする）
- □ パージング＋エンバーゴでラベル重複と系列相関のリークを塞いだか（標準 k-fold は時間順序無視・ラベル重複・系列相関の 3 機構でリークし成績を過大評価する。エンバーゴ幅は全観測の約 1%（h≈0.01T）で通常リークを防げる。合格：検定ラベル区間と重なる訓練観測をパージし、直後の h≈0.01T をエンバーゴ）
- □ ランダム k-fold ではなく時間順序を守る分割（Walk-Forward / Purged K-Fold / CPCV）を使ったか（k-fold は未来で学習し過去を予測する look-ahead を生む。合格：常に「学習済みの過去→未評価の未来」の順で分割）

B. 過学習

- □ 学習曲線・検証曲線でバイアス／バリアンスを診断したか（訓練・検定とも低いなら過小適合、訓練高・検定低なら過学習。合格：learning_curve でギャップの縮小傾向を見て「データ追加で改善する過学習か力不足の過小適合か」を切り分け）
- □ 正則化でバリアンスを削ったか（低 S/N 環境では正則化なしの OLS は月次 OOS R^2=−3.46% と定数ゼロ予測にすら劣る。合格：dropout・early stopping・L1/L2・ノイズ注入を併用し、正則化後の OOS 指標が改善）

- □ 「訓練誤差ゼロ＝過学習」で棄却せず OOS 指標そのもので判定したか（二重降下・複雑さの美徳により過剰パラメータ化域では訓練データを完全補間しても汎化しうる。合格：採否は訓練誤差でなくパージ済み OOS 指標で判断）
- □ 単一経路でなく CPCV でパフォーマンスの分布を得たか（Walk-Forward は市場史の単一経路しか検証できず容易に過学習する。合格：CPCV で複数の純粋 OOS 経路を生成し、点推定でなく分布（特に悪い経路の裾）で評価）
- □ 最終戦略に至るまでの独立試行数 N を記録したか（DSR・PBO・MinBTL はいずれも N を正しく数えられることに依存し、N の過少報告は指標を容易に甘くする。合格：パラメータ探索・特徴量選択・閾値調整を含む全試行を N として記録）

C. 過剰最適化と多重検定

- □ 試行数 N に対しデータ長が最小バックテスト長（MinBTL）を満たすか（MinBTL < $2 \cdot \ln[N] / E[\max_N]^2$。5 年データなら独立構成を 45 個以下、2 年なら 7 個以下に抑えないと、真の OOS Sharpe=0 の無スキル戦略が IS Sharpe=1 で生成される。合格：検証設計の段階で MinBTL を計算し許容試行数の上限を先に決める）
- □ 期待最大 Sharpe による選択バイアスを認識したか（真の SR=0 でも N=10 試行で期待最大 IS Sharpe は 1.57、20 試行で SR>2.0、N=100 で 2.5 超に膨張する。合格：最良戦略の Sharpe を額面で信用せず E[max_N] と比較）
- □ Deflated Sharpe Ratio（DSR）で選択バイアス・非正規性を割り引いても正のままか（DSR は N・歪度・尖度・系列長・試行間分散の 5 変数で観測 Sharpe をデフレートする。合格：DSR>0.95（95% 信頼水準）。0.95 未満なら棄却）
- □ Probability of Backtest Overfitting（PBO）が十分低いか（PBO は IS 最良戦略が OOS で全戦略の中央値を下回る確率。0.5 は IS の選抜が OOS に無情報、1 に近いほど高度の過剰最適化。合格：PBO≤0.05。併せて OOS 損失確率と CSCV の劣化傾き β（通常は負）を確認）
- □ 多重検定を考慮した t 閾値を越えているか（数百のファクターが試された今、単一検定の t>2.0 は甘すぎる。多重性を考慮した 5% 有意の最小 t は約 2.8（単一検定 p≈0.5%）、新規ファクターには実質 t>3.0 が必要。ランダム戦略群では時系列回帰で 3.8、Fama-MacBeth で 3.4。合格：報告 t が 2.8〜3.8 の水準を越える）
- □ 相関試行を実効独立試行数 $\hat{N}$に換算したか（試行が相関すると名目試行数 M は E[max] を過大評価する。$\hat{N} = \rho\hat{} + (1-\rho\hat{}) \cdot M$。合格：平均ペア相関 $\rho\hat{}$から $\hat{N}$を求め DSR/MinBTL に用いる）
- □ ハイカット Sharpe で試行数を織り込んで読み替えたか（割引は非線形で、年率 Sharpe<0.4 ではハイカットがほぼ常に 50% 超、Sharpe>1.0 では最大でも 25%。一律 50% 割引は不適切。合格：試行数 N を数え、報告 Sharpe をハイカット後の値で解釈）
- □ 大量スクリーニングの最良戦略をブートストラップ検定に通したか（hold-out や k-fold は試行数依存の偽陽性を制御できず、95% 水準を 20 回適用すれば偽陽性 1 件は想定内。合格：Reality Check ではなく劣モデル混入に頑健な SPA の p 値で、全戦略の探索を単一の帰無に押し込んで判定）
- □ モデル間の優劣主張に多重検定補正を課したか（12 モデルの Diebold-Mariano 比較を Bonferroni 補正すると片側臨界値が 2.64 まで上がり、優位が「限界的に有意」へ後退しうる。合格：モデル比較

の t 閾値を比較数で補正）

D. ファクター中立性

- □ 複数のファクターモデルに対するスパニング検定で残差アルファが有意に正か（切片が非有意なら既知因子で張られる冗長な因子、有意なら真のアルファ源。単一モデルでは誤指定を吸い込む。合格：FF5・Carhart（UMD 必須）・q-factor に並行して戦略リターンを回帰し、切片 α が t>3.0 で有意に正）
- □ Barra スタイル因子に中立化した残差にアルファが残るか（産業・スタイルのベット由来の見かけの収益を除去するため。合格：Beta・Value・Momentum・Size 等に中立化した残差リターンでアルファを再測定）
- □ 複数シグナルは GRS 検定で全アルファ同時ゼロを棄却できるか（GRS はシグナル群を加えると達成可能な最大 Sharpe 比が有意に上がるかの検定と等価。合格：F 検定で $\alpha_1=\cdots=\alpha_N=0$ を棄却）
- □ 中立化後の残差リターンで rank IC と期待 IR を測ったか（実務のファクターでは IC は 0.02〜0.05 程度、0.10 超は稀。IR≈IC·√BR で breadth により小さい IC でも有意な IR に変換できる（IC=0.05 なら BR=200 で IR≈0.71）。合格：外れ値に頑健な Spearman rank IC を用い、breadth と併せ期待 IR の上限を事前見積もり）

E. リスクと分布

- □ 取引頻度に合ったスケールで歪度・尖度を測り非正規性を定性判定したか（分布はスケール依存で、高頻度ほど非正規性が際立つ（5 分足 S&P500 先物で尖度 $\kappa\approx15.95$）。ただし 4 次以上の標本モーメントは不安定で定量指標には使えない。合格：モデルの取引頻度で歪度・尖度を測り、非正規性の「存在シグナル」として扱う）
- □ 損失側 PnL 残差に GPD を当て裾指数を推定したか（金融日次リターンの形状パラメータ ξ は 0.2〜0.4 で 0 から有意に離れる（重い裾）。合格：$\hat{\xi}$ が 0 から有意に離れ、裾指数 $1/\hat{\xi}$ が 2〜5 の重い裾域に入るかを確認、閾値 u への感度も点検）
- □ リスク集計にコヒーレントな ES／CVaR を用い VaR に依存していないか（VaR は劣加法性を満たさず、分散投資を罰し口座分割で見かけのリスクを減らせる。劣加法性は正規性という強い仮定に依存し重い裾では成り立たない。合格：リスク集計・資本配分に ES/CVaR を使用）
- □ ES を正規でなく GARCH+EVT で条件付けて算出しバックテストしたか（正規は重い裾のもとで ES を系統的に過小評価する。S&P の 0.99 分位バックテストで条件付き正規は期待 74 回に対し 104 回違反（p=0.00）、GARCH+EVT は 73 回（p=0.91）。合格：0.99 分位の違反回数が期待値と整合）
- □ 観測最大ドローダウンを理論期待値と照合したか（正の期待リターンでも最大 DD は log T で増え続け、有限時間で必ず新記録が起こりうる。$\mu=0$ 基準で $E[D]\approx1.2533\cdot\sigma\sqrt{T}$。合格：観測 MDD を μ・σ・T から計算した理論値と比較し、$\sigma\sqrt{T}$ のスケールを大きく超える DD のみを破綻シグナルとみなす。CDaR も併用）
- □ Sharpe 比に標準誤差・95% 信頼区間・MinTRL を添えたか（IID 下でも $SE(SR)=\sqrt{(1+\frac{1}{2}SR^2)/T}$ と大きく、高 Sharpe ほど推定精度が落ちる。合格：95%CI=SR±1.96·SE が 0 を跨がないこと、主張 Sharpe の確立に必要な MinTRL を満たすことを確認）

- □ 年率化を系列相関で補正したか（√頻度 倍が正しいのは IID のときだけ。正の自己相関は Sharpe を過大評価させ、ヘッジファンドで最大 65% 膨らみランキングまで変わる。合格：Ljung-Box で系列相関を検定し、有意なら √12 を捨て $\eta(q)$ で補正）
- □ 指標が突出して良いとき操作可能性を疑い MPPM で検算したか（ほぼ全指標は稀な巨大損失（裾の売り持ち）を隠す動的・オプション的戦略で膨らませられる。合格：ペイオフ分布の裾を点検し、操作不能指標 MPPM で順位を再確認）
- □ 性質の異なる指標を組で読んだか（勝率単独は誤解を招く（勝率 30% でもペイオフ比が高ければ期待値 +$34/トレードで高収益）。合格：下方リスクは Sortino/Omega、経路リスクは MDD/Calmar/Ulcer/テールレシオ、トレード単位は期待値/Profit Factor/Payoff Ratio を併読）

F. コストとキャパシティ

- □ 現実的なコストモデルでネット Sharpe が生き残るか（グロスで統計的に本物でも、回転率の高い短期戦略はコストがアルファを食い潰す。日本のリバーサルは無コスト年率 12% だが 30bps で負に転落する。合格：現実的コスト控除後の期待リターンが正）
- □ 一時インパクトを凹関数でモデル化し指数感度を確認したか（実測の一時インパクトは強く凹で、平方根則 $\Delta = Y \cdot \sigma \cdot \sqrt{(Q/V)}$。指数は文献で 0.5〜0.6 に割れる。線形で置くと大口で過大・小口で過小。合格：$\delta \approx 0.5$ と 3/5 の複数指数でネット損益を再評価し、恒久成分は無裁定条件から線形で置く）
- □ コスト水準の外部妥当性を実約定で検証したか（TAQ ベースの学術推計は実約定より約 1 桁（線形 TAQ で 6〜12 倍）大きい。実測の平均実装ショートフォールは約 11bps。合格：自社の実約定（パッシブファンドの追跡誤差でも可）でコストモデルを較正・検証）
- □ 執行の半減期 θ をシグナルの予測地平と比較したか（$\theta = 1/\kappa$, $\kappa \approx \sqrt{(\lambda \sigma^2/\eta)}$ はボラ・一時インパクト・リスク回避度で決まる。予測地平が θ より短ければ執行が間に合わずインパクトでアルファが溶ける。合格：想定執行サイズ・保有期間・銘柄ボラから θ を計算し、シグナル地平が θ を上回る）
- □ 想定 AUM に対しキャパシティ（損益分岐規模）を算定したか（実コストが従来推計の 1/10 なら損益分岐規模は約 10 倍。短期反転は米国 $9B までしか生き残らないなど回転率の高い戦略はキャパシティが小さくコストモデル依存性が高い。合格：回転率・シグナル減衰・参加率をインパクトモデルに入れ、ネット期待リターンがゼロを割る AUM を算出）
- □ 公開後のアルファ減衰を織り込んだか（公開シグナルは OOS で 26%、公開後で 58% 減衰する（差の 32% は情報を得た投資家の裁定）。合格：公開済みシグナルはグロス値を額面で信用せず約 6 割の減衰を割り引く）

G. レジーム頑健性

- □ 構造変化検定でモデル係数がいつ壊れたかを点検したか（特徴量とリターンの関係は非定常でレジーム間で不連続に変化する。合格：ブレーク既知なら Chow、位置・個数未知なら Bai-Perron、常時監視なら CUSUM を使い分け、ローリング窓で係数の恒常性を検定）
- □ Sharpe 比がレジーム条件付きであることを踏まえたか（マルコフ・スイッチングでは平均・ボラ・相関がレジーム間で異なり、単一の期待値を当てにするのは危険。エッジは有限寿命を持ち景気循環的に盛衰する。合格：現在どのレジームにいるかを推定し、レジーム別の性能を評価）

- □ 価格系列を分数次差分でメモリを保持したまま定常化したか（整数次差分（リターン化, d=1）はメモリをほぼ消去する。合格：ADF の 95% 臨界値を通過する最小の d* を選び、原系列との相関を 0.9 超（多くの流動先物は d<0.6）に保つ）
- □ ウォークフォワードで高原（plateau）を確認し孤立ピークを排除したか（摂動に安定な近最適の広い高原が頑健性の証拠であり、孤立した鋭いピークはノイズへの過学習の兆候。合格：選んだパラメータの隣接も近最適で、walk-forward efficiency（OOS/IS 成績比）が妥当）
- □ crowding（過密化）の進行を監視しているか（公開後、予測子ポートフォリオでは取引量 +18.7%・空売り残高スプレッド拡大・他シグナルとの相関上昇が起こる（アルファ枯渇の先行指標）。合格：シグナルの取引量・空売り残高・他ファンドとの相関を継続監視）

H. 実運用移行

- □ 実弾投入前にペーパートレード／フォワードテストで検証したか（いかなる過去データ上の検証も真の OOS ではない。無スキル戦略では IS Sharpe と OOS Sharpe は無相関。合格：ライブ前にフォワードテストで減衰を実測）
- □ ライブ運用を唯一の真の OOS として最終検定に据えたか（iterated OOS も hold-out も真の OOS ではなく、「真の OOS はライブ運用のみ」。合格：ライブ性能を最終検定とし、稼働後のライブモデルの弄り回しを禁止）
- □ 稼働後も CUSUM 等でパラメータ恒常性を常時監視しているか（レジームが変わればエッジは崩れる。合格：CUSUM/Bai-Perron でブレークを常時監視し、係数一定の棄却を検知したら再検証）
- □ エッジの有限寿命と減衰を運用計画に織り込んだか（独自短期シグナルにも「有限寿命・景気循環的減衰」の定性的含意は及ぶ。合格：減衰を前提にネット期待値を保守的に置き、性能劣化時の撤退基準を事前設定）

参考文献

- **Agrrawal (2023)** Agrrawal, P. (2023). "The Gibbons, Ross, and Shanken Test for Portfolio Efficiency: A Note Based on Its Trigonometric Properties." Mathematics, 11(9), 2198. (GRS の幾何学的・シャープレシオ解釈) (https://www.mdpi.com/2227-7390/11/9/2198)
- **McNeil & Frey (2000)** Alexander J. McNeil & Rüdiger Frey, "Estimation of tail-related risk measures for heteroscedastic financial time series: an extreme value approach," Journal of Empirical Finance, Vol. 7, No. 3–4 (2000), pp. 271–300. (https://doi.org/10.1016/S0927-5398(00)00012-8)
- **Chekhlov, Uryasev & Zabarankin (2005)** Alexei Chekhlov, Stanislav Uryasev & Michael Zabarankin, "Drawdown Measure in Portfolio Optimization," International Journal of Theoretical and Applied Finance, Vol. 8, No. 1 (2005), pp. 13–58. (https://doi.org/10.1142/S0219024905002767)
- **Almgren & Chriss (2000)** Almgren, R., & Chriss, N. (2000). Optimal Execution of Portfolio Transactions. Journal of Risk, 3(2), 5-40. (https://www.cims.nyu.edu/~almgren/papers/optliq.pdf)
- **Almgren, Thum, Hauptmann & Li (2005)** Almgren, R., Thum, C., Hauptmann, E., & Li, H. (2005). Direct Estimation of Equity Market Impact. Risk, 18(7), 58-62. (https://www.cis.upenn.edu/~mkearns/finread/costestim.pdf)
- **米国統計学会 (ASA) の倫理指針 (2016)** American Statistical Association, Committee on Professional Ethics (2016). Ethical Guidelines for Statistical Practice (Discussion #4). ASA (page reflects current 2022 revision). (https://www.amstat.org/your-career/ethical-guidelines-for-statistical-practice)
- **Lo (2002)** Andrew W. Lo (2002), 'The Statistics of Sharpe Ratios', Financial Analysts Journal 58(4), 36-52 (asymptotic distribution of the Sharpe ratio estimator). (https://www.tandfonline.com/doi/abs/10.2469/faj.v58.n4.2453)
- **Ang & Timmermann (2012)** Ang, A., & Timmermann, A. (2012). Regime Changes and Financial Markets. Annual Review of Financial Economics, 4, 313-337. (https://doi.org/10.1146/annurev-financial-110311-101808)
- **Bai & Perron (1998)** Bai, J., & Perron, P. (1998). Estimating and Testing Linear Models with Multiple Structural Changes. Econometrica, 66(1), 47-78. (https://doi.org/10.2307/2998540)
- **Bailey & López de Prado (2012)** Bailey, D. H., & López de Prado, M. (2012). "The Sharpe Ratio Efficient Frontier." Journal of Risk, 15(2), 3–44. (https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1821643)
- **Bailey & López de Prado (2012)** Bailey, D. H., & López de Prado, M. (2012). "The Sharpe Ratio Efficient Frontier." Journal of Risk, 15(2), 3–44. (Introduces the Probabilistic Sharpe

Ratio and Minimum Track Record Length.)
(https://www.davidhbailey.com/dhbpapers/sharpe-frontier.pdf)

- **Bailey & López de Prado (2014)** Bailey, D. H., & López de Prado, M. (2014). "The Deflated Sharpe Ratio: Correcting for Selection Bias, Backtest Overfitting, and Non-Normality." The Journal of Portfolio Management, 40(5), 94-107. (https://ssrn.com/abstract=2460551)
- **Bailey & López de Prado (2014)** Bailey, D. H., & López de Prado, M. (2014). "The Deflated Sharpe Ratio: Correcting for Selection Bias, Backtest Overfitting, and Non-Normality." The Journal of Portfolio Management, 40(5), 94–107. (https://www.davidhbailey.com/dhbpapers/deflated-sharpe.pdf)
- **Bailey & López de Prado (2014)** Bailey, D. H., & López de Prado, M. (2014). The Deflated Sharpe Ratio: Correcting for Selection Bias, Backtest Overfitting, and Non-Normality. Journal of Portfolio Management, 40(5), 94-107. (https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2460551)
- **(Bailey & López de Prado (2014))** Bailey, D. H., & López de Prado, M. (2014). The Deflated Sharpe Ratio: Correcting for Selection Bias, Backtest Overfitting, and Non-Normality. The Journal of Portfolio Management, 40(5), 94-107. (https://doi.org/10.3905/jpm.2014.40.5.094)
- **Bailey, Borwein, López de Prado & Zhu (2014)** Bailey, D. H., Borwein, J. M., López de Prado, M., & Zhu, Q. J. (2014). Pseudo-Mathematics and Financial Charlatanism: The Effects of Backtest Overfitting on Out-of-Sample Performance. Notices of the American Mathematical Society, 61(5), 458-471. (https://www.ams.org/notices/201405/rnoti-p458.pdf)
- **Bailey, Borwein, López de Prado & Zhu (2014)** Bailey, D. H., Borwein, J. M., López de Prado, M., & Zhu, Q. J. (2014). Pseudo-Mathematics and Financial Charlatanism: The Effects of Backtest Overfitting on Out-of-Sample Performance. Notices of the American Mathematical Society, 61(5), 458-471. (https://www.davidhbailey.com/dhbpapers/backtest-pseudo.pdf)
- **Bailey, Borwein, López de Prado & Zhu (2017)** Bailey, D. H., Borwein, J. M., López de Prado, M., & Zhu, Q. J. (2017). "The Probability of Backtest Overfitting." Journal of Computational Finance, 20(4), 39–69. (https://www.davidhbailey.com/dhbpapers/backtest-prob.pdf)
- **Bao (2009)** Bao, Y. (2009). Estimation Risk-Adjusted Sharpe Ratio and Fund Performance Ranking under a General Return Distribution. Journal of Financial Econometrics, 7(2), 152-173. (https://academic.oup.com/jfec/article-abstract/7/2/152/856339)
- **(Belkin et al. (2019))** Belkin, M., Hsu, D., Ma, S., & Mandal, S. (2019). Reconciling modern machine-learning practice and the classical bias-variance trade-off. PNAS, 116(32), 15849-15854. (https://doi.org/10.1073/pnas.1903070116)

- **Benjamini & Hochberg (1995)** Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society, Series B, 57(1), 289-300. (https://doi.org/10.1111/j.2517-6161.1995.tb02031.x)
- **Benjamini & Yekutieli (2001)** Benjamini, Y., & Yekutieli, D. (2001). The Control of the False Discovery Rate in Multiple Testing under Dependency. The Annals of Statistics, 29(4), 1165-1188. (https://doi.org/10.1214/aos/1013699998)
- **(Bishop (1995))** Bishop, C. M. (1995). Training with Noise is Equivalent to Tikhonov Regularization. Neural Computation, 7(1), 108-116. (For sum-of-squares error the regularizer reduces to a positive semi-definite form in the first derivatives of the network mapping.) (https://doi.org/10.1162/neco.1995.7.1.108)
- **Bouchaud (2010)** Bouchaud, J.-P. (2010). Price Impact. In R. Cont (Ed.), Encyclopedia of Quantitative Finance. Wiley. (arXiv:0903.2428) (https://arxiv.org/abs/0903.2428)
- **Bouchaud (2024)** Bouchaud, J.-P. (2024). The Square-Root Law of Market Impact. (https://bouchaud.substack.com/p/the-square-root-law-of-market-impact)
- **Brock-Lakonishok-LeBaron (1992)** Brock, W., Lakonishok, J., & LeBaron, B. (1992). Simple Technical Trading Rules and the Stochastic Properties of Stock Returns. The Journal of Finance, 47(5), 1731-1764. (https://doi.org/10.1111/j.1540-6261.1992.tb04681.x)
- **Brown, Durbin & Evans (1975)** Brown, R. L., Durbin, J., & Evans, J. M. (1975). Techniques for Testing the Constancy of Regression Relationships over Time. Journal of the Royal Statistical Society: Series B, 37(2), 149-192. (https://doi.org/10.1111/j.2517-6161.1975.tb01532.x)
- **(Harvey, Liu & Zhu (2016))** Campbell R. Harvey, Yan Liu & Heqing Zhu, "… and the Cross-Section of Expected Returns," The Review of Financial Studies 29(1), 2016, 5–68. Verified: 313 articles total (250 published + 63 working papers) / 316 factors (p.8); recommended t>3.0 and "most claimed findings likely false" (abstract); two-sided p-value convention (t≈2.8→0.5%, p.26); Table-4 M=10 worked example — single tests 10/10, Bonferroni 3, Holm 4, BHY 6, cutoffs Bonferroni 0.5% / Holm 0.60% / BHY 0.85% (p.21); Holm uniformly dominates Bonferroni; all three admit general dependence. (https://people.duke.edu/~charvey/Research/Published_Papers/P118_and_the_cross.PDF)
- **Carhart (1997)** Carhart, M. M. (1997). "On Persistence in Mutual Fund Performance." The Journal of Finance, 52(1), 57-82. (https://doi.org/10.1111/j.1540-6261.1997.tb03808.x)
- **Acerbi & Tasche (2002)** Carlo Acerbi & Dirk Tasche, "Expected Shortfall: a natural coherent alternative to Value at Risk," Economic Notes, Vol. 31, No. 2 (2002), pp. 379–388 (arXiv preprint cond-mat/0105191, 2001). (https://arxiv.org/abs/cond-mat/0105191)
- **arXiv:2407.17645 (2024)** Carlo Nicolini, Monisha Gopalan, Jacopo Staiano, Bruno Lepri (2024), 'Hopfield Networks for Asset Allocation', arXiv:2407.17645 (CPCV: 21-day

purge/embargo, 45 training combinations, 36 backtest paths, ~3.5yr train / ~2yr test). (https://arxiv.org/abs/2407.17645)

- **CFA Institute review (2021)** CFA Institute, Enterprising Investor (2021). "Book Review: Machine Learning for Asset Managers." (https://rpc.cfainstitute.org/blogs/enterprising-investor/2021/book-review-machine-learning-for-asset-managers)
- **Chordia, Goyal & Saretto (2020)** Chordia, T., Goyal, A., & Saretto, A. (2020). Anomalies and False Rejections. Review of Financial Studies, 33(5), 2134-2179. (https://academic.oup.com/rfs/article-abstract/33/5/2134/5739455)
- **Chow (1960)** Chow, G. C. (1960). Tests of Equality Between Sets of Coefficients in Two Linear Regressions. Econometrica, 28(3), 591-605. (https://doi.org/10.2307/1910133)
- **Clarke, de Silva & Thorley (2002)** Clarke, R., de Silva, H., & Thorley, S. (2002). "Portfolio Constraints and the Fundamental Law of Active Management." Financial Analysts Journal, 58(5), 48-66. (移転係数TCによる制約下の一般化 IR=TC·IC·√BR) (https://papers.ssrn.com/sol3/papers.cfm?abstract_id=290322)
- **Bailey, Borwein, López de Prado & Zhu (2015)** David H. Bailey, Jonathan M. Borwein, Marcos López de Prado, Qiji Jim Zhu, 'The Probability of Backtest Overfitting', Journal of Computational Finance (SSRN working paper 2326253; CSCV and PBO). (https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2326253)
- **Langer et al. (1978)** Ellen J. Langer, Arthur Blank & Benzion Chanowitz, "The Mindlessness of Ostensibly Thoughtful Action: The Role of 'Placebic' Information in Interpersonal Interaction," Journal of Personality and Social Psychology 36(6), 1978, 635–642. The photocopier / 'because' compliance study; year and attribution correct. (https://doi.org/10.1037/0022-3514.36.6.635)
- **Fama & French (1993)** Fama, E. F., & French, K. R. (1993). "Common Risk Factors in the Returns on Stocks and Bonds." Journal of Financial Economics, 33(1), 3-56. (https://doi.org/10.1016/0304-405X(93)90023-5)
- **Fama & French (2015)** Fama, E. F., & French, K. R. (2015). "A Five-Factor Asset Pricing Model." Journal of Financial Economics, 116(1), 1-22. (RMW/CMA追加、標本1963:07-2013:12、クロスセクション分散の71-94%説明、HMLの冗長化、GRS棄却) (https://doi.org/10.1016/j.jfineco.2014.10.010)
- **Fama-MacBeth (1973)** Fama, E. F., & MacBeth, J. D. (1973). Risk, Return, and Equilibrium: Empirical Tests. Journal of Political Economy, 81(3), 607-636. (https://www.journals.uchicago.edu/doi/10.1086/260061)
- **Fama & MacBeth (1973)** Fama, E. F., & MacBeth, J. D. (1973). Risk, Return, and Equilibrium: Empirical Tests. Journal of Political Economy, 81(3), 607-636. (https://doi.org/10.1086/260061)
- **Frazzini, Israel & Moskowitz (2012)** Frazzini, A., Israel, R., & Moskowitz, T. J. (2012). Trading Costs of Asset Pricing Anomalies. Working paper (SSRN 2294498).

(https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2294498)

- **Frazzini, Israel & Moskowitz (2018)** Frazzini, A., Israel, R., & Moskowitz, T. J. (2018). Trading Costs. Working paper (SSRN 3229719). (https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3229719)
- **Kenneth R. French Data Library** French, K. R. "Data Library." Tuck School of Business, Dartmouth College. (FF因子ポートフォリオ構築の一次資料) (https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html)
- **FXStreet Learning Center / Van Tharp** FXStreet Learning Center. "Stability Statistics" (Expectancy = Average Profit × Win Rate − Average Loss × Loss Rate; concept after Van K. Tharp). (https://learningcenter.fxstreet.com/education/learning-center/unit-3/chapter-2/stability-statistics/index.html)
- **(Geman, Bienenstock & Doursat (1992))** Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural Networks and the Bias/Variance Dilemma. Neural Computation, 4(1), 1-58. (https://doi.org/10.1162/neco.1992.4.1.1)
- **Gibbons, Ross & Shanken (1989)** Gibbons, M. R., Ross, S. A., & Shanken, J. (1989). "A Test of the Efficiency of a Given Portfolio." Econometrica, 57(5), 1121-1152. (https://doi.org/10.2307/1913625)
- **Goetzmann, Ingersoll, Spiegel & Welch (2007)** Goetzmann, W. N., Ingersoll, J. E., Spiegel, M. I., & Welch, I. (2007). "Portfolio Performance Manipulation and Manipulation-Proof Performance Measures." Review of Financial Studies, 20(5), 1503–1546. (https://ivo-welch.org/research/journalcopy/2007-rfs.pdf)
- **Goodwin (1998)** Goodwin, T. H. (1998). "The Information Ratio." Financial Analysts Journal, 54(4), 34–43. (https://tsgperformance.com/wp-content/uploads/2020/11/Goodwin-information-ratio.pdf)
- **Grinold & Kahn (1995)** Grinold, R. C., & Kahn, R. N. (1995). Active Portfolio Management: A Quantitative Approach for Producing Superior Returns and Controlling Risk. McGraw-Hill. (https://books.google.com/books/about/Active_Portfolio_Management_A_Quantitati.html?id=a1yB8LTQnOEC)
- **Grinold & Kahn (2000)** Grinold, R. C., & Kahn, R. N. (2000). Active Portfolio Management (2nd ed.). McGraw-Hill. (基本法則 IR=IC√BR、ICの典型レンジ) (https://www.google.com/search?tbm=bks&q=Grinold+Kahn+%22Active+Portfolio+Management%22)
- **(Gu, Kelly & Xiu (2020))** Gu, S., Kelly, B., & Xiu, D. (2020). Empirical Asset Pricing via Machine Learning. The Review of Financial Studies, 33(5), 2223-2273. (Sample 1957-2016, ~30,000 stocks, 94 characteristics; full 920-feature OLS OOS R2 = -3.46%; NN3 peak 0.40%; Diebold-Mariano 12-way Bonferroni critical value 2.64.) (https://doi.org/10.1093/rfs/hhaa009)

- **Arian, Norouzi & Seco (2024)** Hamid R. Arian, Daniel Norouzi M., Luis A. Seco (2024), 'Backtest Overfitting in the Machine Learning Era: A Comparison of Out-of-Sample Testing Methods in a Synthetic Controlled Environment', Knowledge-Based Systems (SSRN 4686376). (https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4686376)
- **Hansen (2005)** Hansen, P. R. (2005). A Test for Superior Predictive Ability. Journal of Business & Economic Statistics, 23(4), 365-380. (https://doi.org/10.1198/073500105000000063)
- **(Hansen (2005))** Hansen, P. R. (2005). A Test for Superior Predictive Ability. Journal of Business & Economic Statistics, 23(4), 365-380. (https://cdr.lib.unc.edu/downloads/zp38wf793)
- **Harvey (2017)** Harvey, C. R. (2017). Presidential Address: The Scientific Outlook in Financial Economics. Journal of Finance, 72(4), 1399-1440. (https://onlinelibrary.wiley.com/doi/abs/10.1111/jofi.12530)
- **Harvey & Liu (2015)** Harvey, C. R., & Liu, Y. (2015). Backtesting. The Journal of Portfolio Management, 42(1), 13-28. (https://people.duke.edu/~charvey/Research/Published_Papers/P120_Backtesting.PDF)
- **Harvey, Liu & Zhu (2016)** Harvey, C. R., Liu, Y., & Zhu, H. (2016). …and the Cross-Section of Expected Returns. Review of Financial Studies, 29(1), 5-68. (https://academic.oup.com/rfs/article/29/1/5/1843824)
- **Harvey, Liu & Zhu (2016)** Harvey, C. R., Liu, Y., & Zhu, H. (2016). ...and the Cross-Section of Expected Returns. The Review of Financial Studies, 29(1), 5-68. (https://doi.org/10.1093/rfs/hhv059)
- **Harvey, Liu & Zhu (2016)** Harvey, C. R., Liu, Y., & Zhu, H. (2016). "…and the Cross-Section of Expected Returns." The Review of Financial Studies, 29(1), 5-68. (313論文・316ファクター、多重検定でt>3.0閾値、Bonferroni/Holm/BHY) (https://www.nber.org/papers/w20592)
- **Holm (1979)** Holm, S. (1979). A Simple Sequentially Rejective Multiple Test Procedure. Scandinavian Journal of Statistics, 6(2), 65-70. (https://www.jstor.org/stable/4615733)
- **Hou, Xue & Zhang (2015)** Hou, K., Xue, C., & Zhang, L. (2015). "Digesting Anomalies: An Investment Approach." The Review of Financial Studies, 28(3), 650-705. (標本1972/1-2012/12: size 0.31%[t=2.12]、投資I/A 0.45%[t=4.95]、ROE 0.58%[t=4.81]、約80アノマリー) (https://global-q.org/uploads/1/2/2/6/122679606/houxuezhang2015rfs.pdf)
- **Hou, Xue & Zhang (2020)** Hou, K., Xue, C., & Zhang, L. (2020). Replicating Anomalies. Review of Financial Studies, 33(5), 2019-2133. (https://academic.oup.com/rfs/article-abstract/33/5/2019/5236964)
- **(Hsu & Kuan (2005))** Hsu, P.-H., & Kuan, C.-M. (2005). Reexamining the Profitability of Technical Analysis with Data Snooping Checks. Journal of Financial Econometrics, 3(4), 606-628. (https://homepage.ntu.edu.tw/~ckuan/pdf/snoop01.pdf)

- **Hsu, Hsu & Kuan (Step-SPA)** Hsu, P.-H., Hsu, Y.-C., & Kuan, C.-M. (2010). Testing the predictive ability of technical analysis using a new stepwise test without data snooping bias. Journal of Empirical Finance, 17(3), 471-484. (https://doi.org/10.1016/j.jempfin.2010.01.001)
- **Huberman & Stanzl (2004)** Huberman, G., & Stanzl, W. (2004). Price Manipulation and Quasi-Arbitrage. Econometrica, 72(4), 1247-1275. (https://doi.org/10.1111/j.1468-0262.2004.00531.x)
- **mlfinlab feature-importance docs** Hudson & Thames, MLFinLab - Feature Importance module (MDI, MDA, SFI), implementing López de Prado, Advances in Financial Machine Learning, Ch. 8. (Former mlfinlab.com documentation domain is defunct; source available on GitHub.) (https://github.com/hudson-and-thames/mlfinlab/blob/master/mlfinlab/feature_importance/importance.py)
- **mlfinlab** Hudson & Thames, mlfinlab — Purged and Embargo cross-validation (documentation; PurgedKFold pct_embargo parameter). (https://www.mlfinlab.com/en/latest/cross_validation/purged_embargo.html)
- **Ioannidis (2005)** Ioannidis, J. P. A. (2005). Why Most Published Research Findings Are False. PLoS Medicine, 2(8), e124. (https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.0020124)
- **Jensen & Bennington (1970)** Jensen, M. C., & Bennington, G. A. (1970). Random Walks and Technical Theories: Some Additional Evidence. The Journal of Finance, 25(2), 469-482. (https://doi.org/10.1111/j.1540-6261.1970.tb00671.x)
- **Keating & Shadwick (2002)** Keating, C., & Shadwick, W. F. (2002). "A Universal Performance Measure." Journal of Performance Measurement, 6(3), 59–84. (https://people.duke.edu/~charvey/Teaching/BA453_2004/Keating_A_universal_performance.pdf)
- **(Kelly, Malamud & Zhou (2024))** Kelly, B., Malamud, S., & Zhou, K. (2024). The Virtue of Complexity in Return Prediction. The Journal of Finance, 79(1), 459-503. (In the high-complexity regime P>T, expected OOS accuracy and portfolio performance are strictly increasing in complexity; holds already for ridgeless/minimum-L2-norm regression, with optimal shrinkage improving it further.) (https://doi.org/10.1111/jofi.13298)
- **Kidd / CFA Institute (2012)** Kidd, D. (2012). "The Sortino Ratio: Is Downside Risk the Only Risk That Matters?" Investment Performance Measurement feature, CFA Institute. (https://rpc.cfainstitute.org/sites/default/files/-/media/documents/code/gips/the-sortino-ratio.pdf)
- **Korajczyk & Sadka (2004)** Korajczyk, R. A., & Sadka, R. (2004). Are Momentum Profits Robust to Trading Costs? Journal of Finance, 59(3), 1039-1082. (https://doi.org/10.1111/j.1540-6261.2004.00656.x)

- **Kyle (1985) (本文では「Kyle のλ」「(Kyle / Huberman-Stanzl)」として言及)** Kyle, A. S. (1985). Continuous Auctions and Insider Trading. Econometrica, 53(6), 1315-1335. (https://doi.org/10.2307/1913210)
- **Lahiri (1999)** Lahiri, S. N. (1999). Theoretical Comparisons of Block Bootstrap Methods. Annals of Statistics, 27(1), 386-404. (https://projecteuclid.org/journals/annals-of-statistics/volume-27/issue-1/Theoretical-comparisons-of-block-bootstrap-methods/10.1214/aos/1018031117.full)
- **(Ledoit & Wolf (2008))** Ledoit, O., & Wolf, M. (2008). Robust performance hypothesis testing with the Sharpe ratio. Journal of Empirical Finance, 15(5), 850-859. (http://www.ledoit.net/jef_2008pdf.pdf)
- **Lo (2002)** Lo, A. W. (2002). "The Statistics of Sharpe Ratios." Financial Analysts Journal, 58(4), 36–52. (https://doi.org/10.2469/faj.v58.n4.2453)
- **Lo (2002)** Lo, A. W. (2002). "The Statistics of Sharpe Ratios." Financial Analysts Journal, 58(4), 36–52. (https://traders.studentorg.berkeley.edu/papers/The-Statistics-of-Sharpe-Ratios.pdf)
- **Lo (2004)** Lo, A. W. (2004). The Adaptive Markets Hypothesis: Market Efficiency from an Evolutionary Perspective. Journal of Portfolio Management, 30(5), 15-29. (https://doi.org/10.3905/jpm.2004.442611)
- **López de Prado (2018)** López de Prado, M. (2018). "The 10 Reasons Most Machine Learning Funds Fail." The Journal of Portfolio Management, 44(6), 120-133. (https://www.garp.org/hubfs/Whitepapers/a1Z1W0000054x6lUAA.pdf)
- **(López de Prado (2018))** López de Prado, M. (2018). Advances in Financial Machine Learning. Wiley. (Ch. 7: Cross-Validation in Finance - purging, embargo, and Combinatorial Purged Cross-Validation.) (https://www.wiley.com/en-us/Advances+in+Financial+Machine+Learning-p-9781119482086)
- **López de Prado (2020)** López de Prado, M. (2020). Machine Learning for Asset Managers. Cambridge Elements in Quantitative Finance. Cambridge University Press. (https://www.cambridge.org/core/elements/abs/machine-learning-for-asset-managers/6D9211305EA2E425D33A9F38D0AE3545)
- **Magdon-Ismail, Atiya, Pratap & Abu-Mostafa (2004)** Malik Magdon-Ismail, Amir F. Atiya, Amrit Pratap & Yaser S. Abu-Mostafa, "On the maximum drawdown of a Brownian motion," Journal of Applied Probability, Vol. 41, No. 1 (2004), pp. 147–161. (https://doi.org/10.1239/jap/1077134674)
- **Martin & McCann (1989)** Martin, P. G., & McCann, B. B. (1989). The Investor's Guide to Fidelity Funds. John Wiley & Sons. (Ulcer Index / Ulcer Performance Index; documented by creator Peter Martin.) (http://www.tangotools.com/ui/ui.htm)
- **McLean & Pontiff (2016)** McLean, R. D., & Pontiff, J. (2016). Does Academic Research Destroy Stock Return Predictability? Journal of Finance, 71(1), 5-32.

(https://onlinelibrary.wiley.com/doi/abs/10.1111/jofi.12365)

- **McLean & Pontiff (2016)** McLean, R. D., & Pontiff, J. (2016). Does Academic Research Destroy Stock Return Predictability? Journal of Finance, 71(1), 5-32. (https://doi.org/10.1111/jofi.12365)
- **MSCI Barra USE4 Methodology (Menchero, Orr & Wang, 2011)** Menchero, J., Orr, D. J., & Wang, J. (2011). "The Barra US Equity Model (USE4) Methodology Notes." MSCI. (USE4スタイル因子、Eigenfactor Risk Adjustment、Volatility Regime Adjustment、最適化バイアス調整) (https://www.top1000funds.com/wp-content/uploads/2011/09/USE4_Methodology_Notes_August_2011.pdf)
- **Mertens (2002)** Mertens, E. (2002). "Comments on the Correct Variance of Estimated Sharpe Ratios in Lo (2002, FAJ) When Returns Are IID." Working note. (http://www.elmarmertens.com/research/discussion/soprano01.pdf)
- **(Mohri et al. (2018))** Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2018). Foundations of Machine Learning, 2nd ed. MIT Press. (Rademacher complexity, VC dimension, Massart's lemma, generalization bounds: Ch. 3.) (https://cs.nyu.edu/~mohri/mlbook/)
- **(Nakkiran et al. (2019))** Nakkiran, P., Kaplun, G., Bansal, Y., Yang, T., Barak, B., & Sutskever, I. (2019). Deep Double Descent: Where Bigger Models and More Data Hurt. arXiv:1912.02292 (ICLR 2020). (Effective Model Complexity; sample-wise double descent where even quadrupling training samples can hurt test performance.) (https://arxiv.org/abs/1912.02292)
- **Opdyke (2007)** Opdyke, J. D. (2007). "Comparing Sharpe Ratios: So Where Are the p-Values?" Journal of Asset Management, 8(5), 308–336. (https://link.springer.com/article/10.1057/palgrave.jam.2250084)
- **Pardo (2008)** Pardo, R. (2008). The Evaluation and Optimization of Trading Strategies (2nd ed.). Wiley. (https://www.wiley.com/en-us/The+Evaluation+and+Optimization+of+Trading+Strategies%2C+2nd+Edition-p-9780470128015)
- **(Patton, Politis & White (2009))** Patton, A., Politis, D. N., & White, H. (2009). Correction to 'Automatic Block-Length Selection for the Dependent Bootstrap' by D. Politis and H. White. Econometric Reviews, 28(4), 372-375. (https://public.econ.duke.edu/~ap172/Patton_Politis_White_2009.pdf)
- **Artzner, Delbaen, Eber & Heath (1999)** Philippe Artzner, Freddy Delbaen, Jean-Marc Eber & David Heath, "Coherent Measures of Risk," Mathematical Finance, Vol. 9, No. 3 (1999), pp. 203–228. (https://doi.org/10.1111/1467-9965.00068)
- **Phipson & Smyth (2010)** Phipson, B., & Smyth, G. K. (2010). Permutation P-values Should Never Be Zero: Calculating Exact P-values When Permutations Are Randomly Drawn. Statistical Applications in Genetics and Molecular Biology, 9(1), Article 39. (https://gksmyth.github.io/pubs/PermPValuesPreprint.pdf)

- **Politis & Romano (1994)** Politis, D. N., & Romano, J. P. (1994). The Stationary Bootstrap. Journal of the American Statistical Association, 89(428), 1303-1313. (https://doi.org/10.1080/01621459.1994.10476870)
- **(Politis & Romano (1994))** Politis, D. N., & Romano, J. P. (1994). The Stationary Bootstrap. Journal of the American Statistical Association, 89(428), 1303-1313. (https://www.tandfonline.com/doi/abs/10.1080/01621459.1994.10476870)
- **(Politis & White (2004))** Politis, D. N., & White, H. (2004). Automatic Block-Length Selection for the Dependent Bootstrap. Econometric Reviews, 23(1), 53-70. (https://ideas.repec.org/a/taf/emetrv/v23y2004i1p53-70.html)
- **(Susan Potter (2024))** Potter, S. (2024). Monte Carlo Permutation Tests for Strategy Significance. susanpotter.net. (https://www.susanpotter.net/quant/monte-carlo-permutation-tests-strategy-significance/)
- **PyQuant News** PyQuant News. "Tail Ratio: A Key to Extreme Returns." (Tail ratio = 95th percentile / |5th percentile|.) (https://www.pyquantnews.com/free-python-resources/tail-ratio-key-to-extreme-returns)
- **Rockafellar & Uryasev (2002)** R. Tyrrell Rockafellar & Stanislav Uryasev, "Conditional value-at-risk for general loss distributions," Journal of Banking & Finance, Vol. 26, No. 7 (2002), pp. 1443–1471. (https://doi.org/10.1016/S0378-4266(02)00271-6)
- **Cont (2001)** Rama Cont, "Empirical properties of asset returns: stylized facts and statistical issues," Quantitative Finance, Vol. 1, No. 2 (2001), pp. 223–236. (https://doi.org/10.1080/713665670)
- **(Riondato / Two Sigma (2018))** Riondato, M. (2018). Sharpe Ratio: Estimation, Confidence Intervals, and Hypothesis Testing. Two Sigma Technical Report Series No. 2018-001. (https://www.twosigma.com/wp-content/uploads/sharpe-tr-1.pdf)
- **RiskLab AI** RiskLab AI, 'Backtesting through Cross-Validation (CPCV)' (path-count formula, variance formula, k=2 sweet spot). (https://www.risklab.ai/research/backtesting/backtesting_cross_validation)
- **(Arnott, Harvey & Markowitz (2019))** Robert D. Arnott, Campbell R. Harvey & Harry Markowitz, "A Backtesting Protocol in the Era of Machine Learning," The Journal of Financial Data Science 1(1), Winter 2019, 64–74. Verified: alpha-bet ticker strategy (NYSE 1963–1988, OOS to 2015, ~50yr, nearly 6% alpha); ~55yr / <700 monthly / just-over-200 quarterly observations; winsorization warning quote; 20 random strategies→one exceeds t=2.0; winner's curse three outcomes; Heisenberg/overcrowding; iterated-OOS-is-not-OOS; the seven protocol categories; "the only true out of sample is the live trading experience." (https://people.duke.edu/~charvey/Research/Published_Papers/P138_A_backtesting_protocol.pdf)
- **Romano & Wolf (2005)** Romano, J. P., & Wolf, M. (2005). Stepwise Multiple Testing as Formalized Data Snooping. Econometrica, 73(4), 1237-1282.

(https://doi.org/10.1111/j.1468-0262.2005.00615.x)

- **(scikit-learn (2024))** scikit-learn developers. Validation curves / Learning curve (User Guide, section 3.4). Diagnostic rules for bias vs. variance from train/validation score patterns. (https://scikit-learn.org/stable/modules/learning_curve.html)
- **scikit-learn docs** scikit-learn, sklearn.model_selection.TimeSeriesSplit (documentation). (https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.TimeSeriesSplit.html)
- **skfolio docs** skfolio, skfolio.model_selection.CombinatorialPurgedCV (documentation; definitions of purging and embargo). (https://skfolio.org/generated/skfolio.model_selection.CombinatorialPurgedCV.html)
- **(Srivastava et al. (2014))** Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. Journal of Machine Learning Research, 15, 1929-1958. (https://jmlr.org/papers/v15/srivastava14a.html)
- **Sullivan, Timmermann & White (1999)** Sullivan, R., Timmermann, A., & White, H. (1999). Data-Snooping, Technical Trading Rule Performance, and the Bootstrap. The Journal of Finance, 54(5), 1647-1691. (https://onlinelibrary.wiley.com/doi/10.1111/0022-1082.00163)
- **Chordia, Goyal & Saretto (2017)** Tarun Chordia, Amit Goyal & Alessio Saretto, "p-Hacking: Evidence from Two Million Trading Strategies," working paper 2017 (publ. as 'Anomalies and False Rejections,' The Review of Financial Studies 33(5), 2020). Cited in AHM (2019): 2.1 million strategies tested, 17 survive the statistical and economic thresholds — verified against AHM (2019), p.67. (https://www.researchgate.net/publication/319578800_p-Hacking_Evidence_from_Two_Million_Trading_Strategies)
- **Tóth et al. (2011)** Tóth, B., Lempérière, Y., Deremble, C., de Lataillade, J., Kockelkoren, J., & Bouchaud, J.-P. (2011). Anomalous Price Impact and the Critical Nature of Liquidity in Financial Markets. Physical Review X, 1(2), 021006. (arXiv:1105.1694) (https://arxiv.org/abs/1105.1694)
- **White (2000)** White, H. (2000). A Reality Check for Data Snooping. Econometrica, 68(5), 1097-1126. (https://doi.org/10.1111/1468-0262.00152)
- **(White (2000))** White, H. (2000). A Reality Check for Data Snooping. Econometrica, 68(5), 1097-1126. (https://users.ssc.wisc.edu/~bhansen/718/White2000.pdf)
- **(Wicklin / SAS (2021))** Wicklin, R. (2021). The stationary block bootstrap in SAS. The DO Loop (SAS Blogs), 2021-01-20. (https://blogs.sas.com/content/iml/2021/01/20/stationary-bootstrap-sas.html)
- **Deflated Sharpe ratio (Wikipedia)** Wikipedia contributors. "Deflated Sharpe ratio." Wikipedia, The Free Encyclopedia. (https://en.wikipedia.org/wiki/Deflated_Sharpe_ratio)

- **Wikipedia (Walk-forward optimization)** Wikipedia, 'Walk forward optimization' (attributes walk-forward analysis to Robert E. Pardo, 'Design, Testing and Optimization of Trading Systems', 1992; expanded 2nd ed. 'The Evaluation and Optimization of Trading Strategies', 2008). (https://en.wikipedia.org/wiki/Walk_forward_optimization)
- **Wikipedia (Drawdown)** Wikipedia. "Drawdown (economics)." (Maximum Drawdown definition.) (https://en.wikipedia.org/wiki/Drawdown_(economics))
- **Wikipedia, Fama–French three-factor model** Wikipedia. "Fama–French three-factor model." (CAPM が平均約70%、3ファクターは標本内で90%超という説明力の記述、およびSMB/HML定義・NYSE中央値・30/40/30ブレークポイント・負のBE除外の出所) (https://en.wikipedia.org/wiki/Fama%E2%80%93French_three-factor_model)
- **(Luo et al. / Deutsche Bank (2014))** Yin Luo et al., "Seven Sins of Quantitative Investing," Deutsche Bank Markets Research — Global Quantitative Strategy, 8 September 2014. Proprietary investment-bank note, not peer-reviewed (the section discloses this in the limitations paragraph); its internal numeric claims (survivorship, look-ahead IC uplift, split-adjusted momentum, Japan reversal, DCBS shorting, etc.) cannot be checked against a peer-reviewed source. (https://hudsonthames.org/wp-content/uploads/2022/01/DB-201409-Seven_Sins_of_Quantitative_Investing.pdf)
- **Young (1991)** Young, T. W. (1991). "Calmar Ratio: A Smoother Tool." Futures, 20(12), October 1991. (Attribution and 36-month definition verified via Wikipedia "Calmar ratio.") (https://en.wikipedia.org/wiki/Calmar_ratio)