Backtesting Reality Checklist

Signal Pilot Education Hub

Backtesting Validation Checklist

From Lesson 24: Backtesting Reality

Use this checklist to validate backtests and detect overfitting before going live.

📋 Pre-Backtest: Strategy Setup

Step 1: Define Strategy Rules

- [ ] **Entry criteria (explicit, not vague):**
-
-
-
- [ ] **Exit criteria (stop + target):**

- Stop loss: _____
- Target(s): _____
- [ ] **Position sizing:**
- Risk per trade: ____% (1-2% recommended)
- ATR-based / Fixed $ / Other: ____

Step 2: Data Split

- [ ] **Total data period:** _ to _
- [ ] **In-sample (60-70%):** _ to _ (develop & optimize here)
- [ ] **Out-of-sample (30-40%):** _ to _ (validate here)
- [ ] Data split completed BEFORE any optimization? ✓

✅ Backtest Execution Checklist

Realistic Cost Modeling

- [ ] **Spread cost modeled:**
- Normal spread: $__ or __ basis points
- Cost per trade (round trip): $__
- [ ] **Slippage modeled:**
- Assumed slippage: 0.05% per side (0.10% round trip)
- Cost per trade: $__
- [ ] **Commissions modeled:**
- Broker commission: $__ per side
- Round trip cost: $__
- [ ] **Total realistic cost per trade:** $__

Look-Ahead Bias Check

- [ ] Strategy uses ONLY data available at decision time? ✓
- [ ] No "future candles" referenced in code? ✓
- [ ] Indicators using historical data only (not peeking forward)? ✓

Survivorship Bias Check

- [ ] Testing on current index constituents only? Yes / No
- [ ] If Yes: Risk of survivorship bias (missing delistings/bankruptcies)
- [ ] Data includes delisted stocks? Yes / No (should be Yes)

🎯 Backtest Results Analysis

In-Sample Results (Development Data)

- [ ] Total trades: __ (min 100 for statistical significance)
- [ ] Win rate: ____% (realistic: 50-65%)
- [ ] Average R-multiple: ____R (realistic: 1.5-3.0R)
- [ ] Profit factor: ____ (realistic: 1.5-2.5)
- [ ] Sharpe ratio: ____ (realistic: 1.0-2.5, suspicious if > 3.0)
- [ ] Max drawdown: ____% (realistic: 10-25%)

Out-of-Sample Results (Validation Data)

- [ ] Total trades: __
- [ ] Win rate: ____% (should be within 10% of in-sample)
- [ ] Average R-multiple: ____R (should be within 20% of in-sample)
- [ ] Profit factor: ____ (should be within 20% of in-sample)
- [ ] Max drawdown: ____% (should be similar to in-sample)

Performance degradation:

- [ ] Out-of-sample vs. in-sample difference: ____%
- [ ] If > 30% degradation: OVERFIT (scrap strategy)
- [ ] If < 20% degradation: Robust (proceed to paper trading)

🚩 Red Flags Detector

Mark any that apply:

- [ ] Sharpe ratio > 3.0 (too good to be true)
- [ ] Average R > 5.0 (unrealistic without extreme execution costs)
- [ ] Only 20-50 trades (sample too small)
- [ ] Perfect equity curve (no drawdowns = suspicious)
- [ ] Tested on 1 asset only (likely fit to that specific regime)
- [ ] 10+ optimized parameters (curve-fit to noise)
- [ ] Out-of-sample much worse than in-sample (overfit)

Red flag count: ____/7

Action:

- 0-1 flags: Proceed to paper trading ✓
- 2-3 flags: Revise strategy, reduce parameters
- 4+ flags: Scrap strategy (too overfit)

🔍 Walk-Forward Analysis (Advanced)

Test across multiple periods:

Period	Train Data	Test Data	Win Rate	Avg R	Profit Factor
1	__	__	____%	____R	____
2	__	__	____%	____R	____
3	__	__	____%	____R	____
4	__	__	____%	____R	____

- [ ] Performance consistent across all periods? Yes / No
- [ ] If No: Strategy is regime-dependent (fails in certain markets)
- [ ] If Yes: Robust strategy ✓

📊 Validation Scorecard (60 Points Max)

Category 1: Sample Size (10 points)

- [ ] < 30 trades: 0 points (statistically meaningless)
- [ ] 30-100 trades: 5 points (marginal)
- [ ] 100-300 trades: 8 points (good)
- [ ] 300+ trades: 10 points (excellent)

Category 2: Out-of-Sample Performance (15 points)

- [ ] Degradation > 50%: 0 points (overfit)
- [ ] Degradation 30-50%: 5 points (concerning)
- [ ] Degradation 20-30%: 10 points (acceptable)
- [ ] Degradation < 20%: 15 points (robust)

Category 3: Realistic Costs (10 points)

- [ ] No costs modeled: 0 points (fantasy)
- [ ] Commission only: 3 points (incomplete)

- [ ] Commission + spread: 7 points (good)
- [ ] Commission + spread + slippage: 10 points (realistic)

Category 4: Look-Ahead Bias (10 points)

- [ ] Code not reviewed: 0 points
- [ ] Reviewed, unclear: 5 points
- [ ] Confirmed no future data: 10 points ✓

Category 5: Parameter Count (5 points)

- [ ] 10+ parameters: 0 points (overfit)
- [ ] 5-9 parameters: 2 points (risky)
- [ ] 2-4 parameters: 4 points (good)
- [ ] 0-1 parameters: 5 points (robust)

Category 6: Sharpe Ratio (5 points)

- [ ] Sharpe < 0.5: 0 points (poor)
- [ ] Sharpe 0.5-1.0: 2 points (marginal)
- [ ] Sharpe 1.0-2.5: 5 points (realistic)
- [ ] Sharpe > 3.0: 0 points (suspicious)

Category 7: Win Rate (5 points)

- [ ] < 40%: 0 points (low, needs high R:R)
- [ ] 40-50%: 3 points (acceptable if R:R > 2)
- [ ] 50-65%: 5 points (realistic)
- [ ] > 70%: 2 points (suspicious unless low R:R)

Category 8: Testing Across Assets (10 points)

- [ ] 1 asset: 2 points (risky)
- [ ] 2-5 assets: 6 points (good)
- [ ] 6+ assets: 10 points (robust)

TOTAL SCORE: ____/60

Action plan:

- [ ] 50-60 points: Proceed to paper trading ✓
- [ ] 40-49 points: Revise and re-test
- [ ] < 40 points: Scrap strategy

🔄 Paper Trading Protocol

Before going live:

- [ ] Paper trade for 30-60 trades minimum
- [ ] Track ACTUAL fill prices (not theoretical)
- [ ] Log slippage on every trade
- [ ] Compare paper results to backtest
- [ ] If paper within 20% of backtest: Go live with small size
- [ ] If paper much worse: Re-validate backtest assumptions

Paper trading results:

- Trades: _
- Win rate: _**% (backtest:** %)
- Avg R: _**R (backtest:** _R)
- Difference: ____%

📊 Post-Backtest Review

Validation quality:

- Did I model realistic costs? Yes / No
- Did I split data before optimization? Yes / No
- Is out-of-sample within 20% of in-sample? Yes / No
- Did I check for look-ahead bias? Yes / No
- Is sample size > 100 trades? Yes / No

What went right:

- _____

What could be improved:

- _____

Lesson learned:

- _____

Decision: Proceed to paper trading / Revise strategy / Scrap

Remember:

- ✅ Split data BEFORE optimization (60% in-sample, 40% out-of-sample)
- ✅ Model ALL costs (spread, slippage, commission)
- ✅ 100+ trades minimum for statistical validity
- ✅ Out-of-sample degradation < 20% = robust
- ✅ Sharpe > 3.0 = suspicious (likely overfit)
- ✅ Paper trade 30-60 trades before going live

This is for educational purposes only. Not financial advice.
