Week 2: Data and Measurement in Finance

Understanding Where Your Numbers Come From

Part I: Data Science Begins with Data

The Fundamental Question

Gelman, Hill, and Vehtari (2020) put it bluntly:

> *"Before fitting a model... it is a good idea to understand where your numbers are coming from."*

This is not preliminary work : it *is* the analysis.

Why Data Understanding Matters

In finance, data problems lead to:

- **Real losses**: Backtests on biased data overestimate returns
- **Regulatory failures**: Risk models trained on crisis-free periods underestimate tails
- **Misallocated capital**: Algorithmic systems that leak future information fail in production

The most sophisticated model built on flawed data produces flawed results : often with false confidence.

Learning Objectives

By the end of this session, you should be able to:

1. **Articulate** the data generating process (DGP) underlying a financial dataset
2. **Distinguish** between measured variables and latent constructs
3. **Identify** common selection biases and their consequences
4. **Apply** EDA techniques appropriate for financial returns
5. **Implement** data validation procedures before analysis

Part II: Data Generating Process & Measurement

What Is a DGP?

Every dataset is the output of some **data generating process** : the mechanism that determines:

- What gets recorded
- When it gets recorded
- How it gets recorded

Understanding the DGP reveals assumptions embedded in your data *before* you add modelling assumptions.

Stock Prices: Not Just "The Market"

The prices you observe are the result of:

Component	What It Determines	Theory Connection
Exchange mechanisms	How orders are matched, which prices are recorded	Market microstructure (order flow, liquidity)
Data vendor processing	How raw tick data is aggregated, cleaned, distributed	Aggregation creates information loss
Selection rules	Which securities are included, for how long	Survivorship bias, sample selection
Timing conventions	When prices are measured (close, VWAP, bid-ask midpoint)	Price discovery process, bid-ask spread

Each choice affects downstream analysis and embeds theoretical assumptions.

The Trader's Dilemma: Make or Take?

Every time a trader wants to buy or sell, they face a strategic choice:

Make liquidity (patience) : post a limit order and *wait*

- "I'll buy at £99.95" $\rightarrow$ your order joins the book as a **bid**
- You get a better price, but you might not get filled
- You are a **maker** : you *add* liquidity to the market

Take liquidity (urgency) : hit an existing order *now*

- "I'll buy at whatever the cheapest seller is offering" $\rightarrow$ you pay the **ask** (£100.05)
- You trade immediately, but you pay the spread as your cost
- You are a **taker** : you *remove* liquidity from the market

The **spread** is the price of this choice. Patient traders earn it; urgent traders pay it.
Every "price" in your dataset is the result of someone choosing to *take*.

First Principles: How Prices Form

Every price you observe in financial data is the outcome of this make-or-take negotiation. The *makers* set the terms:

- **A buyer** wants to pay **as little as possible** $\rightarrow$ posts a **bid**: "I'll buy at most £99.95"
- **A seller** wants to receive **as much as possible** $\rightarrow$ posts an **ask**: "I'll sell at no less than £100.05"

These limit orders accumulate in the **order book** : a live record of supply and demand:

- **Bids stack below** the current price (demand waiting to be filled)
- **Asks stack above** the current price (supply waiting to be taken)
- **The spread** (£100.05 − £99.95 = £0.10) is the gap where **no one yet agrees**

A *taker* closes the gap: a buyer pays the ask price, or a seller accepts the bid price. That crossing *is* the "price" in your dataset.

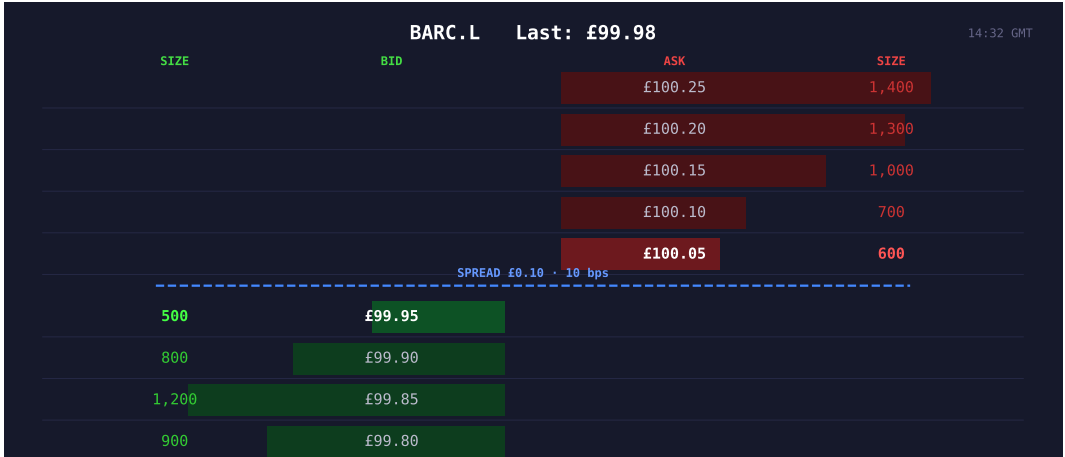
The Order Book: A Concrete Example

Consider a stock trading near £100. The book at one instant:

Side	Price	Volume	Meaning
Ask (sell)	£100.25	1,400	"I'll sell 1,400 shares at £100.25 or higher"
Ask (sell)	£100.15	1,000	"I'll sell 1,000 shares at £100.15 or higher"
Ask (sell)	£100.05	600	← **Best Ask** (cheapest available seller)
	Spread: £0.10		*No one agrees in this gap*
Bid (buy)	£99.95	500	← **Best Bid** (most generous buyer)
Bid (buy)	£99.90	800	"I'll buy 800 shares at £99.90 or lower"
Bid (buy)	£99.85	1,200	"I'll buy 1,200 shares at £99.85 or lower"

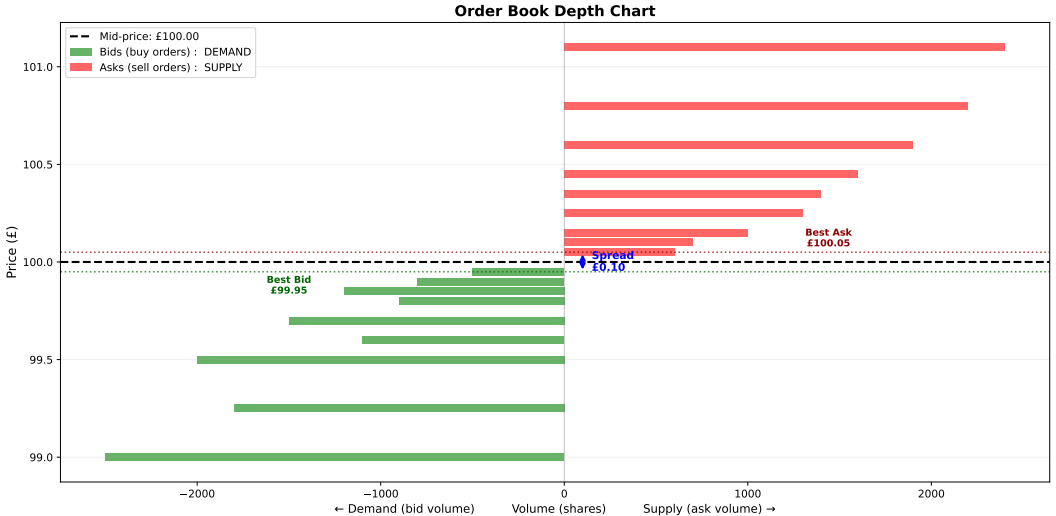
What Traders Actually See: The DOM Ladder

On a trading terminal, the order book appears as a **Depth of Market (DOM) ladder** : a compact vertical strip with bids on the left and asks on the right, converging at the spread.



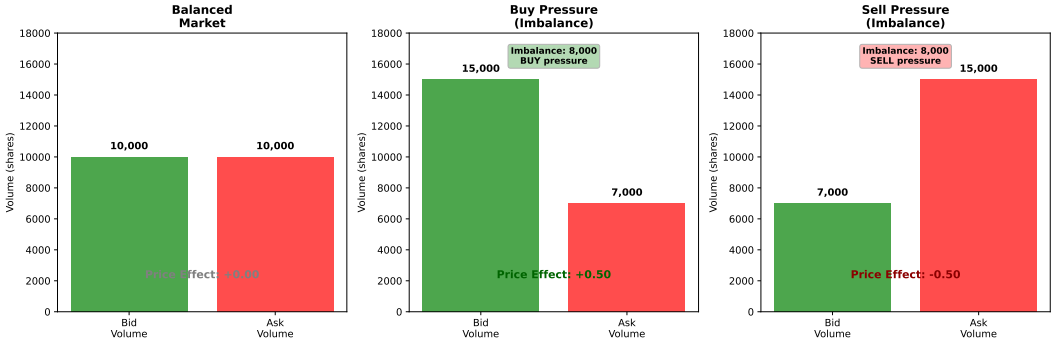
The Depth Chart: Visualising the Order Book

The depth chart is this same order book rendered visually : demand (bids) extending left, supply (asks) extending right, just like the supply-demand diagrams from economics.



Order Imbalance & Price Discovery

Order imbalance: When buy volume sell volume, price adjusts.




```
=== Order Imbalance & Price Impact ===

Scenario 1 (Balanced):
  Bid volume: 10,000 | Ask volume: 10,000 | Imbalance: 0
  → Price change: +0.00 (no pressure)
```

Professional Data Infrastructure

Key principle: Use curated databases, not ad-hoc API calls.

Why this matters:

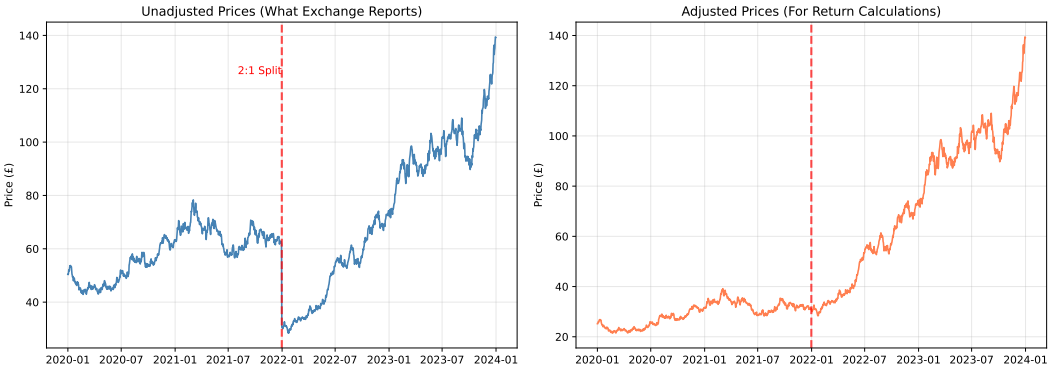
- **Consistency**: Same data across all analyses (reproducibility)
- **Quality**: Pre-validated, survivorship-bias-aware, adjusted for corporate actions
- **Efficiency**: Load once, analyse many times (vs repeated API calls)

In this course: Bloomberg database contains ~30 securities, 2015-2024, daily adjusted prices.

In labs, you'll load from this database using a fallback cascade (local file $\rightarrow$ Colab URL $\rightarrow$ synthetic).

DGP Example: Adjusted vs Unadjusted Prices

Critical distinction: adjusted prices account for corporate actions (splits, dividends).




```
Returns calculated on unadjusted: 176.2%
Returns calculated on adjusted: 452.5%
```

The DGP Affects What Questions You Can Answer

Bloomberg database structure:

- Daily close prices (adjusted for splits/dividends)
- Index constituents at point-in-time
- Survivorship-bias-aware (includes delisted securities)

This DGP enables: Long-term return analysis, portfolio backtests, survivorship studies

This DGP prevents: Intraday patterns, microstructure analysis, order flow studies

Measurement: Observed vs Latent

We rarely observe what we truly want to study:

Latent Construct	Observable Proxy	Measurement Gap
True risk	Historical volatility	Backward-looking; regime-dependent
Information asymmetry	Bid-ask spread	Also reflects inventory, competition
Market sentiment	Text analysis, word counts	Ignores semantic meaning, context
Firm productivity	Accounting ratios (ROA, ROE)	Ignores size, treats firms as homogeneous (Fox Paradox)

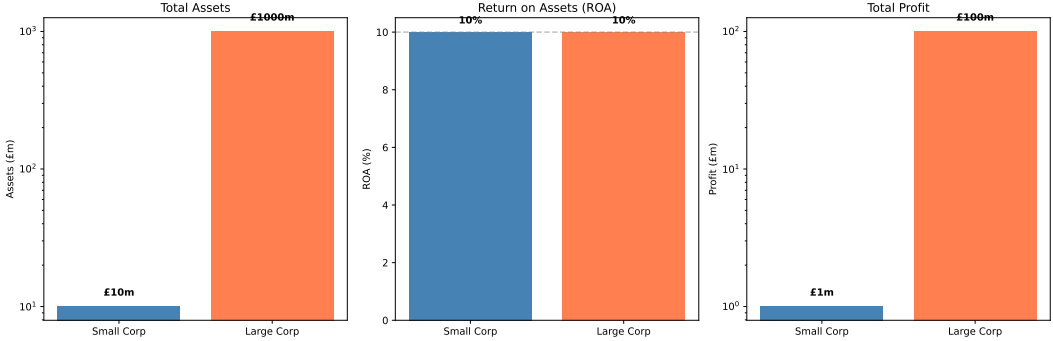
This is the **measurement problem** in statistical science.

Modern example: Sentiment analysis counts words ("bullish", "bearish") but misses context and irony. Transformer models (BERT, GPT) help by capturing semantic

The Fox Paradox: When Ratios Mislead

Problem: Accounting ratios (ROA, ROE) treat all firms as homogeneous, ignoring scale.

Example: Two firms, both with 10% ROA:




```
=== Fox Paradox Demonstration ===
      Firm  Assets (£m)  Profit (£m)  ROA (%)
Small Corp           10            1       10
Large Corp         1000          100       10
```

Classical Measurement Error Model

When proxy x measures true variable x^* with noise:

$$x = x^* + u$$

where $u \perp x^*$ (classical error assumption).

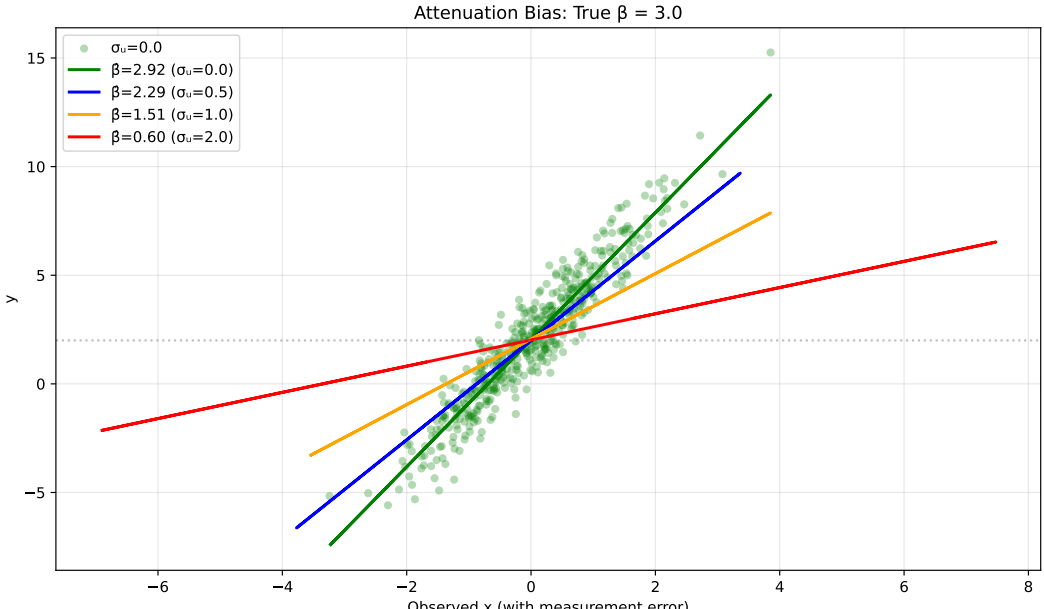
Consequence in regression ($y = \alpha + \beta x^* + \varepsilon$):

$$\hat{\beta}_{OLS} = \beta \cdot \underbrace{\frac{\text{Var}(x^*)}{\text{Var}(x^*) + \text{Var}(u)}}_{\text{signal-to-total ratio}}$$

Since $0 < \text{Var}(x^*)/[\text{Var}(x^*) + \text{Var}(u)] < 1$, we have $|\hat{\beta}_{OLS}| < |\beta|$.

This is **attenuation bias** : measurement error shrinks coefficients toward zero.

Attenuation Bias: Simulation



Validity vs Reliability

Validity: Does your measure capture the construct you intend to study?

Reliability: Are your measurements consistent and reproducible?

Both matter, but **validity** $>$ **reliability** : a reliable but invalid measure is consistently wrong.

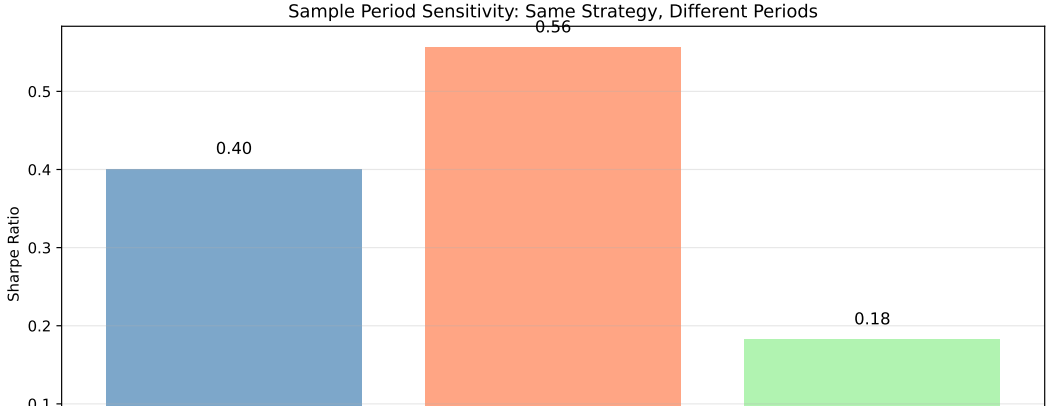
Example: High-frequency volatility estimates are highly reliable (consistent) but may not be valid measures of "risk" (depends on your risk concept).

Reliability in Financial Data

Reliability question: Are findings robust to reasonable variations in measurement?

Example metric: **Sharpe ratio** = (Return - Risk-free rate) / Volatility

This measures risk-adjusted performance : higher is better. But is it reliable across time periods?



The HDI Lesson: Composite Measures Can Mislead

Human Development Index (HDI): UN measure combining life expectancy, education, and income.

Sounds comprehensive, but Gelman, Hill, and Vehtari (2020) show:

> *"The map is pretty much a map of state income with a mysterious transformation and a catchy name."*

Lesson: Most HDI variation comes from income alone : other components add little information.

Finance parallels:

- "Smart beta" funds may be mostly value or momentum tilts with marketing labels
- "Alternative data" signals may proxy publicly available information
- "Risk-adjusted returns" depend entirely on how you define risk (volatility? downside? beta?)

Always ask: What is my variable *actually* measuring?

Part III: Selection Bias & Generalisation

Selection Bias: The Silent Killer

Definition: Your sample differs systematically from the population you want to study.

Four common forms in finance:

Bias Type	What Happens	Direction
Survivorship	Failed entities disappear from data	Upward
Availability	Only easy-to-get data is used	Varies
Reporting	Voluntary disclosure is strategic	Upward
Look-ahead	Future information leaks into past	Inflates performance

Survivorship Bias

Databases of currently listed stocks exclude companies that:

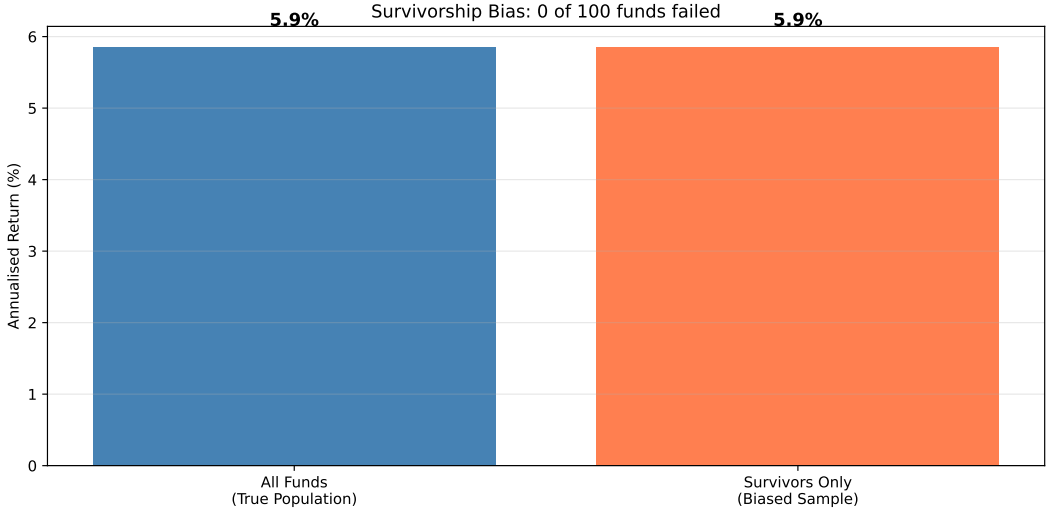
- Failed (bankruptcy, liquidation)
- Were acquired (merger, takeover)
- Delisted (voluntary or involuntary)

Result: The worst performers are removed $\rightarrow$ upward bias in measured returns.

Magnitude from research: Academic studies find survivorship bias of ~0.5-1.5% per year in US mutual funds; **crisis periods show much larger biases** (5-10%+ in UK banking 2008).

Survivorship Bias: Simulation

Simulation setup: 100 funds, 5 years monthly returns (mean 0.5%, volatility 4% monthly). Funds with cumulative loss $> 50\%$ "fail" and exit the database.



Detecting Survivorship Bias

How to check your data source for bias:

1. **Count securities with data ending before present** : these are delistings
2. **Check final returns for delistings** : often large negative returns
3. **Compare returns with/without delistings** : quantifies the bias

Example diagnostic:

- Total securities in database: 100
- Securities with data ending >1 year ago: 15 (15%)
- Average return of survivors: 8% per year
- Average return of all securities: 6% per year
- **Survivorship bias: 2 percentage points per year**

Professional databases (Bloomberg, CRSP, FactSet) maintain delisting information and point-in-time records to enable survivorship-free analysis.

Look-Ahead Bias: The Insidious Form

The most dangerous form in backtesting : using information not available at the time.

Why each type creates bias:

- **Point-in-time failures**: Restated earnings incorporate information revealed *after* trading date → you "know" future information when making past decisions
- **Index membership**: Announcement of S&P 500 addition causes price jump → using post-announcement prices for pre-announcement period inflates returns
- **Universe selection**: "I'll study high-momentum stocks in 2020" → selecting on outcome creates cherry-picking → phantom predictability

All three share common flaw: Decision at time t uses information from time t+1 or later.

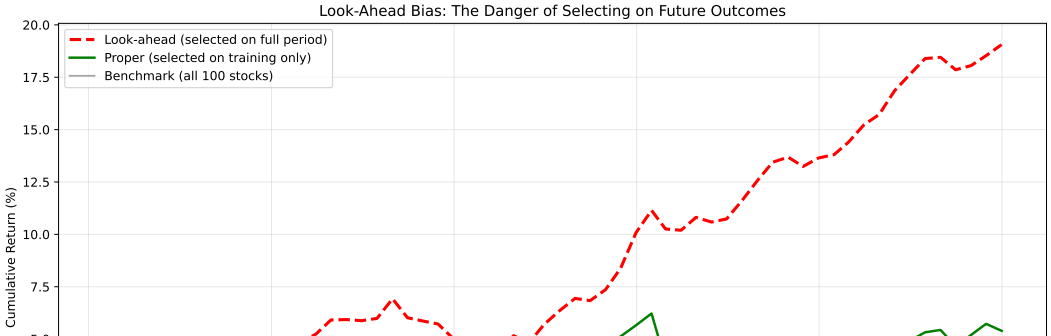
> ⚠ The Golden Rule
>
> At any point in your backtest, you may only use information available at that point in time.

Look-Ahead Bias: Example

Simulation: 100 stocks, 10 years monthly (120 periods), each with random returns (mean 0.05%, vol 2% monthly). We select top 10 performers.

Wrong approach: Select top 10 over *full* 10 years, then evaluate performance in second half $\rightarrow$ uses future information.

Right approach: Select top 10 over *first* 5 years only, then evaluate in second half $\rightarrow$ only uses past information.



Connection to Statistical Science: Generalisation

Selection bias violates **Gelman's Challenge 1** (Sample $\rightarrow$ Population):

- Your sample: survivors, available data, reported outcomes
- Your population: all entities that existed, including failures

Consequence: Inferences don't generalise from sample to population.

Solution: Survivorship-bias-free databases, point-in-time datasets, temporal validation.

Professional tools: Bloomberg maintains **point-in-time databases** that record:

- Which securities existed at each date (including now-delisted)
- Index memberships as of specific dates (S&P 500 constituents in 2010, not today's)
- As-originally-reported financials (not restated)

This enables look-ahead-free backtesting.

Part IV: Exploratory Data Analysis in Practice

The Philosophy of EDA

John Tukey's philosophy: look at data before modelling it.

EDA goals:

1. Understand distributional properties
2. Identify patterns and relationships
3. Detect anomalies and outliers
4. Generate hypotheses (rather than test them)

Gelman, Hill, and Vehtari (2020): "All graphs are comparisons."

Remember: EDA generates questions; modelling provides answers.

Step 1: First Look at Your Data

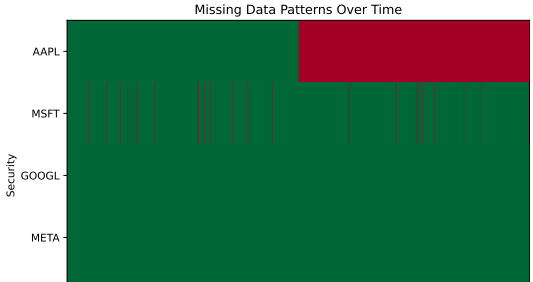
Always start with basic inspection:

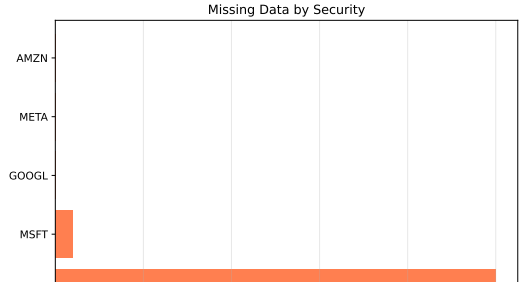
```
=== Data Structure ===
Shape: (1462, 5)
Date range: 2020-01-01 00:00:00 to 2024-01-01 00:00:00
Data types: [dtype('float64')]

=== First 3 rows ===
                AAPL        MSFT        GOOGL        META        AMZN
2020-01-01  120.662789  144.087901  133.109004  107.659211  122.803641
2020-01-02  125.289229  145.103732  131.834117  104.471862  121.400290
2020-01-03  124.609702  142.219353  129.343588  102.419741  124.289865

=== Last 3 rows ===
                AAPL        MSFT       GOOGL       META       AMZN
2023-12-30  887.221891  364.569372  53.245292  61.460758  98.329511
2023-12-31  882.953509  363.957197  53.862049  59.609061  99.013962
```

Step 2: Check for Missing Data

Why check: Missing data patterns reveal systematic issues (delistings, thin trading, data vendor problems).

```
=== Missing Data Summary ===
AAPL: 50.0% missing
MSFT: 2.0% missing
GOOGL: 0.0% missing
META: 0.0% missing
AMZN: 0.0% missing
```





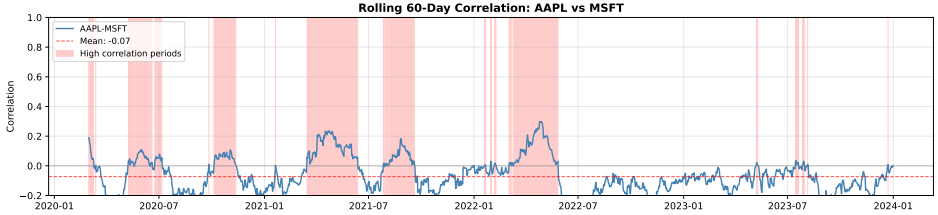
Step 3: Calculate and Examine Returns

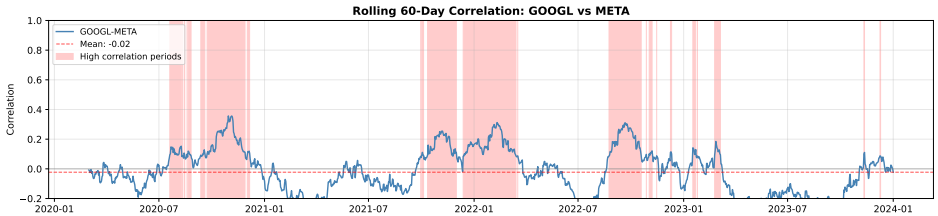
```
=== Return Statistics ===
              AAPL         MSFT        GOOGL         META         AMZN
count   1461.0000    1461.0000    1461.0000    1461.0000    1461.0000
mean       0.0016       0.0008      -0.0004      -0.0002       0.0000
std        0.0243       0.0168       0.0222       0.0222       0.0194
min       -0.0761      -0.0499      -0.0631      -0.0683      -0.0581
25%       -0.0152      -0.0104      -0.0155      -0.0145      -0.0137
50%        0.0015       0.0003      -0.0004      -0.0003      -0.0003
75%        0.0170       0.0118       0.0136       0.0148       0.0136
max        0.0994       0.0695       0.0735       0.0724       0.0685

=== Skewness (negative = left tail longer) ===
AAPL     0.136
MSFT     0.093
GOOGL    0.021
META     0.095
AMZN     0.070
```

Time-Varying Correlations: The Crisis Effect

Static correlations hide non-stationarity: Correlations increase during crises (when you need diversification most).

Rolling correlation: Calculate correlation over moving window (e.g., 60 trading days).





Step 4: Check for Outliers

IQR method (Interquartile Range): Standard statistical approach for outlier detection.

Logic:

1. Calculate Q1 (25th percentile) and Q3 (75th percentile)
2. IQR = Q3 - Q1 (middle 50% of data)
3. Outliers: values beyond Q1 - 1.5×IQR or Q3 + 1.5×IQR
4. **Why 1.5**: Tukey's rule : captures ~99.3% of normal data

```
=== Outlier Detection (IQR Method) ===

Outliers by security:
  AAPL: 11 outliers (0.8%)
  MSFT: 18 outliers (1.2%)
  GOOGL: 11 outliers (0.8%)
  META: 19 outliers (1.3%)
  AMZN: 5 outliers (0.3%)
```

Stylised Facts of Financial Returns

Before deep analysis, check if your data exhibits the known empirical facts:

1. **Returns are approximately unpredictable** : weak autocorrelation
2. **Volatility is predictable and persistent** : clustering
3. **Returns have fat tails** : more extremes than Normal predicts
4. **Returns are negatively skewed** : crashes $>$ melt-ups
5. **The leverage effect** : negative returns increase subsequent volatility

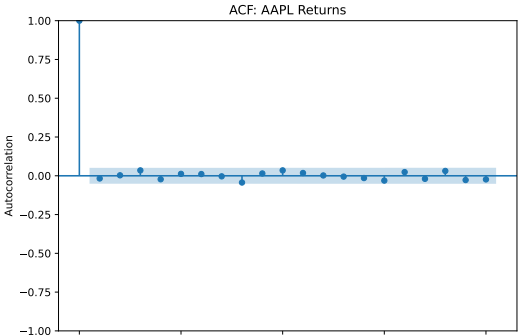
These are **empirical facts**, not assumptions.

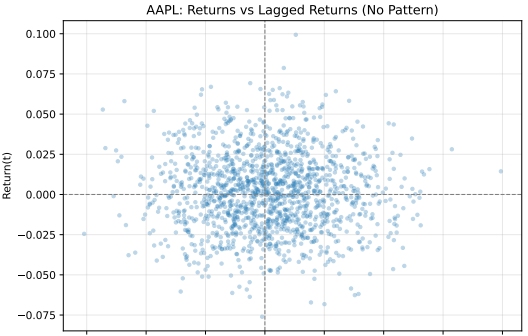
Checking Fact 1: Weak Return Autocorrelation

ACF (Autocorrelation Function): Measures correlation between a series and its lagged values.

- **Blue shaded region**: 95% confidence band for "no correlation"
- **Bars within band**: No significant autocorrelation at that lag
- **Bars outside band**: Significant autocorrelation (predictability)

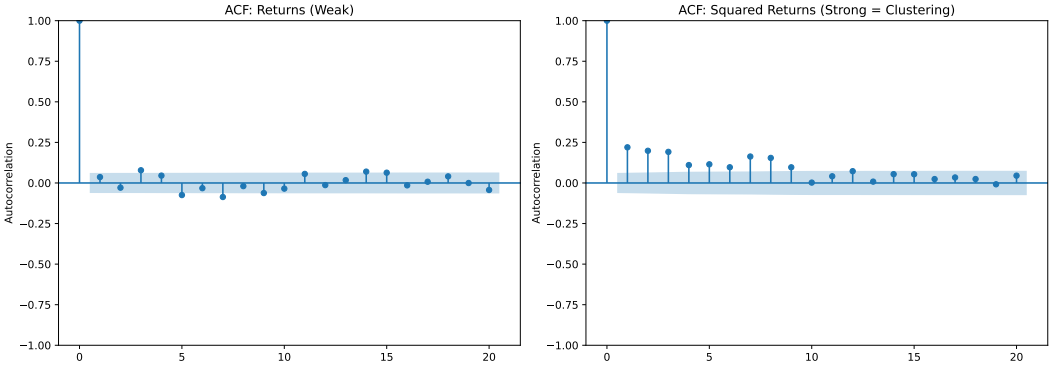
For returns: Most lags should be within band (unpredictable).





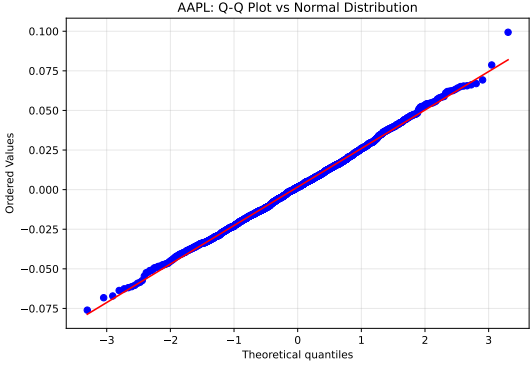
Checking Fact 2: Volatility Clustering

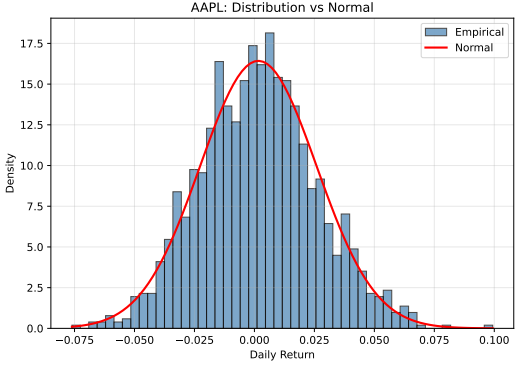
Key insight: Returns unpredictable, but **volatility** (squared returns) shows persistence.




```
Returns ACF(1): 0.0364 (near zero)
Squared returns ACF(1): 0.2198 (strong positive)
```

Checking Fact 3: Fat Tails



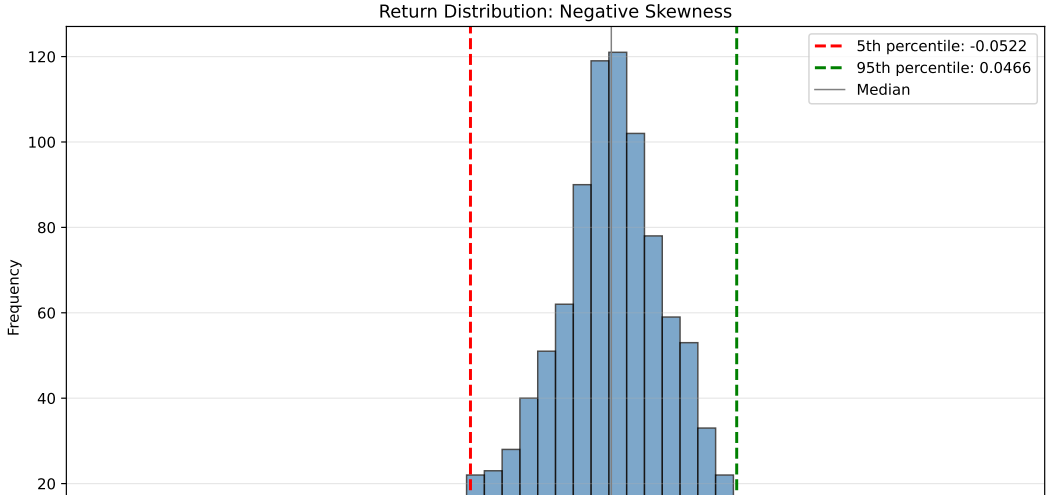



```
Excess kurtosis: 0.03 (Normal = 0)
Normality test p-value: 0.1008
→ Cannot reject normality
```

Checking Fact 4: Negative Skewness

Negative skewness: Left tail (losses) extends further than right tail (gains).

Interpretation: Extreme losses are more likely than extreme gains (crashes vs rallies).



Why Stylised Facts Matter for Modelling

These facts constrain sensible model specifications:

Stylised Fact	Modelling Implication
Weak return autocorrelation	Simple **AR** (AutoRegressive) models won't forecast returns
Volatility clustering	Need **GARCH** (Generalised AutoRegressive Conditional Heteroskedasticity)
Fat tails	Normal-based VaR underestimates risk
Negative skewness	Symmetric distributions mis-specify downside
Leverage effect	Volatility increases when prices fall

Simple linear models assuming i.i.d. normal errors will be severely misspecified.

Next weeks: We'll cover AR models (Week 3) and GARCH models (Week 4) that properly handle these properties.

Part V: Data Quality & Validation

Common Data Quality Issues

Even Bloomberg data can have issues:

Price errors:

- Decimal point mistakes (rare in Bloomberg, common in free sources)
- Stale prices (unchanged for many days during thin trading)
- Missing corporate action adjustments

Timing errors:

- Timestamp misalignments across sources
- Weekend/holiday prices (should be excluded)
- Look-ahead bias from restated data

Building a Validation Pipeline

Automated validation: Run systematic checks before every analysis.

Four key checks:

1. **Missing values**: Flag columns with >5% missing
2. **Extreme returns**: Flag |return| > 50% daily (likely errors)
3. **Stale prices**: Flag unchanged prices for >5 consecutive days
4. **Date gaps**: Flag gaps >10 days (missing trading days)

Pattern: Define validator class with methods for each check, then run all checks and report issues.

In labs: You'll implement this full validation pipeline as a reusable class.

Validation Workflow in Practice

```
=== Data Provenance Log ===
dataset: Bloomberg Sample
date_range: 2020-01-01 00:00:00 to 2024-01-01 00:00:00
shape: (1462, 5)
securities: 5
observations: 1462
missing_values: 0
missing_pct: 0.00%
extreme_returns: 0
max_date_gap_days: 1

  Validation log saved to data_validation_log.json
```

Missing Data Patterns: MCAR, MAR, MNAR

Not all missing data is equal:

Pattern	Description	Implication	Financial Example
MCAR	Missing Completely At Random	Safe to ignore (no bias)	Random data recording glitches
MAR	Missing At Random (given observed)	Imputation may work	Small-cap stocks missing during holidays
MNAR	Missing Not At Random	Serious bias risk	Failed funds stop reporting; illiquid assets gap during stress

In finance, missing data is rarely MCAR:

- **Failed funds stop reporting** (MNAR: failure $\rightarrow$ missing)
- **Illiquid assets have gaps** (MNAR: low liquidity $\rightarrow$ missing)
- **Voluntary disclosure is strategic** (MNAR: poor performance $\rightarrow$ missing)

Quality Gate Pattern

Quality gate: Automated pass/fail check before analysis proceeds.

Logic:

1. Define acceptable thresholds (e.g., $<5\%$ missing, <10 extreme returns)
2. Run validation checks
3. If any check fails $\rightarrow$ halt analysis, investigate issues
4. If all pass $\rightarrow$ proceed with confidence

Example thresholds:

- Missing data: $<5\%$ acceptable, $>10\%$ investigate
- Extreme returns ($|r| > 50\%$): <5 occurrences (corporate actions)
- Data freshness: <30 days since last update

This prevents silent failures: Better to catch issues early than discover them after modelling.

In labs, you'll implement this as a reusable function with customizable thresholds.

Part VI: Statistical Science Foundations

Data Problems Are Uncertainty Problems

Recall from Week 1: **statistical science is the study of variation and uncertainty**.

Data quality issues introduce **systematic uncertainty**:

- **Survivorship bias** $\rightarrow$ overestimate expected returns (upward bias)
- **Look-ahead bias** $\rightarrow$ overestimate predictability (spurious patterns)
- **Measurement error** $\rightarrow$ underestimate relationships (attenuation bias)
- **Missing data (MNAR)** $\rightarrow$ biased sample statistics

These aren't random noise : they're **systematic biases** that invalidate inference.

The Three Challenges: Data Quality Edition

Gelman, Hill, and Vehtari (2020)'s three fundamental challenges apply directly to data work:

Challenge 1: Generalisation (Sample → Population)

- **Question**: Is your sample representative of the population?
- **Data threat**: Survivorship bias, availability bias
- **Example**: Studying "bank performance" using only current survivors misses failures
- **Solution**: Survivorship-bias-free databases, careful sample selection

Challenge 2: Comparison (Treatment → Control)

Challenge 2: Causal Inference

- **Question**: Can you attribute effects to specific causes?
- **Data threat**: Look-ahead bias, confounding from DGP
- **Example**: Using restated earnings to predict stock returns (data includes post-announcement info)
- **Solution**: Point-in-time datasets, temporal validation

Look-ahead bias is a form of post-treatment bias: using information generated after the "treatment" (investment decision) to evaluate outcomes.

Challenge 3: Measurement (Proxy $\rightarrow$ Construct)

Challenge 3: Validity

- **Question**: Does your measurement capture what you intend to study?
- **Data threat**: Proxy variables, latent constructs
- **Example**: Using historical volatility to measure "risk" (backward-looking, regime-dependent)
- **Solution**: Multiple measures, validation studies, acknowledge limitations

Attenuation bias: measurement error systematically underestimates true relationships.

Data Work IS Statistical Work

Key insight: Data cleaning and preparation aren't preliminary tasks : they're integral parts of statistical inference.

Every data decision affects uncertainty:

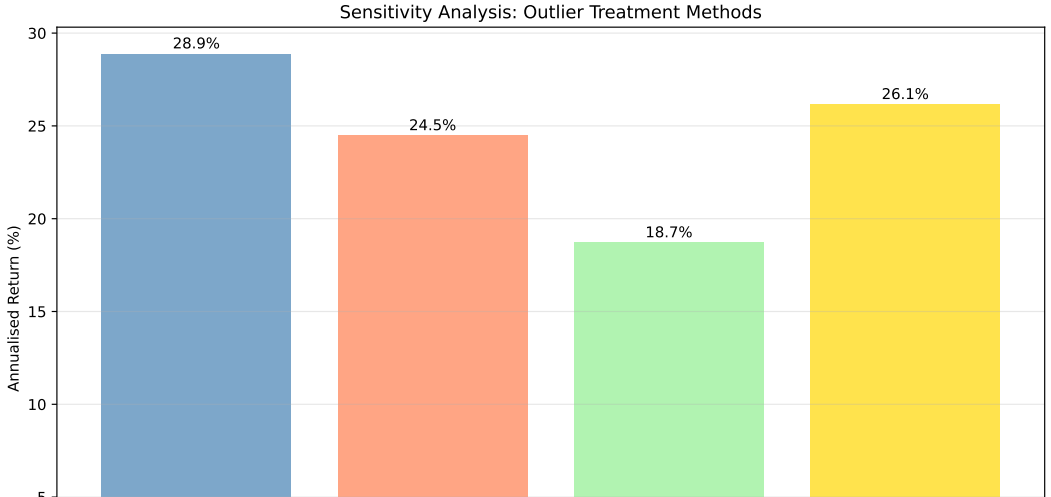
1. **Which data to include/exclude** $\rightarrow$ selection effects on generalisation
2. **How to handle missing values** $\rightarrow$ bias vs variance trade-off
3. **How to measure constructs** $\rightarrow$ validity vs reliability trade-off
4. **How to validate quality** $\rightarrow$ type I vs type II error trade-off

From Week 1: Variation and uncertainty propagate through your entire analysis pipeline.

Quantifying Data Processing Uncertainty

Question: How sensitive are conclusions to data cleaning choices?

Sensitivity analysis: Test multiple reasonable outlier treatments, compare results.



Statistical Principles for Data Work

Six principles from Week 1, applied to data:

1. **Variation**: All data contains measurement variation : quantify it
2. **Uncertainty**: Data quality introduces systematic uncertainty : don't ignore it
3. **Generalisation**: Selection effects limit generalisation : use representative samples
4. **Comparison**: Look-ahead bias creates spurious comparisons : use point-in-time data
5. **Measurement**: Proxies aren't perfect : validate and document limitations
6. **Signal vs Noise**: Data errors add noise : validate to preserve signal

This is what "data science as statistical science" means.

Connection to Assessment

In your coursework, you will:

- Load and validate financial data (CW1, CW2)
- Document data provenance and quality checks (all assignments)
- Handle missing data and outliers (all assignments)
- Avoid survivorship and look-ahead bias (CW2 backtests)
- Justify measurement choices (all assignments)

Quality of data work directly affects quality of inference.

Part VII: This Week's Labs

Lab Structure: Dual-Track Approach

Track 1: Colab/Home (Lab 02 APIs)

- Practice API calls with fallback strategies
- Implement data validation pipeline
- Explore stylised facts with free data
- Build reproducible workflows

Track 2: Bloomberg Terminal (Lab 02 Survivorship Bias)

- Extract data from Bloomberg using Excel formulas
- Quantify survivorship bias with real UK banking crisis data
- Compare survivors vs full population returns
- Professional data workflow experience

What You'll Build This Week

By end of labs, you should have:

1. Working data validation pipeline (Colab lab)
2. Provenance logging system (Colab lab)
3. Quantified estimate of survivorship bias (Bloomberg lab)
4. EDA workflow for financial returns (both labs)
5. Documentation of all data decisions (both labs)

These are portable skills : use them in every future analysis.

Key Takeaways

Five principles for responsible data science:

1. **Understand your DGP** : prices are processed, not raw
2. **Quantify selection bias** : survivorship, look-ahead, availability
3. **Validity** $>$ **Reliability** : measure the right thing, not just consistently
4. **Verify stylised facts** : your data should show known empirical patterns
5. **Document everything** : provenance, quality checks, transformations

Remember: Data understanding is the analysis, not preparation for it.

Directed Learning

Core readings (course textbook):

- **Chapter 02: Data and Measurement in Finance** : covers DGP, selection bias, validity
  - https://quinfer.github.io/financial-data-science/chapters/02_data_measurement.html
- **Complete Lab 02 APIs** (homework version) : hands-on validation pipeline
  - https://quinfer.github.io/financial-data-science/labs/lab02_apis.html
- **Prepare for Bloomberg session** (if applicable) : review Excel formulas

Extension (recommended):

- **Chapter 03: Time Series & Volatility** : extends stylised facts coverage
  - https://quinfer.github.io/financial-data-science/chapters/03_volatility_modelling.html
- Write brief reflection: "What data quality issues affect my coursework area?"
- Implement validation pipeline for your own data

Further reading (if interested): Gelman & Hill (2020) Ch 2; Tsay (2010) Ch 1

Exit Quiz

Quick Check: Data & Measurement Concepts

Answer these three questions before leaving:

Q1: Which type of selection bias is created when you use restated earnings data in a backtest?

a) Survivorship bias
b) Availability bias
c) Look-ahead bias
d) Reporting bias

Q2: What does it mean if ACF of returns is near zero but ACF of squared returns is high and persistent?

a) Returns are predictable
b) Volatility clustering is present
c) Data has errors
d) Distribution is normal

Q3: A hedge fund database shows average returns of 12% per year. Why might this overestimate true performance?

References

Gelman, Andrew, Jennifer Hill, and Aki Vehtari. 2020. *Regression and Other Stories*. Cambridge, UK: Cambridge University Press. https://avehtari.github.io/ROS-Examples/.

Tsay, Ruey S. 2010. *Analysis of Financial Time Series*. 3rd ed. Hoboken, NJ: Wiley.