



Statistics of Financial Markets

Chapter 1: Introduction and Data Sources



Daniel Traian PELE

Bucharest University of Economic Studies

IDA Institute Digital Assets

Blockchain Research Center

AI4EFin Artificial Intelligence for Energy Finance

Romanian Academy, Institute for Economic Forecasting

MSCA Digital Finance

Why Data Matters

- Financial markets generate massive amounts of data every second.
- Understanding how to process, analyze, and interpret this data is key to making informed investment decisions.
- Data quality and selection directly impact model accuracy.

- Understanding the Past
 - ▶ Historical data helps us understand past market behavior.
 - ▶ Identify trends.
 - ▶ Analyze how markets have reacted to different events.
- Predicting the Future
 - ▶ Analyzing historical data can provide insights into potential future market movements.
 - ▶ Inform investment decisions.

Why Data Matters (continued)

- Understanding price data
 - ▶ OHLC (Open, High, Low, Close) prices
 - ▶ Adjusted close prices
- Returns
 - ▶ Simple returns
 - ▶ Log-returns
 - ▶ Cumulative returns
- Volatility estimation
 - ▶ Historical volatility
 - ▶ Range-based estimators (Parkinson, Garman-Klass, Rogers-Satchell, Yang-Zhang)
- Modern data challenges
 - ▶ High-frequency data
 - ▶ Alternative data sources

Examples of Data Usage

- Stock Trading
 - ▶ Price patterns, momentum strategies
 - ▶ Traders use historical price and volume data to identify patterns and trends in the market.
- Risk Management
 - ▶ Value-at-Risk (VaR), stress testing
 - ▶ Data is essential for quantifying and managing financial risks, such as market volatility, credit risk, and liquidity risk.
- Portfolio Construction
 - ▶ Correlation analysis, mean-variance optimization
 - ▶ Quantitative analysts (quants) develop trading strategies and manage risk.
- Algorithmic Trading
 - ▶ Signals from statistical models
 - ▶ Automated trading systems rely on real-time market data to execute trades at high speed.

Data Challenges

- ❑ Missing Data and Outliers
 - ▶ Financial data can be noisy, incomplete, or inaccurate.
 - ▶ Outliers may distort statistical estimates if not handled properly.
- ❑ Survivorship Bias
 - ▶ Databases often include only currently active securities, excluding delisted or bankrupt firms.
 - ▶ This leads to overly optimistic performance estimates.
- ❑ Look-Ahead Bias
 - ▶ Using information that was not available at the time of the decision.
 - ▶ Particularly dangerous in backtesting trading strategies.
- ❑ Non-Stationarity
 - ▶ Financial time series often exhibit time-varying statistical properties.
 - ▶ Structural breaks, regime changes, and evolving market microstructure.
- ❑ High-Frequency Noise
 - ▶ Microstructure noise in tick-by-tick data.
 - ▶ The sheer volume of financial data can be overwhelming.

Outline

1. Motivation ✓
2. **Price Data**
3. Returns
4. Volatility Estimators
5. Risk-Adjusted Returns
6. Conclusions
7. References

OHLC Data

- **Open Price O_t**
 - ▶ The first traded price at the beginning of the time period t .
 - ▶ It reflects the initial market sentiment and can be influenced by news events occurring after the previous close.
- **High Price H_t**
 - ▶ The highest traded price during period t .
- **Low Price L_t**
 - ▶ The lowest traded price during period t .
- **Close Price P_t**
 - ▶ The last traded price at the end of period t .
 - ▶ It is often considered the most significant price because it reflects the final consensus of value for that period.

These four prices are the basic building blocks of financial data analysis.

Adjusted Close Price

- P_t^{adj} : accounts for corporate actions such as dividends, stock splits, and rights offerings that affect a stock's price but not its value.
- Ensures that historical price data reflects the true economic value of the stock over time, allowing for accurate performance analysis and comparison.

Individual Adjustment Factors:

- **Stock Splits and Reverse Splits:** $F_i = \text{Split Ratio}_i$
- **Stock Dividends:** For a stock dividend of $d_i\%$: $F_i = 1 + d_i$
- **Cash Dividends:** $f_{div} = 1 - \frac{D}{P_{close}}$
where D_i is the cash dividend per share and P_{i-1} is the close price before the dividend is paid.
- **Rights Offerings:** Shareholders can buy n new shares at a subscription price P_{rights} for every m shares held:

$$F_i = \frac{P_{i-1} + \left(\frac{n}{m} \times P_{rights}\right)}{P_{i-1}}$$

Adjusted Price: $P_t^{adj} = P_t \times \prod_j f_j$

Simple Returns

- Return measures the gain or loss of an investment over a period, expressed as a percentage of the investment's initial cost.

Definition 1 (Simple Return)

The **simple return** (or arithmetic return) of an asset over one period is defined as:

$$R_t = \frac{P_t - P_{t-1}}{P_{t-1}} = \frac{P_t}{P_{t-1}} - 1$$

where P_t is the price at time t .

- Simple returns are intuitive and directly represent the percentage change in price.
- However, they are **not time-additive**: the multi-period simple return is a product, not a sum.
- Portfolio return: $R_p = \sum_{i=1}^n w_i R_i$ (simple returns aggregate across assets).

Log>Returns

Definition 2 (Log-Return)

The **log-return** (or continuously compounded return) is defined as:

$$r_t = \ln \left(\frac{P_t}{P_{t-1}} \right) = \ln P_t - \ln P_{t-1} = \ln(1 + R_t)$$

Time-additivity property:

$$r_t(k) = r_t + r_{t-1} + \cdots + r_{t-k+1}$$

- For small changes in price: $r_t \approx R_t$.
- Under certain conditions, log-returns are more likely to follow a normal distribution.
- Log-returns simplify the mathematical handling of compounded returns.
- Enable more convenient statistical analysis, especially in the context of portfolio management and risk analysis.

Cumulative Returns

- Cumulative returns represent the total change in the value of an investment over a period.
- Can be calculated using either simple returns or log-returns.

Cumulative Simple Return (CSR):

$$\text{CSR}_T = \prod_{t=1}^T (1 + R_t) - 1$$

Cumulative Log-Return (CLR):

$$\text{CLR}_T = \sum_{t=1}^T r_t$$

- The relationship between CSR and CLR: $\text{CSR}_T = e^{\text{CLR}_T} - 1$.
- CLR is simply the sum of individual log-returns (time-additivity).
- CSR requires multiplicative compounding of $(1 + R_t)$ terms.

Volatility Overview

- ▣ Volatility is a statistical measure of the dispersion of returns for a given security or market index.
- ▣ It represents the degree of variation of an asset's price over time.
- ▣ Several estimators have been developed to capture the complexities of financial market volatility.

Close-to-Close Estimators:

- ▣ Historical Volatility

Range-Based Estimators:

- ▣ Parkinson's Estimator
- ▣ Garman-Klass Estimator
- ▣ Rogers-Satchell Estimator
- ▣ Yang-Zhang Estimator

Model-Based Estimators:

- ▣ Realized Volatility
- ▣ GARCH Models
- ▣ HEAVY Models
- ▣ Stochastic Volatility Models with Jumps
- ▣ Implied Volatility

Historical Volatility

- ▣ Measures the dispersion or variability of returns over a specified period in the past.
- ▣ Uses closing prices only.

Definition 3 (Historical Volatility)

Historical volatility is calculated as the standard deviation of past returns:

$$\hat{\sigma} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (r_i - \bar{r})^2}$$

where r_i are the log-returns and $\bar{r}$ is the sample mean of returns.

- ▣ Limitations of Historical Volatility
 - ▶ Past performance may not predict future volatility.
 - ▶ Does not account for shocks or market regime changes.
 - ▶ Ignores intra-day price information (High, Low, Open).

Square Root of Time Rule

- Fundamental concept in financial risk management and quantitative finance.
- Provides a method for scaling volatility (or risk) over different time horizons, assuming that returns are IID and follow a normal distribution.

Suppose we have an asset with returns $r_t \sim \mathcal{N}(\mu, \sigma^2)$ i.i.d. The total return over n periods is:

$$R_n = r_1 + r_2 + \dots + r_n$$

Since the returns are IID: $\mathbb{E}[R_n] = n\mu$ and $\text{Var}(R_n) = n\sigma^2$.

The standard deviation (volatility) over n periods is therefore:

$$\sigma_n = \sqrt{\text{Var}(R_n)} = \sqrt{n} \sigma$$

Annualized volatility (Bachelier):

$$\sigma_{\text{annual}} = \sigma_{\text{daily}} \times \sqrt{252}$$

where 252 is the typical number of trading days in a year.

Parkinson's Estimator

- ▣ Traditional estimators often rely solely on closing prices, potentially overlooking valuable information contained within the trading period.
- ▣ Parkinson (1980) introduced an estimator that utilizes the high and low prices within a period, offering a more efficient volatility estimate under specific assumptions.
- ▣ Assumptions: continuous trading, zero drift in the price process.

Definition 4 (Parkinson's Estimator)

$$\hat{\sigma}_P^2 = \frac{1}{4n \ln 2} \sum_{i=1}^n (\ln H_i - \ln L_i)^2$$

where H_i is the highest price on day i , L_i is the lowest price on day i , and n is the number of periods.

- ▣ Uses high-low range, more efficient than close-to-close estimators.
- ▣ Captures intra-period price variability more effectively.

Garman-Klass Estimator

- Garman and Klass (1980) extended the Parkinson estimator by incorporating open and close prices in addition to high and low prices.

Definition 5 (Garman-Klass Estimator)

$$\hat{\sigma}_{GK}^2 = \frac{1}{n} \sum_{i=1}^n \left[\frac{1}{2} (h_i - l_i)^2 - (2 \ln 2 - 1) c_i^2 \right]$$

where:

- $h_i = \ln H_i - \ln O_i$ (log high-to-open)
- $l_i = \ln L_i - \ln O_i$ (log low-to-open)
- $c_i = \ln C_i - \ln O_i$ (log close-to-open)

Advantages:

- Most efficient estimator under GBM

Limitations:

- Assumes zero drift
- Assumes continuous trading (no gaps)

Rogers-Satchell Estimator

- Rogers and Satchell (1991) developed an estimator that allows for non-zero drift in the price process, addressing a key limitation of Parkinson and Garman-Klass estimators.

Definition 6 (Rogers-Satchell Estimator)

$$\hat{\sigma}_{RS}^2 = \frac{1}{n} \sum_{i=1}^n [h_i(h_i - c_i) + l_i(l_i - c_i)]$$

where:

- $h_i = \ln H_i - \ln O_i$
 - $l_i = \ln L_i - \ln O_i$
 - $c_i = \ln C_i - \ln O_i$
-
- Allows for non-zero drift (trending markets).
 - Uses all four OHLC prices.
 - More robust than Parkinson and Garman-Klass in trending markets.
 - Still assumes continuous trading (no overnight jumps).

Yang-Zhang Estimator

- Yang and Zhang (2000) proposed the most comprehensive range-based volatility estimator, combining overnight volatility, open-to-close volatility, and the Rogers-Satchell estimator.

Definition 7 (Yang-Zhang Estimator)

$$\hat{\sigma}_{YZ}^2 = \hat{\sigma}_O^2 + k \hat{\sigma}_C^2 + (1 - k) \hat{\sigma}_{RS}^2$$

where:

- $\hat{\sigma}_O^2$: overnight (close-to-open) variance
- $\hat{\sigma}_C^2$: open-to-close variance
- $\hat{\sigma}_{RS}^2$: Rogers-Satchell variance
- $k = \frac{0.34}{1.34 + \frac{n+1}{n-1}}$

- Most comprehensive range-based estimator.

Statistics of Financial Markets

- Handles both drift and opening jumps.
- Minimum variance estimator among range based methods



Sharpe Ratio

- The Sharpe Ratio is the most widely used measure of risk-adjusted return.
- Developed by William F. Sharpe (1966), Nobel Laureate in Economics.

Definition 8 (Sharpe Ratio)

$$SR = \frac{E[R_p] - R_f}{\sigma_p}$$

where:

- $E[R_p]$: expected return of the portfolio
 - R_f : risk-free rate
 - σ_p : standard deviation of the portfolio's excess return
-
- Measures the excess return per unit of risk.
 - Higher Sharpe Ratio indicates better risk-adjusted performance.
 - Assumes returns are normally distributed.
 - Widely used for comparing funds, strategies, and portfolios.

Interpreting the Sharpe Ratio

Sharpe Ratio	Rating	Interpretation
$SR < 0$	Bad	Returns below risk-free rate
$0 \leq SR < 0.5$	Below Average	Insufficient compensation for risk
$0.5 \leq SR < 1.0$	Good	Adequate risk-adjusted return
$1.0 \leq SR < 2.0$	Very Good	Strong risk-adjusted performance
$SR \geq 2.0$	Excellent	Exceptional (rare in practice)

- Most hedge funds and mutual funds have Sharpe Ratios between 0.5 and 1.5.
- A Sharpe Ratio above 2.0 is extremely rare over long time horizons.

Mathematical Framework

Portfolio Expected Return:

For a portfolio with n assets and weights $w_1, w_2, \dots, w_n$:

$$E[R_p] = \sum_{i=1}^n w_i E[R_i]$$

where:

- ▣ w_i : weight of asset i in the portfolio (with $\sum_{i=1}^n w_i = 1$)
- ▣ $E[R_i]$: expected return of asset i

- ▣ The portfolio return is a linear combination of individual asset returns.
- ▣ Weights can be positive (long positions) or negative (short positions).
- ▣ In practice, expected returns are estimated from historical data or forward-looking models.

Portfolio Volatility

Portfolio Variance (scalar form):

$$\sigma_p^2 = \sum_{i=1}^n \sum_{j=1}^n w_i w_j \sigma_i \sigma_j \rho_{ij}$$

where:

- σ_i : standard deviation of asset i
- ρ_{ij} : correlation between assets i and j

Portfolio Volatility:

$$\sigma_p = \sqrt{\sum_{i=1}^n \sum_{j=1}^n w_i w_j \sigma_i \sigma_j \rho_{ij}}$$

Matrix notation:

$$\sigma_p = \sqrt{\mathbf{w}^T \Sigma \mathbf{w}}$$

where $\mathbf{w}$ is the vector of portfolio weights and Σ is the covariance matrix.

- Diversification reduces portfolio volatility when $\rho_{ij} < 1$.

Portfolio Optimization

Efficient Frontier (Markowitz, 1952):

- Maximize return for given risk:

$$\max_w w^T \mu \quad \text{s.t.} \quad w^T \Sigma w = \sigma_0^2, \quad w^T \mathbf{1} = 1$$

- Minimize risk for given return:

$$\min_w w^T \Sigma w \quad \text{s.t.} \quad w^T \mu = \mu_0, \quad w^T \mathbf{1} = 1$$

- The efficient frontier represents the set of portfolios offering the highest expected return for each level of risk.
- Portfolios below the efficient frontier are suboptimal (dominated).
- The minimum variance portfolio lies at the leftmost point of the frontier.
- The tangency portfolio maximizes the Sharpe Ratio.

Efficient Frontier – Example

Asset Universe: SPY, QQQ, GLD, TLT, VNQ

Asset	Ann. Return (%)	Ann. Volatility (%)	Sharpe Ratio
SPY	14.33	19.35	0.6370
QQQ	19.57	24.04	0.7309
GLD	11.62	14.31	0.6725
TLT	-0.37	16.16	-0.1469
VNQ	8.23	22.75	0.2738

- Data period: 2018-01-01 to 2025-03-04 (1801 trading days).
- 10,000 random portfolios generated to illustrate the efficient frontier.
- QQQ has the highest return but also the highest volatility.
- GLD offers the best diversification benefit (lowest volatility, decent Sharpe).
- TLT has a negative Sharpe Ratio, reflecting the bond bear market.
- The optimal portfolio (maximum Sharpe) typically overweights QQQ and GLD.

Summary

- ▣ **Price Data:** OHLC prices and adjusted close prices are the fundamental building blocks of financial data analysis. Adjusted prices account for corporate actions (splits, dividends, rights offerings).
- ▣ **Returns:** Simple returns are intuitive for portfolio aggregation; log-returns offer time-additivity and mathematical convenience. Cumulative returns track total investment performance.
- ▣ **Volatility Estimators:** Range-based estimators (Parkinson, Garman-Klass, Rogers-Satchell, Yang-Zhang) are more efficient than close-to-close historical volatility by utilizing intra-day price information.
- ▣ **Risk-Adjusted Returns:** The Sharpe Ratio provides a standardized measure for comparing risk-adjusted performance across assets and portfolios.
- ▣ **Portfolio Optimization:** The efficient frontier identifies optimal risk-return trade-offs using mean-variance optimization (Markowitz framework).
- ▣ **Data Challenges:** Survivorship bias, look-ahead bias, non-stationarity, and high-frequency noise must be carefully addressed in empirical analysis.

References



Parkinson, M. (1980). The Extreme Value Method for Estimating the Variance of the Rate of Return. *Journal of Business*, 53(1), 61–65.



Garman, M.B. and Klass, M.J. (1980). On the Estimation of Security Price Volatilities from Historical Data. *Journal of Business*, 53(1), 67–78.



Rogers, L.C.G. and Satchell, S.E. (1991). Estimating Variance from High, Low and Closing Prices. *The Annals of Applied Probability*, 1(4), 504–512.



Yang, D. and Zhang, Q. (2000). Drift-Independent Volatility Estimation Based on High, Low, Open, and Close Prices. *Journal of Business*, 73(3), 477–492.



Sharpe, W.F. (1966). Mutual Fund Performance. *Journal of Business*, 39(1), 119–138.



Markowitz, H. (1952). Portfolio Selection. *The Journal of Finance*, 7(1), 77–91.



Bachelier, L. (1900). Théorie de la spéculation. *Annales Scientifiques de l'École Normale Supérieure*, 17, 21–86.



Fiszeder, P. and Perczak, G. (2013). A New Look at Variance Estimation Based on Low, High and Closing Prices. *Statistica Neerlandica*, 67(4), 456–481.