1 Introduction

This book is about the analysis of financial markets data. After this brief introductory chapter, we turn immediately in Chaps. 2 and 3 to the sources of the data, returns on equities and prices and yields on bonds. Chapter 4 develops methods for informal, often graphical, analysis of data. More formal methods based on statistical inference, that is, estimation and testing, are introduced in Chap. 5. The chapters that follow Chap. 5 cover a variety of more advanced statistical techniques: ARIMA models, regression, multivariate models, copulas, GARCH models, factor models, cointegration, Bayesian statistics, and nonparametric regression.

Much of finance is concerned with financial risk. The *return* on an investment is its revenue expressed as a fraction of the initial investment. If one invests at time t_1 in an asset with price P_{t_1} and the price later at time t_2 is P_{t_2}, then the net return for the holding period from t_1 to t_2 is $(P_{t_2} - P_{t_1})/P_{t_1}$. For most assets, future returns cannot be known exactly and therefore are random variables. *Risk* means uncertainty in future returns from an investment, in particular, that the investment could earn less than the expected return and even result in a loss, that is, a negative return. Risk is often measured by the standard deviation of the return, which we also call the volatility. Recently there has been a trend toward measuring risk by value-at-risk (VaR) and expected shortfall (ES). These focus on large losses and are more direct indications of financial risk than the standard deviation of the return. Because risk depends upon the probability distribution of a return, probability and statistics are fundamental tools for finance. Probability is needed for risk calculations, and statistics is needed to estimate parameters such as the standard deviation of a return or to test hypotheses such as the so-called random walk hypothesis which states that future returns are independent of the past.

D. Ruppert, D.S. Matteson, *Statistics and Data Analysis for Financial Engineering*, Springer Texts in Statistics,
DOI 10.1007/978-1-4939-2614-5_1

In financial engineering there are two kinds of probability distributions that can be estimated. Objective probabilities are the true probabilities of events. Risk-neutral or pricing probabilities give model outputs that agree with market prices and reflect the market's beliefs about the probabilities of future events. The statistical techniques in this book can be used to estimate both types of probabilities. Objective probabilities are usually estimated from historical data, whereas risk-neutral probabilities are estimated from the prices of options and other financial instruments.

Finance makes extensive use of probability models, for example, those used to derive the famous Black–Scholes formula. Use of these models raises important questions of a statistical nature such as: Are these models supported by financial markets data? How are the parameters in these models estimated? Can the models be simplified or, conversely, should they be elaborated?

After Chaps. 4–8 develop a foundation in probability, statistics, and exploratory data analysis, Chaps. 12 and 13 look at ARIMA models for time series. Time series are sequences of data sampled over time, so much of the data from financial markets are time series. ARIMA models are stochastic processes, that is, probability models for sequences of random variables. In Chap. 16 we study optimal portfolios of risky assets (e.g., stocks) and of risky assets and risk-free assets (e.g., short-term U.S. Treasury bills). Chapters 9–11 cover one of the most important areas of applied statistics, regression. Chapter 15 introduces cointegration analysis. In Chap. 17 portfolio theory and regression are applied to the CAPM. Chapter 18 introduces factor models, which generalize the CAPM. Chapters 14–21 cover other areas of statistics and finance such as GARCH models of nonconstant volatility, Bayesian statistics, risk management, and nonparametric regression.

Several related themes will be emphasized in this book:

Always look at the data According to a famous philosopher and baseball player, Yogi Berra, "You can see a lot by just looking." This is certainly true in statistics. The first step in data analysis should be plotting the data in several ways. Graphical analysis is emphasized in Chap. 4 and used throughout the book. Problems such as bad data, outliers, mislabeling of variables, missing data, and an unsuitable model can often be detected by visual inspection. *Bad data* refers to data that are outlying because of errors, e.g., recording errors. Bad data should be corrected when possible and otherwise deleted. Outliers due, for example, to a stock market crash are "good data" and should be retained, though the model may need to be expanded to accommodate them. It is important to detect both bad data and outliers, and to understand which is which, so that appropriate action can be taken.

All models are false Many statisticians are familiar with the observation of George Box that "all models are false but some models are useful." This fact should be kept in mind whenever one wonders whether a statistical, economic, or financial model is “true.” Only computer-simulated data have a “true model.” No model can be as complex as the real world, and even if such a model did exist, it would be too complex to be useful.

Bias-variance tradeoff If useful models exist, how do we find them? The answer to this question depends ultimately on the intended uses of the model. One very useful principle is *parsimony* of parameters, which means that we should use only as many parameters as necessary. Complex models with unnecessary parameters increase estimation error and make interpretation of the model more difficult. However, a model that is too simple will not capture important features of the data and will lead to serious biases. Simple models have large biases but small variances of the estimators. Complex models have small biases but large variances. Therefore, model choice involves finding a good tradeoff between bias and variance.

Uncertainty analysis It is essential that the uncertainty due to estimation and modeling errors be quantified. For example, portfolio optimization methods that assume that return means, variances, and correlations are known exactly are suboptimal when these parameters are only estimated (as is always the case). Taking uncertainty into account leads to other techniques for portfolio selection—see Chap. 16. With complex models, uncertainty analysis could be challenging in the past, but no longer is so because of modern statistical techniques such as resampling (Chap. 6) and Bayesian MCMC (Chap. 20).

Financial markets data are not normally distributed Introductory statistics textbooks model continuously distributed data with the normal distribution. This is fine in many domains of application where data are well approximated by a normal distribution. However, in finance, stock returns, changes in interest rates, changes in foreign exchange rates, and other data of interest have many more outliers than would occur under normality. For modeling financial markets data, heavy-tailed distributions such as the t-distributions are much more suitable than normal distributions—see Chap. 5. *Remember:* In finance, the normal distribution is not normal.

Variances are not constant Introductory textbooks also assume constant variability. This is another assumption that is rarely true for financial markets data. For example, the daily return on the market on Black Monday, October 19, 1987, was -23%, that is, the market lost 23% of its value in a single day! A return of this magnitude is virtually impossible under a normal model with a constant variance, and it is still quite unlikely under a t-distribution with constant variance, but much more likely under a t-distribution model with conditional heteroskedasticity, e.g., a GARCH model (Chap. 14).

1.1 Bibliographic Notes

The dictum that “All models are false but some models are useful” is from Box (1976).

References

Box, G. E. P. (1976) Science and statistics, *Journal of the American Statistical Association*, 71, 791–799.