CHAPTER 1

Introduction

It is no surprise that the natural sciences (chemistry, physics, life sciences/biology, astronomy, earth science, and environmental science) and engineering are fields that rely on advanced quantitative methods. One of the toolsets used by professionals in these fields is from the branch of mathematics known as probability and statistics. The social sciences, such as psychology, sociology, political science, and economics, use probability and statistics to varying degrees.

There are branches within each field of the natural sciences and social sciences that utilize probability and statistics more than others. Specialists in these areas not only apply the tools of probability and statistics, but they have also contributed to the field of statistics by developing techniques to organize, analyze, and test data. Let's look at examples from physics and engineering (the study of natural phenomena in terms of basic laws and physical quantities and the design of physical artifacts) and biology (the study of living organisms) in the natural sciences, and psychology (the study of the human mind) and economics (the study of production, resource allocation, and consumption of goods and services) in the social sciences.

Statistical physics is the branch of physics that applies probability and statistics for handling problems involving large populations of particles. One of the first areas of application was the explanation of thermodynamics laws in terms of statistical mechanics. It was an extraordinary scientific achievement with far-reaching consequences. In the field of *engineering*, the analysis of risk, be it natural or industrial, is another area that makes use of statistical methods. This discipline has contributed important innovations especially in the study of rare extreme events. The engineering of electronic communications applied statistical methods early, contributing to the development of fields such as queue theory (used in communication switching systems) and introduced the fundamental innovation of measuring information.

Biostatistics and *biomathematics* within the field of biology include many areas of great scientific interest such as public health, epidemiology, demography, and genetics, in addition to designing biological experiments

(such as clinical experiments in medicine) and analyzing the results of those experiments. The study of the dynamics of populations and the study of evolutionary phenomena are two important fields in biomathematics. *Biometry* and *biometrics* apply statistical methods to identify quantities that characterize living objects.

Psychometrics, a branch of psychology, is concerned with designing tests and analyzing the results of those tests in an attempt to measure or quantify some human characteristic. Psychometrics has its origins in personality testing, intelligence testing, and vocational testing, but is now applied to measuring attitudes and beliefs and health-related tests.

Econometrics is the branch of economics that draws heavily on statistics for testing and analyzing economic relationships. Within econometrics, there are theoretical econometricians who analyze statistical properties of estimators of models. Several recipients of the Nobel Prize in Economic Sciences received the award as a result of their lifetime contribution to this branch of economics. To appreciate the importance of econometrics to the discipline of economics, when the first Nobel Prize in Economic Sciences was awarded in 1969, the corecipients were two econometricians, Jan Tinbergen and Ragnar Frisch (who is credited for first using the term econometrics in the sense that it is known today). Further specialization within econometrics, and the area that directly relates to this book, is *financial econometrics*. As Jianqing Fan (2004) writes, financial econometrics

> uses statistical techniques and economic theory to address a variety of problems from finance. These include building financial models, estimation and inferences of financial models, volatility estimation, risk management, testing financial economics theory, capital asset pricing, derivative pricing, portfolio allocation, risk-adjusted returns, simulating financial systems, hedging strategies, among others.

Robert Engle and Clive Granger, two econometricians who shared the 2003 Nobel Prize in Economics Sciences, have contributed greatly to the field of financial econometrics.

Historically, the core probability and statistics course offered at the university level to undergraduates has covered the fundamental principles and applied these principles across a wide variety of fields in the natural sciences and social sciences. Universities typically offered specialized courses within these fields to accommodate students who sought more focused applications. The exceptions were the schools of business administration that early on provided a course in probability and statistics with applications to business decision making. The applications cut across finance, marketing, management, and accounting. However, today, each of these areas in busi-

ness requires specialized tools for dealing with real-world problems in their respective disciplines.

This brings us to the focus of this book. Finance is an area that relies heavily on probability and statistics. The quotation above by Jianqing Fan basically covers the wide range of applications within finance and identifies some of the unique applications. Two examples may help make this clear. First, in standard books on statistics, there is coverage of what one might refer to as "probability distributions with appealing properties." A distribution called the "normal distribution," referred to in the popular press as a "bell-shaped curve," is an example. Considerable space is devoted to this distribution and its application in standard textbooks. Yet, the overwhelming historical evidence suggests that real-world financial data commonly used in financial applications are not normally distributed. Instead, more focus should be on distributions that deal with extreme events, or, in other words, what are known as the "tails" of a distribution. In fact, many market commentators and regulators view the failure of financial institutions and major players in the financial markets to understand non-normal distributions as a major reason for the recent financial debacles throughout the world. This is one of the reasons that, in certain areas in finance, extreme event distributions (which draw from extreme value theory) have supplanted the normal distribution as the focus of attention. The recent financial crisis has clearly demonstrated that because of the highly leveraged position (i.e., large amount of borrowing relative to the value of equity) of financial institutions throughout the world, these entities are very sensitive to extreme events. This means that the management of these financial institutions must be aware of the nature of the tails of distributions, that is, the probability associated with extreme events.

As a second example, the statistical measure of correlation that measures a certain type of association between two random variables may make sense when the two random variables are normally distributed. However, correlation may be inadequate in describing the link between two random variables when a portfolio manager or risk manager is concerned with extreme events that can have disastrous outcomes for a portfolio or a financial institution. Typically models that are correlation based will underestimate the likelihood of extreme events occurring simultaneously. Alternative statistical measures that would be more helpful, the copula measure and the tail dependence, are typically not discussed in probability and statistics books.

It is safe to say that the global financial system has been transformed since the mid-1970s due to the development of models that can be used to value derivative instruments. Complex derivative instruments such as options, caps, floors, and swaptions can only be valued (i.e., priced) using tools from probability and statistical theory. While the model for such pric-

ing was first developed by Black and Scholes (1976) and known as the Black-Scholes option pricing model, it relies on models that can be traced back to the mathematician Louis Bachelier (1900).

In the remainder of this introductory chapter, we do two things. First, we briefly distinguish between the study of probability and the study of statistics. Second, we provide a roadmap for the chapters to follow in this book.

PROBABILITY VS. STATISTICS

Thus far, we have used the terms "probability" and "statistics" collectively as if they were one subject. There is a difference between the two that we distinguish here and which will become clearer in the chapters to follow.

Probability models are theoretical models of the occurrence of uncertain events. At the most basic level, in probability, the properties of certain types of probabilistic models are examined. In doing so, it is assumed that all parameter values that are needed in the probabilistic model are known. Let's contrast this with statistics. Statistics is about empirical data and can be broadly defined as a set of methods used to make inferences from a known sample to a larger population that is in general unknown. In finance and economics, a particular important example is making inferences from the past (the known sample) to the future (the unknown population). In statistics. we apply probabilistic models and we use data and eventually judgment to estimate the parameters of these models. We do not assume that all parameter values in the model are known. Instead, we use the data for the variables in the model to estimate the value of the parameters and then to test hypotheses or make inferences about their estimated values.

Another way of thinking about the study of probability and the study of statistics is as follows. In studying probability, we follow much the same routine as in the study of other fields of mathematics. For example, in a course in calculus, we prove theorems (such as the fundamental theory of calculus that specifies the relationship between differentiation and integration), perform calculations given some function (such as the first derivative of a function), and make conclusions about the characteristics of some mathematical function (such as whether the function may have a minimum or maximum value). In the study of probability, there are also theorems to be proven (although we do not focus on proofs in this book), we perform calculations based on probability models, and we reach conclusions based on some assumed probability distribution. In deriving proofs in calculus or probability theory, deductive reasoning is utilized. For this reason, probability can be considered as a fundamental discipline in the field of mathematics, just as we would view algebra, geometry, and trigonometry. In contrast,

statistics is based on inductive reasoning. More specifically, given a sample of data (i.e., observations), we make generalized probabilistic conclusions about the population from which the data are drawn or the process that generated the data.

OVERVIEW OF THE BOOK

The 21 chapters that follow in this book are divided into four parts covering descriptive statistics, probability theory, inductive statistics, and multivariate linear regression.

Part One: Descriptive Statistics

The six chapters in Part One cover descriptive statistics. This topic covers the different tasks of gathering data and presenting them in a more concise yet as informative as possible way. For example, a set of 1,000 observations may contain too much information for decision-making purposes. Hence, we need to reduce this amount in a reasonable and systematic way.

The initial task of any further analysis is to gather the data. This process is explained in Chapter 2. It provides one of the most essential—if not the most essential—assignment. Here, we have to be exactly aware of the intention of our analysis and determine the data type accordingly. For example, if we wish to analyze the contributions of the individual divisions of a company to the overall rate of return earned by the company, we need a completely different sort of data than when we decompose the risk of some investment portfolio into individual risk factors, or when we intend to gain knowledge of unknown quantities in general economic models. As part of the process of retrieving the essential information contained in the data, we describe the methods of presenting the distribution of the data in comprehensive ways. This can be done for the data itself or, in some cases, it will be more effective after the data have been classified.

In Chapter 3, methodologies for reducing the data to a few representative quantities are presented. We refer to these representative quantities as statistics. They will help us in assessing where certain parts of the data are positioned as well as how the data disperse relative to particular positions. Different data sets are commonly compared based on these statistics that, in most cases, proves to be very efficient.

Often, it is very appealing and intuitive to present the features of certain data in charts and figures. In Chapter 4, we explain the particular graphical tools suitable for the different data types discussed in Chapter 2. In general, a graphic uses the distributions introduced in Chapter 2 or the statistics

from Chapter 3. By comparing graphics, it is usually a simple task to detect similarities or differences among different data sets.

In Chapters 2, 3, and 4, the analysis focuses only on one quantity of interest and in such cases we say that we are looking at univariate (i.e., one variable) distributions. In Chapter 5, we introduce multivariate distributions; that is, we look at several variables of interest simultaneously. For example, portfolio analysis relies on multivariate analysis. Risk management in general considers the interaction of several variables and the influence that each variable exerts on the others. Most of the aspects from the one-dimensional analysis (i.e., analysis of univariate distributions) can be easily extended to higher dimensions while concepts such as dependence between variables are completely new. In this context, in Chapter 6 we put forward measures to express the degree of dependence between variables such as the covariance and correlation. Moreover, we introduce the conditional distribution, a particular form of distribution of the variables given that some particular variables are fixed. For example, we may look at the average return of a stock portfolio given that the returns of its constituent stocks fall below some threshold over a particular investment horizon.

When we assume that a variable is dependent on some other variable, and the dependence is such that a movement in the one variable causes a known constant shift in the other, we model the set of possible values that they might jointly assume by some straight line. This statistical tool, which is the subject of Chapter 6, is called a *linear regression*. We will present measures of goodness-of-fit to assess the quality of the estimated regression. A popular application is the regression of some stock return on the return of a broad-based market index such as the Standard and Poor's 500. Our focus in Chapter 6 is on the univariate regression, also referred to as a *simple linear regression*. This means that there is one variable (the independent variable) that is assumed to affect some variable (the dependent variable). Part Four of this book is devoted to extending the bivariate regression model to the multivariate case where there is more than one independent variable.

An extension of the regression model to the case where the data set in the analysis is a time series is described in Chapter 7. In time series analysis we observe the value of a particular variable over some period of time. We assume that at each point in time, the value of the variable can be decomposed into several components representing, for example, seasonality and trend. Instead of the variable itself, we can alternatively look at the changes between successive observations to obtain the related difference equations. In time series analysis we encounter the notion of noise in observations. A well-known example is the so-called random walk as a model of a stock price process. In Chapter 7, we will also present the error correction model for stock prices.

Part Two: Basic Probability Theory

The basics of probability theory are covered in the nine chapters of Part Two. In Chapter 8, we briefly treat the historical evolution of probability theory and its main concepts. To do so, it is essential that mathematical set operations are introduced. We then describe the notions of outcomes, events, and probability distributions. Moreover, we distinguish between countable and uncountable sets. It is in this chapter, the concept of a random variable is defined. The concept of random variables and their probability distributions are essential in models in finance where, for example, stock returns are modeled as random variables. By giving the associated probability distribution, the random behavior of a stock's return will then be completely specified.

Discrete random variables are introduced in Chapter 9 where some of their parameters such as the mean and variance are defined. Very often we will see that the intuition behind some of the theory is derived from the variables of descriptive statistics. In contrast to descriptive statistics, the parameters of random variables no longer vary from sample to sample but remain constant for all drawings. We conclude Chapter 9 with a discussion of the most commonly used discrete probability distributions: binomial, hypergeometric, multinomial, Poisson, and discrete uniform. Discrete random variables are applied in finance whenever the outcomes to be modeled consist of integer numbers such as the number of bonds or loans in a portfolio that might default within a certain period of time or the number of bankruptcies over some period of time.

In Chapter 10, we introduce the other type of random variables, continuous random variables and their distributions including some location and scale parameters. In contrast to discrete random variables, for continuous random variables any countable set of outcomes has zero probability. Only entire intervals (i.e., uncountable sets) can have positive probability. To construct the probability distribution function, we need the probability density functions (or simply density functions) typical of continuous random variables. For each continuous random variable, the density function is uniquely defined as the marginal rate of probability for any single outcome. While we hardly observe true continuous random variables in finance, they often serve as an approximation to discretely distributed ones. For example, financial derivatives such as call options on stocks depend in a completely known fashion on the prices of some underlying random variable such as the underlying stock price. Even though the underlying prices are discrete, the theoretical derivative pricing models rely on continuous probability distributions as an approximation.

Some of the most well-known continuous probability distributions are presented in Chapter 11. Probably the most popular one of them is the nor-

mal distribution. Its popularity is justified for several reasons. First, under certain conditions, it represents the limit distribution of sums of random variables. Second, it has mathematical properties that make its use appealing. So, it should not be a surprise that a vast variety of models in finance are based on the normal distribution. For example, three central theoretical models in finance—the Capital Asset Pricing Model, Markowitz portfolio selection theory, and the Black-Scholes option pricing model—rely upon it. In Chapter 11, we also introduce many other distributions that owe their motivation to the normal distribution. Additionally, other continuous distributions in this chapter (such as the exponential distribution) that are important by themselves without being related to the normal distribution are discussed. In general, the continuous distributions presented in this chapter exhibit pleasant features that act strongly in favor of their use and, hence, explain their popularity with financial model designers even though their use may not always be justified when comparing them to real-world data.

Despite the use of the widespread use of the normal distribution in finance, it has become a widely accepted hypothesis that financial asset returns exhibit features that are not in agreement with the normal distribution. These features include the properties of asymmetry (i.e., skewness), excess kurtosis, and heavy tails. Understanding skewness and heavy tails is important in dealing with risk. The skewness of the distribution of say the profit and loss of a bank's trading desk, for example, may indicate that the downside risk is considerably greater than the upside potential. The tails of a probability distribution indicate the likelihood of extreme events. If adverse extreme events are more likely than what would be predicted by the normal distribution, then a distribution is said to have a heavy (or fat) tail. Relying on the normal distribution to predict such unfavorable outcomes will underestimate the true risk. For this reason, in Chapter 12 we present a collection of continuous distributions capable of modeling asymmetry and heavy tails. Their parameterization is not quite easily accessible to intuition at first. But, in general, each of the parameters of some distribution has a particular meaning with respect to location and overall shape of the distribution. For example, the Pareto distribution that is described in Chapter 12 has a tail parameter governing the rate of decay of the distribution in the extreme parts (i.e., the tails of the distribution).

The distributions we present in Chapter 12 are the generalized extreme value distributions, the log-normal distribution, the generalized Pareto distribution, the normal inverse Gaussian distribution, and the α-stable (or alpha-stable) distribution. All of these distributions are rarely discussed in introductory statistics books nor covered thoroughly in finance books. However, as the overwhelming empirical evidence suggests, especially during volatile periods, the commonly used normal distribution is unsuitable for modeling

financial asset returns. The α-stable distributions, a more general class of limiting distributions than the normal distribution, qualifies as a candidate for modeling stock returns in very volatile market environments such as during a financial crisis. As we will explain, some distributions lack analytical closed-form solutions of their density functions, requiring that these distributions have to be approximated using their characteristic functions, which is a function, as will be explained, that is unique to every probability distribution.

In Chapter 13, we introduce parameters of location and spread for both discrete and continuous probability distributions. Whenever necessary, we point out differences between their computation in the discrete and the continuous cases. Although some of the parameters are discussed in earlier chapters, we review them in Chapter 13 in greater detail. The parameters presented in this chapter include quantiles, mean, and variance. Moreover, we explain the moments of a probability distribution that are of higher order (i.e., beyond the mean and variance), which includes skewness and kurtosis. Some distributions, as we will see, may not possess finite values of all of these quantities. As an example, the α-stable distributions only has a finite mean and variance for certain values of their characteristic function parameters. This attribute of the α-stable distribution has prevented it from enjoying more widespread acceptance in the finance world, because many theoretical models in finance rely on the existence of all moments.

The chapters in Part Two thus far have only been dealing with one-dimensional (univariate) probability distributions. However, many fields of finance deal with more than one variable such as a portfolio consisting of many stocks and/or bonds. In Chapter 14, we extend the analysis to joint (or multivariate) probability distributions, the theory of which will be introduced separately for discrete and continuous probability distributions. The notion of random vectors, contour lines, and marginal distributions are introduced. Moreover, independence in the probabilistic sense is defined. As measures of linear dependence, we discuss the covariance and correlation coefficient and emphasize the limitations of their usability. We conclude the chapter with illustrations using some of the most common multivariate distributions in finance.

Chapter 15 introduces the concept of conditional probability. In the context of descriptive statistics, the concept of conditional distributions was explained earlier in the book. In Chapter 15, we give the formal definitions of conditional probability distributions and conditional moments such as the conditional mean. Moreover, we discuss Bayes' formula. Applications in finance include risk measures such as the expected shortfall or conditional value-at-risk, where the expected return of some portfolio or trading position is computed conditional on the fact that the return has already fallen below some threshold.

The last chapter in Part Two, Chapter 16, focuses on the general structure of multivariate distributions. As will be seen, any multivariate distribution can be decomposed into two components. One of these components, the copula, governs the dependence between the individual elements of a random vector and the other component specifies the random behavior of each element individually (i.e., the so-called marginal distributions of the elements). So, whenever the true distribution of a certain random vector representing the constituent assets of some portfolio, for example, is unknown, we can recover it from the copula and the marginal distributions. This is a result frequently used in modeling market, credit, and operational risks. In the illustrations, we demonstrate the different effects various choices of copulae (the plural of copula) have on the multivariate distribution. Moreover, in this chapter, we revisit the notion of probabilistic dependence and introduce an additional dependence measure. In previous chapters, the insufficiency of the correlation measure was pointed out with respect to dependence between asset returns. To overcome this deficiency, we present a measure of tail dependence, which is extremely valuable in assessing the probability for two random variables to jointly assume extremely negative or positive values, something the correlation coefficient might fail to describe.

Part Three: Inductive Statistics

Part Three concentrates on statistical inference as the method of drawing information from sample data about unknown parameters. In the first of the three chapters in Part Three, Chapter 17, the point estimator is presented. We emphasize its random character due to its dependence on the sample data. As one of the easiest point estimators, we begin with the sample mean as an estimator for the population mean. We explain why the sample mean is a particular form of the larger class of linear estimators. The quality of some point estimators as measured by their bias and their mean square error is explained. When samples become very large, estimators may develop certain behavior expressed by their so-called *large sample criteria*. Large sample criteria offer insight into an estimator's behavior as the sample size increases up to infinity. An important large sample criterion is the consistency needed to assure that the estimators will eventually approach the unknown parameter. Efficiency, another large sample criterion, guarantees that this happens faster than for any other unbiased estimator. Also in this chapter, retrieving the best estimator for some unknown parameter, which is usually given by the so-called *sufficient statistic* (if it should exist), is explained. Point estimators are necessary to specify all unknown distributional parameters of models in finance. For example, the return volatility of some portfolio measured by the standard deviation is not automatically known

even if we assume that the returns are normally distributed. So, we have to estimate it from a sample of historical data.

In Chapter 18, we introduce the confidence interval. In contrast to the point estimator, a confidence interval provides an entire range of values for the unknown parameter. We will see that the construction of the confidence interval depends on the required confidence level and the sample size. Moreover, the quality criteria of confidence intervals regarding the trade-off between precision and the chance to miss the true parameter are explained. In our analysis, we point out the advantages of symmetric confidence intervals, as well as emphasizing how to properly interpret them. The illustrations demonstrate different confidence intervals for the mean and variance of the normal distribution as well as parameters of some other distributions, such as the exponential distribution, and discrete distributions, such as the binomial distribution.

The final chapter in Part Two, Chapter 19, covers hypothesis testing. In contrast to the previous two chapters, the interest is not in obtaining a single estimate or an entire interval of some unknown parameter but instead in verifying whether a certain assumption concerning this parameter is justified. For this, it is necessary to state the hypotheses with respect to our assumptions. With these hypotheses, one can then proceed to develop a decision rule about the parameter based on the sample. The types of errors made in hypothesis testing—type I and type II errors—are described. Tests are usually designed so as to minimize—or at least bound—the type I error to be controlled by the test size. The often used *p*-value of some observed sample is introduced in this chapter. As quality criteria, one often focuses on the power of the test seeking to identify the most powerful test for given hypotheses. We explain why it is desirable to have an unbiased and consistent test. Depending on the problem under consideration, a test can be either a one-tailed test or a two-tailed test. To test whether a pair of empirical cumulative relative frequency distributions stem from the same distribution, we can apply the Kolmogorov-Smirnov test. The likelihood-ratio test is presented as the test used when we want to find out whether certain parameters of the distribution are zero or not. We provide illustrations for the most common test situations. In particular, we illustrate the problem of having to find out whether the return volatility of a certain portfolio has increased or not, or whether the inclusion of new stocks into some portfolio increased the overall portfolio return or not.

Part Four: Multivariate Linear Regression

One of the most commonly used statistical tools in finance is regression analysis. In Chapter 6, we introduced the concept of regression for one independent and one dependent variable (i.e., univariate regression or simple

linear regression). However, much more must be understand about regression analysis and for this reason in the three chapters in Part Four we extend the coverage to the multivariate linear regression case.

In Chapter 20, we will give the general assumptions of the multivariate linear regression model such as normally and independently distributed errors. Relying on these assumptions, we can lay out the steps of estimating the coefficients of the regression model. Regression theory will rely on some knowledge of linear algebra and, in particular, matrix and vector notation. (This will be provided in Appendix B.) After the model has been estimated, it will be necessary to evaluate its quality through diagnostic checks and the model's statistical significance. The analysis of variance is introduced to assess the overall usefulness of the regression. Additionally, determining the significance of individual independent variables using the appropriate *F*-statistics is explained. The two illustrations presented include the estimation of the duration of certain sectors of the financial market and the prediction of the 10-year Treasury yield.

In Chapter 21, we focus on the design and the building process of multivariate linear regression models. The three principal topics covered in this chapter are the problem of multicollinearity, incorporating dummy variables into a regression model and model building techniques using stepwise regression analysis. Multicollinearity is the problem that is caused by including in a multivariate linear regression independent variables that themselves may be highly correlated. Dummy variables allow the incorporation of independent variables that represent a characteristic or attribute such as industry sector or a time period within which an observation falls. Because the value of a variable is either one or zero, dummy variables are also referred to as binary variables. A stepwise regression is used for determining the suitable independent variables to be included in the final regression model. The three methods that can be used in a stepwise regression—stepwise inclusion method, stepwise exclusion method, and standard stepwise regression method—are described.

In the introduction to the multivariate linear regression in Chapter 21, we set forth the assumptions about the function form of the model (i.e., that it is linear) and assumptions about the residual or error term in the model (normally distribution, constant variance, and uncorrelated). These assumptions must be investigated. Chapter 22 describes these assumptions in more detail and how to test for any violations. The tools for correcting any violation are briefly described.

Appendixes

Statistics draws on other fields in mathematics. For this reason, we have included two appendices that provide the necessary theoretical background in

mathematics to understand the presentations in some of the chapters. In Appendix A, we present important mathematical functions and their features that are needed primarily in the context of Part Two of this book. These functions include the continuous function, indicator function, and monotonic function. Moreover, important concepts from differential and integral calculus are explained. In Appendix B, we cover the fundamentals of matrix operations and concepts needed to understand the presentation in Part Four.

In Appendix C, we explain the construction of the binomial and multinomial coefficients used in some discrete probability distributions covered in Chapter 9. In Appendix D, we present an explicit computation of the price formula for European-style call options when stock prices are assumed to be log-normally distributed.