# Introduction and Overview

This book provides a self-contained guide to probability and statistics, and their application to data science. We present probability theory intertwined with statistics, as opposed to first covering probability and then statistics, as in most existing texts. The goal is to highlight the connections between probabilistic concepts and the statistical techniques used to estimate them from data. Throughout the text, and from the very beginning, computational examples illustrate the application of the material to real-world data, extracted from the datasets listed in the Appendix. Code in Python reproducing these examples is available on the book website (www.ps4ds.net). The website also contains additional supporting material, including videos, slides, and solutions to all exercises. In the remainder of this section, we provide an overview of the contents of each chapter.

Chapter 1 introduces **probability**. We begin with an informal definition, which enables us to build intuition about the properties of probability. Then, we present a more rigorous definition, based on the mathematical framework of probability spaces. Next, we describe conditional probability, which makes it possible to update probabilities when additional information is revealed. In our first encounter with statistics, we explain how to estimate probabilities and conditional probabilities from data, as illustrated by an analysis of votes in the United States Congress. Building upon the concept of conditional probability, we define independence and conditional independence, which are critical concepts in probabilistic modeling. The chapter ends with a surprising twist: In practice, probabilities are often impossible to compute analytically! Fortunately, the Monte Carlo method provides a pragmatic solution to this challenge, allowing us to approximate probabilities very accurately using computer simulations. We apply the method to model a 3x3 basketball tournament from the 2020 Tokyo Olympics.

Chapter 2 introduces **random variables**, and explains how to use them to model uncertain numerical quantities that are **discrete**. We first provide a mathematical definition of random variables, building upon the framework of probability spaces. Then, we explain how to manipulate discrete random variables in practice, using their probability mass function (pmf), and describe the main properties of the pmf. Motivated by an example where we analyze Kevin Durant's free-throw shooting, we define the empirical pmf, a nonparametric estimator of the pmf that does not make strong assumptions about the data. Next, we define several popular discrete parametric distributions (Bernoulli, binomial, geometric, and Poisson), which yield parametric estimators of the pmf, and explain how to fit them to data via maximum-likelihood estimation. We conclude the chapter by comparing the advantages and disadvantages of nonparametric and parametric models, illustrated by a real-data example, where we model the number of calls arriving at a call center.

https://doi.org/10.1017/9781009180108.002 Published online by Cambridge University Press

Chapter 3 introduces **continuous random variables**, which enable us to model uncertain continuous quantities. We again begin with a formal definition, but quickly move on to describe how to manipulate continuous random variables in practice. We define the cumulative distribution function and quantiles (including the median), and explain how to estimate them from data. We then introduce the concept of probability density and describe its main properties. We present two approaches to obtain nonparametric models of probability densities from data: The histogram and kernel density estimation. Next, we define two celebrated continuous parametric distributions – the exponential and the Gaussian – and show how to fit them to data using maximum-likelihood estimation. We use these distributions to model the interarrival time of calls at a call center, and height in a population, respectively. Finally, we discuss how to simulate continuous random variables via inverse transform sampling.

Chapter 4 describes how to jointly model the interactions between several uncertain discrete quantities. Mathematically, this is achieved by representing the quantities as **multiple discrete random variables** within the same probability space. In practice, such variables are characterized using their joint pmf. We explain how to estimate the joint pmf from data and use it to model precipitation at three locations in Oregon. Then, we introduce marginal and conditional distributions, utilizing the real-world Oregon precipitation data as a running example. Marginal distributions describe the individual behavior of each variable in a model. Conditional distributions describe the behavior of a variable, when the values of other variables are fixed. Next, we generalize the concepts of independence and conditional independence to random variables. At this point, we discuss the problem of causal inference, which seeks to identify causal relationships between variables. Causal inference enables us to understand why a relatively unknown NBA player can have a better three-point shooting percentage than the best shooter in history. We then turn our attention to a fundamental challenge in statistics and data science: It is impossible to completely characterize the dependence between the variables of a probabilistic model, unless they are very few. This phenomenon, known as the curse of dimensionality, is the reason why independence and conditional independence assumptions are needed to make probabilistic models tractable. We conclude the chapter by describing two popular models based on such assumptions: Naive Bayes and Markov chains.

In Chapter 5 we describe how to jointly model continuous quantities, by representing them as **multiple continuous random variables** within the same probability space. We define the joint cumulative distribution function and the joint probability density function, and explain how to estimate the latter from data using a multivariate generalization of kernel density estimation. Next, we introduce marginal and conditional distributions of continuous variables and also discuss independence and conditional independence. Throughout, we model real-world temperature data as a running example. Then, we explain how to jointly simulate multiple random variables, in order to correctly account for the dependence between them. Finally, we define Gaussian random vectors, which are the most popular multidimensional parametric models for continuous data, and apply them to model anthropometric data.

Chapter 6 discusses how to build probabilistic models that include both **discrete and continuous variables**. Mathematically, this is achieved by defining them as random variables within the same probability space. In practice, the variables are manipulated using their marginal and conditional distributions. We define the conditional pmf of a discrete random variable given a continuous variable, and the conditional probability density of a continuous random variable given a discrete variable. We use these objects to build mixture models and

https://doi.org/10.1017/9781009180108.002 Published online by Cambridge University Press

apply them to model height in a population. Next, we describe Gaussian discriminant analysis, a classification method based on mixture models with Gaussian conditional distributions, and apply it to diagnose Alzheimer's disease. Then, we explain how to perform clustering using Gaussian mixture models and leverage the approach to cluster NBA players. Finally, we introduce the framework of Bayesian statistics, which enables us to explicitly encode our uncertainty about model parameters, and use it to analyze poll data from the 2020 United States presidential election.

Chapter 7 focuses on **averaging**, which is a fundamental operation in probability and statistics. We begin by defining an averaging procedure for random variables, known as the mean. We show that the mean is linear, and also that the mean of the product of independent variables equals the product of their means. Then, we derive the mean of popular parametric distributions. Next, we caution that the mean can be severely distorted by extreme values, as illustrated by an analysis of NBA salaries. In addition, we define the mean square, which is the average squared value of a random variable, and the variance, which is the mean square deviation from the mean. We explain how to estimate the variance from data and use it to describe temperature variability at different geographic locations. Then, we define the conditional mean, which represents the average of a variable when other variables are fixed. We prove that the conditional mean is an optimal solution to the problem of regression, where the goal is to estimate a quantity of interest as a function of other variables. We end the chapter by studying how to estimate average causal effects, motivated by two real-world causal-inference questions: *Do all-caps titles attract more views on YouTube?* and *Do private lessons improve students' grades?*

Chapter 8 focuses on **correlation**, a key metric in data science, which quantifies to what extent two quantities are linearly related. We begin by defining the correlation between normalized and centered random variables. Then, we generalize the definition to all random variables and introduce the concept of covariance, which measures the average joint variation of two random variables. Next, we explain how to estimate correlation from data and analyze the correlation between the height of NBA players and different basketball stats. In addition, we study the connection between correlation and simple linear regression. We then discuss the differences between uncorrelation and independence. In order to gain better intuition about the properties of correlation, we provide a geometric interpretation of correlation, where the covariance is an inner product between random variables. Finally, we show that correlation does not imply causation, as illustrated by the spurious correlation between temperature and unemployment in Spain.

Chapter 9 explains how to **estimate population parameters** from data. As running examples, we consider the problems of estimating the mean height in a population and the prevalence of COVID-19 in New York City. We begin by introducing random sampling, a simple yet powerful approach that enables us to obtain accurate estimates from limited data. We then define the bias and the standard error, which quantify the average error of an estimator and how much it varies, respectively. In order to gain a deeper understanding of the properties of random sampling, we derive deviation bounds, which characterize the probabilistic behavior of a random variable just based on its mean and variance. We use these bounds to prove the celebrated law of large numbers, which states that averaging many independent samples from a distribution yields an accurate estimate of its mean. An important consequence of this law is that random sampling provides an arbitrarily precise estimate of means and proportions, as the number of data grows. However, we also caution that this is

https://doi.org/10.1017/9781009180108.002 Published online by Cambridge University Press

not necessarily the case, if the underlying data contain extreme values, as demonstrated by a real-world economic dataset. Next, we discuss another fundamental mathematical phenomenon, the central limit theorem (CLT), according to which averages of independent quantities tend to have Gaussian distributions. We again provide a cautionary tale, inspired by the 2008 Financial Crisis, warning that the CLT does not hold in the absence of independence. Then, we explain how to use the CLT to build confidence intervals, which quantify the uncertainty of estimates obtained from finite data. Finally, we introduce the bootstrap, a popular computational technique to estimate standard errors and build confidence intervals.

Chapter 10 presents the framework of **hypothesis testing**, which can be used to evaluate whether the available data provide sufficient evidence to support a certain hypothesis. We consider two questions as running examples: *Is a toy die unfair?* and *Is Giannis Antetokounmpo's free-throw shooting worse in away games than in home games?* The main idea behind hypothesis testing is to play devil's advocate and assume a null hypothesis, which contradicts our hypothesis of interest. We explain how to use parametric modeling to implement this idea and define the $p$-value. A small $p$-value indicates that the data cannot be explained by the null hypothesis, which is evidence in favor of the original hypothesis. We prove that thresholding the $p$-value is guaranteed to control the probability of endorsing a false finding. In addition, we define the power of a test, which quantifies the test's ability to identify positive findings. Next, we show how to perform hypothesis testing without a parametric model, focusing on the permutation test. Then, we discuss multiple testing, a setting of great practical interest where many tests are performed simultaneously. Using real data from NBA players, we demonstrate that avoiding false findings in such situations is very challenging, but can be achieved by adjusting the $p$-value threshold. To end the chapter, we provide three reasons why hypothesis testing should not be used as the only stamp of approval for scientific discoveries. First, hypothesis testing does not necessarily identify causal effects; it is complementary to causal inference. Second, small $p$-values do not imply practical significance. Third, relying on $p$-values to validate findings produces a strong incentive to cherry-pick results, a practice known as p-hacking.

Chapter 11 covers **principal component analysis and low-rank models**, which are popular techniques to process high-dimensional datasets with many features. We begin by defining the mean of random vectors and random matrices. Then, we introduce the covariance matrix, which encodes the variance of any linear combination of the entries in a random vector, and explain how to estimate it from data. We model the geographic location of Canadian cities as a running example. Next, we present principal component analysis (PCA), a method to extract the directions of maximum variance in a dataset. We explain how to use PCA to find optimal low-dimensional representations of high-dimensional data, and apply it to a dataset of human faces. Then, we introduce low-rank models for matrix-valued data and describe how to fit them using the singular-value decomposition. We show that this approach is able to automatically identify meaningful patterns in real-world weather data. Finally, we explain how to estimate missing entries in a matrix under a low-rank assumption and apply this methodology to predict movie ratings via collaborative filtering.

Chapter 12 delves deeper into the problems of **regression and classification**, where the goal is to estimate a certain quantity of interest (the response) from observed features. In regression, the response is modeled as a numerical variable. In classification, the response belongs to a finite set of predetermined classes. We begin with a comprehensive description of linear regression models, which are ubiquitous in data science and statistics, because

https://doi.org/10.1017/9781009180108.002 Published online by Cambridge University Press

of their simplicity and interpretability. As a running example, we build a linear model of premature mortality in United States counties. Then, we discuss how to leverage linear regression to perform causal inference. In addition, we explain under what conditions linear models tend to overfit, and under what conditions they generalize robustly to held-out data. Motivated by the threat of overfitting, we introduce the concept of regularization. First, we provide a theoretical analysis of $\ell_2$-norm regularization (a.k.a. ridge regression) and show that it can mitigate overfitting in practice. Second, we explain how to leverage $\ell_1$-norm regularization (a.k.a. the lasso) to perform sparse regression, where the goal is to fit a linear model that only depends on a small subset of the available features. Next, we introduce two popular linear models for binary and multiclass classification: Logistic and softmax regression. We apply these methods to several classification tasks involving real data: Diagnosis of Alzheimer's disease, digit recognition, and identification of wheat varieties. At this point, we turn our attention to nonlinear models, which are cornerstones of modern machine learning. First, we present regression and classification trees and explain how to combine them to build complex nonlinear models via bagging, random forests, and boosting. Second, we describe the framework of deep learning and explain how to train neural networks to perform regression and classification. Finally, we end the chapter (and the book) by discussing how to evaluate classification models.

https://doi.org/10.1017/9781009180108.002 Published online by Cambridge University Press