3

Continuous Variables

Overview

Physical quantities such as temperature, time, or distance are typically modeled as being continuous. In this chapter we explain how to model uncertain continuous quantities probabilistically using random variables. We define continuous random variables in Section 3.1. Section 3.2 describes the cumulative distribution function and the quantiles of a distribution, and explains how to estimate them from data. Section 3.3 introduces the probability density, a fundamental tool to describe and manipulate continuous random variables. Section 3.4 discusses functions of continuous random variables, and explains how to derive their probability density. In Section 3.5 we present two nonparametric approaches for estimating probability densities: the histogram and kernel density estimation. Section 3.6 describes two popular continuous parametric distributions: the exponential and the Gaussian distribution. Section 3.7 explains how to fit continuous parametric models to data using maximum-likelihood estimation. Finally, in Section 3.8 we discuss how to simulate continuous random variables.

3.1 Continuous Random Variables

In Chapter 2 we explain how to model discrete uncertain quantities using random variables. Mathematically, discrete random variables are functions in a probability space. We define continuous random variables analogously, as functions mapping outcomes in the sample space of a probability space to the real line. The only difference is that the range of the functions is continuous (and therefore uncountably infinite), as opposed to discrete.

In practice, we manipulate discrete random variables using their probability mass function, which encodes the probability that the random variable equals any specific value. Similarly, when using continuous random variables for probabilistic modeling, we almost never define the underlying probability space explicitly. Instead, we describe the random variable in terms of the probability that it belongs to different intervals of the real line. In order for this probability to be well defined, every interval must belong to the collection of events in the probability space associated to the random variable. This is imposed in the mathematical definition of continuous random variables.

Definition 3.1 (Mathematical definition of continuous random variable)**.** *Let*

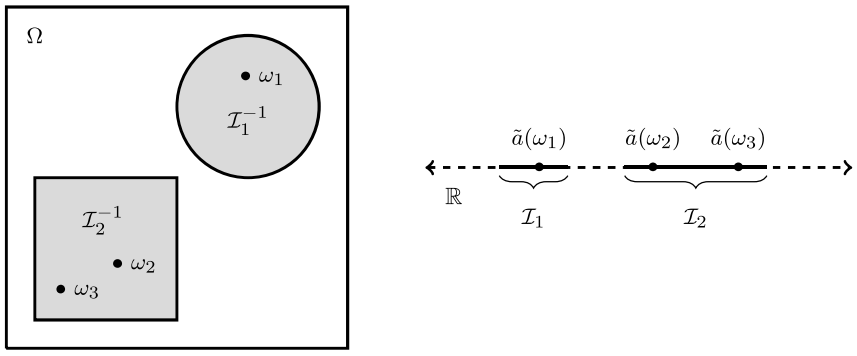

Figure 3.1 Mathematical definition of a continuous random variable. The continuous random variable $\tilde{a}$ maps outcomes in the sample space Ω, represented by the Venn diagram on the left, to the real line $\mathbb{R}$ depicted on the right. The events $\mathcal{I}_1^{-1}$ and $\mathcal{I}_2^{-1}$ contain all outcomes mapping to the two intervals $\mathcal{I}_1$ and $\mathcal{I}_2$ respectively. For $i = 1, 2$, the probability that $\tilde{a}$ belongs to $\mathcal{I}_i$ is equal to $P(\mathcal{I}_i^{-1})$, represented by the area of $\mathcal{I}_i^{-1}$ in the Venn diagram. If $\mathcal{I}_1$ and $\mathcal{I}_2$ are disjoint, then so are $\mathcal{I}_1^{-1}$ and $\mathcal{I}_2^{-1}$ as explained in the proof of Theorem 3.2.

$(\Omega, \mathcal{C}, \mathrm{P})$ be a probability space and $\tilde{a} : \Omega \to \mathbb{R}$ a function mapping elements in the sample space Ω to the real line $\mathbb{R}$. The function $\tilde{a}$ is a valid random variable if for any interval $\mathcal{I} := [a, b] \subseteq \mathbb{R}$, $a \leq b$, the event

$$\mathcal{I}^{-1} := \{\omega : \tilde{a}(\omega) \in \mathcal{I}\}, \tag{3.1}$$

containing the outcomes mapping to the interval, belongs to the collection $\mathcal{C}$. This means that the probability

$$\mathrm{P}(\tilde{a} \in \mathcal{I}) = \mathrm{P}(\mathcal{I}^{-1}) \tag{3.2}$$

is well defined. Such functions are called measurable. The random variable is said to be continuous if the probability that it equals any single point $a \in \mathbb{R}$ is zero,

$$\mathrm{P}(\tilde{a} = a) = 0. \tag{3.3}$$

Since we usually ignore the underlying probability space, we often denote the event that a random variable $\tilde{a}$ belongs to a set S by

$$\{\tilde{a} \in S\} := \{\omega : \tilde{a}\,(\omega) \in S\}. \tag{3.4}$$

Figure 3.1 illustrates the mathematical definition of continuous random variables. It depicts two events in the underlying probability space, each mapping to an interval.

Notice that the probability that a continuous random variable is equal to a specific value is zero. Although this might seem a bit strange at first, it is a natural consequence of modeling a quantity as being continuous. A single point

on a line has zero length. A single point inside an object has zero mass. In fact, it is impossible to assign nonzero probability to every point in an interval of the real line, because the number of points in any interval is uncountably infinite, so the sum of the probabilities would explode to infinity (see Exercise 3.1).

An important consequence of Definition 3.1 is that the probability that a continuous random variable belongs to any union of disjoint intervals is the sum of the probabilities assigned to each of the intervals.

Theorem 3.2. *Let $\tilde{a}$ be a continuous random variable satisfying Definition 3.1. For any n disjoint intervals of $\mathbb{R}$ $\mathcal{I}_1, \mathcal{I}_2, \ldots, \mathcal{I}_n$,*

$$\mathrm{P}(\tilde{a} \in \cup_{i=1}^{n} \mathcal{I}_i) = \sum_{i=1}^{n} \mathrm{P}(\tilde{a} \in \mathcal{I}_i). \tag{3.5}$$

Similarly, if $\mathcal{I}_1, \mathcal{I}_2, \ldots$ is a countably infinite sequence of disjoint intervals of $\mathbb{R}$,

$$\mathrm{P}(\tilde{a} \in \cup_{i=1}^{\infty} \mathcal{I}_i) = \sum_{i=1}^{\infty} \mathrm{P}(\tilde{a} \in \mathcal{I}_i). \tag{3.6}$$

Proof We prove the finite case, the infinite case follows from the same argument. For all i, let us denote by

$$\mathcal{I}_i^{-1} := \{\omega : \tilde{a}(\omega) \in \mathcal{I}_i\} \tag{3.7}$$

the event containing the outcomes mapped to $\mathcal{I}_i$. The key insight is that if $\mathcal{I}_1, \ldots, \mathcal{I}_n$ are disjoint, then so are $\mathcal{I}_1^{-1}, \ldots, \mathcal{I}_n^{-1}$. This is illustrated by Figure 3.1: since $\mathcal{I}_1$ and $\mathcal{I}_2$ are disjoint, so are $\mathcal{I}_1^{-1}$ and $\mathcal{I}_2^{-1}$, because otherwise there would be an outcome $\omega \in \mathcal{I}_1^{-1} \cap \mathcal{I}_2^{-1}$, which would map to a point $\tilde{a}(\omega) \in \mathcal{I}_1 \cap \mathcal{I}_2$. In addition, the union of all outcomes mapping to $\cup_{i=1}^{n} \mathcal{I}_i$, which we denote by $(\cup_{i=1}^{n} \mathcal{I}_i)^{-1}$ is equal to $\cup_{i=1}^{n} \mathcal{I}_i^{-1}$. Consequently,

$$\begin{aligned}
\mathrm{P}(\tilde{a} \in \cup_{i=1}^{n} \mathcal{I}_i) &= \mathrm{P}(\{\omega : \tilde{a}\,(\omega) \in \cup_{i=1}^{n} \mathcal{I}_i\}) && (3.8)\\
&= \mathrm{P}(\{\omega : \omega \in (\cup_{i=1}^{n} \mathcal{I}_i)^{-1}\}) && (3.9)\\
&= \mathrm{P}(\{\omega : \omega \in \cup_{i=1}^{n} \mathcal{I}_i^{-1}\}) && (3.10)\\
&= \sum_{i=1}^{n} \mathrm{P}(\{\omega : \omega \in \mathcal{I}_i^{-1}\}) && (3.11)\\
&= \sum_{i=1}^{n} \mathrm{P}(\tilde{a} \in \mathcal{I}_i). && (3.12)
\end{aligned}$$

■

The reason why Theorem 3.2 is important, is that it *liberates* us from having to refer to the probability space associated to a random variable in order to describe its behavior. Instead, we just need to know the probability that it belongs to any interval of the real line, which is encoded in the cumulative distribution function defined in Section 3.2 (see Lemma 3.5) or in the probability density function defined in Section 3.3 (see Theorem 3.16).

Since the probability assigned to any single point is zero, when describing

continuous random variables it does not matter whether we consider closed or open intervals. For any continuous random variable $\tilde{a}$ and any interval $[a, b] \subseteq \mathbb{R}$, $a \leq b$,

$$\begin{aligned} \mathrm{P}(\tilde{a} \in [a, b]) &= \mathrm{P}(\tilde{a} = a) + \mathrm{P}(\tilde{a} \in (a, b)) + \mathrm{P}(\tilde{a} = b) && (3.13) \\ &= \mathrm{P}(\tilde{a} \in (a, b)). && (3.14) \end{aligned}$$

Strictly speaking, given our definition of continuous random variables, we can only consider the probability that they belong to *Borel sets*, which are sets that can be described as countable unions of intervals. Incredibly enough, there exist subsets of $\mathbb{R}$ that are not Borel sets. However, it is extremely unlikely that you will ever come across them, unless you happen to do very theoretical research in measure theory, which is the area of mathematics that deals with these issues.

3.2 The Cumulative Distribution Function

3.2.1 Definition

According to Theorem 3.2, in order to describe the behavior of a continuous random variable, we just need the probability that it belongs to any interval. This information can be captured using the cumulative distribution function, which encodes the probability that the random variable is less than or equal to any real value.

Definition 3.3 (Cumulative distribution function)**.** *Let $(\Omega, \mathcal{C}, \mathrm{P})$ be a probability space and $\tilde{a} : \Omega \to \mathbb{R}$ a random variable. The cumulative distribution function (cdf) of $\tilde{a}$ is defined as*

$$F_{\tilde{a}}(a) := \mathrm{P}(\tilde{a} \leq a). \tag{3.15}$$

In words, $F_{\tilde{a}}(a)$ is the probability that $\tilde{a}$ is less than or equal to a.

As established in the following lemma, the cdf always has the same overall structure. It starts at zero, because the probability that the random variable is smaller than an arbitrarily small number is zero. It ends at one, because the probability that the random variable is smaller than an arbitrarily large number is one. It is also nondecreasing.

Lemma 3.4 (Properties of the cdf)**.** *For any random variable $\tilde{a}$ with cdf $F_{\tilde{a}}$*

$$\lim_{a \to -\infty} F_{\tilde{a}}(a) = 0, \tag{3.16}$$

$$\lim_{a \to \infty} F_{\tilde{a}}(a) = 1. \tag{3.17}$$

In addition, $F_{\tilde{a}}$ is nondecreasing,

$$F_{\tilde{a}}(b) \geq F_{\tilde{a}}(a) \quad \text{if } b > a. \tag{3.18}$$

Proof Intuitively, $\lim_{a \to -\infty} F_{\tilde{a}}(a) = 0$ because the random variable $\tilde{a}$ is fixed, so if we make a arbitrarily small, eventually the probability that $\tilde{a}$ is less than

or equal to a becomes zero. To make this argument mathematically rigorous, we apply Theorem 3.2,

$$\lim_{a\to-\infty} F_{\tilde{a}}(a) = 1 - \lim_{a\to-\infty} \mathrm{P}(\tilde{a} > a) \tag{3.19}$$

$$= 1 - \mathrm{P}(\tilde{a} > 0) - \lim_{n\to\infty} \sum_{i=0}^{n} \mathrm{P}(-i \geq \tilde{a} > -(i+1)) \tag{3.20}$$

$$= 1 - \mathrm{P}\left(\lim_{n\to\infty} \{\tilde{a} > 0\} \cup \cup_{i=0}^{n} \{-i \geq \tilde{a} > -(i+1)\}\right) \tag{3.21}$$

$$= 1 - \mathrm{P}(\tilde{a} \in \mathbb{R}) = 0. \tag{3.22}$$

$\mathrm{P}(\tilde{a} \in \mathbb{R}) = \mathrm{P}(\{\omega : \omega \in \Omega\}) = 1$ because all outcomes of the original sample space are mapped to the real line.

Similarly, $\lim_{a\to\infty} F_{\tilde{a}}(a) = 1$, because if we make a arbitrarily large, eventually the probability that $\tilde{a}$ is less than or equal to a is one. The proof is essentially the same as that of (3.16).

The inequality in (3.18) is a consequence of Lemma 1.12, because $\{\tilde{a} \leq a\}$ is a subset of $\{\tilde{a} \leq b\}$, so

$$F_{\tilde{a}}(b) = \mathrm{P}(\tilde{a} \leq b) \geq \mathrm{P}(\tilde{a} \leq a) = F_{\tilde{a}}(a). \tag{3.23}$$

■

We can compute the probability that a random variable belongs to any interval from its cdf, just by subtracting its value at the end and the beginning of the interval. Consequently, the cdf completely determines the behavior of a continuous random variable, just like the pmf completely determines the behavior of a discrete random variable.

Lemma 3.5. *For any a, $b \in \mathbb{R}$, $a \leq b$, and any continuous random variable $\tilde{a}$,*

$$\mathrm{P}(a < \tilde{a} \leq b) = F_{\tilde{a}}(b) - F_{\tilde{a}}(a). \tag{3.24}$$

Proof By Theorem 3.2

$$\mathrm{P}(\tilde{a} \leq b) = \mathrm{P}(\tilde{a} \in (-\infty, a] \cup (a, b]) \tag{3.25}$$

$$= \mathrm{P}(a < \tilde{a} \leq b) + \mathrm{P}(\tilde{a} \leq a). \tag{3.26}$$

■

Example 3.6 (Using the cdf to compute probabilities)**.** Consider a continuous random variable $\tilde{a}$ with the following cdf:

$$F_{\tilde{a}}(a) := \begin{cases} 0 & \text{for } a < 0, \\ 0.5\,a & \text{for } 0 \leq a \leq 1, \\ 0.5 & \text{for } 1 \leq a \leq 2, \\ 0.5\left(1 + (a-2)^2\right) & \text{for } 2 \leq a \leq 3, \\ 1 & \text{for } a > 3. \end{cases} \tag{3.27}$$

Figure 3.2 shows the cdf. You can check that it satisfies the properties in Lemma 3.4.

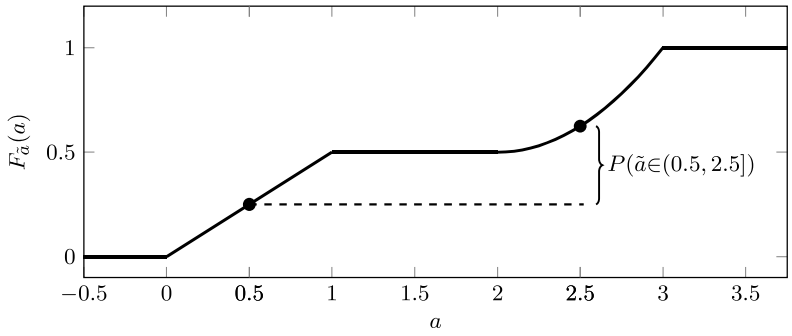

Figure 3.2 Example of cumulative distribution function. The graph shows the cdf of the random variable $\tilde{a}$ in Example 3.6, and illustrates how to compute the probability that $\tilde{a}$ belongs to an interval using the cdf.

We can apply Lemma 3.5 to compute the probability that $\tilde{a}$ belongs to any interval. For example,

$$\mathrm{P}\left(0.5 < \tilde{a} \leq 2.5\right) = F_{\tilde{a}}\left(2.5\right) - F_{\tilde{a}}\left(0.5\right) = 0.375, \tag{3.28}$$

as illustrated in Figure 3.2.

It is often useful to model quantities that are *uniformly* distributed in an interval, meaning that the probability that they lie in a certain interval is proportional to the length of the interval. The following example shows that this can be captured by defining a random variable with a linear cdf. We define these random variables more formally in Definition 3.15 using their probability density.

Example 3.7 (Linear cdf)**.** Consider the continuous random variable $\tilde{u}$ with cdf

$$F_{\tilde{u}}\left(u\right) := \begin{cases} 0 & \text{for } u < 0, \\ u & \text{for } 0 \leq u \leq 1, \\ 1 & \text{for } u > 1. \end{cases} \tag{3.29}$$

The cdf is shown in Figure 3.3. By Lemma 3.5, the probability of the random variable belonging to any interval $[a, b] \subseteq [0, 1]$ within the unit interval is equal to the length of the interval,

$$\mathrm{P}\left(a < \tilde{u} \leq b\right) = F_{\tilde{u}}\left(b\right) - F_{\tilde{u}}\left(a\right) \tag{3.30}$$

$$= b - a. \tag{3.31}$$

We mainly use the cdf to manipulate continuous random variables, but discrete random variables also have cdfs, as we show in the following example.

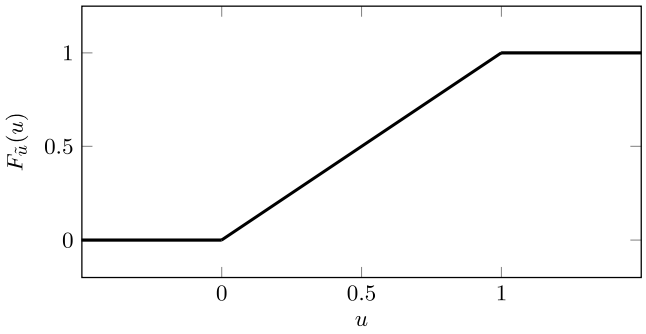

Figure 3.3 Uniform probability in the unit interval. Cumulative distribution function of a random variable $\tilde{u}$ that is uniformly distributed in $[0, 1]$. Due to the linear shape of the cdf, the probability that $\tilde{u}$ belongs to any interval in $[0, 1]$ is proportional to its length.

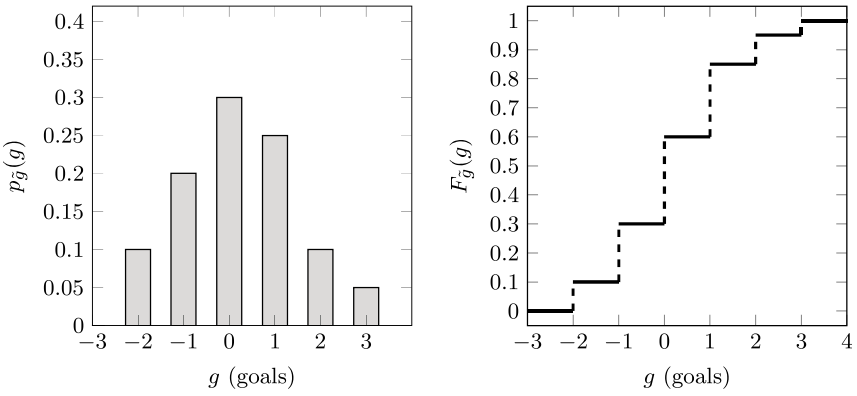

Figure 3.4 Cumulative distribution function of a discrete random variable. The graphs show the pmf (left) and the cdf (right) of the random variable $\tilde{g}$ from Example 2.7. The random variable is discrete, so the cdf is piecewise constant. Each jump is equal to the probability that the random variable equals that value.

Example 3.8 (Cumulative distribution function of a discrete random variable)**.** We consider the discrete random variable $\tilde{g}$ from Example 2.7. By definition, the cdf of the random variable equals the probability that it is less than or equal to

any real number. We can compute that probability using the pmf. This yields

$$F_{\tilde{g}}(g) := \mathrm{P}(\tilde{g} \leq g) = \begin{cases} 0 & \text{for } g < -2, \\ 0.1 & \text{for } -2 \leq g < -1, \\ 0.3 & \text{for } -1 \leq g < 0, \\ 0.6 & \text{for } 0 \leq g < 1, \\ 0.85 & \text{for } 1 \leq g < 2, \\ 0.95 & \text{for } 2 \leq g < -3, \\ 1 & \text{for } g \geq 3. \end{cases} \tag{3.32}$$

The cdf is piecewise constant because there are only six values where the probability is nonzero. The cdf has a *jump* at each of these values equal to the corresponding probability, as illustrated in Figure 3.4.

As illustrated by Example 3.8, the cdf of any discrete random variable is discontinuous, because the probability that the random variable equals individual values is nonzero. In contrast, for continuous random variables, the cdf is continuous because the probability of each single point is zero. In fact, we could have defined continuous random variables as those random variables that have a continuous cdf.

Lemma 3.9 (Continuous cdf)**.** *The cdf $F_{\tilde{a}}$ of a random variable $\tilde{a}$ is continuous if and only if the probability that it equals any single point $a \in \mathbb{R}$ is zero.*

Proof Recall that $F_{\tilde{a}}$ is continuous if and only if for any a we have

$$F_{\tilde{a}}(a) = \lim_{\epsilon \to 0} F_{\tilde{a}}(a - \epsilon). \tag{3.33}$$

By Lemma 3.5

$$\mathrm{P}(\tilde{a} = a) = F_{\tilde{a}}(a) - \lim_{\epsilon \to 0} F_{\tilde{a}}(a - \epsilon), \tag{3.34}$$

so the result follows. ■

3.2.2 The Quantiles Of A Distribution

The distribution of a continuous random variables can be described in terms of its *quantiles*, which are points that divide the real line into regions with equal probability. When the number of regions is four the quantiles are called *quartiles*. The cdf makes it very simple to define quantiles.

Definition 3.10 (Quantiles)**.** *Let m be an integer and $\tilde{a}$ a random variable. The m-quantiles of $\tilde{a}$ are $m - 1$ points q_1, q_2, $\ldots$, q_{m-1} such that*

$$\mathrm{P}(\tilde{a} \leq q_1) = \mathrm{P}(q_1 \leq \tilde{a} \leq q_2) = \cdots = \mathrm{P}(\tilde{a} \geq q_{m-1}), \tag{3.35}$$

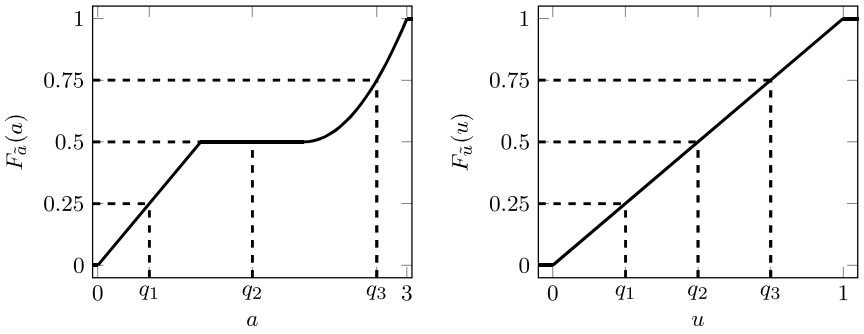

Figure 3.5 Quartiles. The graph shows the quartiles of the random variables in Examples 3.6 (left plot) and 3.7 (right plot). They can be found by *inverting* the cdf. To find the median, for example, we look for the point where the cdf equals 1/2. For the random variable in Example 3.6 any point between 1 and 2 could be considered the median; we arbitrarily choose 1.5.

or equivalently

$$F_{\tilde{a}}(q_i) = \mathrm{P}(\tilde{a} \leq q_i) = \frac{i}{m}, \qquad i = 1, 2, \ldots, m-1. \tag{3.36}$$

In the case that several points satisfy the equality, we usually choose their midpoint. In words, the quantiles partition the real line so that $\tilde{a}$ has the same probability of being in each interval of the partition. When $m := 4$, the quantiles are called quartiles. When $m := 10$, the quantiles are called deciles. When $m := 100$, they are called percentiles.

Consider the three quartiles q_1, q_2, and q_3 of a distribution. The difference between the third and first quartile $q_3 - q_1$ is called the *interquartile range*. It provides a measure for how spread out the distribution of the random variable is. The second quartile q_2 is called the *median*. It separates the possible values of the random variable into two intervals with equal probability and can therefore be interpreted as the midpoint of the distribution.

Definition 3.11 (Median)**.** *The median q_2 of a continuous random variable $\tilde{a}$ satisfies*

$$\mathrm{P}(\tilde{a} \leq q_2) = \mathrm{P}(\tilde{a} > q_2) = \frac{1}{2}, \tag{3.37}$$

or equivalently

$$F_{\tilde{a}}(q_2) = \frac{1}{2}. \tag{3.38}$$

Figure 3.5 shows the quartiles of the random variables in Examples 3.6 and 3.7, illustrating their connection to the cdf.

3.2.3 Estimating The CDF And Quantiles From Data

In order to model an uncertain quantity as a continuous random variable, we can estimate its cdf from data. Since the cdf represents a probability, it is natural to approximate it using empirical probabilities. Figure 3.6 shows the empirical cdf for two real-world datasets: height of men in the US army, extracted from Dataset 5, and national gross domestic products (GDPs), extracted from Dataset 6.

Definition 3.12 (Empirical cdf)**.** *Let $X := \{x_1, x_2, \ldots, x_n\}$ denote a real-valued dataset. The empirical cumulative distribution function $F_X : \mathbb{R} \to [0, 1]$ maps each value $a \in \mathbb{R}$ to the fraction of data that are smaller or equal to a,*

$$F_X(a) := \frac{1}{n} \sum_{i=1}^{n} 1(x_i \leq a), \tag{3.39}$$

where $1(x_i \leq a)$ is an indicator function that is equal to one if $x_i \leq a$ and to zero otherwise.

Similarly, we can use empirical probabilities to estimate the quantiles.

Definition 3.13 (Quantile estimation)**.** *Let $X := \{x_1, x_2, \ldots, x_n\}$ denote a real-valued dataset, interpreted as realizations from a random variable $\tilde{a}$. The m-quantiles of the data are $m - 1$ points $\hat{q}_1$, $\hat{q}_2$, $\ldots$, $\hat{q}_{m-1}$ such that*

$$\mathrm{P}_X(\tilde{a} \leq \hat{q}_1) = \mathrm{P}_X(\hat{q}_1 \leq \tilde{a} \leq \hat{q}_2) = \cdots = \mathrm{P}_X(\tilde{a} \geq \hat{q}_{m-1}), \tag{3.40}$$

where P_X denotes an empirical probability computed from X following Definition 1.22. Equivalently,

$$F_X(\hat{q}_i) = \frac{i}{m}, \qquad i = 1, 2, \ldots, m-1, \tag{3.41}$$

where F_X is the empirical cdf. In the case that several points satisfy the equality, we choose their midpoint. In practice, the quantiles can be computed efficiently by sorting the data. For example, the median is the midpoint of the sorted dataset.

The quartiles of a dataset provide a very concise description of the overall distribution of the data. They can be visualized using a *box plot*, which shows the median $\hat{q}_2$ of the data enclosed in a box. The bottom and top of the box are the first $\hat{q}_2$ and third $\hat{q}_3$ quartiles. This way of visualizing a dataset was proposed by the mathematician John Tukey. Tukey's box plot also includes *whiskers*, which depend on the interquartile range (IQR, defined as $\hat{q}_3 - \hat{q}_1$). The lower whisker is a line extending from the bottom of the box to the smallest value within 1.5 IQR of the first quartile. The higher whisker extends from the top of the box to the highest value within 1.5 IQR of the third quartile. Values beyond the whiskers are considered *outliers* and are plotted separately.

By looking at the box plots in Figure 3.6, we can immediately see that the two datasets have very different distributions. The height data is evenly spread out, so the box plot is almost symmetric around the median. In stark contrast, the GDP dataset is very skewed. The number of countries decreases dramatically as

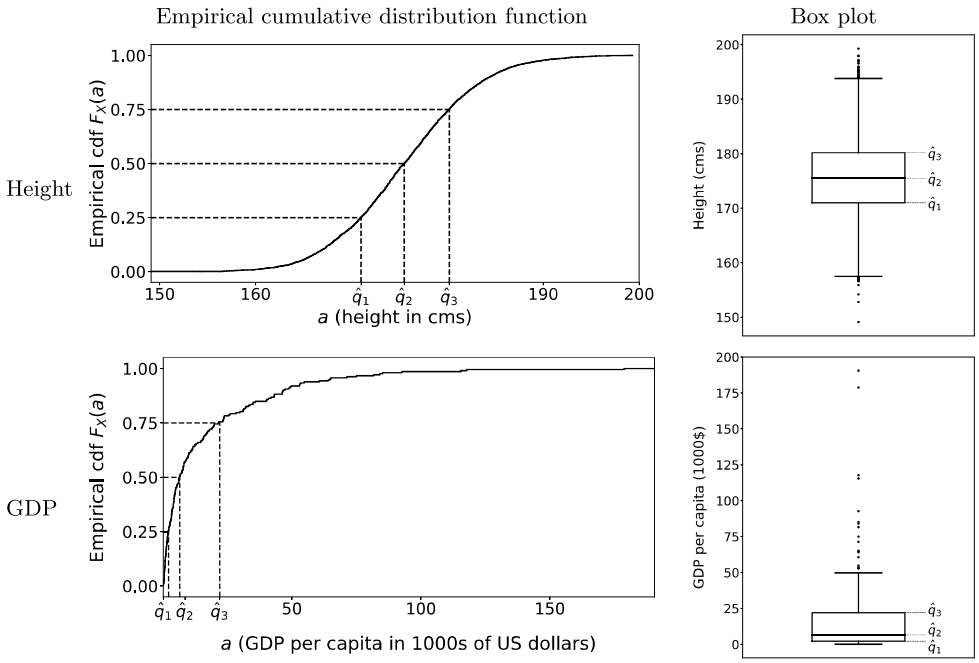

Figure 3.6 Empirical cdf and box plot. The top row shows the empirical cdf (left) of the heights of 4,082 men in the United States army and the corresponding box plot (right). Both plots are annotated to indicate the three quartile estimates $\hat{q}_1$ (171.0 cm), $\hat{q}_2$ (175.5 cm), and $\hat{q}_3$ (180.2 cm). The bottom row shows the empirical cdf (left) of the gross domestic product per capita of 212 countries in 2019, and the corresponding box plot (right). Both plots are again annotated to indicate the three quartile estimates $\hat{q}_1$ ($2,181), $\hat{q}_2$ ($6,520), and $\hat{q}_3$ ($21,988).

the GDP per capita increases. The difference between the median and the first quartile is around $2,400, whereas the difference between the third quartile and the median is around $15,500! Some outliers such as Monaco and Liechstenstein have GDPs per capita above $175,000.

Figure 3.7 further illustrates how box plots can be used to compare the distributions of different continuous quantities. It shows boxplots corresponding to the maximum monthly temperatures recorded at a weather station in Oxford over 150 years, extracted from Dataset 7. We immediately see that the summer months are warmer, the winter months are colder, and there is more intra-month variation in the summer and winter than in the spring and fall.

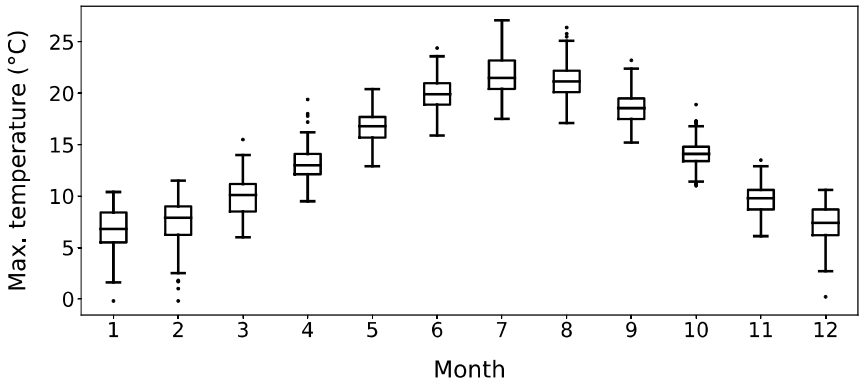

Figure 3.7 Temperature box plots. Box plots of the maximum monthly temperature recorded at a weather station in Oxford over 150 years.

3.3 The Probability Density Function

The cdf is a global quantity: its value at a specific point equals the probability of the random variable from $-\infty$ up until that point. In order to describe the local behavior of a continuous random variables, it is more intuitive and convenient to use their *probability density* instead. The probability density of a random variable $\tilde{a}$ at a point a, denoted by $f_{\tilde{a}}(a)$, captures the instantaneous rate of change in probability at a. Mathematically, we define $f_{\tilde{a}}(a)$ as the ratio between the probability that $\tilde{a}$ is in a neighborhood of a and the length of the neighborhood when the neighborhood becomes arbitrarily small:

$$f_{\tilde{a}}(a) = \lim_{\epsilon \to 0} \frac{\mathrm{P}(a - \epsilon \leq \tilde{a} \leq a)}{\epsilon}. \tag{3.42}$$

Intuitively, for a very small interval of length ϵ around a,

$$\mathrm{P}(a - \epsilon < \tilde{a} \leq a) \approx \epsilon f_{\tilde{a}}(a), \tag{3.43}$$

as illustrated in Figure 3.8. Note that the random variable is more likely to take values in intervals where the probability density is high (assuming the intervals have the same length). By Lemma 3.5, the probability density at a is equal to the derivative of the cdf at a:

$$f_{\tilde{a}}(a) = \lim_{\epsilon \to 0} \frac{\mathrm{P}(a - \epsilon \leq \tilde{a} \leq a)}{\epsilon} \tag{3.44}$$

$$= \lim_{\epsilon \to 0} \frac{F(a) - F(a - \epsilon)}{\epsilon} \tag{3.45}$$

$$= \frac{\mathrm{d}F_{\tilde{a}}(a)}{\mathrm{d}a}. \tag{3.46}$$

Thus, the density is only defined at points where the cdf is differentiable. We typically describe random variables with differentiable cdfs using their probability

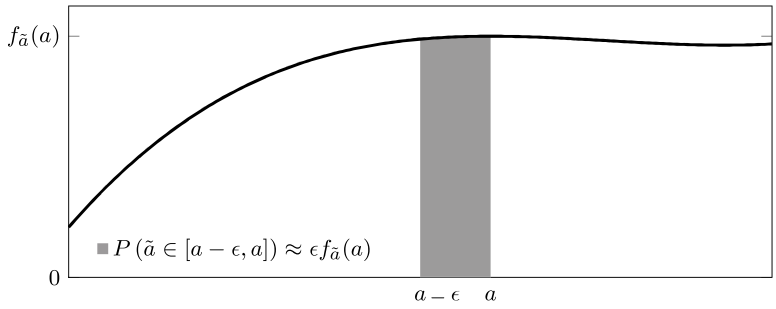

Figure 3.8 Probability density. The probability density of a random variable is equal to the local rate of change in probability. For small ϵ the area of the shaded region, which equals $P\left(\tilde{a} \in [a-\epsilon, a]\right)$, is approximately given by $\epsilon f_{\tilde{a}}(a)$.

density function f, which encodes the probability density of a random variable over the whole real line.

Definition 3.14 (Probability density function)**.** *Let $\tilde{a} : \Omega \to \mathbb{R}$ be a random variable with cdf $F_{\tilde{a}}$. If $F_{\tilde{a}}$ is differentiable, then the pdf of $\tilde{a}$ is the derivative of its cdf,*

$$f_{\tilde{a}}\left(a\right) := \lim_{\epsilon \to 0} \frac{\mathrm{P}(a-\epsilon \leq \tilde{a} \leq a)}{\epsilon} \tag{3.47}$$

$$= \frac{\mathrm{d}F_{\tilde{a}}\left(a\right)}{\mathrm{d}a}. \tag{3.48}$$

In Example 3.7 we consider a random variable with a linear cdf, and show that the probability that it belongs to intervals of different lengths is proportional to the length of the interval. The pdf of the random variable is constant, because it is the derivative of a linear function. In fact, this is how uniform random variables are typically defined.

Definition 3.15 (Uniform distribution)**.** *A uniform random variable $\tilde{u}$ on the interval $[a, b]$ has a constant pdf within the interval,*

$$f_{\tilde{u}}(u) = \frac{1}{b-a}, \qquad \text{if } a \leq u \leq b, \tag{3.49}$$

and zero otherwise.

The pdf of a continuous random variable can be larger than one. For example, if we consider a uniform distribution in the interval $[0, 0.5]$, the pdf equals 2. This reminds us that the pdf does not encode a probability, like the pmf or the cdf, but rather a probability density.

The pdf of a random variable $\tilde{a}$ can be integrated to obtain the probability that $\tilde{a}$ belongs to any union of intervals B (or more formally, to any Borel set). To show why, let us decompose B into a partition of very small intervals of length ϵ

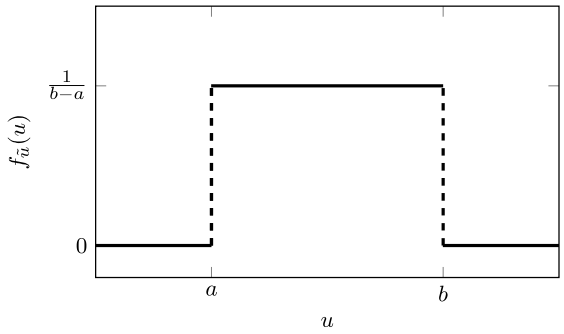

Figure 3.9 Uniform distribution. Probability density function of a uniform random variable $\tilde{u}$ in the interval $[a, b]$.

between a grid of points $a_1, \ldots, a_n$. The probability that $\tilde{a}$ belongs to B is the sum of the probabilities that it belongs to the intervals in the partition, which can be approximated using the probability density at the points on the grid by (3.43)

$$\mathrm{P}(\tilde{a} \in B) = \sum_{i=1}^{n} \mathrm{P}(\tilde{a} \in [a_i - \epsilon, a_i]) \tag{3.50}$$

$$\approx \sum_{i=1}^{n} f(a_i)\epsilon. \tag{3.51}$$

If the cdf of $\tilde{a}$ is differentiable, the approximation becomes an equality when we take the limit $\epsilon \to 0$. Moreover, the sum is a Riemann sum of the pdf. The pdf is integrable, because it is the derivative of the cdf. Consequently, as $\epsilon \to 0$, the sum converges to an integral, so that

$$\mathrm{P}(\tilde{a} \in B) = \int_{a \in B} f(a) \, \mathrm{d}a. \tag{3.52}$$

The probability of $\tilde{a}$ belonging to B is the area under the pdf restricted to B, as illustrated in Figure 3.10. There is a direct analogy with mass. If we know the density of a one-dimensional rod, we can compute the mass of any subset of the rod by integrating the density over the subset.

Theorem 3.16. *Let $\tilde{a}$ be a continuous random variable with pdf $f_{\tilde{a}}$, for any Borel set $B \subseteq \mathbb{R}$, which can be expressed as a countable union of intervals,*

$$\mathrm{P}\left(\tilde{a} \in B\right) = \int_{B} f_{\tilde{a}}\left(a\right) \, \mathrm{d}a. \tag{3.53}$$

Proof By the fundamental theorem of calculus and the definition of pdf, for any

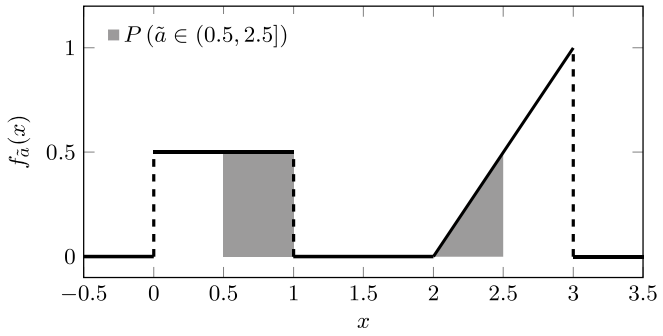

Figure 3.10 Example of probability density function. The plot shows the pdf of the random variable in Examples 3.6 and 3.17. The probability that the random variable belongs to the set of interest is equal to the shaded area under the pdf.

$a \leq b$,

$$\mathrm{P}\left(a < \tilde{a} \leq b\right) = F_{\tilde{a}}\left(b\right) - F_{\tilde{a}}\left(a\right) \tag{3.54}$$

$$= \int_{a}^{b} f_{\tilde{a}}\left(a\right) \, \mathrm{d}a. \tag{3.55}$$

Since B can be decomposed into a countable union of disjoint intervals (we can merge any intervals that are not disjoint), $B = \cup_i \mathcal{I}_i$,

$$\mathrm{P}\left(\tilde{a} \in B\right) = \mathrm{P}\left(\tilde{a} \in \cup_i \mathcal{I}_i\right) \tag{3.56}$$

$$= \sum_{i=1}^{n} \mathrm{P}\left(\tilde{a} \in \mathcal{I}_i\right) \tag{3.57}$$

$$= \sum_{i=1}^{n} \int_{\mathcal{I}_i} f_{\tilde{a}}\left(a\right) \, \mathrm{d}a \tag{3.58}$$

$$= \int_{B} f_{\tilde{a}}\left(a\right) \, \mathrm{d}a. \tag{3.59}$$

■

Example 3.17 (Using the pdf to compute probabilities)**.** To compute the pdf of the random variable in Example 3.6, we differentiate its cdf to obtain

$$f_{\tilde{a}}\left(a\right) = \begin{cases} 0 & \text{for } a < 0, \\ 0.5 & \text{for } 0 \leq a \leq 1 \\ 0 & \text{for } 1 \leq a \leq 2 \\ a - 2 & \text{for } 2 \leq a \leq 3 \\ 0 & \text{for } a > 3. \end{cases} \tag{3.60}$$

Figure 3.10 shows the pdf. To determine the probability that $\tilde{a}$ is between 0.5 and 2.5, we integrate over that interval to obtain the same answer as in Example 3.6,

$$\begin{aligned} \mathrm{P}\left(0.5 < \tilde{a} \leq 2.5\right) &= \int_{a=0.5}^{2.5} f_{\tilde{a}}\left(a\right) \, \mathrm{d}a & (3.61) \\ &= \int_{a=0.5}^{1} 0.5 \, \mathrm{d}a + \int_{a=2}^{2.5} (a-2) \, \mathrm{d}a & (3.62) \\ &= 0.375. & (3.63) \end{aligned}$$

Just like a pmf completely characterizes a discrete random variable, the pdf completely determines the behavior of a continuous random variable. Indeed, by Theorem 3.16, it encodes the probability that the random variable belongs to any finite or countable union of intervals of the real line. Because of this, we often define random variables by stating that they are distributed according to a certain pdf, without ever mentioning the underlying probability space.

The following theorem shows that pdfs are always nonnegative and integrate to one, and also that any such function is a valid pdf.

Theorem 3.18. *A function $f : \mathbb{R} \to \mathbb{R}$ is the pdf of a continuous random variable if and only if it is nonnegative and*

$$\int_{\mathbb{R}} f(a) \, \mathrm{d}a = 1. \tag{3.64}$$

Proof If f is the pdf of a continuous random variable $\tilde{a}$, then it must be nonnegative because the cdf is nondecreasing by Lemma 3.4, and the pdf is the derivative of the cdf. In addition, since all points in the original probability space map to a real number,

$$\begin{aligned} \int_{\mathbb{R}} f_{\tilde{a}}\left(a\right) \, \mathrm{d}a &= \mathrm{P}\left(\tilde{a} \in \mathbb{R}\right) & (3.65) \\ &= \mathrm{P}\left(\omega \in \Omega\right) & (3.66) \\ &= 1. & (3.67) \end{aligned}$$

We sketch a proof of the converse. We build a probability space where the sample space is the real line and the collection of events is the collection of Borel sets (which can be easily shown to satisfy Definition 1.7). If f is nonnegative and integrates to one, we can use it to define the probability measure. For any Borel set B we set $\mathrm{P}\left(B\right) := \int_B f\left(a\right) \, \mathrm{d}a$. You can check that the probability measure satisfies Definition 1.9. We can then define the continuous random variable associated to the pdf as the identity function. ■

3.4 Functions Of Random Variables

As discussed in Section 2.1.4, in probabilistic modeling it is often useful to characterize the behavior of a deterministic function of uncertain quantities represented

by random variables. In contrast to the discrete case, not all deterministic functions applied to a random variable yield a valid random variable. Any function of a random variable is obviously a function itself, but to satisfy Definition 3.1 we need to guarantee that the pre-image of any interval is measurable. Fortunately, most reasonable functions, including all continuous functions, result in valid random variables, so we don't need to worry too much about this in practice.

Lemma 3.19 (Function of a continuous random variable)**.** *Let* $(\Omega, \mathcal{C}, \mathrm{P})$ *be a probability space, and* $\tilde{a} : \Omega \to \mathbb{R}$ *a continuous random variable associated to the probability space. Let* $g : \mathbb{R} \to \mathbb{R}$ *be any function. Then* $\tilde{b} := g \circ \tilde{a}$*, also denoted by* $g(\tilde{a})$*, is a continuous random variable as long as for any interval* (a, b)*,* $a \leq b$*, the set* $\{x : g(x) \in (a, b)\}$ *of real values mapped by* g *to the interval is a Borel set (i.e. a countable union of intervals).*

Proof If $\{x : g(x) \in (a, b)\}$ is a Borel set, then the event $\{\omega : g(\tilde{a}(\omega)) \in (a, b)\}$ is in the collection $\mathcal{C}$ because $\tilde{a}$ is a random variable. ■

The following theorem characterizes how the distribution of a random variable is transformed when we scale the variable by a multiplicative constant. The pdf is stretched proportionally to the scaling factor, as illustrated in Example 3.21.

Theorem 3.20 (Multiplying a random variable by a constant)**.** *Let* $\tilde{a}$ *be a continuous random variable with pdf* $f_{\tilde{a}}$*. For any positive constant* $\alpha > 0$*, the pdf of the scaled random variable* $\tilde{b} = \alpha\tilde{a}$ *is*

$$f_{\tilde{b}}(b) = \frac{1}{\alpha} f_{\tilde{a}}\left(\frac{b}{\alpha}\right). \tag{3.68}$$

Proof To derive the pdf of $\tilde{b}$, we compute its cdf and then differentiate it. The cdf can be computed by expressing it in terms of the pdf of $\tilde{a}$:

$$\begin{aligned} F_{\tilde{b}}(b) &:= \mathrm{P}\left(\tilde{b} \leq b\right) && (3.69)\\ &= \mathrm{P}\left(\alpha\tilde{a} \leq b\right) && (3.70)\\ &= \mathrm{P}\left(\tilde{a} \leq \frac{b}{\alpha}\right) && (3.71)\\ &= F_{\tilde{a}}\left(\frac{b}{\alpha}\right). && (3.72) \end{aligned}$$

■

Example 3.21 (Current and voltage)**.** Your friend Mariana, who is an electrical engineer, is modeling the voltage across a resistor in a certain device. She wants to know how the pdf of the voltage, represented by the random variable $\tilde{v}$, depends on the pdf $f_{\tilde{c}}$ of the current, represented by the random variable $\tilde{c}$. Since the voltage is equal to the current multiplied by the resistance r, which is a deterministic quantity, we have $\tilde{v} = r\tilde{c}$. By Theorem 3.20, the pdf of the voltage equals

$$f_{\tilde{v}}(v) = \frac{1}{r} f_{\tilde{c}}\left(\frac{v}{r}\right). \tag{3.73}$$

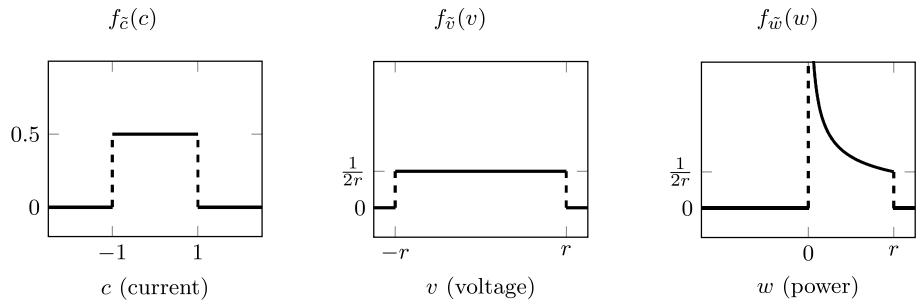

Figure 3.11 Current, voltage and power. As derived in Example 3.21, if the pdf of the current across a resistor is uniform between -1 and 1 amperes (left), then the pdf of the voltage is uniform between $-r$ and r (center), where r denotes the resistance. The pdf of the power (right) is not uniform, because the power is a quadratic function of the current. Instead, it exhibits a square-root decay between 0 and r, as derived in Example 3.22.

To test out the formula, Mariana uses a source that generates current uniformly at random between -1 and 1 amperes (the pdf is shown in the left plot of Figure 3.11). The corresponding pdf of the voltage, depicted in the middle plot of Figure 3.11, is uniform between $-r$ and r volts, as predicted by (3.73).

In the proof of Theorem 3.20, we derive the pdf of a function of a random variable by first deriving its cdf, and then differentiating it. This is an effective strategy in general, as illustrated by the following example.

Example 3.22 (Current and power)**.** Mariana now wants to derive the pdf of the power dissipated in the resistor, modeled as a random variable $\tilde{w}$. The power is equal to the square of the current multiplied by the resistance r, so $\tilde{w} = r\tilde{c}^2$. For $w < 0$, $F_{\tilde{w}}(w) = \mathrm{P}(\tilde{w} \leq w) = 0$, because the power is nonnegative. For $w \geq 0$

$$F_{\tilde{w}}(w) = \mathrm{P}(\tilde{w} \leq w) \tag{3.74}$$
$$= \mathrm{P}\left(r\tilde{c}^2 \leq w\right) \tag{3.75}$$
$$= \mathrm{P}\left(-\sqrt{\frac{w}{r}} \leq \tilde{c} \leq \sqrt{\frac{w}{r}}\right) \tag{3.76}$$
$$= F_{\tilde{c}}\left(\sqrt{\frac{w}{r}}\right) - F_{\tilde{c}}\left(-\sqrt{\frac{w}{r}}\right). \tag{3.77}$$

To compute the pdf, we differentiate the cdf:

$$f_{\tilde{w}}(w) = \frac{\mathrm{d}}{\mathrm{d}w}\left(F_{\tilde{c}}\left(\sqrt{\frac{w}{r}}\right) - F_{\tilde{c}}\left(-\sqrt{\frac{w}{r}}\right)\right) \tag{3.78}$$
$$= \frac{1}{2\sqrt{rw}}\left(f_{\tilde{c}}\left(\sqrt{\frac{w}{r}}\right) + f_{\tilde{c}}\left(-\sqrt{\frac{w}{r}}\right)\right), \tag{3.79}$$

if $w \geq 0$ and 0 otherwise.

The right plot in Figure 3.11 shows the pdf of the power when the current is uniformly distributed between -1 and 1 amperes. In contrast to the voltage, the power is not uniformly distributed. By (3.79) it equals

$$f_{\tilde{w}}(w) = \frac{1}{2\sqrt{rw}}, \tag{3.80}$$

between 0 and r, because $f_{\tilde{c}}\left(\sqrt{\frac{w}{r}}\right)$ and $f_{\tilde{c}}\left(-\sqrt{\frac{w}{r}}\right)$ equal $1/2$ for $0 \leq w \leq r$.

To end the section, we study an interesting phenomenon. If we feed a continuous random variable $\tilde{a}$ into its cdf $F_{\tilde{a}}$, the resulting random variable $F_{\tilde{a}}(\tilde{a})$ always has the same distribution. Regardless of the distribution of $\tilde{a}$, $F_{\tilde{a}}(\tilde{a})$ is uniformly distributed in the unit interval. This is known as the *probability integral transform*, which turns out to be very useful for the analysis of p values in hypothesis testing (see Theorem 10.13).

Theorem 3.23 (Probability integral transform)**.** *Let $\tilde{a}$ be a continuous random variable with cdf $F_{\tilde{a}}$. The random variable*

$$\tilde{b} := F_{\tilde{a}}(\tilde{a}) \tag{3.81}$$

is uniformly distributed in the unit interval $[0, 1]$.

Proof By definition, the cdf of $\tilde{b}$ equals

$$F_{\tilde{b}}(b) = \mathrm{P}\left(\tilde{b} \leq b\right) \tag{3.82}$$

$$= \mathrm{P}\left(F_{\tilde{a}}\left(\tilde{a}\right) \leq b\right). \tag{3.83}$$

Let us assume for a moment that $F_{\tilde{a}}$ is invertible. Since cdfs are nondecreasing by Lemma 3.4, $a \leq b$ is equivalent to $F_{\tilde{a}}(a) \leq F_{\tilde{a}}(b)$. Consequently, the event $F_{\tilde{a}}(\tilde{a}) \leq b$ is the same as the event $\tilde{a} \leq F_{\tilde{a}}^{-1}(b)$, which implies

$$F_{\tilde{b}}(b) = \mathrm{P}\left(\tilde{a} \leq F_{\tilde{a}}^{-1}(b)\right) \tag{3.84}$$

$$= F_{\tilde{a}}\left(F_{\tilde{a}}^{-1}(b)\right) \tag{3.85}$$

$$= b, \quad 0 \leq b \leq 1, \tag{3.86}$$

establishing that $\tilde{b}$ is indeed uniformly distributed in $[0, 1]$.

If $F_{\tilde{a}}$ is not invertible, we can define the generalized inverse of the cdf as

$$F_{\tilde{a}}^{-1}(b) := \min_{x} \left\{x : F_{\tilde{a}}(x) = b\right\}. \tag{3.87}$$

The generalized inverse is well defined, as long as the cdf is continuous (otherwise the set $\{x : F_{\tilde{a}}(x) = b\}$ could be empty). Since (3.84) and (3.85) hold for the generalized inverse, $\tilde{b}$ is uniformly distributed in $[0, 1]$ even if the cdf of $\tilde{a}$ is not invertible. ■

3.5 Nonparametric Probability-Density Estimation

In this section we consider *nonparametric* approaches to estimate a pdf from data, which means that they do not assume an underlying parametric model.

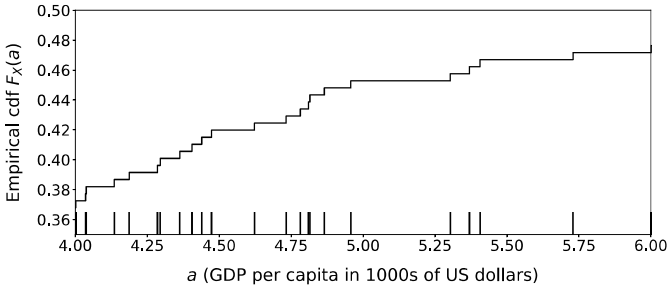

Figure 3.12 The empirical cdf is piecewise constant. Zoomed in plot of the empirical cdf in the bottom row of Figure 3.6, which corresponds to the gross domestic product per capita of 212 countries in 2019. The small vertical lines at the bottom represent the observed data. The empirical cdf at a point a is equal to the fraction of data smaller than a, so its value only changes if a is one of the data points. As a result it is piecewise constant.

3.5.1 The Histogram

In Section 3.2.3 we show how to estimate the cdf from data using the empirical cdf. In order to estimate the corresponding pdf, you might be tempted to differentiate the empirical cdf, but unfortunately this is impossible. The empirical cdf is piecewise constant, as illustrated in Figure 3.12, and therefore not differentiable.

To derive a statistical estimator for the pdf, let us consider its intuitive definition. For any pdf $f_{\tilde{a}}$ of a continuous random variable $\tilde{a}$,

$$\mathrm{P}(a - \epsilon < \tilde{a} \leq a) \approx \epsilon f_{\tilde{a}}(a), \tag{3.88}$$

when ϵ is small enough. Inspired by this, we can apply the following strategy:

1 Divide the possible values of the random variable into short segments, which we call bins or buckets.
2 Approximate the probability of the random variable being in each bin using the empirical-probability estimator.
3 Compute the density in each bin dividing the estimated probability by its length (assuming that the density is constant within each bin).

This strategy is equivalent to computing the histogram of the data and then normalizing it, in order to ensure that it integrates to one and is therefore a valid pdf.

Definition 3.24 (Histogram of continuous data)**.** *Let $X := \{x_1, x_2, \ldots, x_n\}$ be a dataset with values in an interval $[m, m + \ell] \subseteq \mathbb{R}$ of the real line with length ℓ (we can choose m to be any real number smaller than the minimum value in X). To build a histogram of the data we divide the interval into b* bins *or* buckets *of*

length ℓ/b:

$$\mathcal{B}_i := \left[m + \frac{(i-1)\ell}{b}, m + \frac{i\ell}{b} \right), \quad 1 \leq i \leq b. \tag{3.89}$$

We then count how many elements of X are in each bin. The count for the ith bin equals

$$c_i := \sum_{j=1}^{n} 1\left(x_j \in \mathcal{B}_i\right), \quad 1 \leq i \leq b, \tag{3.90}$$

where $1\left(x_j \in \mathcal{B}_i\right)$ is an indicator function that is equal to one if x_j is in $\mathcal{B}_i$ and to zero otherwise.

The histogram can be normalized to provide an estimate of the pdf of the data f_{hist}. For any $t \in \mathcal{B}_i$ and any $1 \leq i \leq b$,

$$f_{\text{hist}}(t) := \frac{b}{n\ell} \sum_{j=1}^{n} 1\left(x_j \in \mathcal{B}_i\right). \tag{3.91}$$

The following lemma confirms that we have chosen the right normalization, and the normalized histogram indeed integrates to one.

Lemma 3.25. *The normalized histogram defined in Definition 3.24 is a valid pdf.*

Proof From the definition, the estimated pdf f_{hist} is nonnegative, so by Theorem 3.18 we only need to check that it integrates to one. The width of each bin $\int_{t \in \mathcal{B}_i} \mathrm{d}t$ is ℓ/b, which implies

$$\int_{t \in \mathbb{R}} f_{\text{hist}}(t)\, \mathrm{d}t = \sum_{i=1}^{b} \int_{t \in \mathcal{B}_i} \frac{b}{n\ell} \sum_{j=1}^{n} 1\left(x_j \in \mathcal{B}_i\right) \mathrm{d}t \tag{3.92}$$

$$= \frac{b}{n\ell} \sum_{i=1}^{b} \sum_{j=1}^{n} 1\left(x_j \in \mathcal{B}_i\right) \int_{t \in \mathcal{B}_i} \mathrm{d}t \tag{3.93}$$

$$= \frac{1}{n} \sum_{j=1}^{n} \sum_{i=1}^{b} 1\left(x_j \in \mathcal{B}_i\right) \tag{3.94}$$

$$= \frac{1}{n} \sum_{i=1}^{n} 1 = 1, \tag{3.95}$$

where $\sum_{i=1}^{b} 1\left(x_j \in \mathcal{B}_i\right) = 1$ because each data point is in exactly one of the bins. ■

A key consideration when building a histogram is how to choose the number of bins, or equivalently the length of the bins. If the bins are too large, then the density estimate will be very coarse, and therefore not very informative. If the bins are too small, then they will contain very few data points, resulting in a poor estimate of the probability in each bin with many noisy fluctuations. This is

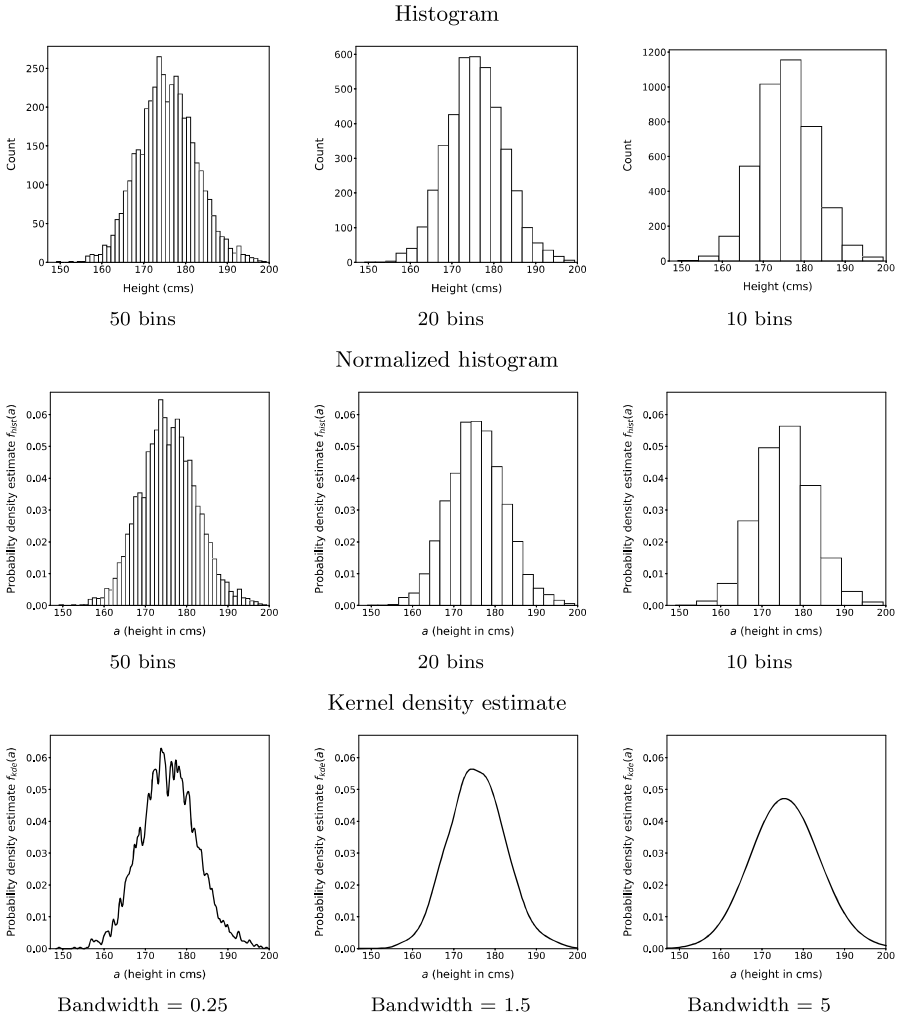

Figure 3.13 Density estimation for height in US army. The top row shows shows histograms of the heights of 4,082 men in the United States army with different number of bins. The corresponding pdf estimates obtained by normalizing the histograms are shown on the second row. The third row shows pdf estimates obtained via kernel-density estimation using Gaussian kernels with different bandwidths. Using many bins in the histogram and a small bandwidth in KDE results in noisy estimates. Decreasing the number of bins and increasing the bandwidth smooths the pdf estimates, but may also eliminate informative structure.

illustrated by the top two rows in Figures 3.13 and 3.14, which show histograms of heights in the United States army and of GDP per capita (using the same data from Datasets 5 and 6 as in Figure 3.6) for different numbers of bins.

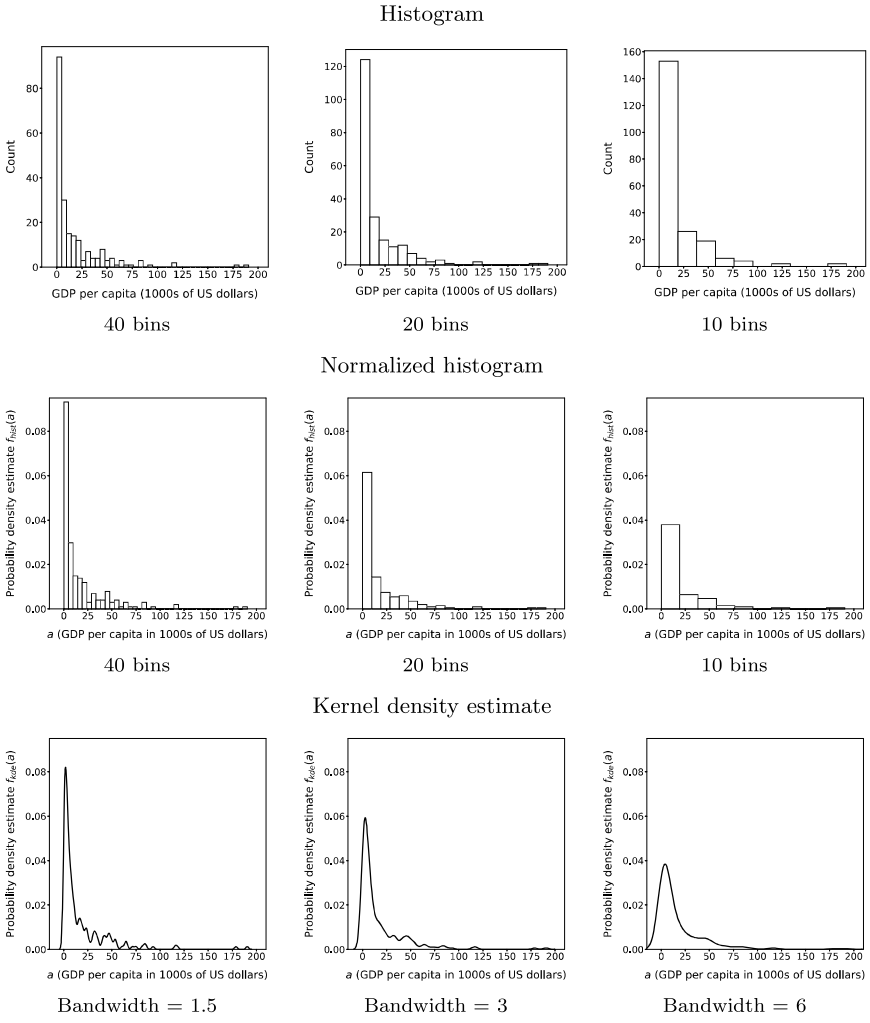

Figure 3.14 Density estimation of GDP per capita. The top row shows shows histograms of the gross domestic product per capita of 212 countries in 2019 with different number of bins. The corresponding pdf estimates obtained by normalizing the histograms are shown on the second row. The third row shows pdf estimates obtained via kernel-density estimation using Gaussian kernels with different bandwidths. Using many bins in the histogram and a small bandwidth in KDE results in noisy estimates. Decreasing the number of bins and increasing the bandwidth smooths the pdf estimates, but may also eliminate informative structure.

3.5.2 Kernel Density Estimation

The pdf estimate provided by the normalized histogram can be decomposed into the contributions of each data point. Given a dataset $X := \{x_1, x_2, \ldots, x_n\}$,

$$f_{\text{hist}}(t) = \sum_{j=1}^{n} \Pi_j(t), \tag{3.96}$$

where Π_j is a rectangle of length ℓ/b that is nonzero only within the bin $\mathcal{B}_j$ where the data point x_j is located

$$\Pi_j(t) = \begin{cases} \frac{b}{n\ell} & \text{for } t \in \mathcal{B}_j, \\ 0 & \text{otherwise.} \end{cases} \tag{3.97}$$

There are n such rectangles, so to ensure that their total area adds up to one, their height must be $\frac{b}{n\ell}$. You can check that (3.96) is exactly equivalent to Definition 3.24. In words, the normalized histogram estimates the pdf through a superposition of small rectangles centered at the bin centers.

You might be wondering why the rectangles are centered at the bin centers. This has a clear drawback: when a point is near the edge of the bin, then it is used to estimate the density over an interval that is shifted with respect to the point. To address this, we can center the rectangle at the data point. This approach is known as *kernel density estimation* (KDE). Here the *kernel* is the function Π_j used to decompose the pdf. There is no need for the kernel to be rectangular. In fact, it makes sense for it to decay away from its center, so that the influence of each data point on the density estimate is higher the closer we are to the point.

Definition 3.26 (Kernel density estimator)**.** *Let $X := \{x_1, x_2, \ldots, x_n\}$ denote a real-valued dataset. The corresponding kernel density estimate at a point $a \in \mathbb{R}$ is*

$$f_{X,h}(a) := \frac{1}{n\,h} \sum_{i=1}^{n} K\left(\frac{a - x_i}{h}\right), \tag{3.98}$$

where $K : \mathbb{R} \to \mathbb{R}$ is a kernel function centered at the origin that satisfies

$$K(a) \geq 0 \quad \text{for all } a \in \mathbb{R}, \tag{3.99}$$

$$\int_{\mathbb{R}} K(a)\, \mathrm{d}a = 1. \tag{3.100}$$

The parameter $h \geq 0$ determines the width of the kernel function.

A popular choice for the kernel is the Gaussian function

$$K(a) := \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{a^2}{2}\right), \tag{3.101}$$

which is smooth and decays rapidly away from its center. Figure 3.15 shows a simple example that illustrates KDE with rectangular and Gaussian kernels.

KDE estimates the density based on a weighted local average of the data points.

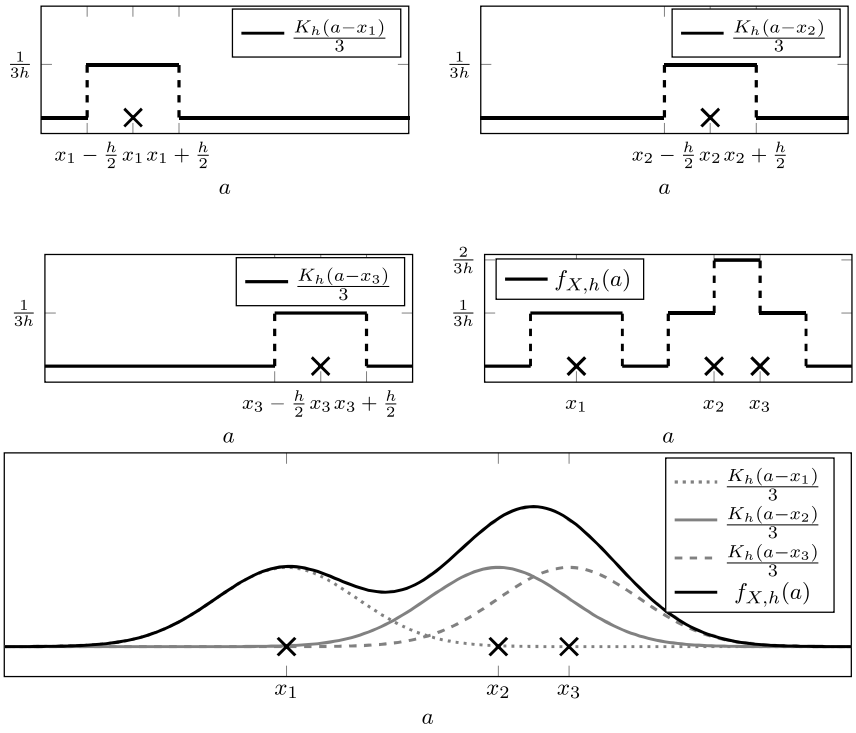

Figure 3.15 Kernel density estimation. In this example the dataset $X := \{x_1, x_2, x_3\}$ consists of three points. The top two rows show a KDE estimate using rectangular kernels. The three first plots show the separate components corresponding to each individual data point. The fourth plot shows their sum, which is the resulting pdf estimate. Below, a Gaussian kernel is applied to obtain a pdf estimate from the same data.

As in the case of the histogram, an important consideration is the spatial extent of this average, which is governed by the bandwidth h of the kernel. Increasing h dilates the kernel, so that the density estimate takes into account more points. If the bandwidth is very small, individual samples have a large influence on the density estimate. This enables us to capture irregular structure more easily, but may also overfit spurious fluctuations if we don't have a lot of data. Increasing the bandwidth smooths out such fluctuations and yields more stable estimates. However making h too large results in over-smoothing, and can eliminate meaningful structure from the estimate. The tradeoff is illustrated in the bottom row of Figures 3.13 and 3.14.

3.6 Continuous Parametric Distributions

Directly estimating the cdf or the pdf from data can be inaccurate when the number of available data are limited. As discussed in Section 2.3, we can address this issue by building parametric models that incorporate assumptions about the quantity of interest. In the case of continuous random variables, we design a pdf that only depends on a small number of parameters, which are then estimated from the data as described in Section 3.7. In this section, we describe two of the most popular continuous parametric models: the exponential distribution (Section 3.6.1) and the Gaussian distribution (Section 3.6.2).

3.6.1 The Exponential Distribution

Exponential parametric models are often used to represent the time between intermittent phenomena such as earthquakes, telephone calls, radioactive decay of particles, or neuronal impulses. We describe the assumptions underlying the exponential model and derive its pdf in the following example.

Example 3.27 (Time until the next earthquake)**.** We consider the problem of modeling earthquakes occurring in the San Francisco Bay Area, as in Example 2.21. We are interested in the distribution of the time until the next earthquake.

In our derivation of the Poisson distribution, one of the main assumptions is that for small enough ϵ, the probability of an earthquake occurring in a period of length ϵ is equal to $\lambda\epsilon$ (and the probability of more than one earthquake is negligible), where λ is a fixed parameter quantifying the rate at which earthquakes occur. Let $\tilde{t}$ be a continuous random variable that represents the time until the next earthquake. We can interpret the assumption in the following way: If no earthquake occurs by time t, then the probability that the earthquake occurs between t and $t + \epsilon$ is $\lambda\epsilon$. More formally,

$$\mathrm{P}\left(t \leq \tilde{t} \leq t + \epsilon \,|\, \tilde{t} > t\right) \approx \lambda\epsilon, \tag{3.102}$$

with equality when $\epsilon \to 0$. We write the conditional probability in terms of the *survival function* $S(t) := 1 - F_{\tilde{t}}(t)$:

$$\mathrm{P}\left(t < \tilde{t} \leq t + \epsilon \,|\, \tilde{t} > t\right) = \frac{\mathrm{P}\left(t < \tilde{t} \leq t + \epsilon, \tilde{t} > t\right)}{\mathrm{P}\left(\tilde{t} > t\right)} \tag{3.103}$$

$$= \frac{F_{\tilde{t}}(t + \epsilon) - F_{\tilde{t}}(t)}{1 - F_{\tilde{t}}(t)} \tag{3.104}$$

$$= \frac{S(t) - S(t + \epsilon)}{S(t)} \approx \lambda\epsilon. \tag{3.105}$$

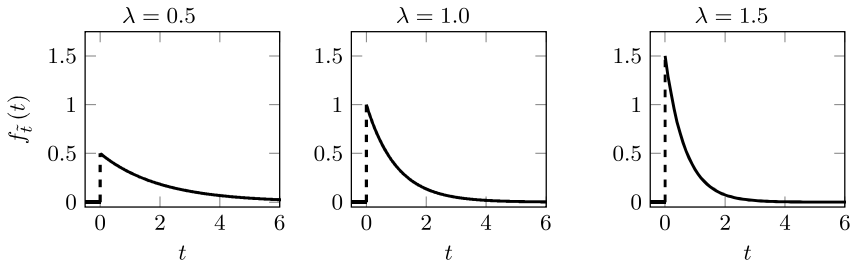

Figure 3.16 Exponential distribution. Probability density functions of exponential random variables with different parameter values.

Reordering the terms in the equation and taking the limit when $\epsilon \to 0$, we have

$$-\lambda = \frac{1}{S(t)} \lim_{\epsilon \to 0} \frac{S(t+\epsilon) - S(t)}{\epsilon} \tag{3.106}$$

$$= \frac{1}{S(t)} \frac{\mathrm{d}S(t)}{\mathrm{d}t} \tag{3.107}$$

$$= \frac{\mathrm{d}\log S(t)}{\mathrm{d}t}. \tag{3.108}$$

Now, integrating on both sides and taking exponentials we obtain

$$c \exp(-\lambda t) = S(t) \tag{3.109}$$

$$= 1 - F_{\tilde{a}}(t), \tag{3.110}$$

for some constant c. Since we are measuring time starting at zero, $F_{\tilde{a}}(0) = \mathrm{P}(\tilde{t} \leq 0) = 0$, which implies $c = 1$. Differentiating the cdf to derive the pdf of the time to the next earthquake reveals that it is an exponential function of t:

$$F_{\tilde{a}}(t) = 1 - \exp(-\lambda t), \tag{3.111}$$

$$f_{\tilde{a}}(t) = \frac{\mathrm{d}F_{\tilde{a}}(t)}{\mathrm{d}t} = \lambda \exp(-\lambda t). \tag{3.112}$$

Random variables with the pdf derived in Example 3.27 are called exponential random variables. Figure 3.16 shows several examples.

Definition 3.28 (Exponential distribution)**.** *An exponential random variable $\tilde{t}$ with parameter λ has a pdf of the form*

$$f_{\tilde{t}}(t) = \begin{cases} \lambda e^{-\lambda t}, & \text{if } t \geq 0, \\ 0, & \text{otherwise.} \end{cases} \tag{3.113}$$

Example 3.29 (Interarrival times at a call center)**.** In this example we study Dataset 3, which consists of calls arriving at the telephone call-center of an anonymous bank in Israel. Our goal is to estimate the distribution of the times between

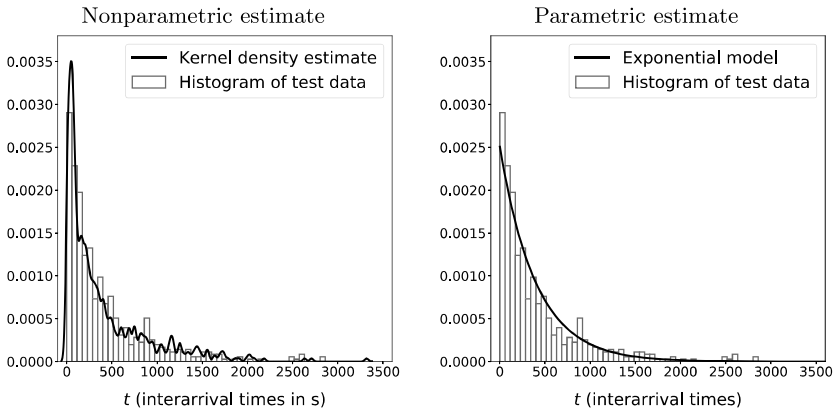

Figure 3.17 Parametric vs. nonparametric model for interarrival times at a call center. The left graph shows a nonparametric estimate of the pdf obtained by applying KDE to a training set containing data from January to June, as well as the histogram of a test set containing data from July to December. The right column shows a parametric estimate obtained via maximum-likelihood estimation based on an exponential model (see Section 3.7 and Figure 3.22), again compared to the histogram of the test data.

calls that arrive between 6 am and 7 am on weekdays. We model the interarrival time as a continuous random variable. We apply kernel density estimation to obtain a nonparametric estimate of the pdf, as well as maximum-likelihood estimation to obtain a parametric estimate based on an exponential model (see Section 3.7 and Figure 3.22). In order to evaluate these estimates, we build a training set using data from January to June and use the remaining months (July to December) as a test set.

Figure 3.17 shows the results. Both models provide a good approximation to the test data. It turns out that there is a surprisingly large number of short interarrival times in the training data. The parametric exponential model is not able to fit this pattern while remaining consistent with the rest of the interarrival times, so it underestimates the density in that region. This is an example of how parametric models may underfit the data when their underlying assumptions do not hold. Interestingly, the test set does not have as many short interarrival times as the training data, so the KDE estimator is also not accurate in that region.

An important property of the exponential distribution is that it is *memoryless*. Let $\tilde{t}$ be an exponential random variable, which we know is larger than a certain value t_0. Then, conditioned on the event $\tilde{t} > t_0$, it turns out that the distribution of $t - t_0$ is also exponential (starting at t_0 instead of at zero). Assume $\tilde{t}$ represents the time you wait until someone answers the phone. If $\tilde{t}$ is memoryless, then no matter how long you have waited, the distribution of the time you need to wait

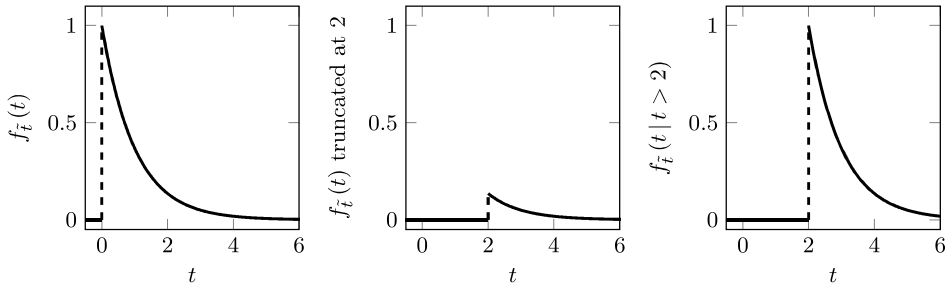

Figure 3.18 The exponential distribution is memoryless. Graphical explanation of why the exponential distribution is memoryless. Truncating the pdf at t_0 yields an exponential curve. Normalizing the curve so that it is a valid density results in a shifted exponential pdf.

from that moment is always the same (if you have ever tried to call the customer services of an airline to obtain reimbursement for a cancelled flight, you know the feeling).

Lemma 3.30 (The exponential distribution is memoryless)**.** *Let $\tilde{t}$ be a random variable distributed according to an exponential pdf with parameter λ. We define the cdf of $\tilde{t}$ conditioned on the event $\tilde{t} > t_0$ as*

$$F_{\tilde{t}}(t \,|\, \tilde{t} > t_0) := \mathrm{P}(\tilde{t} \leq t \,|\, \tilde{t} > t_0) \tag{3.114}$$

and the pdf conditioned on the same event as

$$f_{\tilde{t}}(t \,|\, \tilde{t} > t_0) := \frac{\mathrm{d}F_{\tilde{t}}(t \,|\, \tilde{t} > t_0)}{\mathrm{d}t}. \tag{3.115}$$

For any $t_0 \geq 0$, this conditional pdf is a copy of $f_{\tilde{t}}$ shifted to t_0,

$$f_{\tilde{t}}(t \,|\, \tilde{t} > t_0) = \lambda \exp(-\lambda(t - t_0)). \tag{3.116}$$

Proof The conditional cdf of $\tilde{t}$ given $\tilde{t} \geq t_0$ evaluated at $t > t_0$ is

$$\begin{aligned} F_{\tilde{t}\,|\,\tilde{t}>t_0}(t) &= \mathrm{P}\left(\tilde{t} \leq t \,|\, \tilde{t} > t_0\right) && (3.117)\\ &= \frac{\mathrm{P}\left(t_0 < \tilde{t} \leq t\right)}{\mathrm{P}\left(\tilde{t} > t_0\right)} && (3.118)\\ &= \frac{F_{\tilde{t}}(t) - F_{\tilde{t}}(t_0)}{1 - F_{\tilde{t}}(t_0)} && (3.119)\\ &= \frac{e^{-\lambda t_0} - e^{-\lambda t}}{e^{-\lambda t_0}} && (3.120)\\ &= 1 - e^{-\lambda(t-t_0)}. && (3.121) \end{aligned}$$

Differentiating with respect to t yields an exponential pdf $f_{\tilde{t}|\tilde{t}>t_0}(t) = \lambda e^{-\lambda(t-t_0)}$ starting at t_0. ■

Figure 3.18 provides a graphical explanation of the memoryless property. If we truncate the exponential pdf at t_0, we obtain an exponential curve, but it is not a pdf because it does not integrate to one. Renormalizing it ensures that it does, and is therefore a valid pdf.

Figure 3.19 shows that memoryless structure can arise in real data. We select calls from Dataset 3 that occurred between 9 and 10 am during weekdays, and consider the interarrival times between calls. We approximate the pdf of the interarrival times that are larger than t_0 for different values of t_0, via kernel-density estimation. The resulting pdfs are quite similar, indicating that their distribution is indeed approximately memoryless.

3.6.2 The Gaussian Distribution

The Gaussian or normal model is arguably the most popular parametric model in all of probability and statistics. It is used all over the place in the natural sciences and in engineering. The reason is that variables that are sums of independent quantities tend to have a Gaussian distribution. We describe this phenomenon, known as the central limit theorem, in more detail in Section 9.7.

Definition 3.31 (Gaussian distribution)**.** *A Gaussian or normal random variable with mean μ and standard deviation $\sigma \geq 0$ has a pdf of the form*

$$f_{\tilde{a}}(a) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(a-\mu)^2}{2\sigma^2}}. \tag{3.122}$$

The squared standard deviation σ^2 is called the variance. A Gaussian distribution with mean μ and variance σ^2 is usually denoted by $\mathcal{N}(\mu, \sigma^2)$.

We have not yet defined the mean and variance of a random variable. They represent the average of the distribution and the average squared deviation from the mean, respectively, as explained in Sections 7.1 and 7.7. Figure 3.20 shows the pdfs of Gaussian random variables with different values of μ and σ. The bell-shaped pdf is centered at the mean μ. The standard deviation parameter σ determines how concentrated the density is around μ.

If we shift and scale a Gaussian random variable, it remains Gaussian. Shifting modifies the mean, and scaling modifies the standard deviation.

Theorem 3.32 (Shifting and scaling a Gaussian random variable)**.** *Let $\tilde{a}$ be a Gaussian random variable with mean μ and variance σ^2. The random variable*

$$\tilde{b} := \alpha\tilde{a} + \beta \tag{3.123}$$

is a Gaussian random variable with mean $\alpha\mu + \beta$ and variance $\alpha^2\sigma^2$.

Proof For simplicity, let us assume that $\alpha > 0$ (the same argument can be

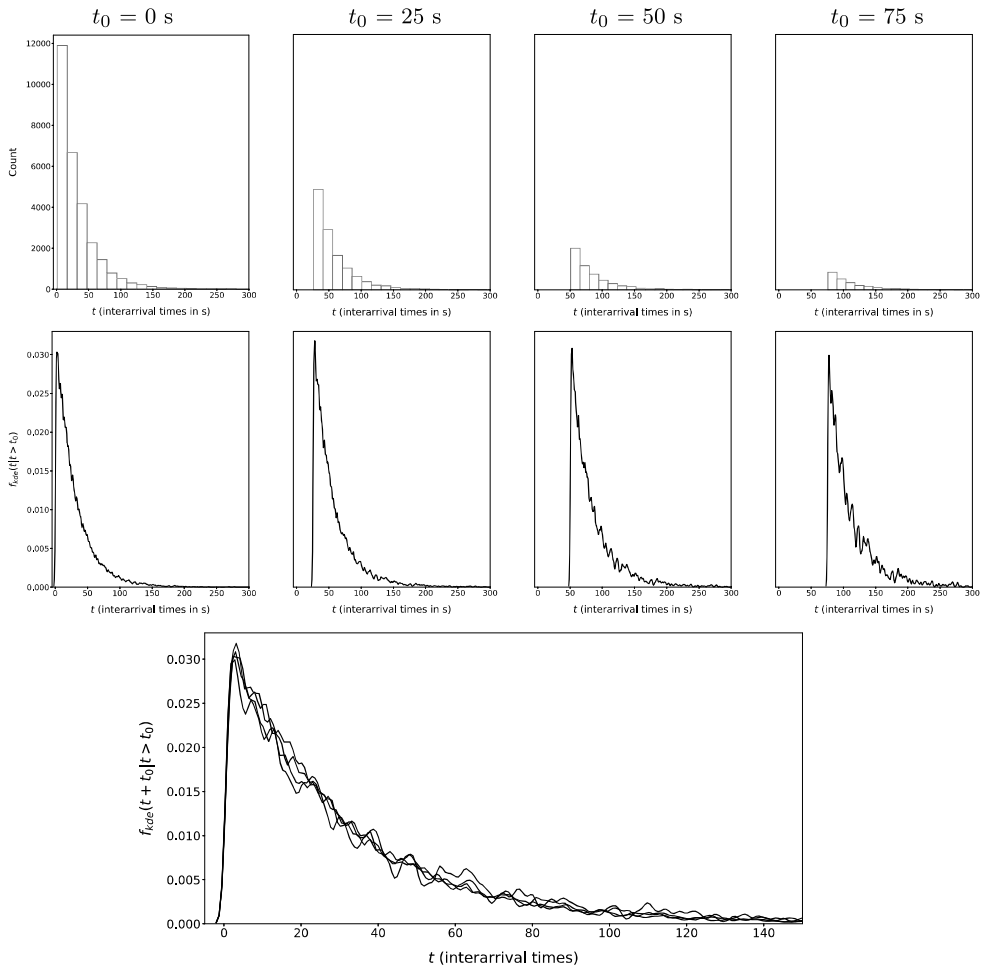

Figure 3.19 Memoryless property in real data. We consider calls arriving at the bank call center from Dataset 3 on weekdays from 9 am to 10 am. The top row shows the histograms of interarrival times larger than t_0 for different values of t_0. The corresponding estimates of the conditional pdf obtained via kernel-density estimation are shown below. The plot at the bottom shows the superposition of the four conditional pdfs shifted to lie on top of each other, demonstrating that they are very similar, which indicates that the interarrival times in that range are approximately memoryless.

applied if $\alpha < 0$ with minor modifications). By the definition of cdf,

$$F_{\tilde{b}}(b) = \mathrm{P}(\alpha\tilde{a} + \beta \leq b) \tag{3.124}$$

$$= \mathrm{P}\left(\tilde{a} \leq \frac{b-\beta}{\alpha}\right) \tag{3.125}$$

$$= \int_{-\infty}^{\frac{b-\beta}{\alpha}} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(a-\mu)^2}{2\sigma^2}\right) \mathrm{d}a. \tag{3.126}$$

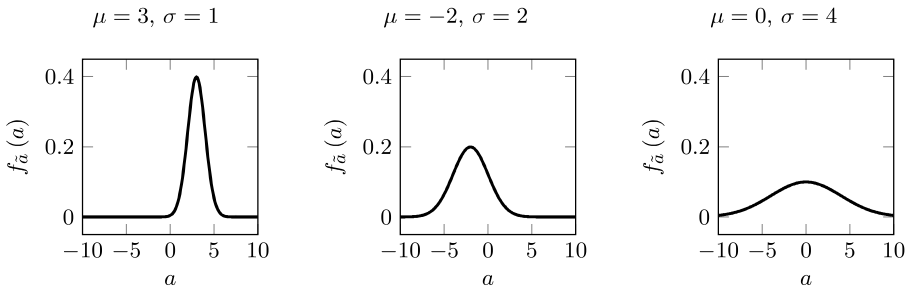

Figure 3.20 Gaussian distribution. Probability density functions of Gaussian random variables with different means and standard deviations.

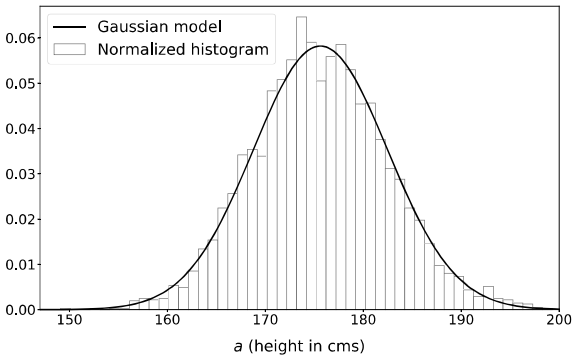

Figure 3.21 Gaussian parametric model for height in US army. The graph shows a Gaussian parametric pdf estimate obtained by fitting the heights of 4,082 men in the United States army using maximum-likelihood estimation. The pdf is shown superimposed onto the histogram of the data used to fit the model.

Differentiating with respect to b yields

$$f_{\tilde{b}}(b) = \frac{1}{\sqrt{2\pi}\alpha\sigma} \exp\left(-\frac{(\frac{b-\beta}{\alpha} - \mu)^2}{2\sigma^2}\right) \tag{3.127}$$

$$= \frac{1}{\sqrt{2\pi}\alpha\sigma} \exp\left(-\frac{(b - \alpha\mu - \beta)^2}{2\alpha^2\sigma^2}\right). \tag{3.128}$$

■

Gaussian random variables are often used to model continuous quantities that have approximately bell-shaped histograms, such as the height data in Figure 3.13. Figure 3.21 shows the result of fitting a Gaussian parametric model to these data via maximum-likelihood estimation (see Section 3.7). The fit is quite good, although the underlying data is not completely symmetric around the center of the distribution, as can be seen from the histogram.

3.7 Maximum-Likelihood Estimation

In this section we explain how to fit parametric models based on a predefined pdf to data. Let f_θ be a nonnegative real-valued function depending on a parameter vector θ. We assume that for any fixed θ in a certain set S the function integrates to one, $\int_{\mathbb{R}} f_\theta(a)\, \mathrm{d}a = 1$, and can therefore be interpreted as a pdf. Given a data point a, the pdf $f_\theta(a)$ at a quantifies how likely we are to observe a data point in that vicinity under the parametric model (the probability of a data point in $[a - \epsilon, a]$ for small ϵ is approximately $f_\theta(a)\epsilon$). It is therefore reasonable to choose θ to make the density as high as possible. This is analogous to our derivation of maximum likelihood for discrete models, although here we maximize the density with respect to the parameters instead of the probability.

In order to perform an estimate based on n data points, $x_1, x_2, \ldots, x_n$, we need to make assumptions about the data-generating process. As in the discrete setting, we assume that the observations are mutually independent, and identically distributed according to the parametric model f_θ. For continuous distributions, the i.i.d. assumption implies that the probability density of the data under the parametric model is $\prod_{i=1}^n f_\theta(x_i)$, as we explain in detail in Section 5.7 (see Definition 5.17). This product is often very small, so to avoid numerical instabilities we typically compute its logarithm instead.

Definition 3.33 (Likelihood function)**.** *Let $f_\theta : \mathbb{R} \to \mathbb{R}^+$ be a parametric pdf model, and $X := \{x_1, x_2, \ldots, x_n\}$ a real-valued dataset. The likelihood of the model given these data under i.i.d. assumptions is*

$$\mathcal{L}_X(\theta) := \prod_{i=1}^n f_\theta\left(x_i\right). \tag{3.129}$$

The log-likelihood function is equal to the logarithm of the likelihood function,

$$\log \mathcal{L}_X(\theta) = \sum_{i=1}^n \log f_\theta\left(x_i\right). \tag{3.130}$$

Maximum-likelihood estimation selects the value of the parameters that maximizes the likelihood or the log-likelihood, and therefore the density of the observed data according to the parametric model under i.i.d. assumptions. Maximizing the likelihood or the log-likelihood is equivalent because the logarithm is a monotone function.

Definition 3.34 (Maximum-likelihood estimator)**.** *Let $f_\theta : \mathbb{R} \to \mathbb{R}^+$ be a parametric model dependent on a parameter vector θ, $X := \{x_1, x_2, \ldots, x_n\}$ a real-valued dataset, and S the set of parameter values for which f_θ is a valid pmf. The maximum-likelihood estimate of θ is*

$$\theta_{\mathrm{ML}} := \arg\max_{\theta \in S} \mathcal{L}_X(\theta) \tag{3.131}$$

$$= \arg\max_{\theta \in S} \log \mathcal{L}_X(\theta). \tag{3.132}$$

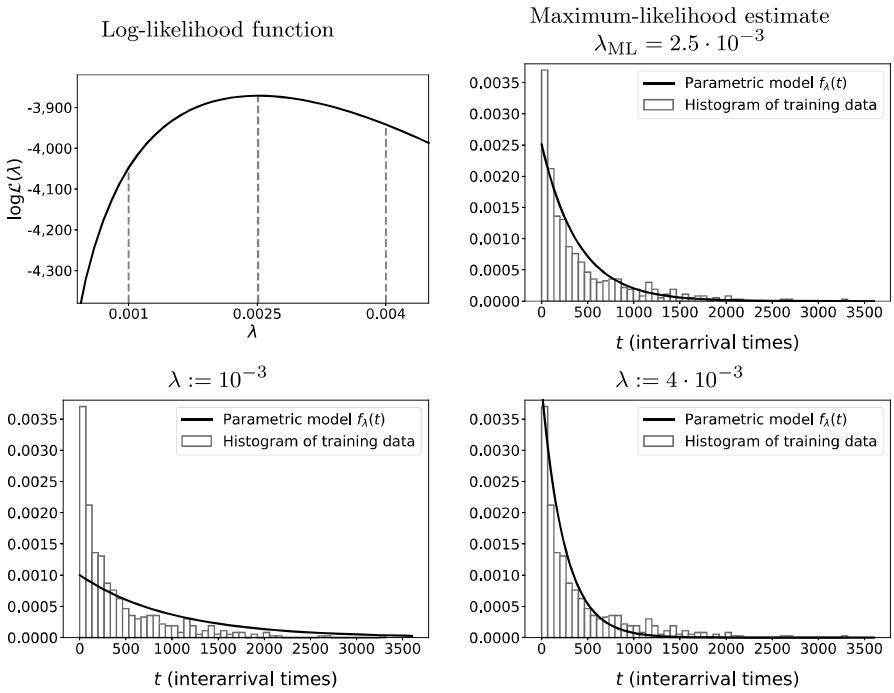

Figure 3.22 Exponential model for interarrival times at a call center. Log-likelihood function of an exponential parametric model applied to the real dataset described in Example 3.29 (top left) and corresponding fit to the data obtained via maximum likelihood, superimposed onto the histogram of the data used to fit the model (top right). The two bottom graphs show the fits corresponding to two other choices of the parameter λ, which produce much worse fits. The parametric model is not able to fit the high number of calls with very short durations (leftmost bin of the histogram), because it is not consistent with the rest of the data under the assumptions of the exponential model.

Theorem 3.35 (Maximum-likelihood estimator for the exponential distribution)**.** *Let $X := \{x_1, x_2, \ldots, x_n\}$ denote a real-valued dataset. The maximum-likelihood estimator of the parameter of the exponential distribution under i.i.d assumptions equals*

$$\lambda_{\text{ML}} = \frac{1}{\frac{1}{n}\sum_{i=1}^{n} x_i}. \tag{3.133}$$

Proof The log-likelihood is

$$\log \mathcal{L}_X(\lambda) = \sum_{i=1}^{n} \log f_\lambda(x_i) \tag{3.134}$$

$$= \sum_{i=1}^{n} \log \lambda \exp(-\lambda x_i) \tag{3.135}$$

$$= n \log \lambda - \lambda \sum_{i=1}^{n} x_i. \tag{3.136}$$

The derivative and second derivative of the log-likelihood function are

$$\frac{\mathrm{d} \log \mathcal{L}_{x_1,\ldots,x_n}(\lambda)}{\mathrm{d}\lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i, \tag{3.137}$$

$$\frac{\mathrm{d}^2 \log \mathcal{L}_{x_1,\ldots,x_n}(\lambda)}{\mathrm{d}\lambda^2} = -\frac{n}{\lambda^2} < 0 \quad \text{for all } \lambda > 0. \tag{3.138}$$

The function is concave, as the second derivative is negative, so there cannot be different local maxima. The maximum is obtained by setting the first derivative equal to zero. ■

Figure 3.22 shows the log-likelihood of an exponential model applied to the same call-center data from Dataset 3 used in Example 3.29. The figure also shows the corresponding maximum-likelihood fit and the fits produced by other values of the λ parameter. The fit achieved via maximum likelihood is much better adapted to the data.

The following theorem derives maximum-likelihood estimators for the parameters of a Gaussian distribution. The estimate of the mean parameter is the average of the data, which is also known as the sample mean, and is a standard estimator for the mean of a distribution (see Section 7.1.4). The estimate for the variance is obtained by averaging the squared deviation from the mean, which yields an estimator that is very close to the sample variance, as defined in Section 7.7.2.

Theorem 3.36 (Maximum-likelihood estimator for the Gaussian distribution)**.** *Let $X := \{x_1, x_2, \ldots, x_n\}$ denote a real-valued dataset. The maximum-likelihood estimators of the parameters of the Gaussian distribution under i.i.d assumptions equal*

$$\mu_{ML} = \frac{1}{n} \sum_{i=1}^{n} x_i, \tag{3.139}$$

$$\sigma^2_{ML} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu_{ML})^2. \tag{3.140}$$

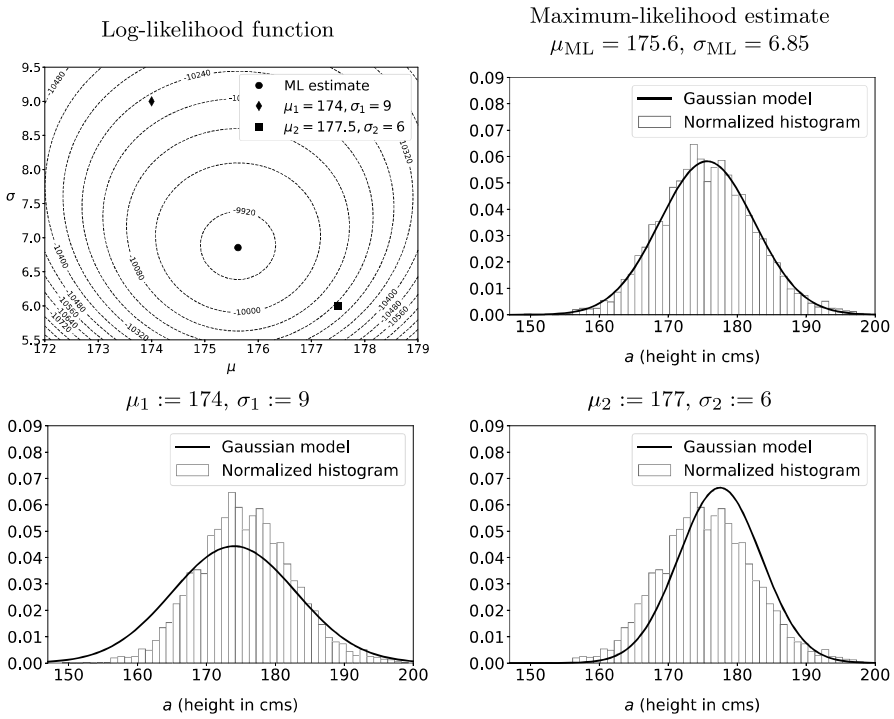

Figure 3.23 Gaussian model for height in US army. The contour plot in the top-left corner shows the log-likelihood function of a Gaussian parametric model applied to height data from 4,082 men in the United States army (see Figure 3.13). The graph in the top-right corner shows the corresponding parametric fit obtained via maximum likelihood, superimposed onto the histogram of the data used to fit the model. The two bottom graphs show the fits corresponding to two other choices of the parameter λ.

Proof The likelihood function is equal to

$$\mathcal{L}_{\{x_1,\ldots,x_n\}}(\mu,\sigma) = \prod_{i=1}^{n} f_{\mu,\sigma}(x_i) \tag{3.141}$$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \tag{3.142}$$

and the log-likelihood to

$$\log \mathcal{L}_{\{x_1,\ldots,x_n\}}(\mu,\sigma) = -\frac{n\log(2\pi)}{2} - n\log\sigma - \sum_{i=1}^{n}\frac{(x_i-\mu)^2}{2\sigma^2}. \tag{3.143}$$

The maximum-likelihood estimators of the parameters μ and σ are

$$\{\mu_{\text{ML}}, \sigma_{\text{ML}}\} = \arg\max_{\{\mu,\sigma\}} \log \mathcal{L}_{\{x_1,\ldots,x_n\}}(\mu, \sigma) \tag{3.144}$$

$$= \arg\max_{\{\mu,\sigma\}} -n \log \sigma - \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2}. \tag{3.145}$$

The first and second partial derivative of the log-likelihood function with respect to μ equal

$$\frac{\partial \log \mathcal{L}_{\{x_1,\ldots,x_n\}}(\mu, \sigma)}{\partial \mu} = \sum_{i=1}^{n} \frac{x_i - \mu}{\sigma^2}, \tag{3.146}$$

$$\frac{\partial^2 \log \mathcal{L}_{\{x_1,\ldots,x_n\}}(\mu, \sigma)}{\partial \mu^2} = -\frac{n}{\sigma^2}. \tag{3.147}$$

For a fixed value of σ, the function is concave with respect to μ, so we can maximize it by setting the first partial derivative to zero. Regardless of the value of σ, the maximum is at $\mu_{\text{ML}} = \frac{1}{n}\sum_{i=1}^{n} x_i$. We can therefore plug this value into the log likelihood and maximize with respect to σ. The derivative of the resulting function with respect to σ is

$$\frac{\partial \log \mathcal{L}_{\{x_1,\ldots,x_n\}}(\mu_{\text{ML}}, \sigma)}{\partial \sigma} = -\frac{n}{\sigma} + \sum_{i=1}^{n} \frac{(x_i - \mu_{\text{ML}})^2}{\sigma^3}. \tag{3.148}$$

For $\sigma > 0$, the derivative is zero if σ^2 is equal to $\sigma^2_{\text{ML}} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_{\text{ML}})^2$. If σ^2 is between 0 and σ_{ML}, the derivative is positive; if it is larger, it is negative. The maximum is therefore at σ_{ML}. ■

Figure 3.23 shows the log-likelihood of a Gaussian model applied to the height data from Dataset 5 used in Figure 3.13). The figure also shows the maximum-likelihood fit, corresponding to the maximum of the log-likelihood function, comparing it to the fits produced by other choices of model parameters.

3.8 Inverse-Transform Sampling

Simulation is a fundamental tool in probabilistic modeling. For example, it facilitates the computation of complicated probabilities via the Monte Carlo method (see Section 1.7). The main strategy for generating simulated samples from a random variable decouples the process into two steps:

1 Producing uniform samples in the unit interval $[0, 1]$.
2 Transforming the uniform samples so that they have the desired distribution.

Here we focus on the second step, assuming that we have access to a random-number generator that generates independent samples following a uniform distribution in $[0, 1]$. Such generators can be based on actual random physical phenomena, or on algorithms for generating sequences of numbers that have statistical

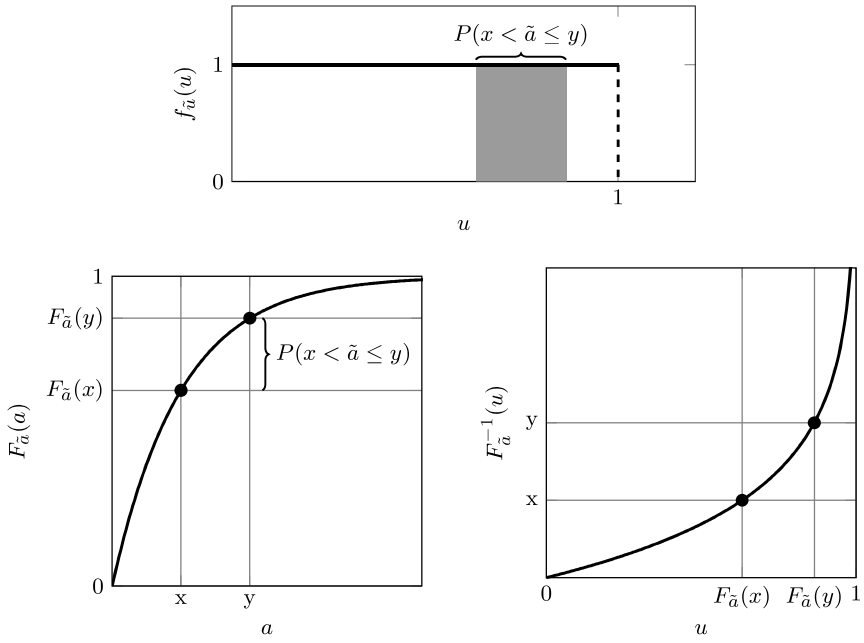

Figure 3.24 Inverse-transform sampling. Let $\tilde{u}$ be a uniform random variable in the unit interval $[0, 1]$. The top plot shows that the probability of obtaining a sample of $\tilde{u}$ belonging to an interval of length $\mathrm{P}(x < \tilde{a} \leq y)$ is equal to $\mathrm{P}(x < \tilde{a} \leq y)$. The bottom left plot shows that $\mathrm{P}(x < \tilde{a} \leq y) = F_{\tilde{a}}(y) - F_{\tilde{a}}(x)$. The bottom right plot shows that for any $a \leq b$, the inverse cdf maps an interval of length $\mathrm{P}(x < \tilde{a} \leq y)$ to $(a, b]$. This means that uniform samples mapped in this way have the same distribution as $\tilde{a}$.

properties resembling truly random sequences (this is known as pseudorandom-number generation).

Let us consider the problem of transforming uniform random samples so that they have the same distribution as a random variable $\tilde{a}$. We fix an interval $(x, y]$. The probability of the uniform samples mapped to the interval should equal $\mathrm{P}(x < \tilde{a} \leq y)$. This can be achieved by choosing an interval of length $\mathrm{P}(x < \tilde{a} \leq y)$ and mapping it to $(x, y]$. The probability of obtaining a sample in that interval using a uniform pdf is exactly $\mathrm{P}(x < \tilde{a} \leq y)$, as shown in the top plot of Figure 3.24.

Recall that by Lemma 3.5, $\mathrm{P}(x < \tilde{a} \leq y) = F_{\tilde{a}}(y) - F_{\tilde{a}}(x)$ (bottom left plot in Figure 3.24). If we map each point u in the unit interval to $F_{\tilde{a}}^{-1}(u)$, then $F_{\tilde{a}}(x)$ is mapped to x, $F_{\tilde{a}}(y)$ is mapped to y and the values in between are mapped to $(x, y]$, because the cdf (and therefore its inverse) is nondecreasing (bottom right plot in Figure 3.24). This is precisely what we want: the width of the interval mapping to $(x, y]$ is $\mathrm{P}(x < \tilde{a} \leq y)$! This holds for any x and y, so the inverse cdf transforms the uniform samples to yield the desired distribution.

Definition 3.37 (Inverse-transform sampling)**.** *Let $\tilde{a}$ be a continuous random variable with cdf $F_{\tilde{a}}$ and $\tilde{u}$ a random variable that is uniformly distributed in $[0, 1]$ and independent of $\tilde{a}$.*

1 Obtain a sample u of $\tilde{u}$.
2 Set $a := F^{-1}_{\tilde{a}}(u)$.

The careful reader will point out that $F_{\tilde{a}}$ may not be invertible at every point. In such cases we use the generalized inverse of the cdf defined in the proof of Theorem 3.23:

$$F^{-1}_{\tilde{a}}(u) := \min_x \{F_{\tilde{a}}(x) = u\}. \tag{3.149}$$

The following theorem provides a formal proof that inverse-transform sampling works.

Theorem 3.38 (Inverse-transform sampling works)**.** *Let $\tilde{a}$ be a continuous random variable with cdf $F_{\tilde{a}}$ and $\tilde{u}$ a random variable that is uniformly distributed in $[0, 1]$ and independent of $\tilde{a}$. The distribution of $\tilde{b} = F^{-1}_{\tilde{a}}(\tilde{u})$ is the same as the distribution of $\tilde{a}$.*

Proof We just need to show that the cdf of $\tilde{b}$ is equal to $F_{\tilde{a}}$. As in the proof of Theorem 3.23, we use the fact that the events $\left\{F^{-1}_{\tilde{a}}(\tilde{u}) \leq y\right\}$ and $\{\tilde{u} \leq F_{\tilde{a}}(y)\}$ are equivalent, because cdfs are nondecreasing. We have

$$\begin{aligned} F_{\tilde{b}}(y) &= \mathrm{P}(\tilde{b} \leq y) && (3.150)\\ &= \mathrm{P}\left(F^{-1}_{\tilde{a}}(\tilde{u}) \leq y\right) && (3.151)\\ &= \mathrm{P}(\tilde{u} \leq F_{\tilde{a}}(y)) && (3.152)\\ &= \int_{u=0}^{F_{\tilde{a}}(y)} \mathrm{d}u && (3.153)\\ &= F_{\tilde{a}}(y). && (3.154) \end{aligned}$$

■

Example 3.39 (Sampling from an exponential distribution)**.** Let $\tilde{a}$ be an exponential random variable with parameter λ. Its cdf $F_{\tilde{a}}(x) := 1 - e^{-\lambda x}$ is invertible in $[0, \infty]$. Its inverse equals

$$F^{-1}_{\tilde{a}}(u) = \frac{1}{\lambda} \log\left(\frac{1}{1-u}\right). \tag{3.155}$$

$F^{-1}_{\tilde{a}}(\tilde{u})$ is an exponential random variable with parameter λ by Theorem 3.38. The random variable $\tilde{a}$ in Figure 3.24 is distributed according to an exponential distribution. Figure 3.25 provides a numerical demonstration that applying (3.155) reshapes the uniform distribution to have an exponential shape.