1

Introduction

> The top 1% of the population controls 35% of the wealth. On Twitter, the top 2% of users send 60% of the messages. In the health care system, the treatment for the most expensive fifth of patients create four-fifths of the overall cost. These figures are always reported as shocking, as if the normal order of things has been disrupted, as if [it] is a surprise of the highest order. It's not. Or rather, it shouldn't be.
> – Clay Shirky, in response to the question "What scientific concept would improve everybody's cognitive toolkit?" [194]

Introductory probability courses often leave the impression that the Gaussian distribution is what we should expect to see in the world around us. It is referred to as the "Normal" distribution after all! As a result, statistics like the ones in the quote above tend to be treated as aberrations, since they would never happen if the world were Gaussian. The Gaussian distribution has a "scale," a typical value (the mean) around which individual measurements are centered and do not deviate from by too much. For example, if we consider human heights, which are approximately Gaussian, the average height of an adult male in the US is 5 feet 9 inches and most people's heights do not differ by more than 10 inches from this. In contrast, there are order-of-magnitude differences between individuals in terms of wealth, Twitter followers, health care costs, and so on.

However, order-of-magnitude differences like those just mentioned are not new and should not be surprising. Over a century ago, Italian economist Vilfredo Pareto discovered that the richest 20 percent of the population controlled 80 percent of the property in Italy. This is now termed the "Pareto Principle," aka the "80-20" rule and variations of this principle have shown up repeatedly in widely disparate areas in the time since Pareto's discovery. For example, in 2002 Microsoft reported that 80 percent of the errors in Windows are caused by 20 percent of the bugs [188], and similar versions of the Pareto principle apply (though not always with 80/20) to many aspects of business, for example, most of the profit is made from a small percentage of the customers and most of the sales are made by a small percentage of the sales team.

Statistics related to the Pareto principle make for compelling headlines, but they are typically an indication of something deeper. When we see such figures, it is likely that there is not a Gaussian distribution underlying them, but rather a heavy-tailed distribution is the reason for the "surprising" statistics. The most celebrated such distribution again carries Vilfredo Pareto's name: *the Pareto distribution*. Heavy-tailed distributions such as the Pareto distribution are just as prominent as (if not more so than) the Gaussian distribution and have been observed in hundreds of applications in physics, biology, computer science, the

https://doi.org/10.1017/9781009053730.002 Published online by Cambridge University Press

social sciences, and beyond over the past century. Some examples include the sizes of cities [92, 163], the file sizes in computer systems and networks [52, 146], the size of avalanches and earthquakes [109, 144], the length of protein sequences in genomes [130, 145], the size of meteorites [13, 162], the degree distribution of the web graph [36, 116], the returns of stocks [49, 94], the number of copies of books sold [14, 110], the number of households affected during blackouts in power grids [114], the frequency of word use in natural language [77, 227], and many more.

Given the breadth of areas where heavy-tailed phenomena have been observed, one might guess that, by now, observations of heavy-tailed phenomena in new areas are expected – that heavy tails are treated as *more normal than the Normal*. After all, Pareto's work has been widely known for more than a century. However, despite a century of experience, statistics related to the Pareto Principle and, more broadly, heavy-tailed distributions are still typically presented as surprising curiosities – anomalies that could not have been anticipated. Even in scientific communities, observations of heavy-tailed phenomena are often presented as mysteries to be explained rather than something to be expected a priori. In many cases, there is even a significant amount of controversy and debate that follows the identification of heavy-tailed phenomena in data.

Surprising? Mysterious? Controversial?

Given the century of mathematical and statistical work around heavy tails, it certainly should not be the case that heavy tails are surprising, mysterious, and controversial. In fact, there are many reasons why one should *expect* to see heavy-tailed distributions arise. Perhaps the main reason why they are still viewed as surprising is that the version of the central limit theorem taught in introductory probability courses gives the impression that the Gaussian will occur everywhere. However, this introductory version of the central limit theorem does not tell the whole story. There is a "generalized" version of the central limit theorem that states that either the Gaussian *or a heavy-tailed distribution* will emerge as the limit of sums of random variables. Unfortunately, the technical nature of this result means it rarely features in introductory courses, which leads to unnecessary surprises about the presence of heavy-tailed distributions. Going beyond sums of random variables, when random variables are combined in other natural ways (e.g., products or max/min) heavy tails are even more likely to emerge, whereas the Gaussian distribution is not.

So heavy-tailed phenomena should not be considered surprising. What about mysterious? The view of heavy tails as mysterious is, to some extent, a consequence of unfamiliarity. People are familiar with the Gaussian distribution because of its importance in introductory probability courses, and when something emerges that has qualitatively and quantitatively different properties it seems mysterious and counter-intuitive. The Pareto Principle is one illustration of the counterintuitive properties that make heavy-tailed distributions seem mysterious, but there are many others. For example, while the Gaussian distribution has a clear "scale" – most samples will be close to the mean – samples from heavy-tailed distributions frequently differ by orders of magnitude and may even be "scale free" (e.g., in the case of the Pareto distribution). Another example is that, while the moments (the mean, variance, etc.) of the Gaussian distribution are all finite, it is not uncommon to see data that fits a heavy-tailed distribution having an infinite variance, or even an infinite mean! For example, the degree distribution of many complex networks tends to have a tail that matches that of

a Pareto with infinite variance (see, for example, [23]). This can potentially lead to mind-bending challenges when trying to apply statistical tools, which often depend on averages and variances.

The combination of surprise and mystery that surrounds heavy-tailed phenomena means that there is often considerable excitement that follows the discovery of data that fits a heavy-tailed distribution in a new field. Unfortunately, this excitement often sparks debate and controversy – often enough that an unfortunate pattern has emerged. A heavy-tailed phenomenon is discovered in a new field. The excitement over the discovery leads researchers to search for heavy tails in other parts of the field. Heavy tails are then discovered in many settings and are claimed to be a universal property. However, the initial excitement of discovery and lack of previous background in statistics related to heavy tails means that the first wave of research identifying heavy tails uses intuitive but flawed statistical tools. As a result, a controversy emerges – which settings where heavy tails have been observed really have heavy tails? Are they really universal? Over time, more careful statistical analyses are used, showing that some places really do exhibit heavy tails while others were false discoveries. By the end, a mature view of heavy tails emerges, but the whole process can take decades.

At this point, the pattern just described has been replicated in many areas, including computer science [68], biology [119], chemistry [160], ecology [10], and astronomy [216]. Maybe the most prominent example of this story is still ongoing in the area of *network science*. Near the turn of the century, the study of complex networks began to explode in popularity due to the growing importance of networks in our lives and the increasing ease of gathering data about large networks. Initial results in the area were widely celebrated and drove an enormous amount of research to look at the universality of scale-free networks. However, as the field matured and the statistical tools became more sophisticated, it became clear that many of the initial results were flawed. For example, claims that the internet graph [80] and the power network [24] are heavy-tailed were refuted [4, 222], among others. This led to a controversy in the area that continues to this day, 20 years later [37, 212].

Demystifying Heavy Tails

The goal of this book is to demystify heavy-tailed phenomena. Heavy tails are not anomalies – and their emergence should not be surprising or controversial either! Heavy tails are an unavoidable part of our lives, and viewing statistics like the ones that started this chapter as anomalies prevents us from thinking clearly about the world around us. Further, while properties of heavy-tailed phenomena like the Pareto Principle may initially make heavy-tailed distributions seem counterintuitive, they need not be. This book strives to provide tools and techniques that can make heavy tails as easy and intuitive to reason about as the Gaussian, to highlight when one should expect the emergence of heavy-tailed phenomena, and to help avoid controversy when identifying heavy tails in data.

Because of the ubiquitousness and seductive nature of heavy-tailed phenomena, they are a topic that has permeated wide ranging fields, from astronomy and physics, to biology and physiology, to social science and economics. However, despite their ubiquity, they are also, perhaps, one of the most misused and misunderstood mathematical areas, shrouded in both excitement and controversy. It is easy to get excited about heavy-tailed phenomena as you start to realize the important role they play in the world around us and become exposed to the beautiful and counterintuitive properties they possess. However, as you start to dig into

the topic, it quickly becomes difficult. The mathematics that underlie the analysis of heavy-tailed distributions are technical and advanced, often requiring prerequisites of graduate-level probability and statistics courses. This is the reason why introductory probability courses typically do not present much, if any, material related to heavy-tailed distributions. If they are mentioned, they are typically used as examples illustrating that "strange" things can happen (e.g., distributions can have an infinite mean). Thus, a scientist or researcher in a field outside of mathematics who is interested in learning more about heavy tails may find it difficult, if not impossible, to learn from the classical texts on the topic.

It is exactly this difficulty that led us to write this book. In this book we hope to introduce the fundamentals of heavy-tailed distributions using only tools that one learns in an introductory probability course. The book intentionally does not spend much time on describing the settings where heavy tails arise – there are simply too many different areas to do justice to even a small subset of them. Instead, we assume that if you have found your way to this book, then heavy tails are important to you. Given that, our goal is to provide an introduction to how to think about heavy tails both intuitively and mathematically.

The book is divided into three parts, which focus on three foundational guiding questions.

- **Part I: Properties.** *What leads to the counterintuitive properties of heavy-tailed phenomena?*
- **Part II: Emergence.** *Why do heavy-tailed phenomena occur so frequently in the world around us?*
- **Part III: Estimation.** *How can we identify and estimate heavy-tailed phenomena using data?*

In Part I of the book we provide insight into some of most mysterious and elegant properties of heavy-tailed distributions, connecting these properties to formal definitions of subclasses of heavy-tailed distributions. We focus on three foundational properties: "scale-invariance" (aka, scale-free), the "catastrophe principle," and "increasing residual life." We illustrate that these properties provide qualitatively different behaviors than what is seen under light-tailed distributions like the Gaussian, and provide intuition underlying the properties. The three chapters that make up Part I strive to demystify some of the particularly exotic properties of heavy-tailed distributions and to provide a clear view of how these properties interact with each other and with the broader class of heavy-tailed distributions.

In Part II of the book we explore simple laws that can "explain" the emergence of heavy-tailed distributions in the same way that the central limit theorem "explains" the prominence of the Gaussian distribution. We study three foundational stochastic processes in order to understand when one should expect the emergence of heavy-tailed distributions as opposed to light-tailed distributions. Our discussions in the three chapters that make up Part II highlight that heavy-tailed distributions should not be viewed as anomalies. In fact, heavy tails should not be surprising at all; in many cases they should be treated as something as natural as, if not more natural than, the emergence of the Gaussian distribution.

In Part III of this book we focus on the statistical tools used for the estimation of heavy-tailed phenomena. Unfortunately, there is no perfect recipe for "properly" detecting and estimating heavy-tailed distributions in data. Our treatment, therefore, seeks to highlight a

handful of important approaches and to provide insight into when each approach is appropriate and when each may be misleading. Combined, the chapters that make up Part III highlight a crucial point: one must proceed carefully when estimating heavy-tailed phenomena in real-world data. It is naive to expect to estimate *exact* heavy-tailed distributions in data. Instead, a realistic goal is to estimate the *tail* of heavy-tailed phenomena. Even in doing this, one should not rely on a single method for estimation. Instead, it is a necessity to build confidence through the use of multiple, complementary estimation approaches.

1.1 Defining Heavy-Tailed Distributions

Before we tackle our guiding questions, we start with the basic question: *What is a heavy-tailed distribution?*

One of the reasons for the mystique that surrounds heavy-tailed distributions is that if you ask five people from different communities this question, you are likely to get five different answers. Depending on the community, the term heavy-tailed may be used interchangeably with terms like scale-free, power-law, fat-tailed, long-tailed, subexponential, self-similar, stable, and others. Further, the same names may mean different things to different communities!

Sometimes the term "heavy-tailed" is used to refer to a specific distribution such as the Pareto or the Zipf distribution. Other times, it is used to identify particular properties of a distribution, such as the fact that it is scale-free, has an infinite (or very large) variance, a decreasing failure rate, and so on. As a result, there is often a language barrier when discussing heavy-tailed distributions that stems from different associations with the same terms across communities.

Hopefully, reading this book will equip you to navigate the zoo of terminology related to heavy-tailed distributions. Each of the terms mentioned earlier does have a concrete, precise, established mathematical definition. It is just that these terms are often used carelessly, which leads to confusion. It will take us most of the book to get through the definitions of all the terms mentioned in the previous paragraph, but we start in this section by laying the foundation – defining the term "heavy-tailed" and discussing some of the most celebrated examples.

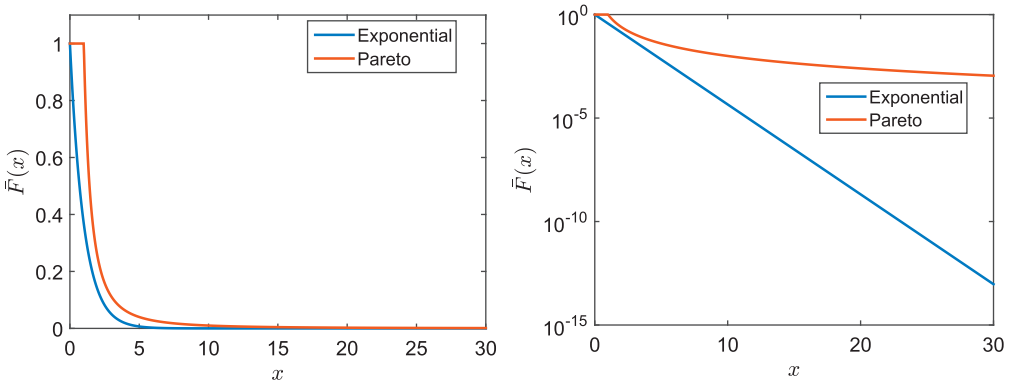
The term "heavy-tailed" is inherently relative – heavier than what? A Gaussian distribution has a heavier tail than a Uniform distribution, and an Exponential distribution has a heavier tail than a Gaussian distribution, but neither of these is considered "heavy-tailed." Thus, the key feature of the definition is the comparison point chosen.

The comparison point that is used to define the class of heavy-tailed distributions is the Exponential distribution. That is, a distribution is considered to be heavy-tailed if it has a heavier tail than any Exponential distribution. Formally, this is stated in terms of the cumulative distribution function (c.d.f.) F of a random variable X, that is, $F(x) = \Pr(X \leq x)$, and the complementary cumulative distribution function (c.c.d.f.) $\bar{F}$, that is, $\bar{F}(x) = 1 - F(x)$.

Definition 1.1 A distribution function F is said to be heavy-tailed if and only if, for all $\mu > 0$,

$$\limsup_{x\to\infty} \frac{1 - F(x)}{e^{-\mu x}} = \limsup_{x\to\infty} \frac{\bar{F}(x)}{e^{-\mu x}} = \infty.$$

https://doi.org/10.1017/9781009053730.002 Published online by Cambridge University Press



(a) Linear scale (b) Log scale

Figure 1.1 Contrasting heavy-tailed and light-tailed distributions: The plots show the c.c.d.f. of the exponential distribution (with mean 1) and a heavy-tailed Pareto distribution (with minimal value $x_m = 1$, scale parameter $\alpha = 2$). While the contrast in tail behavior is difficult to discern on a linear scale (Fig. (a)), it is quite evident when the probabilities are plotted on a logarithmic scale (Fig. (b)).

Otherwise, F is light-tailed. A random variable X is said to be heavy-tailed (light-tailed) if its distribution function is heavy-tailed (light-tailed).

Note that the definition of heavy-tailed distributions given above applies to the *right* tail of the distribution, that is, it is concerned with the behavior of the probability of taking values larger than x as $x \to \infty$. In some applications, one might also be interested in the *left* tail. In such cases, the definition of heavy-tailed can be applied to both the right tail (without change) and the left tail (by considering the right tail of $-X$).

The definition of heavy-tailed is, in some sense, natural. It looks explicitly at the "tail" of the distribution (i.e., the c.c.d.f. $\bar{F}(x)$), and it is easy to see from the definition that the tails of distributions that are heavy-tailed are "heavier" (i.e., decay more slowly) than the tails of distributions that are light-tailed; see Figure 1.1.

The particular choice of the Exponential distribution as the boundary between heavy-tailed and light-tailed may, at first, seem arbitrary. In fact, without detailed study of the class of heavy-tailed distributions, it is difficult to justify this particular choice. But, as we will see throughout this book, the Exponential distribution serves to separate two classes of distributions that have qualitatively different behavioral properties and require fundamentally different mathematical tools to work with.

To begin to examine the distinction between heavy-tailed and light-tailed distributions, it turns out to be useful to consider two alternative, but equivalent, definitions of "heavy-tailed."

Lemma 1.2 *Consider a random variable X. The following statements are equivalent.*

(i) X *is heavy-tailed.*
(ii) The moment generating function $M(s) := \mathbb{E}\left[e^{sX}\right] = \infty$ *for all* $s > 0$.
(iii) $\liminf_{x\to\infty} -\frac{\log \Pr(X>x)}{x} = 0$.

https://doi.org/10.1017/9781009053730.002 Published online by Cambridge University Press

The proof of this lemma provides useful intuition about heavy-tailed distribution; however, before proving this result, let us interpret the two new, equivalent definitions of heavy-tailed that it provides.

First, consider *(ii)*, which states that a random variable is heavy-tailed if and only if its moment generating function $M(s) := \mathbb{E}\left[e^{sX}\right]$ is infinite for all $s > 0$. This definition highlights that heavy-tailed distributions require a different analytic approach than light-tailed distributions. For light-tailed distributions the moment generating function often provides an important tool for characterizing the distribution. It can be used to derive the moments of the distribution, but it also can be inverted to characterize the distribution itself. Further, it is a crucial tool for analysis because of the simplicity of handling convolutions via the moment generating function, for example, when deriving concentration inequalities such as Chernoff bounds. In contrast, the definition given by *(ii)* shows that such techniques are not applicable for heavy-tailed distributions.

Next, consider *(iii)*, which states that a random variable X is heavy-tailed if and only if the log of its tail, $\log \Pr(X > x)$, decays sublinearly. This again highlights that heavy-tailed distributions require a different analytic approach than light-tailed distributions. In particular, when studying the tail of light-tailed distributions it is common to use concentration inequalities such as Chernoff bounds, which inherently have an exponential decay. As a result, such bounds focus on determining the optimal decay rate, which is characterized by deriving a maximal μ such that $\Pr(X > x) \leq Ce^{-\mu x}$. However, the definition given by *(iii)* highlights that the maximum possible μ for heavy-tailed distributions is zero, and so fundamentally different analytic approaches must be used.

To build more intuition on the relationship between these three equivalent definitions of "heavy-tailed," as well as to get practice working with the definitions, it is useful to consider the proof of Lemma 1.2.

Proof of Lemma 1.2 To prove Lemma 1.2, we need to show the equivalence of each of the three definitions of heavy-tailed. We do this by showing that *(i)* implies *(ii)*, that *(ii)* implies *(iii)*, and finally that *(iii)* implies *(i)*.

$(i) \Rightarrow (ii)$. Suppose that X is heavy-tailed, with distribution F. By definition, this implies that for any $s > 0$, there exists a strictly increasing sequence $(x_k)_{k\geq 1}$ satisfying $\lim_{k\to\infty} x_k = \infty$, such that

$$\lim_{k\to\infty} e^{sx_k}\bar{F}(x_k) = \infty. \tag{1.1}$$

We can now bound $\mathbb{E}\left[e^{sX}\right]$ as follows.

$$\begin{aligned}\mathbb{E}\left[e^{sX}\right] &= \int_0^\infty e^{sx} dF(x) \\ &\geq \int_{x_k}^\infty e^{sx} dF(x) \\ &\geq e^{sx_k}\bar{F}(x_k).\end{aligned}$$

Since the above inequality holds for all k, it now follows from (1.1) that $\mathbb{E}\left[e^{sX}\right] = \infty$. Therefore, Condition *(i)* implies Condition *(ii)*.

https://doi.org/10.1017/9781009053730.002 Published online by Cambridge University Press

$(ii) \Rightarrow (iii)$. Suppose that X satisfies Condition *(ii)*. For the purpose of obtaining a contradiction, let us assume that Condition *(iii)* does not hold. Since $-\frac{\log \Pr(X>x)}{x} \geq 0$, this means that

$$\liminf_{x\to\infty} -\frac{\log \Pr(X>x)}{x} > 0.$$

The above statement implies that there exist $\mu > 0$ and $x_0 > 0$ such that

$$-\frac{\log \Pr(X>x)}{x} \geq \mu \iff \Pr(X>x) \leq e^{-\mu x} \quad \forall\, x \geq x_0. \tag{1.2}$$

Now, pick s such that $0 < s < \mu$. We may now bound the moment generating function of X at s as follows:

$$\begin{aligned} M(s) = \mathbb{E}\left[e^{sX}\right] &= \int_0^\infty \Pr\left(e^{sX} > x\right) dx \\ &= \int_0^{e^{sx_0}} \Pr\left(e^{sX} > x\right) dx + \int_{e^{sx_0}}^\infty \Pr\left(X > \frac{\log(x)}{s}\right) dx. \end{aligned}$$

Here, we have used the following representation for the expectation of a nonnegative random variable Y: $\mathbb{E}[Y] = \int_0^\infty \Pr(Y > y)\, dy$. While the first term above can be bounded from above by e^{sx_0}, we may bound the second using (1.2), since $x \geq e^{sx_0}$ is equivalent to $\log(x)/s \geq x_0$.

$$\begin{aligned} M(s) &\leq e^{sx_0} + \int_{e^{sx_0}}^\infty e^{-\mu\frac{\log(x)}{s}} dx \\ &= e^{sx_0} + \int_{e^{sx_0}}^\infty x^{-\mu/s} dx. \end{aligned}$$

Since $\mu/s > 1$, we have $\int_1^\infty x^{-\mu/s} dx < \infty$, which implies that $M(s) < \infty$, giving us a contradiction. Therefore, Condition *(ii)* implies Condition *(iii)*.

$(iii) \Rightarrow (i)$. Suppose that the random variable X, having distribution F, satisfies Condition *(iii)*. Thus, there exists a strictly increasing sequence $(x_k)_{k\geq 1}$ satisfying $\lim_{k\to\infty} x_k = \infty$, such that

$$\lim_{k\to\infty} -\frac{\log \bar{F}(x_k)}{x_k} = 0.$$

Given $\mu > 0$, this in turn implies that there exists $k_0 \in \mathbb{N}$ such that

$$\begin{aligned} & -\tfrac{\log \bar{F}(x_k)}{x_k} < e^{-\frac{\mu}{2}} && \forall\, k > k_0 \\ \iff\quad & \bar{F}(x_k) > e^{-\frac{\mu x_k}{2}} && \forall\, k > k_0 \\ \iff\quad & \tfrac{\bar{F}(x_k)}{e^{-\mu x_k}} > e^{\frac{\mu x_k}{2}} && \forall\, k > k_0. \end{aligned}$$

The last assertion above implies that $\lim_{k\to\infty} \frac{\bar{F}(x_k)}{e^{-\mu x_k}} = \infty$, which implies $\limsup_{x\to\infty} \frac{\bar{F}(x)}{e^{-\mu x}} = \infty$. Since this is true for any $\mu > 0$, we conclude that Condition *(iii)* implies Condition *(i)*.

□

https://doi.org/10.1017/9781009053730.002 Published online by Cambridge University Press

1.2 Examples of Heavy-Tailed Distributions

We now have three equivalent definitions of heavy-tailed distributions and, through the proof, we understand how these three definitions are related. But, even with these restatements, the definition of heavy-tailed is still opaque. It is difficult to get behavioral intuition about the properties of heavy-tailed distributions from any of the definitions. Further, it is very hard to see much about what makes heavy-tailed distributions have the mysterious properties that are associated with them using these definitions alone.

In part, this is due to the breadth of the definition of heavy-tailed. The important properties commonly associated with heavy-tailed distributions, such as scale invariance, infinite variance, the Pareto principle, etc., do not hold for all heavy-tailed distributions; they hold only for certain subclasses of heavy-tailed distributions.

As a result, it is important to build intuition for the class of heavy-tailed distributions by looking at specific examples. That is the goal of the remainder of this chapter. In particular, we focus in detail on the Pareto distribution, the Weibull distribution, and the LogNormal distribution with the goal of providing both the mathematical formalism for these distributions and some insight in their important properties and applications. Additionally, we briefly introduce some of the other important examples of heavy-tailed distributions that come up frequently in applications, including the Cauchy, Fréchet, Lévy, Burr, and Zipf distributions.

Perhaps the most important thing to keep in mind as you read these sections is the contrast between the properties of the heavy-tailed distributions that we discuss and the properties of light-tailed distributions, such as the Gaussian and Exponential distributions, with which you are likely more familiar. To set the stage, we summarize the important formulas for these two distributions next.

The Gaussian Distribution

The Gaussian distribution, also called the Normal distribution or the bell curve, is perhaps the most widely recognized distribution and is extremely important in statistics and beyond. It is defined using two parameters, the mean μ and the variance σ^2, and is expressed most conveniently through its probability density function (p.d.f.), $f(x)$, or its moment generating function (m.g.f.), $M(s)$. Given a random variable $Z \sim \text{Gaussian}(\mu, \sigma)$, we have

$$f_Z(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}},$$
$$M_Z(s) = E[e^{sZ}] = e^{\mu s + \frac{1}{2}\sigma^2 s^2}.$$

Since $M_Z(s) < \infty$ for all $s > 0$, it follows that the Gaussian distribution is light-tailed. The light-tailedness of the Gaussian distribution can also be deduced directly by bounding its c.c.d.f. (see Exercise 2).

The particular Gaussian distribution with zero mean and unit variance ($\mu = 0$, $\sigma = 1$) is commonly referred to as the *standard Gaussian*.

The Exponential Distribution

The Exponential distribution is a widely known and broadly applicable distribution that serves as the light-tailed distribution on the boundary between light-tailed and heavy-tailed distributions. It is a nonnegative distribution defined in terms of one parameter: λ, which is

https://doi.org/10.1017/9781009053730.002 Published online by Cambridge University Press

referred to as the "rate" since the mean of the distribution is $1/\lambda$. Given a random variable $X \sim \text{Exponential}(\lambda)$, the p.d.f., c.c.d.f., and m.g.f., can be expressed as

$$\begin{aligned} f_X(x) &= \lambda e^{-\lambda x} \qquad (x \geq 0), \\ \bar{F}(x) &= e^{-\lambda x} \qquad (x \geq 0), \\ M_X(s) &= \frac{1}{(1 - s/\lambda)} \qquad (s < \lambda). \end{aligned}$$

Note that the tail of the Exponential distribution is heavier than that of the Gaussian because e^{-x} goes to zero more slowly than e^{-x^2}. Additionally, unlike the Gaussian, the moment generating function is not finite everywhere.

1.2.1 The Pareto Distribution

Vilfredo Pareto originally presented the Pareto distribution, and introduced the idea of the Pareto Principle, in the study of the allocation of wealth. But since then, it has been used as a model in numerous other settings, including the sizes of cities, the file sizes in computer systems and networks, the price returns of stocks, the size of meteorites, casualties and damages due to natural disasters, frequency of words, and many more. It is perhaps the most celebrated example of a heavy-tailed distribution, and as a result, the term Pareto is sometimes, unfortunately, used interchangeably with the term heavy-tailed.

Formally, a random variable X follows a Pareto(x_m, α) distribution if

$$\Pr(X \geq x) = \bar{F}(x) = \left(\frac{x}{x_m}\right)^{-\alpha}, \text{ for } \alpha > 0, x \geq x_m > 0.$$

Here, α is the shape parameter of the distribution and is also commonly referred to as the *tail index*, while x_m is the minimum value of the distribution, that is, $X \geq x_m$. Given the c.c.d.f. above, it is straightforward to differentiate and obtain the p.d.f.

$$f(x) = \frac{\alpha x_m^\alpha}{x^{\alpha+1}}, \quad x \geq x_m.$$

It is easy to see from the c.c.d.f. that the Pareto is heavy-tailed. In particular, using Definition 1.1, we can compute

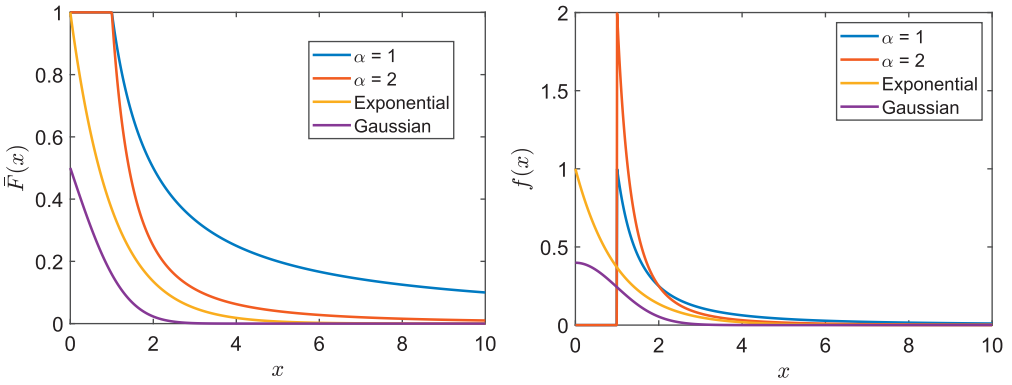
$$\limsup_{x\to\infty} \frac{\bar{F}(x)}{e^{-\mu x}} = \limsup_{x\to\infty} \left(\frac{x_m}{x}\right)^\alpha e^{\mu x} = \infty, \tag{1.3}$$

since the exponential $e^{\mu x}$ grows more quickly than the polynomial x^α.

This highlights the key contrast between the Pareto distribution and common light-tailed distributions like the Gaussian and Exponential distributions: the Pareto tail decays *polynomially*, as $x^{-\alpha}$, instead of *exponentially* (as $e^{-\mu x}$) in the case of the Exponential, or *superexponentially* (as $e^{-x^2/2\sigma^2}$) in the case of the Gaussian. As a consequence, large values are much more likely to occur under a Pareto distribution than under a Gaussian or Exponential distribution. For example, you are much more likely to meet someone whose income is 10 times the average than someone whose height is 10 times the average.

This contrast is present visually too. Figure 1.2 shows that the tail of the Pareto is considerably heavier. The figure illustrates the p.d.f. and c.c.d.f. of the Pareto for different values of

https://doi.org/10.1017/9781009053730.002 Published online by Cambridge University Press



(a) c.c.d.f. (b) p.d.f.

Figure 1.2 Contrasting Pareto distribution with the Exponential and the Gaussian: The plots show (a) the c.c.d.f., and (b) the p.d.f., corresponding to Pareto distributions with $x_m = 1$ with different values of α, alongside the Exponential distribution (with unit mean) and the standard Gaussian.

the tail index α, which is typically the parameter of interest since it controls the degree of the polynomial decay of the p.d.f. and c.c.d.f., and thus determines the "weight" of the tail. As α decreases, the tail becomes heavier, while as $\alpha \to \infty$ the Pareto distribution approaches the Dirac delta function centered at x_m.

While Figure 1.2 already contrasts the Pareto, Gaussian, and Exponential distributions, we can better emphasize this contrast by presenting the figure in a different way, that is, by rescaling its axes. In particular, Figure 1.3 shows the same c.c.d.f.s but presents the data on a log-log scale, that is, with logarithmic horizontal and vertical axes. With this change, a remarkable pattern emerges – the Pareto c.c.d.f. becomes a straight line, while the Gaussian and Exponential distributions quickly drop off a cliff and disappear. This image viscerally highlights the heaviness of the Pareto's tail as compared to the tails of the Exponential and the Gaussian.

To understand why the Pareto is linear when viewed on a log-log scale, let us do a quick calculation. Letting $C_1 = x_m^\alpha$ we can write

$$\bar{F}(x) = \left(\frac{x}{x_m}\right)^{-\alpha} = C_1 x^{-\alpha}.$$

Taking logarithms of both sides then gives

$$\underbrace{\log \bar{F}(x)}_{'y'} = \underbrace{\log C_1}_{y\text{-intercept}} + \underbrace{(-\alpha)}_{\text{slope}} \underbrace{\log x}_{'x'},$$

which reveals that, on a log-log scale, the c.c.d.f. is simply a linear function with y-intercept $\log C_1$ and slope $-\alpha$. Not only that, the p.d.f. is also of the same form, that is, $f(x) = C_2 x^{-(\alpha+1)}$ where $C_2 = \alpha x_m^\alpha$ and so it also is linear in the log-log scaling.

This property – being (approximately) linear on a log-log scale – is important enough that it has received a few different names from different communities. Distributions of the form $\bar{F}(x) = Cx^{-\alpha}$ for some constant C are referred to as *power law* distributions. A related set of distributions are *fat-tailed distributions*, which are distributions with $\bar{F}(x) \sim x^{-\alpha}$ as

https://doi.org/10.1017/9781009053730.002 Published online by Cambridge University Press

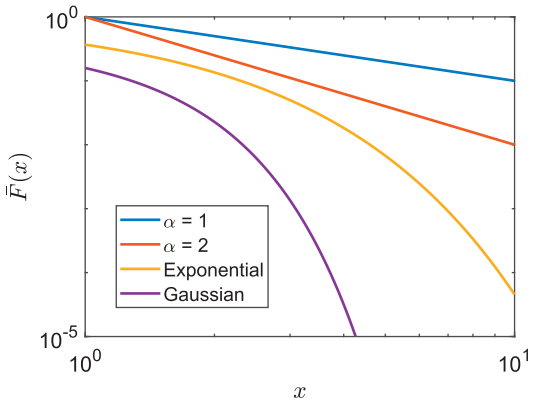


Figure 1.3 A clearer contrast between the Pareto distribution and the Exponential and Gaussian: The plots show the c.c.d.f.s corresponding to different Pareto distributions with $x_m = 1$ and different values of α, alongside the Exponential distribution (with unit mean) and the standard Gaussian, on a log-log scale. This scaling demonstrates clearly how the Pareto tail (linear on a log-log plot) is heavier than those of the Exponential and the Gaussian.

$x \to \infty$, where we use $a(x) \sim b(x)$ as $x \to \infty$ as shorthand for $\lim_{x\to\infty} a(x)/b(x) = 1$. Finally, the class of *regularly varying* distributions, which we introduce in Chapter 2, generalizes both power law and fat-tailed distributions and has strong connections to the concept of scale-invariance.

The fact that Pareto distributions, and more generally power law distributions, are approximately linear on a log-log plot has a number of important consequences. Maybe the most prominent one is that it provides an intuitive exploratory tool for identifying power laws in data. Specifically, when presented with data, one can look at the empirical p.d.f. and c.c.d.f. on a log-log scale and check whether they are approximately linear. If so, then there is the potential that the data comes from a power-law distribution. One can even go further and hope to estimate the tail index α using linear regression on the empirical p.d.f. and c.c.d.f. This is a common approach across fields, which we illustrate in Figure 1.4 using population data for US cities as per the 2010 census. Notice that the empirical c.c.d.f. (on a log-log scale) looks roughly linear for large populations. It is therefore tempting to postulate that the distribution of city populations (asymptotically) follows a power law, and further to estimate the tail index by fitting a least squares regression line to the empirical c.c.d.f. beyond, say 10^4 (since the tail "looks linear" beyond this point), as shown in Figure 1.4. However, as we discuss in Chapter 8, this approach is not statistically sound and may lead to incorrect conclusions. In fact, the temptation to make conclusions based on such naive analyses is one of the most common reasons for the controversy that often surrounds the identification of heavy-tailed phenomena.

Moments

One of the biggest contrasts between the Pareto distribution and light-tailed distributions such as the Gaussian and Exponential is the fact that the Pareto distribution can have infinite moments. In fact, for $X \sim \text{Pareto}(x_m, \alpha)$, $\mathbb{E}[X^n] = \infty$ if $n \geq \alpha$. More specifically, the mean of the Pareto distribution is

https://doi.org/10.1017/9781009053730.002 Published online by Cambridge University Press

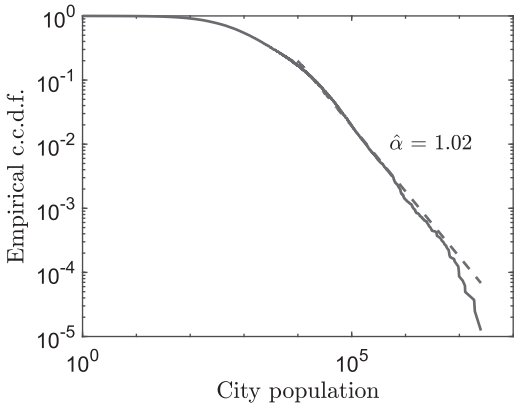


Figure 1.4 Visualizing data on a log-log plot: The figure shows the empirical c.c.d.f. of the populations of U.S. cities as per the 2010 census (data sourced from [2]). Note that on a log-log scale, the data beyond population 10^4 looks approximately linear. The least squares regression line on this data yields an estimate $\hat{\alpha} = 1.02$ of the power law exponent.

$$\mathbb{E}[X] = \begin{cases} \infty, & \alpha \leq 1; \\ \frac{\alpha x_m}{\alpha - 1}, & \alpha > 1. \end{cases}$$

The variance is

$$\mathrm{Var}[X] = \begin{cases} \infty, & \alpha \in (1, 2]; \\ \left(\frac{x_m}{\alpha - 1}\right)^2 \frac{\alpha}{\alpha - 2}, & \alpha > 2. \end{cases}$$

And, in general, the nth moment is

$$\mathbb{E}[X^n] = \begin{cases} \frac{\alpha x_m^n}{\alpha - n}, & n < \alpha; \\ \infty, & n \geq \alpha. \end{cases}$$

Importantly, it is not just a curiosity that the Pareto distribution can have infinite moments. In many cases where data has been modeled using the Pareto distribution, the distribution that is fit has infinite variance and/or mean. For example, file sizes in computer systems and networks [52] and the degree distributions of complex networks such as the web [68] appear to have infinite variance. Additionally, the logarithmic returns on stocks in finance tend to have finite variance, but infinite fourth moment [75], leading to values of α in the range $(2, 4)$.

The Pareto Principle

We began this chapter with a quote about the Pareto principle, so it is important to return to it now that we have formally introduced the Pareto distribution. The classical version of the Pareto principle is that the wealthiest 20 percent of the population holds 80 percent of the wealth. Mathematically, we can ask a more general question about what fraction of the wealth the largest P fraction of the population holds.

https://doi.org/10.1017/9781009053730.002 Published online by Cambridge University Press

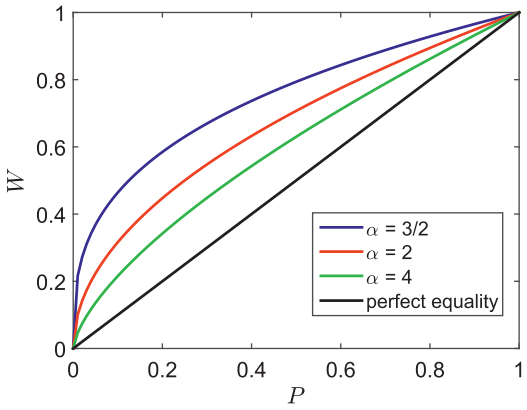


Figure 1.5 Lorenz curves for the Pareto distribution: Lorenz curves for different values of α. The smaller the value of α, the more pronounced the concentration of wealth within a small fraction of the population. The black line represents perfect equality, that is, the utopian scenario in which all individuals have exactly the same wealth.

To compute an analytic version of the Pareto principle, we consider the fraction of the population whose wealth exceeds x. Call this fraction $P(x)$, and then we can calculate $P(x)$ in the case of a Pareto distribution as follows:

$$P(x) = \int_x^\infty f(t)dt = \alpha x_m^\alpha \int_x^\infty t^{-(\alpha+1)}dt = \left(\frac{x}{x_m}\right)^{-\alpha}.$$

Then, the fraction of wealth that is in the hands of such people, which we denote by $W(x)$, is

$$W(x) = \frac{\int_x^\infty t f(t)dt}{\int_{x_m}^\infty t f(t)dt} = \frac{\alpha x_m^\alpha \int_x^\infty t^{-\alpha}dt}{\alpha x_m^\alpha \int_{x_m}^\infty t^{-\alpha}dt} = \left(\frac{x}{x_m}\right)^{-\alpha+1},$$

assuming that $\alpha > 1$. Combining the above equations then gives that, regardless of x, the fraction of wealth W owned by the richest P fraction of the population is

$$W = P^{(\alpha-1)/\alpha}.$$

We illustrate the curve of W as a function of P in Figure 1.5. It is always concave and increasing, and when α is close to 1, it indicates that wealth is concentrated in a very small fraction of the population. Such extreme concentration is an example of a more general phenomenon called the "catastrophe principle," which we discuss in detail in Chapter 3.

Curves like those in Figure 1.5 are referred to as Lorenz curves, after Max Lorenz, who developed them in 1905 as a way to represent the inequality of wealth distribution. The Gini coefficient, which is typically used to quantify wealth inequality today, is the ratio of the area between the line of perfect equality (the 45 degree line) and the Lorenz curve, and the area above the line of perfect equality. The greater the value of the Gini coefficient, the more pronounced the asymmetry in wealth distribution. Understanding properties of the Gini coefficient is still an area of active research, for example, [86] and the references therein.

https://doi.org/10.1017/9781009053730.002 Published online by Cambridge University Press

Relationship to the Exponential Distribution

While heavy-tailed distributions often behave qualitatively differently than light-tailed distributions, there are still some connections between the two that can be useful. In particular, a heavy-tailed distribution can often be viewed as an exponential transformation of a light-tailed distribution. In the case of the Pareto, this connection is to the Exponential distribution. Specifically,

$$X \sim \text{Pareto}(x_m, \alpha) \quad \iff \quad \log(X/x_m) \sim \text{Exponential}(\alpha).$$

Or, equivalently,

$$Y \sim \text{Exponential}(\alpha) \quad \iff \quad x_m e^Y \sim \text{Pareto}(x_m, \alpha).$$

To see why this is true requires a simple change of variables. In particular, let $Y = \log(X/x_m)$ where $X \sim \text{Pareto}(x_m, \alpha)$. Then,

$$\Pr(Y > y) = \Pr(\log(X/x_m) > y) = \Pr(X > x_m e^y) = \left(\frac{x_m e^y}{x_m}\right)^{-\alpha} = e^{-\alpha y},$$

where the last expression is the c.c.d.f. of an Exponential distribution with rate α. This transformation turns out to be a powerful analytic tool, and we make use of it on multiple occasions in this book (e.g., Chapter 6, to study multiplicative processes, and Chapter 8, to derive properties of the maximum likelihood estimator for data from a Pareto distribution).

1.2.2 The Weibull Distribution

We just saw that the Pareto distribution has an intimate connection to the Exponential distribution – it is an *exponential* of the Exponential. The second heavy-tailed distribution we introduce has a similar connection to the Exponential distribution – it is a *polynomial* of the Exponential. Specifically, for $\alpha, \beta > 0$,

$$X \sim \text{Exponential}(1) \quad \iff \quad \frac{1}{\beta} X^{1/\alpha} \sim \text{Weibull}(\alpha, \beta).$$

From this relationship, one would expect that when $0 < \alpha < 1$, the Weibull distribution has a heavier tail than the Exponential (though lighter than the Pareto), making it heavy-tailed. On the other hand, when $\alpha > 1$, one would expect that the Weibull has a lighter tail than the Exponential.

It is straightforward to see what this transformation means for the c.c.d.f. of the Weibull distribution. In particular, a random variable follows a Weibull(α, β) distribution if

$$\bar{F}(x) = e^{-(\beta x)^\alpha}, \text{ for } x \geq 0. \tag{1.4}$$

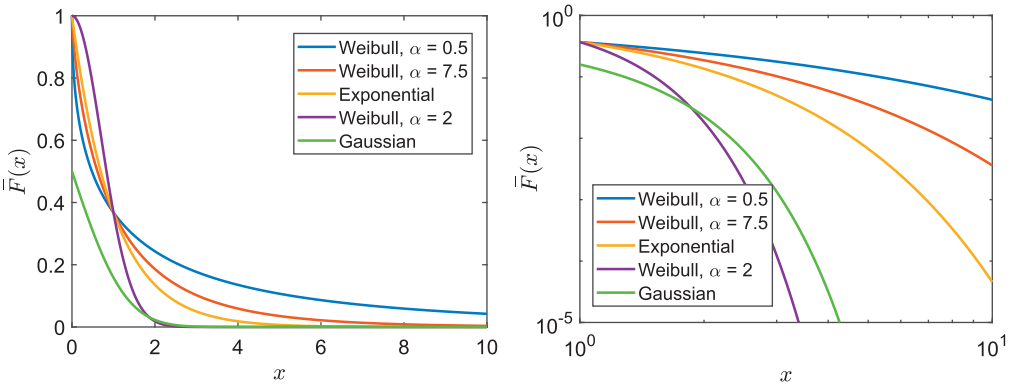
Differentiating the c.c.d.f. gives us the p.d.f.

$$f(x) = \alpha\beta\,(\beta x)^{\alpha-1}\, e^{-(\beta x)^\alpha}, \text{ for } x \geq 0.$$

In these expressions, α is referred to as the *shape parameter* of the distribution, and β is the *scale parameter*. Note that when $\alpha = 1$ the Weibull is equivalent to an Exponential(β).

In fact, the Weibull distribution is an especially helpful distribution when seeking to contrast heavy tails with light tails because it can be either heavy-tailed or light-tailed depending

https://doi.org/10.1017/9781009053730.002 Published online by Cambridge University Press



(a) c.c.d.f. on linear scale

(b) c.c.d.f. on log-log scale

Figure 1.6 Illustration of Weibull c.c.d.f.: Plots show the c.c.d.f. of the Weibull distribution with scale parameter $\beta = 1$ and different values of shape parameter α, alongside the Exponential distribution (with unit mean, which corresponds to $\alpha = \beta = 1$) and the standard Gaussian. Part (a) shows the c.c.d.f.s on a linear scale, while (b) plots them on a log-log scale.

on the shape parameter α. If $\alpha < 1$, then the Weibull is heavy-tailed, while if $\alpha \geq 1$, the Weibull is light-tailed. Mathematically, this can be verified by a quick calculation based on the definition of heavy-tailed distributions:

$$\limsup_{x\to\infty} \frac{\bar{F}(x)}{e^{-\mu x}} = \limsup_{x\to\infty} e^{\mu x - (\beta x)^\alpha}.$$

If $\alpha < 1$, this limit equals ∞ for any $\mu > 0$, while if $\alpha > 1$, it is 0 for any $\mu > 0$. (If $\alpha = 1$, the Weibull is equivalent to an Exponential distribution, which is, of course, light-tailed.)

Figure 1.6(a) illustrates the tail of the Weibull distribution for different values of α, contrasting the c.c.d.f. with those of the Gaussian and Exponential distributions. As was the case with the Pareto, the heaviness of the tail is clearly visible when we look at the log-log plot of the distribution; see Figure 1.6(b). While the Weibull looks nearly linear on a log-log plot when α is small (i.e., when the tail is heaviest) it is not perfectly linear like the Pareto distribution. To see why, we can take logarithms of both sides of (1.4) to obtain

$$\log \bar{F}(x) = -(\beta x)^\alpha.$$

While x^α gets close to $\log x$ as α shrinks to zero, it never entirely matches. However, if we move the negative sign to the other side and take logarithms again, we see that the Weibull looks linear according to a different scaling:

$$\underbrace{\log(-\log \bar{F}(x))}_{'y'} = \underbrace{\alpha \log \beta}_{y\text{-intercept}} + \underbrace{\alpha}_{\text{slope}} \underbrace{\log x}_{'x'}.$$

This tells us that the Weibull c.c.d.f. is linear on a $\log(-\log \bar{F}(x))$ versus $\log x$ plot. As with the Pareto distribution, this is a useful tool for exploratory analysis of data but one that must be used with care and should not be relied on for estimation.

https://doi.org/10.1017/9781009053730.002 Published online by Cambridge University Press

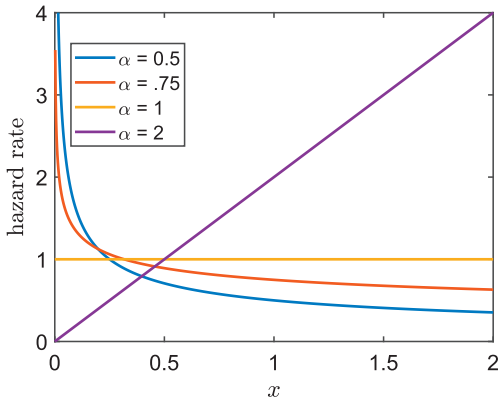


Figure 1.7 Weibull hazard rate for scale parameter $\beta = 1$ and different values of shape parameter α.

The Hazard Rate

The shape parameter α influences not just the tail behavior, but other properties of the Weibull distribution as well. One property that is of particular interest is the *hazard rate* (aka, failure rate) of the distribution. We study the hazard rate in detail in Chapter 4, and the Weibull is a particularly important distribution for that chapter because its hazard rate can have widely varying behaviors.

The hazard rate is defined as $q(t) = f(t)/\bar{F}(t)$ and has the following interpretation. Thinking of the distribution as capturing the *time to failure* (lifetime) of a component, $q(t)$ captures the instantaneous likelihood of a failure at time t of a component that entered into use at time 0, given that failure has not occurred until time t. Interestingly, when $\alpha > 1$, the hazard rate of the Weibull is increasing, meaning the likelihood of an impending failure increases with the age of the component; when $\alpha < 1$, the hazard rate is decreasing, meaning the likelihood of an impending failure actually decreases with the age of the component; and when $\alpha = 1$, the hazard rate is constant. We illustrate this in Figure 1.7.

The properties of the Weibull with respect to its hazard rate make it an extremely important distribution for survival analysis, reliability analysis, and failure analysis in a variety of areas. Additionally, the Weibull plays an important role in weather forecasting, specifically related to wind speed distributions and rainfall. As we discuss in Chapter 7, the Weibull (specifically, the mirror image of the Weibull distribution defined here) is an "extreme value distribution," which means that it is deeply connected to extreme events, such as the maximal rainfall in a day or year, the maximal overvoltage in an electrical system, or the maximal size of insurance claims. However, it was first used to describe the particle size distribution from milling and crushing operations in the 1930s [189]. Interestingly, though the distribution is named after Waloddi Weibull, who studied it in detail in the 1950s, it was introduced much earlier by Fréchet in 1927 in the context of extreme value theory [91].

Moments

An important difference between the Weibull distribution and the Pareto distribution is that all the moments of the Weibull distribution are finite. They can be large, especially when α is small, but they are not infinite.

https://doi.org/10.1017/9781009053730.002 Published online by Cambridge University Press

To express the moments, we need to use the gamma function, Γ, which is a continuous extension of the factorial function. Specifically, $\Gamma(n) = (n-1)!$, for integer n. More generally,

$$\Gamma(z) = \int_0^\infty x^{z-1} e^{-x} dx, \text{for } z > 0.$$

Using the gamma function, we can write the mean and variance of the Weibull as

$$\mathbb{E}[X] = \left(\frac{1}{\beta}\right) \Gamma(1 + 1/\alpha),$$
$$\text{Var}[X] = \left(\frac{1}{\beta}\right)^2 \left[\Gamma(1 + 2/\alpha) - (\Gamma(1 + 1/\alpha))^2\right].$$

Notice that the mean grows quickly as $\alpha \to 0$: it grows like the factorial of $1/\alpha$. More generally, the raw moments of the Weibull are given by

$$\mathbb{E}[X^n] = \left(\frac{1}{\beta}\right)^n \Gamma(1 + n/\alpha).$$

1.2.3 The LogNormal distribution

While both the Pareto and the Weibull can be viewed as transformations of the Exponential distribution, as its name would suggest, the LogNormal distribution is a transformation of the Normal (aka Gaussian) distribution. In fact, the transformation of the Gaussian distribution that produces the LogNormal distribution is the same transformation that creates the Pareto from the Exponential – the LogNormal distribution is an *exponential* of the Gaussian distribution. Specifically,

$$X \sim \text{LogNormal}(\mu, \sigma^2) \quad \iff \quad \log(X) \sim \text{Gaussian}(\mu, \sigma^2).$$

Or, equivalently,

$$Z \sim \text{Gaussian}(\mu, \sigma^2) \quad \iff \quad e^Z \sim \text{LogNormal}(\mu, \sigma^2).$$

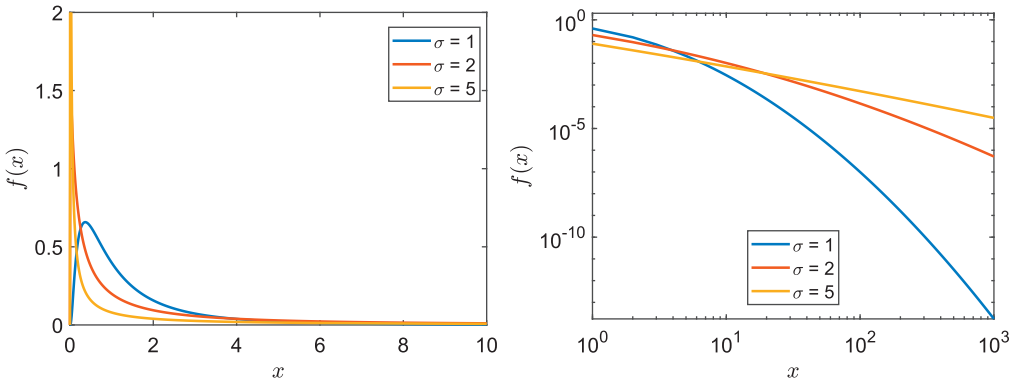
This means that the LogNormal distribution can be specified in terms of the Gaussian distribution via a logarithmic transformation. For example, the p.d.f. of a LogNormal(μ, σ^2) distribution is

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma x} e^{-(\log x - \mu)^2/(2\sigma^2)}. \tag{1.5}$$

Note that the change of variables of $\log x$ introduces a $1/x$ term outside of the exponential in the p.d.f. as compared to the Gaussian distribution. This connection with the Gaussian distribution can also be used to show that the LogNormal distribution is heavy-tailed; this is left as an exercise for the reader (see Exercise 3).

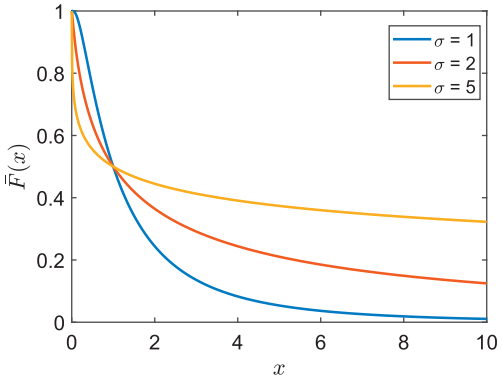
While it may not be evident from the functional form of the p.d.f., the LogNormal distribution has a shape that is quite similar to that of the Pareto distribution. We illustrate the p.d.f. and c.c.d.f. in Figure 1.8. In fact, even when viewed on a log-log plot, the LogNormal and the Pareto can look similar. Specifically, when the variance parameter σ^2 is large, the

https://doi.org/10.1017/9781009053730.002 Published online by Cambridge University Press

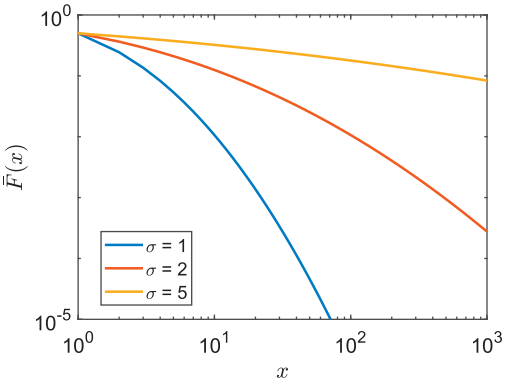


(a) p.d.f. on linear scale

(b) p.d.f. on log-log scale



(c) c.c.d.f. on linear scale



(d) c.c.d.f. on log-log scale

Figure 1.8 Illustration of LogNormal distribution: The c.c.d.f. and p.d.f. of LogNormal distributions with $\mu = 0$ and different values of σ are depicted. The p.d.f.s are plotted on a linear scale in (a), and on a log-log scale in (b). The corresponding c.c.d.f.s are plotted on a linear scale in (c), and on a log-log scale in (d). Note that the p.d.f. as well as the c.c.d.f. appears nearly linear on a log-log plot when σ is large.

LogNormal p.d.f. looks nearly linear on the log-log plot. To see why, let us take logarithms of both sides of (1.5):

$$\begin{aligned}\underbrace{\log f(x)}_{'y'} &= -\log x - \log(\sigma\sqrt{2\pi}) - \frac{(\log x - \mu)^2}{2\sigma^2} \\ &= -\frac{(\log x)^2}{2\sigma^2} + \left(\frac{\mu}{\sigma^2} - 1\right)\underbrace{\log x}_{'x'} - \log(\sigma\sqrt{2\pi}) - \frac{\mu^2}{2\sigma^2}.\end{aligned}$$

This calculation shows that when σ is sufficiently large, the quadratic term above will be small for a large range of x and so the log-log plot will look nearly linear. Consequently, it is nearly impossible to distinguish the LogNormal from a Pareto using the log-log plot. Hence, one should be very careful when using the log-log plot as a statistical tool. We emphasize this point further in Chapter 8.

https://doi.org/10.1017/9781009053730.002 Published online by Cambridge University Press

Properties

The LogNormal inherits many of the useful properties of the Gaussian distribution, with suitable adjustments owing to the exponential transformation between the distributions.

Perhaps the most important property of the Gaussian is that the sum of independent Gaussians is a Gaussian. Because of the exponential transformation, for the LogNormal, this property holds for the product rather than the sum. In particular, suppose $Y_i \sim \text{LogNormal}(\mu_i, \sigma_i^2)$ are n independent random variables, Then

$$Y = \prod_{i=1}^{n} Y_i \quad \Longrightarrow \quad Y \sim \text{LogNormal}\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right).$$

This suggests that the LogNormal distribution is intimately tied to the growth of *multiplicative processes*. In particular, if a process grows multiplicatively, then it is additive on a logarithmic scale and, by the central limit theorem, it is likely to be Gaussian on the logarithmic scale. This, in turn, means that it is a LogNormal in the original scale. As a consequence, the LogNormal is a very common distribution in nature and human behavior. It has been used to model phenomena in finance, computer networks, hydrology, biology, medicine, and more. In fact, the LogNormal distribution was first studied by Robert Gibrat in the context of deriving a multiplicative version of the central limit theorem, which is sometimes termed "Gibrat's Law." Gibrat formulated this law during his study of the dynamics of firm sizes and industry structure [203]. We devote Chapter 6 to a discussion of multiplicative versions of the central limit theorem and their connections to heavy-tailed distributions.

Beyond products, LogNormal distributions also behave pleasantly with respect to other transformations. An important example is that

$$X \sim \text{LogNormal}(\mu, \sigma^2) \quad \Longrightarrow \quad X^a \sim \text{LogNormal}(a\mu, a^2\sigma^2) \text{ for } a \neq 0. \tag{1.6}$$

Moments

Like the Weibull distribution, the moments of the LogNormal are always finite. They can be quite large but are never infinite. Perhaps the most counterintuitive thing about the moments of the LogNormal distribution is that, while we adopt the same parameter names as for the Gaussian, μ and σ^2 do not refer to the mean and variance of the LogNormal. Instead, they refer to the mean and variance of the Gaussian that is obtained by taking the log of the LogNormal. The mean and variance of the LogNormal are as follows:

$$\begin{aligned} \mathbb{E}[X] &= e^{\mu+\sigma^2/2}, \\ \text{Var}[X] &= e^{2\mu+\sigma^2}\left(e^{\sigma^2} - 1\right). \end{aligned}$$

The fact that mean and variance are exponentials of the distribution's parameters emphasizes that one should expect them to be large. More generally, the raw moments of the LogNormal distribution are given by

$$\mathbb{E}[X^n] = e^{n\mu+\frac{1}{2}n^2\sigma^2}.$$

Interestingly, though all the moments of the LogNormal distribution are finite, the distribution is not uniquely determined by its (integral) moments.

https://doi.org/10.1017/9781009053730.002 Published online by Cambridge University Press

1.2.4 Other Heavy-Tailed Distributions

The Pareto, Weibull, and LogNormal are the most commonly used heavy-tailed distributions, but there are also other heavy-tailed distributions that appear frequently. We end this chapter by briefly introducing a few other distributions that come up later in the book as important examples of the concepts we discuss.

The Cauchy Distribution

The Cauchy distribution is an important distribution in statistics and is strongly connected to the central limit theorem, as we discuss in Chapter 5. However, it is most often used as a pathological example as a result of the fact that it does not have a well-defined mean (or variance). In fact, though it is named after Cauchy, the first explicit analysis of it was conducted by Poisson in 1824 in order to provide a counterexample showing that the variance condition in the central limit theorem cannot be dropped.

The c.d.f. and p.d.f. of a Cauchy(x_0, γ) distribution are given, for $x \in \mathbb{R}$, by

$$F(x) = \frac{1}{\pi} \arctan\left(\frac{x - x_0}{\gamma}\right) + \frac{1}{2},$$
$$f(x) = \frac{1}{\pi\gamma}\left(\frac{\gamma^2}{(x - x_0)^2 + \gamma^2}\right),$$

with location parameter $x_0 \in \mathbb{R}$ and scale parameter $\gamma > 0$. The distribution is plotted in Figure 1.9.

While the distribution function looks complicated, the Cauchy has a simple representation as the ratio of two Gaussian random variables. Specifically, if U and V are independent Gaussian random variables with mean 0 and variance 1, then $U/V \sim$ Cauchy$(0, 1)$ (see Exercise 12). A Cauchy$(0, 1)$ is referred to as the *standard Cauchy* and is important in its own right because it coincides with the *Student's t-distribution*, which is crucially important for estimating the mean and variance of a Gaussian distribution from data.

The Cauchy distribution's emergence in the context of the central limit theorem is a result of the fact that sums of Cauchy distributions have a property similar to sums of Gaussian distributions: if $X_1, \ldots, X_n$ are i.i.d. Cauchy$(0, 1)$ random variables, then the sum is also a Cauchy. Specifically, $\frac{1}{n}\sum_{i=1}^{n} X_i \sim$ Cauchy$(0, 1)$; we prove this property in Chapter 5 using characteristic functions.

Finally, a related distribution to the Cauchy is the LogCauchy, which has the same relationship to the Cauchy distribution that the LogNormal has to the Gaussian distribution, that is,

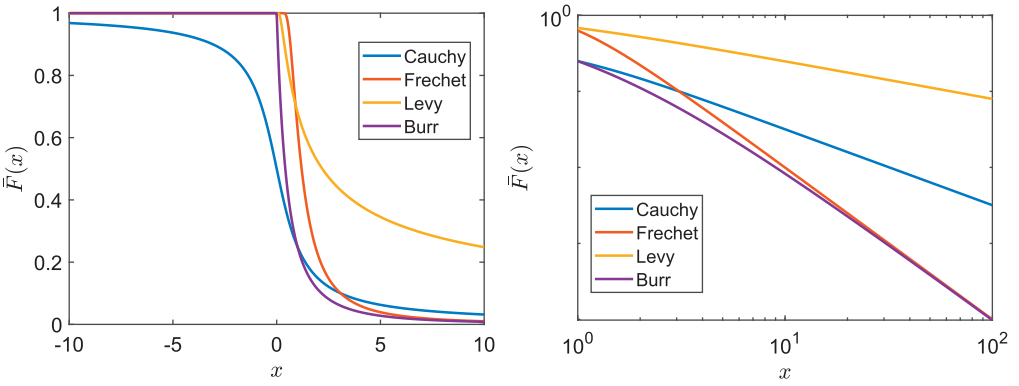
$$X \sim \text{Cauchy}(x_0, \gamma) \quad \iff \quad e^X \sim \text{LogCauchy}(x_0, \gamma),$$

or, equivalently,

$$Y \sim \text{LogCauchy}(x_0, \gamma) \quad \iff \quad \log(Y) \sim \text{Cauchy}(x_0, \gamma).$$

The LogCauchy is one of the few common distributions that has a heavier tail than the Pareto distribution – it has a logarithmically decaying tail. For this reason, it is sometimes referred to as a *super-heavy-tailed distribution*.

https://doi.org/10.1017/9781009053730.002 Published online by Cambridge University Press



(a) c.c.d.f.s (b) c.c.d.f.s on log-log scale

Figure 1.9 The plots show the c.c.d.f. of the standard Cauchy ($x_0 = 0, \gamma = 1$), the Fréchet (with $x_m = 0, \beta = 1, \alpha = 2$), the Lévy (with $\mu = 0, c = 1$), and the Burr distribution (with $c = \lambda = 1, k = 2$). Part (a) shows the plots on a linear scale, and (b) on a log-log scale. Note that all c.c.d.f.s look (asymptotically) linear on a log-log scale; we will formalize this property in Chapter 2.

The Fréchet Distribution

The Fréchet distribution plays a central role in extreme value theory, as we discuss in Chapter 7. It is commonly used in hydrology when studying the extremes of rainfall distributions.

The distribution is named after Maurice Fréchet, who introduced it in 1927; however, it is also referred to as the inverse Weibull distribution, which is a much more descriptive name since it is defined as exactly that. Specifically,

$$X \sim \text{Weibull}(\alpha, \beta) \quad \iff \quad 1/X \sim \text{Fréchet}(\alpha, \beta, 0).$$

More generally, the c.d.f. and p.d.f. of the Fréchet(α, β, x_m) distribution are given, for $x > x_m$, by

$$F(x) = e^{-(\beta(x-x_m))^{-\alpha}},$$
$$f(x) = \alpha\beta\left(\beta(x-x_m)\right)^{-1-\alpha} e^{-(\beta(x-x_m))^{-\alpha}}.$$

Here, $\alpha > 0$ is the shape parameter, $\beta > 0$ is the scale parameter, and x_m is the minimum value taken by the distribution. The distribution is plotted in Figure 1.9.

The Lévy Distribution

The Lévy distribution is used most prominently in the study of financial models to explain stylized phenomena such as volatility clustering. Within mathematics and physics, it also plays an important role in the study of Brownian motion: the hitting time of a single point at a fixed distance from the starting point of a Brownian motion has a Lévy distribution. But perhaps its most prominent use is in the context of the generalized central limit theorem, which we discuss in Chapter 5.

Like the Cauchy distribution and the LogNormal distribution, the Lévy distribution is most conveniently defined as a transformation of the Gaussian distribution. In particular, a

https://doi.org/10.1017/9781009053730.002 Published online by Cambridge University Press

Lévy distribution coincides with the square of the inverse of a Gaussian distribution (and is therefore sometimes also called the inverse Gaussian):

$$Z \sim \text{Gaussian}(\mu, \sigma^2) \quad \Longrightarrow \quad \frac{1}{(Z-\mu)^2} \sim \text{Lévy}(0, 1/\sigma^2).$$

More directly, the p.d.f. of the Lévy(μ, c) distribution is given, for $x \in \mathbb{R}$, by

$$f(x) = \sqrt{\frac{c}{2\pi(x-\mu)^3}}\, e^{-\frac{c}{2(x-\mu)}}. \tag{1.7}$$

From this equation, it is straightforward to see that the Lévy distribution is not just heavy-tailed, it is more specifically a "power law" distribution, since $f(x) \sim \sqrt{\frac{c}{2\pi}} x^{-3/2}$. A plot of the distribution is shown in Figure 1.9.

The Burr Distribution

The Burr distribution is a generalization of the Pareto distribution that often appears in statistics and econometrics. It is most frequently used in the study of household incomes and related wealth distributions. It was introduced by Irving Burr in 1942 as one of a family of 12 distributions, of which it is the *Burr Type XII distribution*. The c.c.d.f. and p.d.f. of a Burr(c, k, λ) distribution are given, for $x > 0$, by

$$\bar{F}(x) = (1 + \lambda x^c)^{-k}, \tag{1.8}$$

$$f(x) = \frac{ckx^{c-1}}{(1+x^c)^{k+1}}, \tag{1.9}$$

where $c, k, \lambda > 0$. The distribution is illustrated in Figure 1.9. When $c = 1$, the Burr corresponds to a so-called Type II Pareto distribution, and it is easy to see that it is a power law distribution, like the Cauchy, Fréchet, and Lévy distributions. In Chapter 7, we discuss an interesting connection between the Burr distribution and the residual life of the Pareto distribution. The hazard rate of the Burr distribution itself serves as an important counterexample in the same chapter.

The Zipf Distribution

We conclude by mentioning the Zipf distribution, which is a discrete version of the Pareto distribution. The Zipf distribution rose to prominence because of "Zipf's law," which states that, given a natural language corpus, the frequency of any word is inversely proportional to its rank in the frequency table of the corpus. That is, the most common word occurs twice as often as the second most common word, three times more than the third most common word, and so on. This law is named after George Zipf, who popularized it in 1935; however, the observation of the phenomenon predated his work by more than fifty years.

The Zipf(s, N) distribution is one example of a distribution that would explain Zipf's law, and is defined in terms of its probability mass function (p.m.f.):

$$p(n; s, N) = \frac{1/n^s}{\sum_{i=1}^{N} 1/i^s}, \tag{1.10}$$

where N can be thought of as the number of elements in the corpus, and s is the exponent characterizing the power law.

https://doi.org/10.1017/9781009053730.002 Published online by Cambridge University Press

While the Zipf distribution is not heavy-tailed, given that it has a finite support, its generalization to the case $N = \infty$, which is called the Zeta distribution, is heavy-tailed (for $s > 1$).

1.3 What's Next

In this chapter we have introduced the definition of the class of heavy-tailed distributions, along with a few examples of common heavy-tailed distributions. Through these examples, you have already seen some illustrations of how heavy-tailed distributions behave differently from light-tailed distributions. But we have not yet sought to build intuition about these differences or to explain why heavy-tailed distributions are so common in the world around us. We have mentioned that controversy often surrounds heavy-tailed distributions because intuitive statistical approaches for identifying heavy tails in data are flawed, but we have not yet provided tools for correct identification and estimation of heavy-tailed phenomena.

The remainder of this book is organized to first provide intuition, both qualitative and mathematical, for the defining properties of heavy-tailed distributions (Part I: Properties), then explain why heavy-tailed distributions are so common in the world around us (Part II: Emergence), and finally develop the statistical tools for the estimation of heavy-tailed distributions (Part III: Estimation).

Given the mystique and excitement that surrounds the discovery of heavy-tailed phenomena, the detection and estimation of heavy tails in data is a task that is often (over)zealously pursued. While reading this book, you may be tempted to skip directly to Part III on estimation. However, the book is written so that the tools used in Part II are developed in Part I, and the tools used in Part III are developed in Parts I and II. Thus, we encourage readers to work through the book in order. That said, we have organized the material in each chapter so that there is a main body that presents the core ideas that are important for later chapters, followed by sections that present examples and/or variations of the main topic. These later sections can be viewed as enrichment opportunities that can be skipped as desired if the goal is to move quickly to Part III. However, if one is looking for the quickest path to understand the background needed before digging into Part III, then we recommend focusing on Chapter 2 from Part I, then Chapters 5 and 7 from Part II before moving to Part III.

Our goal for this book is that, through reading it, heavy-tailed distributions will be demystified for you. That their properties will be intuitive, not mysterious. That their emergence will be expected, not surprising. And that you will have the proper statistical tools for studying heavy-tailed phenomena and so will be able to resolve (or avoid) controversies rather than feed them. Happy reading!

1.4 Exercises

1. For a standard Gaussian random variable Z, show that for $x > 0$,

$$\Pr(Z > x) \leq \frac{e^{-x^2/2}}{\sqrt{2\pi}x}.$$

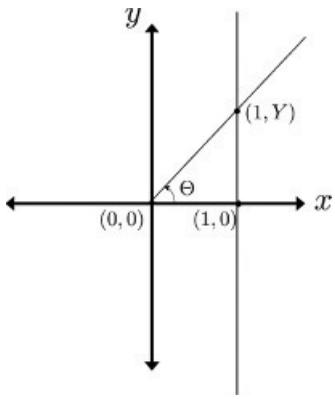
Note: In fact, the above bound can be shown to be asymptotically tight, that is, it can be shown that $\Pr(Z > x) \sim \frac{e^{-x^2/2}}{\sqrt{2\pi}x}$; *see [81, Chapter 7].*

https://doi.org/10.1017/9781009053730.002 Published online by Cambridge University Press

2. Use the bound of Exercise 1 to prove that the Gaussian(μ, σ^2) distribution is light-tailed.
3. Prove that the LogNormal(μ, σ^2) distribution is heavy-tailed.
4. Consider a distribution F over $\mathbb{R}_+$ with finite mean μ. The *excess* distribution F_e corresponding to F is defined as

$$\bar{F}_e(x) = \frac{1}{\mu} \int_x^\infty \bar{F}(y) dy.$$

   Prove that F is heavy-tailed if and only if F_e is heavy-tailed.
5. Let $X \sim$ Exponential(μ) and $Y = 1/X$. Prove that Y is heavy-tailed.
6. The random variable N takes values in $\mathbb{N}$. The distribution of N, conditioned on a uniformly distributed random variable U taking values in $(0, 1)$, is given by $\Pr(N > n \mid U) = U^n$. Assuming U is uniformly distributed, show that N is heavy-tailed.
   Note: Even though the conditional distribution of N given the value of U is light-tailed (in fact, Geometrically distributed), N itself is heavy-tailed!
7. In Exercise 6, you do not need U to be uniformly distributed for N to be heavy-tailed. Prove that, so long as $\Pr(U > x) > 0$ for all $x \in (0, 1)$, N is heavy-tailed.
8. Derive an expression for the Gini coefficient corresponding to the Pareto distribution. Show that the Gini coefficient converges to 1 as tail index $\alpha \downarrow 1$.
9. Compute the Lorenz curve corresponding to the Exponential distribution. Prove that the Gini coefficient in this case equals $1/2$.
10. Prove property (1.6) of the LogNormal distribution.
11. The goal of this exercise is to prove the following geometric interpretation of the standard Cauchy distribution. On the Cartesian plane, draw a random line passing through the origin, making an angle Θ with the x-axis as shown in the following figure, where Θ is uniformly distributed over $(-\pi/2, \pi/2)$. Let $(1, Y)$ denote the point where this random line intersects the vertical line $x = 1$. Prove that Y is a standard Cauchy random variable.



12. Prove that if U and V are independent, standard Gaussian random variables, then U/V is a standard Cauchy.

https://doi.org/10.1017/9781009053730.002 Published online by Cambridge University Press

Hint: The geometric interpretation from Exercise 11 might help. Interpreting (U, V) *as the Cartesian coordinates of a random point on the Cartesian plane, what is the joint distribution of the polar coordinates* (R, Θ)*?*

13. The goal of this exercise is to compare the "heaviness" of the tails of the Pareto, Weibull, and LogNormal distributions. Let $X \sim \text{Pareto}(x_m, \alpha_1)$, $Y \sim \text{LogNormal}(\mu, \sigma^2)$, and $Z \sim \text{Weibull}(\alpha_2, \beta_2)$. Prove that

$$\lim_{x\to\infty} \frac{\Pr(Z > x)}{\Pr(Y > x)} = 0, \quad \lim_{x\to\infty} \frac{\Pr(Y > x)}{\Pr(X > x)} = 0.$$

Note: This exercise shows that the Pareto has a heavier tail than the LogNormal, which has a heavier tail than the Weibull.

https://doi.org/10.1017/9781009053730.002 Published online by Cambridge University Press