1

Probability theory: basic notions

> All epistemological value of the theory of probability is based on this: that large scale random phenomena in their collective action create strict, non random regularity.
>
> (Gnedenko and Kolmogorov, *Limit Distributions for Sums of Independent Random Variables*.)

1.1 Introduction

Randomness stems from our incomplete knowledge of reality, from the lack of information which forbids a perfect prediction of the future. Randomness arises from complexity, from the fact that causes are diverse, that tiny perturbations may result in large effects. For over a century now, Science has abandoned Laplace's deterministic vision, and has fully accepted the task of deciphering randomness and inventing adequate tools for its description. The surprise is that, after all, randomness has many facets and that there are many levels to uncertainty, but, above all, that a new form of predictability appears, which is no longer deterministic but *statistical*.

Financial markets offer an ideal testing ground for these statistical ideas. The fact that a large number of participants, with divergent anticipations and conflicting interests, are simultaneously present in these markets, leads to unpredictable behaviour. Moreover, financial markets are (sometimes strongly) affected by external news – which are, both in date and in nature, to a large degree unexpected. The statistical approach consists in drawing from past observations some information on the frequency of possible price changes. If one then assumes that these frequencies reflect some intimate mechanism of the markets themselves, then one may hope that these frequencies will remain stable in the course of time. For example, the mechanism underlying the roulette or the game of dice is obviously always the same, and one expects that the frequency of all possible outcomes will be invariant in time – although of course each individual outcome is random.

This 'bet' that probabilities are stable (or better, stationary) is very reasonable in the case of roulette or dice;[†] it is nevertheless much less justified in the case of financial markets – despite the large number of participants which confer to the system a certain

[†] The idea that science ultimately amounts to making the best possible guess of reality is due to R. P. Feynman (Seeking New Laws, in *The Character of Physical Laws*, MIT Press, Cambridge, MA, 1965).

regularity, at least in the sense of Gnedenko and Kolmogorov. It is clear, for example, that financial markets do not behave now as they did 30 years ago: many factors contribute to the evolution of the way markets behave (development of derivative markets, world-wide and computer-aided trading, etc.). As will be mentioned below, 'young' markets (such as emergent countries markets) and more mature markets (exchange rate markets, interest rate markets, etc.) behave quite differently. The statistical approach to financial markets is based on the idea that whatever evolution takes place, this happens sufficiently *slowly* (on the scale of several years) so that the observation of the recent past is useful to describe a not too distant future. However, even this 'weak stability' hypothesis is sometimes badly in error, in particular in the case of a crisis, which marks a sudden change of market behaviour. The recent example of some Asian currencies indexed to the dollar (such as the Korean won or the Thai baht) is interesting, since the observation of past fluctuations is clearly of no help to predict the amplitude of the sudden turmoil of 1997, see Figure 1.1.

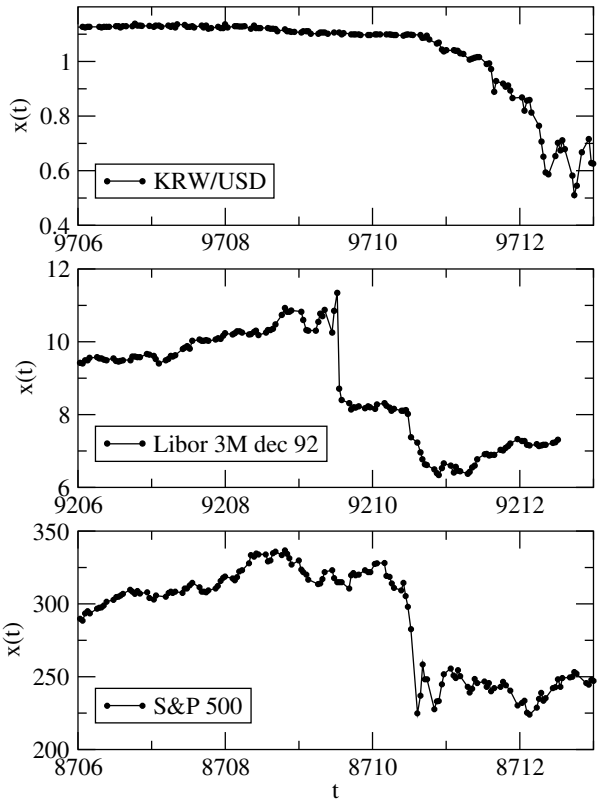

Fig. 1.1. Three examples of statistically unforeseen crashes: the Korean won against the dollar in 1997 (top), the British 3-month short-term interest rates futures in 1992 (middle), and the S&P 500 in 1987 (bottom). In the example of the Korean won, it is particularly clear that the distribution of price changes before the crisis was extremely narrow, and could not be extrapolated to anticipate what happened in the crisis period.

Hence, the statistical description of financial fluctuations is certainly imperfect. It is nevertheless extremely helpful: in practice, the 'weak stability' hypothesis is in most cases reasonable, at least to describe *risks*.[†]

In other words, the amplitude of the possible price changes (but not their sign!) is, to a certain extent, predictable. It is thus rather important to devise adequate tools, in order to *control* (if at all possible) financial risks. The goal of this first chapter is to present a certain number of basic notions in probability theory which we shall find useful in the following. Our presentation does not aim at mathematical rigour, but rather tries to present the key concepts in an intuitive way, in order to ease their empirical use in practical applications.

1.2 Probability distributions

Contrarily to the throw of a dice, which can only return an integer between 1 and 6, the variation of price of a financial asset[‡] can be arbitrary (we disregard the fact that price changes cannot actually be smaller than a certain quantity – a 'tick'). In order to describe a random process X for which the result is a real number, one uses a probability density $P(x)$, such that the probability that X is within a small interval of width $\mathrm{d}x$ around $X = x$ is equal to $P(x)\,\mathrm{d}x$. In the following, we shall denote as $P(\cdot)$ the probability density for the variable appearing as the argument of the function. This is a potentially ambiguous, but very useful notation.

The probability that X is between a and b is given by the integral of $P(x)$ between a and b,

$$\mathcal{P}(a < X < b) = \int_a^b P(x)\,\mathrm{d}x. \tag{1.1}$$

In the following, the notation $\mathcal{P}(\cdot)$ means the probability of a given event, defined by the content of the parentheses $(\cdot)$.

The function $P(x)$ is a density; in this sense it depends on the units used to measure X. For example, if X is a length measured in centimetres, $P(x)$ is a probability density per unit length, i.e. per centimetre. The numerical value of $P(x)$ changes if X is measured in inches, but the probability that X lies between two specific values l_1 and l_2 is of course independent of the chosen unit. $P(x)\,\mathrm{d}x$ is thus invariant upon a change of unit, i.e. under the change of variable $x \to \gamma x$. More generally, $P(x)\,\mathrm{d}x$ is invariant upon any (monotonic) change of variable $x \to y(x)$: in this case, one has $P(x)\,\mathrm{d}x = P(y)\,\mathrm{d}y$.

In order to be a probability density in the usual sense, $P(x)$ must be non-negative ($P(x) \geq 0$ for all x) and must be normalized, that is that the integral of $P(x)$ over the whole range of possible values for X must be equal to one:

$$\int_{x_m}^{x_M} P(x)\,\mathrm{d}x = 1, \tag{1.2}$$

[†] The prediction of *future returns* on the basis of past returns is however much less justified.

[‡] **Asset** is the generic name for a financial instrument which can be bought or sold, like stocks, currencies, gold, bonds, etc.

where x_m (resp. x_M) is the smallest value (resp. largest) which X can take. In the case where the possible values of X are not bounded from below, one takes $x_m = -\infty$, and similarly for x_M. One can actually always assume the bounds to be $\pm\infty$ by setting to zero $P(x)$ in the intervals $]-\infty, x_m]$ and $[x_M, \infty[$. Later in the text, we shall often use the symbol $\int$ as a shorthand for $\int_{-\infty}^{+\infty}$.

An equivalent way of describing the distribution of X is to consider its cumulative distribution $\mathcal{P}_<(x)$, defined as:

$$\mathcal{P}_<(x) \equiv \mathcal{P}(X < x) = \int_{-\infty}^{x} P(x')\,\mathrm{d}x'. \tag{1.3}$$

$\mathcal{P}_<(x)$ takes values between zero and one, and is monotonically increasing with x. Obviously, $\mathcal{P}_<(-\infty) = 0$ and $\mathcal{P}_<(+\infty) = 1$. Similarly, one defines $\mathcal{P}_>(x) = 1 - \mathcal{P}_<(x)$.

1.3 Typical values and deviations

It is quite natural to speak about 'typical' values of X. There are at least three mathematical definitions of this intuitive notion: the **most probable** value, the **median** and the **mean**. The most probable value x^* corresponds to the maximum of the function $P(x)$; x^* needs not be unique if $P(x)$ has several equivalent maxima. The median x_{med} is such that the probabilities that X be greater or less than this particular value are equal. In other words, $\mathcal{P}_<(x_{\text{med}}) = \mathcal{P}_>(x_{\text{med}}) = \frac{1}{2}$. The mean, or **expected value** of X, which we shall note as m or $\langle x \rangle$ in the following, is the average of all possible values of X, weighted by their corresponding probability:

$$m \equiv \langle x \rangle = \int x\,P(x)\,\mathrm{d}x. \tag{1.4}$$

For a unimodal distribution (unique maximum), symmetrical around this maximum, these three definitions coincide. However, they are in general different, although often rather close to one another. Figure 1.2 shows an example of a non-symmetric distribution, and the relative position of the most probable value, the median and the mean.

One can then describe the fluctuations of the random variable X: if the random process is repeated several times, one expects the results to be scattered in a cloud of a certain 'width' in the region of typical values of X. This width can be described by the **mean absolute deviation** (MAD) E_{abs}, by the **root mean square** (RMS) σ (or, **standard deviation**), or by the 'full width at half maximum' $w_{1/2}$.

The mean absolute deviation from a given reference value is the average of the distance between the possible values of X and this reference value,[†]

$$E_{\text{abs}} \equiv \int |x - x_{\text{med}}|\,P(x)\,\mathrm{d}x. \tag{1.5}$$

[†] One chooses as a reference value the median for the MAD and the mean for the RMS, because for a fixed distribution $P(x)$, these two quantities minimize, respectively, the MAD and the RMS.

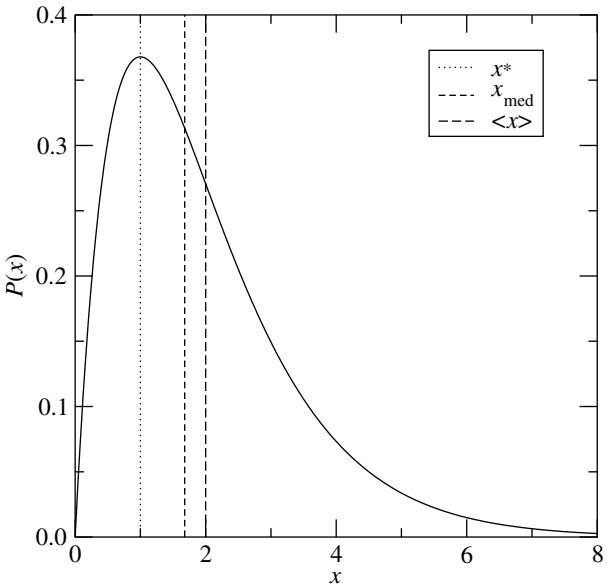

Fig. 1.2. The 'typical value' of a random variable X drawn according to a distribution density $P(x)$ can be defined in at least three different ways: through its mean value $\langle x \rangle$, its most probable value x^* or its median x_{med}. In the general case these three values are distinct.

Similarly, the **variance** (σ^2) is the mean distance squared to the reference value m,

$$\sigma^2 \equiv \langle (x-m)^2 \rangle = \int (x-m)^2 P(x)\,\mathrm{d}x. \tag{1.6}$$

Since the variance has the dimension of x squared, its square root (the RMS, σ) gives the order of magnitude of the fluctuations around m.

Finally, the full width at half maximum $w_{1/2}$ is defined (for a distribution which is symmetrical around its unique maximum x^*) such that $P(x^* \pm (w_{1/2})/2) = P(x^*)/2$, which corresponds to the points where the probability density has dropped by a factor of two compared to its maximum value. One could actually define this width slightly differently, for example such that the total probability to find an event outside the interval $[(x^* - w/2), (x^* + w/2)]$ is equal to, say, 0.1. The corresponding value of w is called a quantile. This definition is important when the distribution has very fat tails, such that the variance or the mean absolute deviation are infinite.

The pair mean–variance is actually much more popular than the pair median–MAD. This comes from the fact that the absolute value is not an analytic function of its argument, and thus does not possess the nice properties of the variance, such as additivity under convolution, which we shall discuss in the next chapter. However, for the empirical study of fluctuations, it is sometimes preferable to use the MAD; it is more *robust* than the variance, that is, less sensitive to rare extreme events, which may be the source of large statistical errors.

1.4 Moments and characteristic function

More generally, one can define higher-order **moments** of the distribution $P(x)$ as the average of powers of X:

$$m_n \equiv \langle x^n \rangle = \int x^n P(x)\,\mathrm{d}x. \tag{1.7}$$

Accordingly, the mean m is the first moment ($n = 1$), and the variance is related to the second moment ($\sigma^2 = m_2 - m^2$). The above definition, Eq. (1.7), is only meaningful if the integral converges, which requires that $P(x)$ decreases sufficiently rapidly for large $|x|$ (see below).

From a theoretical point of view, the moments are interesting: if they exist, their knowledge is often equivalent to the knowledge of the distribution $P(x)$ itself.† In practice however, the high order moments are very hard to determine satisfactorily: as n grows, longer and longer time series are needed to keep a certain level of precision on m_n; these high moments are thus in general not adapted to describe empirical data.

For many computational purposes, it is convenient to introduce the **characteristic function** of $P(x)$, defined as its Fourier transform:

$$\hat{P}(z) \equiv \int \mathrm{e}^{\mathrm{i}zx} P(x)\,\mathrm{d}x. \tag{1.8}$$

The function $P(x)$ is itself related to its characteristic function through an inverse Fourier transform:

$$P(x) = \frac{1}{2\pi} \int \mathrm{e}^{-\mathrm{i}zx} \hat{P}(z)\,\mathrm{d}z. \tag{1.9}$$

Since $P(x)$ is normalized, one always has $\hat{P}(0) = 1$. The moments of $P(x)$ can be obtained through successive derivatives of the characteristic function at $z = 0$,

$$m_n = (-\mathrm{i})^n \left.\frac{\mathrm{d}^n}{\mathrm{d}z^n} \hat{P}(z)\right|_{z=0}. \tag{1.10}$$

One finally defines the **cumulants** c_n of a distribution as the successive derivatives of the logarithm of its characteristic function:

$$c_n = (-\mathrm{i})^n \left.\frac{\mathrm{d}^n}{\mathrm{d}z^n} \log \hat{P}(z)\right|_{z=0}. \tag{1.11}$$

The cumulant c_n is a polynomial combination of the moments m_p with $p \leq n$. For example $c_2 = m_2 - m^2 = \sigma^2$. It is often useful to normalize the cumulants by an appropriate power of the variance, such that the resulting quantities are dimensionless. One thus defines the **normalized cumulants** λ_n,

$$\lambda_n \equiv c_n/\sigma^n. \tag{1.12}$$

† This is not rigorously correct, since one can exhibit examples of different distribution densities which possess exactly the same moments, see Section 1.7 below.

One often uses the third and fourth normalized cumulants, called the **skewness** (ς) and **kurtosis** (κ),[†]

$$\varsigma \equiv \lambda_3 = \frac{\langle (x-m)^3 \rangle}{\sigma^3} \qquad \kappa \equiv \lambda_4 = \frac{\langle (x-m)^4 \rangle}{\sigma^4} - 3. \tag{1.13}$$

The above definition of cumulants may look arbitrary, but these quantities have remarkable properties. For example, as we shall show in Section 2.2, the cumulants simply add when one sums independent random variables. Moreover a Gaussian distribution (or the normal law of Laplace and Gauss) is characterized by the fact that all cumulants of order larger than two are identically zero. Hence the cumulants, in particular κ, can be interpreted as a measure of the distance between a given distribution $P(x)$ and a Gaussian.

1.5 Divergence of moments – asymptotic behaviour

The moments (or cumulants) of a given distribution do not always exist. A necessary condition for the nth moment (m_n) to exist is that the distribution density $P(x)$ should decay faster than $1/|x|^{n+1}$ for $|x|$ going towards infinity, or else the integral, Eq. (1.7), would diverge for $|x|$ large. If one only considers distribution densities that are behaving asymptotically as a power-law, with an exponent $1 + \mu$,

$$P(x) \sim \frac{\mu A_{\pm}^{\mu}}{|x|^{1+\mu}} \text{ for } x \to \pm\infty, \tag{1.14}$$

then all the moments such that $n \geq \mu$ are infinite. For example, such a distribution has no finite variance whenever $\mu \leq 2$. [Note that, for $P(x)$ to be a normalizable probability distribution, the integral, Eq. (1.2), must converge, which requires $\mu > 0$.]

> *The characteristic function of a distribution having an asymptotic power-law behaviour given by Eq. (1.14) is non-analytic around $z = 0$. The small z expansion contains regular terms of the form z^n for $n < \mu$ followed by a non-analytic term $|z|^{\mu}$ (possibly with logarithmic corrections such as $|z|^{\mu} \log z$ for integer μ). The derivatives of order larger or equal to μ of the characteristic function thus do not exist at the origin ($z = 0$).*

1.6 Gaussian distribution

The most commonly encountered distributions are the 'normal' laws of Laplace and Gauss, which we shall simply call **Gaussian** in the following. Gaussians are ubiquitous: for example, the number of *heads* in a sequence of a thousand coin tosses, the exact number of oxygen molecules in the room, the height (in inches) of a randomly selected individual,

[†] Note that it is sometimes $\kappa + 3$, rather than κ itself, which is called the kurtosis.

are all approximately described by a Gaussian distribution.† The ubiquity of the Gaussian can be in part traced to the central limit theorem (CLT) discussed at length in Chapter 2, which states that a phenomenon resulting from a large number of small independent causes is Gaussian. There exists however a large number of cases where the distribution describing a complex phenomenon is *not* Gaussian: for example, the amplitude of earthquakes, the velocity differences in a turbulent fluid, the stresses in granular materials, etc., and, as we shall discuss in Chapter 6, the price fluctuations of most financial assets.

A Gaussian of mean m and root mean square σ is defined as:

$$P_{\rm G}(x) \equiv \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-m)^2}{2\sigma^2}\right). \tag{1.15}$$

The median and most probable value are in this case equal to m, whereas the MAD (or any other definition of the width) is proportional to the RMS (for example, $E_{\rm abs} = \sigma\sqrt{2/\pi}$). For $m = 0$, all the odd moments are zero and the even moments are given by $m_{2n} = (2n-1)(2n-3)\ldots\sigma^{2n} = (2n-1)!!\,\sigma^{2n}$.

All the cumulants of order greater than two are zero for a Gaussian. This can be realized by examining its characteristic function:

$$\hat{P}_{\rm G}(z) = \exp\left(-\frac{\sigma^2 z^2}{2} + {\rm i}mz\right). \tag{1.16}$$

Its logarithm is a second-order polynomial, for which all derivatives of order larger than two are zero. In particular, the kurtosis of a Gaussian variable is zero. As mentioned above, the kurtosis is often taken as a measure of the distance from a Gaussian distribution. When $\kappa > 0$ (**leptokurtic** distributions), the corresponding distribution density has a marked peak around the mean, and rather 'thick' tails. Conversely, when $\kappa < 0$, the distribution density has a flat top and very thin tails. For example, the uniform distribution over a certain interval (for which tails are absent) has a kurtosis $\kappa = -\frac{6}{5}$. Note that the kurtosis is bounded from below by the value -2, which corresponds to the case where the random variable can only take two values $-a$ and a with equal probability.

A Gaussian variable is peculiar because 'large deviations' are extremely rare. The quantity $\exp(-x^2/2\sigma^2)$ decays so fast for large x that deviations of a few times σ are nearly impossible. For example, a Gaussian variable departs from its most probable value by more than 2σ only 5% of the times, of more than 3σ in 0.2% of the times, whereas a fluctuation of 10σ has a probability of less than 2×10^{-23}; in other words, it never happens.

1.7 Log-normal distribution

Another very popular distribution in mathematical finance is the so-called **log-normal** law. That X is a log-normal random variable simply means that $\log X$ is normal, or Gaussian. Its use in finance comes from the assumption that the *rate of returns*, rather than the absolute

† Although, in the above three examples, the random variable cannot be negative. As we shall discuss later, the Gaussian description is generally only valid in a certain neighbourhood of the maximum of the distribution.

change of prices, are independent random variables. The increments of the logarithm of the price thus asymptotically sum to a Gaussian, according to the CLT detailed in Chapter 2. The log-normal distribution density is thus defined as:[†]

$$P_{\rm LN}(x) \equiv \frac{1}{x\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\log^2(x/x_0)}{2\sigma^2}\right), \tag{1.17}$$

the moments of which being: $m_n = x_0^n e^{n^2\sigma^2/2}$.

From these moments, one deduces the skewness, given by $\varsigma_3 = (e^{3\sigma^2} - 3e^{\sigma^2} + 2)/(e^{\sigma^2} - 1)^{3/2}$, ($\simeq 3\sigma$ for $\sigma \ll 1$), and the kurtosis $\kappa = (e^{6\sigma^2} - 4e^{3\sigma^2} + 6e^{\sigma^2} - 3)/(e^{\sigma^2} - 1)^2 - 3$, ($\simeq 19\sigma^2$ for $\sigma \ll 1$).

In the context of mathematical finance, one often prefers log-normal to Gaussian distributions for several reasons. As mentioned above, the existence of a random rate of return, or random interest rate, naturally leads to log-normal statistics. Furthermore, log-normals account for the following symmetry in the problem of exchange rates:[‡] if x is the rate of currency A in terms of currency B, then obviously, $1/x$ is the rate of currency B in terms of A. Under this transformation, $\log x$ becomes $-\log x$ and the description in terms of a log-normal distribution (or in terms of any other even function of $\log x$) is independent of the reference currency. One often hears the following argument in favour of log-normals: since the price of an asset cannot be negative, its statistics cannot be Gaussian since the latter admits in principle negative values, whereas a log-normal excludes them by construction. This is however a red-herring argument, since the description of the fluctuations of the price of a financial asset in terms of Gaussian or log-normal statistics is in any case an *approximation* which is only valid in a certain range. As we shall discuss at length later, these approximations are totally unadapted to describe extreme risks. Furthermore, even if a price drop of more than 100% is in principle possible for a Gaussian process,[§] the error caused by neglecting such an event is much smaller than that induced by the use of either of these two distributions (Gaussian or log-normal). In order to illustrate this point more clearly, consider the probability of observing n times 'heads' in a series of N coin tosses, which is exactly equal to $2^{-N}C_N^n$. It is also well known that in the neighbourhood of $N/2$, $2^{-N}C_N^n$ is very accurately approximated by a Gaussian of variance $N/4$; this is however not contradictory with the fact that $n \geq 0$ by construction!

Finally, let us note that for moderate volatilities (up to say 20%), the two distributions (Gaussian and log-normal) look rather alike, especially in the 'body' of the distribution (Fig. 1.3). As for the tails, we shall see later that Gaussians substantially underestimate their weight, whereas the log-normal predicts that large positive jumps are more frequent

[†] A log-normal distribution has the remarkable property that the knowledge of all its moments is not sufficient to characterize the corresponding distribution. One can indeed show that the following distribution: $\frac{1}{\sqrt{2\pi}}x^{-1}\exp[-\frac{1}{2}(\log x)^2][1 + a\sin(2\pi\log x)]$, for $|a| \leq 1$, has moments which are independent of the value of a, and thus coincide with those of a log-normal distribution, which corresponds to $a = 0$.

[‡] This symmetry is however not always obvious. The dollar, for example, plays a special role. This symmetry can only be expected between currencies of similar strength.

[§] In the rather extreme case of a 20% annual volatility and a zero annual return, the probability for the price to become negative after a year in a Gaussian description is less than one out of 3 million.

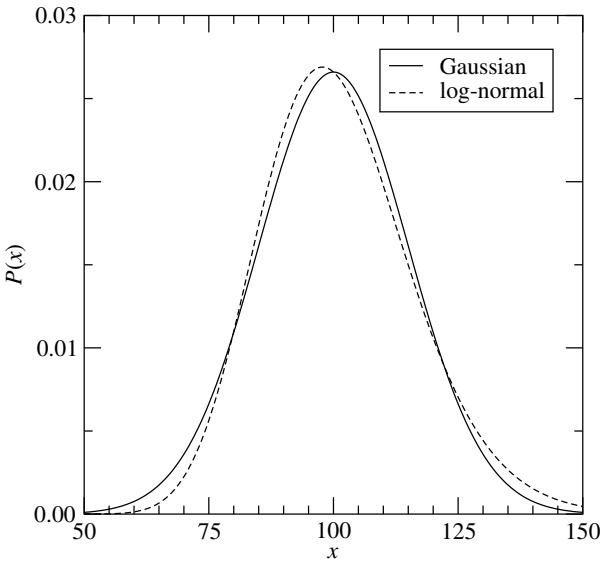

Fig. 1.3. Comparison between a Gaussian (thick line) and a log-normal (dashed line), with $m = x_0 = 100$ and σ equal to 15 and 15% respectively. The difference between the two curves shows up in the tails.

than large negative jumps. This is at variance with empirical observation: the distributions of absolute stock price changes are rather symmetrical; if anything, large negative draw-downs are more frequent than large positive draw-ups.

1.8 Lévy distributions and Paretian tails

Lévy distributions (noted $L_\mu(x)$ below) appear naturally in the context of the CLT (see Chapter 2), because of their stability property under addition (a property shared by Gaussians). The tails of Lévy distributions are however much 'fatter' than those of Gaussians, and are thus useful to describe multiscale phenomena (i.e. when both very large and very small values of a quantity can commonly be observed – such as personal income, size of pension funds, amplitude of earthquakes or other natural catastrophes, etc.). These distributions were introduced in the 1950s and 1960s by Mandelbrot (following Pareto) to describe personal income and the price changes of some financial assets, in particular the price of cotton. An important constitutive property of these Lévy distributions is their power-law behaviour for large arguments, often called **Pareto tails**:

$$L_\mu(x) \sim \frac{\mu A_\pm^\mu}{|x|^{1+\mu}} \text{ for } x \to \pm\infty, \tag{1.18}$$

where $0 < \mu < 2$ is a certain exponent (often called α), and $A_\pm^\mu$ two constants which we call **tail amplitudes**, or **scale parameters**: $A_\pm$ indeed gives the order of magnitude of the