# Why Stock Prices Have a Lognormal Distribution

Walter A. Rosenkrantz
Department of Mathematics and Statistics
University of Massachusetts at Amherst

March 27, 2003

## 1 The Normal Distribution

Because the future price of a stock at time $t$ cannot be predicted with certainty, we model it as a random variable, denoted here by $S(t)$. Since random variables are characterised by their distribution functions it is useful to have a notation to express this concept.

**Definition 1.1** *We use the symbol* $X \stackrel{\mathcal{D}}{=} Y$ *to mean that the random variables* $X$ *and* $Y$ *have the same distribution,i.e.,*

$$P(X \leq t) = P(Y \leq t),\ -\infty < t < \infty$$

Of the many distributions that the reader is likely to encounter in finance it is the normal and log normal distributions defined below, that are the most important.

**Definition 1.2** *A random variable* $X$ *with distribution function given by*

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}}\, e^{-(t-\mu)^2/2\sigma^2}\, dt$$

*is said to have a* normal distribution *with* $E(X) = \mu$ *and* $V(X) = \sigma^2$. *We express this by writing* $X \stackrel{\mathcal{D}}{=} N(\mu, \sigma^2)$.

We will prove later in these notes that $\ln(S(t)\S_0) \stackrel{\mathcal{D}}{=} N(t\mu, t\sigma^2)$, that is, the stock price has a log normal distribution in the sense of the following distribution.

**Definition 1.3** *We say that the non negative random variable* $W$ *has a log normal distribution if* $\ln(W) \stackrel{\mathcal{D}}{=} N(\mu, \sigma^2)$.

It can be shown, we omit the details, that

$$E(W) = e^{(\mu+\sigma^2/2)} \tag{1}$$

$$V(W) = e^{(2\mu+\sigma^2)}(e^{\sigma^2} - 1) \tag{2}$$

The probability density function of the *normal distribution* depends on two parameters $\mu$ and $\sigma$ and is defined by the equation

$$f(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-(x-\mu)^2/2\sigma^2};\ -\infty < x < \infty, \tag{3}$$

where

$$-\infty < \mu < \infty;\ 0 < \sigma < \infty.$$

Using techniques from advanced calculus one can verify that

$$\int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}}\, e^{-(x-\mu)^2/2\sigma^2}\, dx = 1,$$

and that $E(X) = \mu$, $V(X) = \sigma^2$; we omit the details.

The graph of the normal pdf is a *bell shaped* curve called the *normal curve.* Its importance for statistics and finance rests on the fact that the relative frequency histograms of many data sets, including stock market returns, can be approximated by a normal curve. In the physics and engineering literature the normal distribution is also called the *Gaussian distribution.*

**Definition 1.4** *A normally distributed random variable with* $\mu = 0$ *and* $\sigma = 1$ *is said to have the* standard normal distribution. *We denote it by the letter* $Z$.

The pdf of $Z$ is given by:

$$f(x) = \frac{1}{\sqrt{2\pi}}\, e^{-x^2/2};\ -\infty < x < \infty. \tag{4}$$

Its df $\Phi(z)$ is defined by the definite integral

$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}}\, e^{-x^2/2}\, dx. \tag{5}$$

Remark: It is easy to verify, we omit the details, that the random variable

$$X = \mu + \sigma Z \overset{\mathcal{D}}{=} N(\mu, \sigma^2)$$

In figure 1.1 we display the graphs of two normal curves with the same mean ($\mu_1 = \mu_2 = 3$) but different variances ( $\sigma_1 = 2$ and $\sigma_2 = 0.5$.)

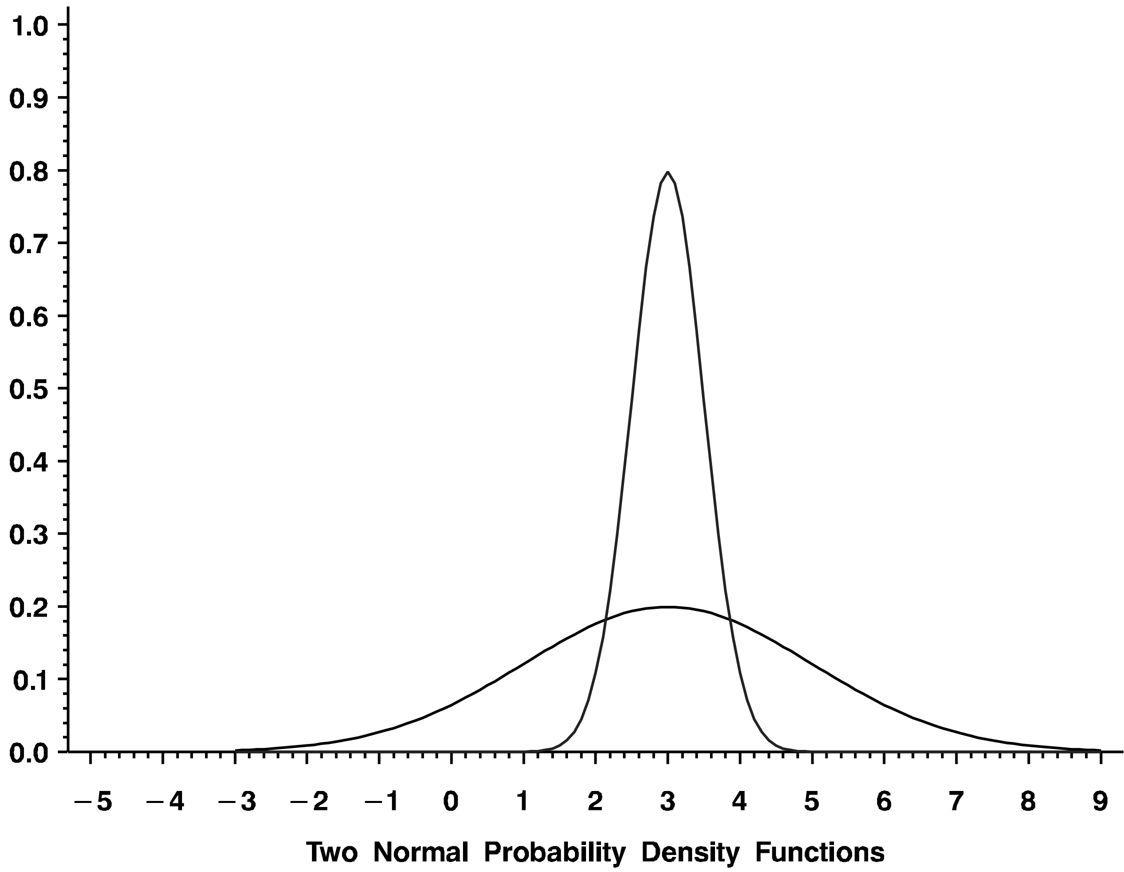


**Figure 1.1**

In figure 1.2 we display the graphs of two normal curves with different means ($\mu_1 = 1.5 < 2.5 = \mu_2$) but the same variance ( $\sigma_1 = \sigma_2 = 0.5$.)

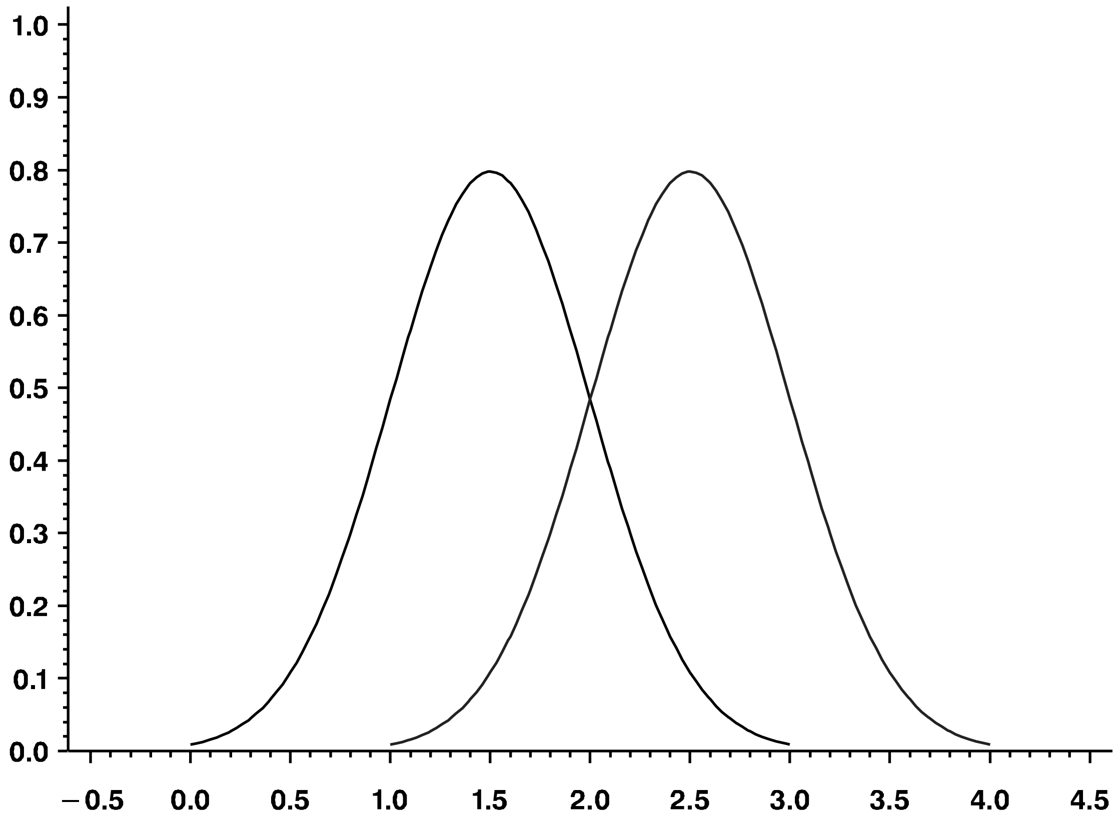


**Figure 1.2**

Figures 1.1 and 1.2 give us some insight into the significance of the constants $\mu$ and $\sigma$.

1. The maximum value of $f(x; \mu, \sigma)$ occurs at $x = \mu$ and its graph is *symmetric* about $\mu$.
2. Looking at figure 1.1 we see the *spread* of the distribution is determined by $\sigma$. A small value of $\sigma$ produces a sharp peak at $x = \mu$; consequently, most of the area under this normal curve is close to $\mu$. A large value of $\sigma$, on the other hand, produces a smaller, more rounded bulge at $x = \mu$. The area under this normal curve is less concentrated about $\mu$.

## 2 The Central Limit Theorem (CLT)

Let $X_1, \ldots, X_n$ denote a random sample of size $n$, taken from the same (parent) distribution $F(x)$. This means that the $X_i's$ are independent and identically distributed (abbreviated iid), with $P(X_j \leq x) = F(x)$, $E(X_j) = \mu$, $V(X_j) = \sigma^2$, $j = 1, \ldots, n,$. We are interested is describing the distribution of the sample mean $\overline{X}$ and sample total $T$ defined as follows:

$$\overline{X} = \frac{\sum_{1 \leq j \leq n} X_j}{n} \text{ and } T = \sum_{1 \leq j \leq n} X_j$$

assuming only that our random sample comes from a general distribution $F(x)$ with (population) mean $\mu$ and (population) variance $\sigma^2$.

**Proposition 2.1** *Let $X_1, \ldots, X_n$ denote a random sample of size $n$, with common mean $\mu$ and variance $\sigma^2$. Then*

$$E(\overline{X}) = \mu; \quad (6)$$

$$\sigma(\overline{X}) = \frac{\sigma}{\sqrt{n}}; \quad (7)$$

$$E(T) = n\mu; \quad (8)$$

$$\sigma(T) = \sigma\sqrt{n}. \quad (9)$$

**Proof of proposition 2.1:** Since $\overline{X} = \frac{1}{n}\sum_{1 \leq i \leq n} X_i$, it follows from the addition rule that

$$E(\overline{X}) = \frac{1}{n} \sum_{1 \leq i \leq n} E(X_i) = \frac{1}{n} \times n\mu = \mu.$$

Similarly, it follows from the addition rule for the variance of a sum of independent random variables that

$$\sigma^2(\overline{X}) = V\left(\sum_{1 \leq i \leq n} \frac{X_i}{n}\right) = n \times \frac{\sigma^2}{n^2} = \frac{\sigma^2}{n}, \text{ and therefore } \sigma(\overline{X}) = \frac{\sigma}{\sqrt{n}}.$$

The computations of $E(T)$ and $\sigma(T)$ proceed along the same lines and are left to the reader.

**Theorem 2.1** *The Central Limit Theorem(CLT): Let $X_1, \ldots, X_n$ be a random sample from a distribution with mean $\mu$ and variance $\sigma^2$. Then, for $n$*

*sufficiently large, the distribution of $T$ is approximately normal with mean $n\mu$ and standard deviation $\sigma(T) = \sigma\sqrt{n}$. More precisely,*

$$\lim_{n\to\infty} P\left(a \le \frac{T - n\mu}{\sigma\sqrt{n}} \le b\right) = P(a \le Z \le b) = \Phi(b) - \Phi(a) \qquad (10)$$

*The distribution of the sample mean $\overline{X}$ is also approximately normal with $E(\overline{X}) = \mu$ and $\sigma(\overline{X}) = \sigma/\sqrt{n}$.*

The CLT is particularly useful because it yields a computable approximation to the distribution of the sample mean in terms of the normal distribution. It is a far reaching and remarkable generalization of the normal approximation to the binomial. The only conditions imposed are : (i) the random variables $X_1, \ldots, X_i, ...$ are iid and (ii) that their means and variances are finite. A sequence of iid Bernoulli random variables is just an important special case.

**Example (A volatile stock):**
Consider a stock with initial price $S_0 = \$100$ and

$$u = 1.04,\ d = \frac{1}{u} = 0.96,\ p = 0.53;\ \ln 1.04 = 0.04,\ \ln 0.96 = -0.04$$

**Note:** Some computations are rounded to two decimal places. According to the binomial lattice model, the price $S_i$ at the beginning of the $i$th week $(i = 1, 2, \ldots n)$ is either $uS_{i-1}$ ($u$ for up) with probability $p$, or $dS_{i-1}$ ($d$ for down) with probability $1 - p$, i.e.,

$$P(S_i = uS_{i-1}) = p \text{ and } P(S_i = dS_{i-1}) = 1 - p$$

Assume that the weekly returns $X_i = \ln(S_i/S_{i-1})$, $i = 1, 2, \ldots 52$ are iid random variables.

1. Show that $E(X_j) = 0.0024$ and $\sigma(X_i) = 0.040$ (some numerical results rounded to 2 decimal places).

   **Solution:** According to the binomial lattice model $X_i$ has the following distribution:

   $$P(X_i = 0.04) = 0.53,\ P(X_i = -0.04) = 0.47$$

   Therefore $E(X_i) = 0.04 \times 0.53 - 0.04 \times 0.47 = 0.0024$. Similarly, $V(X_i) = E(X_i^2) - 0.0024^2 = 0.04^2 - 0.0024^2 = 0.001594$, so $\sigma(X_j) = \sqrt{0.001594} = 0.040$.

2. Estimate the probability that the year-end price of the stock is less than the initial price $S_0 = \$100$.
   **Solution:** Let $S$ denote the year-end price, which corresponds to $n = 52$ periods. Applying the central limit theorem to $T = \ln(S/100)$ with $\mu = 0.0024$, $\sigma = 0.040$, $n = 52$ we have to calculate

$$\begin{aligned} P(\ln(S/100) < 0) &= P(T < 0) = P\left(\frac{T - 52 \times 0.0024}{0.04 \times \sqrt{52}} < \frac{-0.1248}{0.288}\right) \\ &= P(Z < -0.43) = 0.3336 \end{aligned}$$

# 3 An Application of the CLT to the random walk model of Stock Prices

Empirical studies of actual returns for many stocks suggest that the rate of return for the next period is obtained from the current one by first flipping a coin and then adding to the current return an amount $U > 0$ if the coin come up heads, or or an amount $D < 0$ if the coin comes up tails. This is the *random walk model* for the price fluctuations of stocks, a model that apparently fits the facts, but not the prejudices of investors, most of whom believe they are more intelligent than the "market". The argument for the random walk model of stock market prices is based on the assumption that the current price of a stock always reflects all available information, and will change only with the arrival of new information. Since new information cannot be predicted, neither can a future change in price. This, in a nutshell, is the *efficient market hypothesis.*

It is important to note, however, that even though we cannot predict in advance the result of a coin toss, we can give an explicit formula for the distribution of the number of heads in $n$ tosses of a coin (binomial distribution), and, for $n$ large an approximate formula given by the central limit theorem. An interesting consequence of our model is that the logarithm of the stock price at time $t$ has a normal distribution with mean value $\mu t$ and variance $t\sigma^2$. We turn now to the mathematical details, which are of independent interest.

Let $S_0$ denote the initial price of a stock and $S(t)$ its price at a future time $t$. We divide the time axis into units of length equal to $1/n$ and assume that the stock price at time $t_j = j/n$ increases by a factor $u > 1$ with probability $p$, or decreases by a factor $d < 1$ with probability $1 - p$. That is

$$P(S(j/n) = uS((j-1)/n) = p;\ P(S(j/n) = dS((j-1)/n) = 1 - p$$

Denote the rate of return over the time interval $t_j - t_{j-1} = 1/n$ by $X_j^n$, defined in the usual way as

$$X_j^n = \frac{S(j/n) - S((j-1)/n)}{S((j-1)/n)}$$

We claim that for $n$ large enough, so the interval of time is small enough, that

$$X_j^n \doteq \ln\left(\frac{S(j/n)}{S(j-1/n)}\right) \tag{11}$$

where the symbol $a \doteq b$ means that $a$ is approximately equal to $b$ and $\ln(t) = \log_e(t)$, the natural logarithm of $t$. To see this we use the approximation

$$\ln(t) \doteq 1 + t,\ t \text{ small} \tag{12}$$

Examples of the approximation:

$$\begin{aligned}
\ln(1.02) &= \ln(1 + 0.02) = 0.01980 \doteq 0.02 \\
\ln(0.97) &= \ln(1 - 0.03) = -0.03046 \doteq -0.03 \\
\ln(1.05) &= \ln(1 + 0.05) = 0.04879 \doteq 0.05
\end{aligned}$$

Derivation of Equation (11):

$$\begin{aligned}
X_j^n &= \frac{S(j/n) - S((j-1)/n)}{S((j-1)/n)} \\
&\doteq \ln\left(1 + \frac{S(j/n) - S(j-1/n)}{S((j-1/n))}\right) \text{(using Eq. (12))} \\
&\doteq \ln\left(\frac{S(j/n)}{S((j-1)/n)}\right)
\end{aligned}$$

**A CLT for** $\ln S(t)$

We are going to show, using the CLT (Proposition (2.1) and the efficient market hypothesis, that the distribution of $\ln(S(t)/S_0)$ is normal with mean $t\mu$ and variance $t\sigma^2$. In other words, $S(t)/S_0$ has a log normal distribution. Since $S_0$ is a constant it follows that $S(t)$ itself has a log normal distribution.

The basic idea of the proof is to represent $\ln(S(t)/S_0)$ as a sum of independent, identically distributed random variables so that the central limit theorem can be brought to bear. We begin with the algebraic identity

$$\frac{S(t)}{S_0} = \frac{S(1/n)}{S_0} \times \frac{S(2/n)}{S(1/n)} \cdots \frac{S(k/n)}{S((k-1)/n)} \times \frac{S(t)}{S(k/n)},$$

where $k/n \le t < (k+1)/n$; that is $k = [nt]$, and $[x]$ is the *floor function*, or the *greatest integer function*. Note that $[nt] \le nt < [nt] + 1$ and therefore $[nt]/n \le t < ([nt]+1)/n$. Consequently,

$$\lim_{n\to\infty} \frac{[nt]}{n} = t \tag{13}$$

Since the rate of return over a time interval of length $1/n$ goes to zero as $n \to \infty$ it follows that $\lim_{n\to\infty} \ln(S(t)/S([nt]/n)) = 0$. Taking logarithms of both sides we see that

$$\begin{aligned}
\ln\left(\frac{S(t)}{S_0}\right) &= \sum_{1\le j\le [nt]} \ln\left(\frac{S(j/n)}{S((j-1)/n)}\right) + \ln\left(\frac{S(t)}{S([nt]/n)}\right) \\
&\doteq \sum_{1\le j\le [nt]} \ln\left(\frac{S(j/n)}{S((j-1)/n)}\right) \\
&= \sum_{1\le j\le [nt]} X_j^n \text{ (see Eq. (11))}
\end{aligned}$$

Under the assumptions of the random walk model the returns $X_j^n$ are independent, identically distributed random variables with $E(X_j^n) = \mu_n$, $V(X_j^n) = \sigma_n^2$; that is, the mean returns and risks are functions of the length of the time period $1/n$. Thus,

$$\begin{aligned}
E\left(\ln\left(\frac{S(t)}{S_0}\right)\right) &= E\left(\sum_{1\le j\le [nt]} X_j^n\right) = [nt]\mu_n \\
V\left(\ln\left(\frac{S(t)}{S_0}\right)\right) &= V\left(\sum_{1\le j\le [nt]} X_j^n\right) = [nt]\sigma_n^2
\end{aligned}$$

Using the fact that $\lim_{n\to\infty}[nt]/n = t$ it follows that

$$\lim_{n\to\infty}[nt]\mu_n = \lim_{n\to\infty}([nt]/n)(n\mu_n) = t\lim_{n\to\infty}(n\mu_n) \tag{14}$$

$$\lim_{n\to\infty}[nt]\sigma_n^2 = \lim_{n\to\infty}([nt]/n)(n\sigma_n^2) = t\lim_{n\to\infty}(n\sigma_n^2) \tag{15}$$

Looking at Equations (14, 15) we see that these limits exist only if the limits $\lim_{n\to\infty} n\mu_n = \mu$ and $\lim_{n\to\infty} n\sigma_n^2 = \sigma$ exist. We therefore make the reasonable assumption that the rate of return and its variance are proportional to the length of the time interval $1/n$; that is we assume

$$\mu_n = \frac{\mu}{n};\ \sigma_n^2 = \frac{\sigma^2}{n},\ \sigma_n = \frac{\sigma}{\sqrt{n}}$$

Consequently,

$$\begin{aligned}\lim_{n\to\infty}[nt]\mu_n &= t\mu \\ \lim_{n\to\infty}[nt]\sigma_n^2 &= t\sigma^2\end{aligned}$$

We now apply the central limit theorem (see Eq. (10) to

$$T = \sum_{1\le j\le[nt]} X_j^n \doteq \ln(S(t)/S_0)$$

with $[nt]$ instead of $n$, $\mu/n$ and $\sigma/\sqrt{n}$ instead of $\mu$ and $\sigma^2$, respectively. Therefore,

$$\begin{aligned}\lim_{n\to\infty} P\left(\frac{\sum_{1\le j\le[nt]} X_j^n - [nt](\mu/n)}{\sqrt{[nt]}(\sigma/\sqrt{n})} \le z\right) &= \lim_{n\to\infty} P\left(\frac{\sum_{1\le j\le[nt]} X_j^n - t\mu}{\sigma\sqrt{t}} \le z\right) \\ &= P(Z \le z)\end{aligned}$$

In other words, for large $n$ we have

$$\frac{\sum_{1\le j\le[nt]} X_j^n - t\mu}{\sigma\sqrt{t}} \doteq Z$$

in the sense that the distribution of the random variable on the left is approximately the same as a standard normal distribution. This is equivalent to the assertion that

$$\ln(S(t)/S_0) \doteq \sum_{1\le j\le[nt]} X_j^n \doteq t\mu + \sigma\sqrt{t}Z$$

In particular, as $\lim_{n\to\infty}$, we have equality, that is the distribution of $\ln(S(t)/S_0)$ is equal to that of $t\mu + \sigma\sqrt{t}Z$. This completes the proof.