“Normality” of Stock Prices

Bo Shi

Abstract. The Black-Scholes Model, often simply called Black-Scholes, models the varying price of financial instruments over time: stocks in particular. This model assumes that returns on the underlying stock are lognormally distributed, which can be reasonable for many assets that offer options. However, from a selection of 100 stock histories, I found that at least 45 were not lognormally distributed with very high confidence of $100(1\text{-}10^{-10})\%$. Most of these exceptional histories covered a long period of time.

1. Introduction

The normal family of distributions, by far the single most important family of probability distributions in statistics, is commonly associated with Bell-shaped curves (Figure 1).

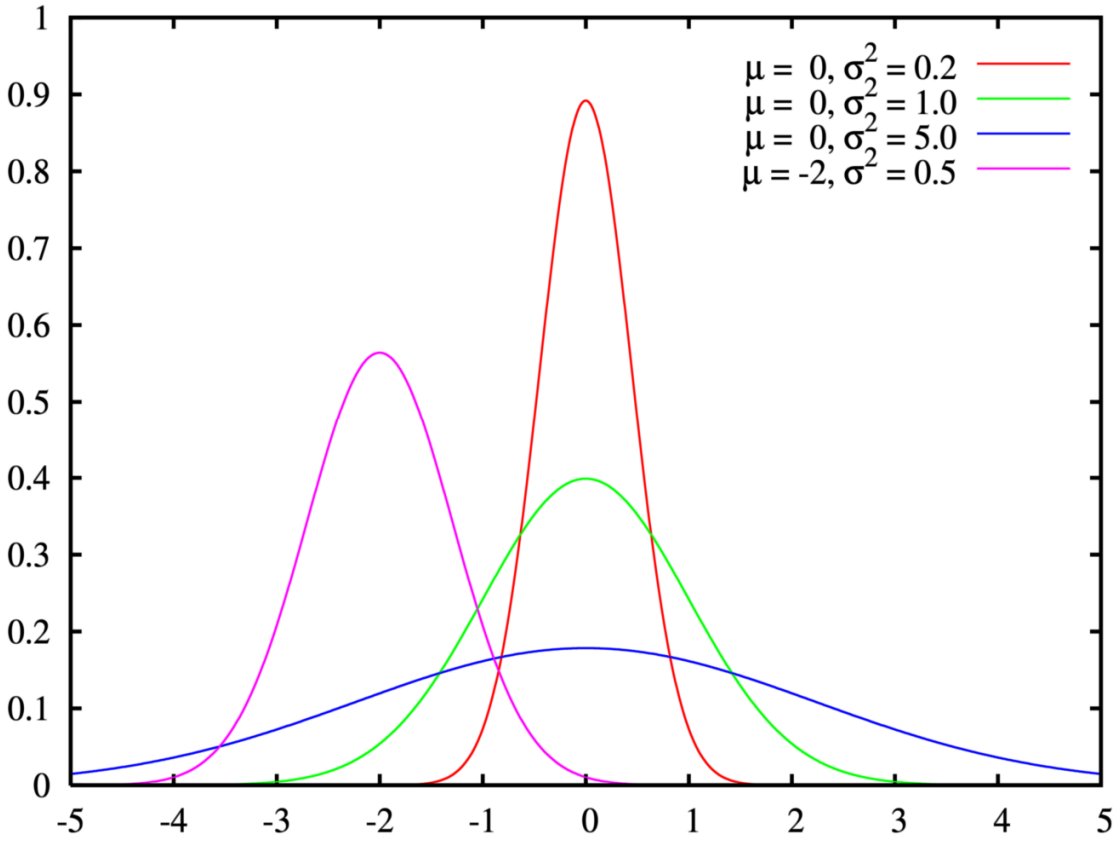


Figure 1: Bell Curves belonging to the family of Normal Distributions

Normal distributions pervade much of science and also finance. It is common to assume that data come from a normal distribution.

In the early 1970s, Fischer Black, Myron Scholes, and Robert Merton made a major breakthrough in the pricing of stock options. This involved the development of what

has become known as the Black-Scholes (or the Black-Scholes-Merton) model, which assumes that stock prices are sample paths from a Geometric Brownian Motion (see [Jo], p.215):

Suppose that we are interested in the price of some stock as it evolves over time. Let the initial time be time 0, and let S(i) be the price of the stock at time i. We say that the collection of prices S(i), $0 \le i < \infty$, follows the Geometric Brownian Motion with drift parameter μ and volatility parameter σ if, for all nonnegative values of i and t, the random variable $\frac{S(i+t)}{S(i)}$ is independent of all prices prior to time i; and if, in addition, $\log(\frac{S(i+t)}{S(i)})$ is a normal random variable with mean μt and variance $t\sigma^2$.

In other words, the sequence of prices is a sample path from a Geometric Brownian Motion if the ratio of the price at time t in the future to the present price will be independent of the past history of prices and this ratio has a lognormal probability distribution with parameters μt and $t\sigma^2$.

A consequence of assuming that a stock's prices follow a geometric Brownian motion is that the present price, and not the price history, affects future price probabilities once μ and σ are determined. Furthermore, probabilities concerning the ratio of the price at time t in the future to the present price will not depend on the present price. For instance, the model implies that the probability a given stock doubles in price in the next month is the same regardless of whether its present price is $10 or $25.

The assumptions of the Black-Scholes model and its extensions are not fully followed by traders, who assume the probability distribution of an equity price has a heavier left tail and a less heavy right tail than the lognormal. My research examined the accuracy of the lognormal assumption in the Black-Scholes model. I used three different methods to test that assumption. For a null hypothesis, I assumed that all stocks' prices follow the Geometric Brownian Motion. For statistical convenience:
1. Weekends and holidays are ignored;
2. I used prices on my.yahoo.com where "Adjusted Price" matches sequences of consecutive "Closing Price ", which avoids dividends, stock splits and similar events.

I tested whether the set of numbers, $\log(S_{i+1}) - \log(S_i)$ could come from a normal distribution, where *i* represents the trading day and S_i represents the price at the end of the trading day *i*.

I begin by establishing notation that will be used, presenting some basic theorems, and then explaining three distinct methods that I used. Finally I present my results and observations.

2. Preliminaries

Let n be the number of log ratios observed;

α : the probability of incorrectly rejecting a null hypothesis;

$a = \chi^2_{1-\alpha/4}(n-1)$, $b = \chi^2_{\alpha/4}(n-1)$, where $P(\chi^2 > \chi^2_\alpha) = \alpha$ defines the critical value χ^2_α;

$\{X_i\}_{i=1}^n = \{\log(S_{i+1}) - \log(S_i)\}_{i=1}^n$;

$\{Y_i\}_{i=1}^n = \{\text{Ordered } X_i \text{ with } Y_1 \le Y_2 \le Y_3 \le \ldots \le Y_n\}$;

$\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$ is the sample mean;

$S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})^2$ is the sample variance;

μ is the mean of the hypothesized normal distribution;

σ^2 is the variance of the hypothesized normal distribution;

$H_0(\mu, \sigma^2)$: the X_i's are independent normal random variables with mean μ and variance σ^2. We also say that each X_i is $N(\mu, \sigma^2)$.

$\Phi(z) = P(Z \le z)$, Z is $N(0,1)$.

[HT] provides the following definitions and theorems.

Definition 1: The gamma function is defined by

$$\Gamma(t) = \int_0^\infty y^{t-1} e^{-y} dy \qquad 0 < t,$$

(see [HT], p.186).

Definition 2: The random variable X has a gamma distribution if its probability density function(p.d.f) is defined by

$$f(x) = \frac{1}{\Gamma(\alpha)\theta^\alpha} x^{\alpha-1} e^{-x/\theta}, \qquad 0 < x < \infty.$$

Here $\alpha > 0$ and $\theta > 0$ are constants with E(X)=$\alpha\theta$ and Var(X)=$\alpha\theta^2$ (see [HT],

p.187).

Definition 3: Let X have a gamma distribution with $\theta = 2$ and $\alpha = r/2$, where r is a positive integer. The p.d.f of X is

$$f(x) = \frac{1}{\Gamma(r/2)2^{r/2}} x^{r/2-1} e^{-x/2}, \qquad 0 < x < \infty.$$

We say that X has a chi-square distribution with r degrees of freedom, which we abbreviate by saying X is $\chi^2(r)$ (see [HT], p.189)

Theorem 1: If the random variable X is N(μ, σ^2), then the random variable V=$\frac{(X-\mu)^2}{\sigma^2} = Z^2$ is $\chi^2(1)$ (see [HT], p.199).

Proof: Because V=Z^2, where Z = $\frac{X-\mu}{\sigma}$ is N(0,1), the distribution function F(v) of V is, for $v \geq 0$,

$$\text{F}(v) \quad = \quad \text{P}(Z^2 \leq v) \quad = \quad \text{P}(-\sqrt{v} \leq Z \leq \sqrt{v}).$$

That is, with $v \geq 0$,

$$\text{F}(v) \quad = \quad \int_{-\sqrt{v}}^{\sqrt{v}} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz \quad = \quad 2\int_{0}^{\sqrt{v}} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz.$$

If we change the variable of integration by writing z = $\sqrt{y}$, then, since $D_y(z) = \frac{1}{2\sqrt{y}}$, we have

$$\text{F}(v) \quad = \quad \int_{0}^{v} \frac{1}{\sqrt{2\pi y}} e^{-\frac{y}{2}} dy, \qquad 0 \leq v.$$

Of course, F(v) = 0, when v<0. Hence the p.d.f f(v) = $F'(v)$ of the continuous type random variable V is, by one form of the fundamental theorem of calculus,

$$\text{f}(v) \quad = \quad \frac{1}{\sqrt{2\pi}} v^{\frac{1}{2}-1} e^{-\frac{v}{2}}, \qquad 0 < v < \infty.$$

Since f(v) is a p.d.f., it must be true that

$$\int_{0}^{\infty} \frac{1}{\sqrt{2\pi}} v^{\frac{1}{2}-1} e^{-\frac{v}{2}} dv = 1.$$

Then the change of variable x = $\frac{v}{2}$ yields

$$1 = \frac{1}{\sqrt{\pi}} \int_{0}^{\infty} x^{\frac{1}{2}-1} e^{-x} dx \quad = \quad \frac{1}{\sqrt{\pi}} \Gamma(\frac{1}{2}).$$

Hence $\Gamma(\frac{1}{2}) = \sqrt{\pi}$, and thus V is $\chi^2(1)$.

Theorem 2: Let the distribution of $X_1, X_2, ..., X_k$ be $\chi^2(r_1), \chi^2(r_2), ..., \chi^2(r_k)$, respectively. If $X_1, X_2, ..., X_k$ are independent, then Y = $X_1 + X_2 + \cdots + X_k$ is $\chi^2(r_1 + r_2 + \cdots + r_k)$ (see [HT], p.301).

Proof: I give the proof for k=2 but it is similar for a general k (by induction). The moment-generating function of Y is given by

$$M_Y(t) = E[e^{tY}] = E[e^{t(X_1+X_2)}] = E[e^{tX_1}e^{tX_2}].$$

Because X_1 and X_2 are independent, this last expectation can be factored so that

$$M_Y(t) = E[e^{tX_1}]E[e^{tX_2}] = (1-2t)^{-r_1/2}(1-2t)^{-r_2/2}, \qquad t < \frac{1}{2},$$

since X_1 and X_2 have chi-square distributions. Thus

$$M_Y(t) = E[e^{tX_1}]E[e^{tX_2}] = (1-2t)^{-(r_1+r_2)/2}, \qquad t < \frac{1}{2},$$

the moment-generating function for a chi-square distribution with $r = r_1 + r_2$ degrees of freedom. The uniqueness of the moment-generating function implies that Y is $\chi^2(r_1 + r_2)$. Thus, it follows that theorem holds for k=2. A direct induction argument proves the theorem for arbitrary k.

Theorem 3: Let $Z_1, Z_2, ..., Z_n$ have standard normal distributions, N(0,1). If these random variables are independent, then V = $Z_1^2 + Z_2^2 + \cdots + Z_n^2$ has a distribution that is $\chi^2(n)$ (see [HT], p.301).

Proof: By Theorem1, Z_i^2 is $\chi^2(1)$ for i =1,2,…,n. From Theorem 2, with k=n, Y=V, and $r_i = 1$, we see that V is $\chi^2(n)$.

Theorem 4: If $X_1, X_2, ..., X_n$ are observations of a random sample of size n from the normal distribution N(μ, σ^2),

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

and

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 \ ,$$

then

$$\frac{(n-1)S^2}{\sigma^2} \quad \text{is} \quad \chi^2(n-1) \text{ (see [HT], p.302).}$$

Proof:

$$\begin{aligned}
\text{V} = \sum_{i=1}^{n}(\frac{X_i - \mu}{\sigma})^2 &= \sum_{i=1}^{n}(\frac{(X_i - \overline{X}) + (\overline{X} - \mu)}{\sigma})^2 \\
&= \sum_{i=1}^{n}[(\frac{(X_i - \overline{X})}{\sigma})^2 + \frac{2(X_i - \overline{X})(\overline{X} - \mu)}{\sigma^2} + \frac{(\overline{X} - \mu)^2}{\sigma^2}] \\
&= \sum_{i=1}^{n}(\frac{(X_i - \overline{X})}{\sigma})^2 + 2\sum_{i=1}^{n}\frac{(X_i - \overline{X})(\overline{X} - \mu)}{\sigma^2} + \sum_{i=1}^{n}\frac{(\overline{X} - \mu)^2}{\sigma^2} \\
&= \sum_{i=1}^{n}(\frac{(X_i - \overline{X})}{\sigma})^2 + \frac{2(\overline{X} - \mu)}{\sigma^2}\sum_{i=1}^{n}(X_i - \overline{X}) + \frac{n(\overline{X} - \mu)^2}{\sigma^2} \\
&= \sum_{i=1}^{n}(\frac{(X_i - \overline{X})}{\sigma})^2 + \frac{n(\overline{X} - \mu)^2}{\sigma^2} \qquad (*)
\end{aligned}$$

since $\sum_{i=1}^{n}(X_i - \overline{X}) = 0.$

But $Y_i = \frac{X_i - \mu}{\sigma}$, i = 1,2,…n, are standardized normal variables that are independent.

Hence V = $\sum_{i=1}^{n} Y_i^2$ is $\chi^2(n)$ by Theorem 3. Moreover, since $\overline{X}$ is N(μ, $\frac{\sigma^2}{n}$), then

$$Z^2 = (\frac{\overline{X} - \mu}{\sigma/\sqrt{n}})^2 = \frac{n(\overline{X} - \mu)^2}{\sigma^2}$$

is $\chi^2(1)$ by theorem 1. In this notation, Equation (*) becomes

$$\text{V} = \frac{(n-1)S^2}{\sigma^2} + Z^2 .$$

Furthermore, we know $\overline{X}$ and S^2 are independent (see [Mo], p.326); thus Z^2 and S^2 are also independent. In the moment-generating function of V, this independence permits us to write

$$\mathrm{E}[\,e^{tV}\,]=\mathrm{E}[\,e^{t\{(n-1)S^2/\sigma^2+z^2\}}\,]=\mathrm{E}[\,e^{t(n-1)S^2/\sigma^2}e^{tZ^2}\,]=\mathrm{E}[\,e^{t(n-1)S^2/\sigma^2}\,]\,\mathrm{E}[\,e^{tZ^2}\,].$$

Since V and Z^2 have chi-square distributions, we can substitute their moment-generating functions to obtain

$$(1-2t)^{-n/2}=E[e^{t(n-1)S^2/\sigma^2}](1-2t)^{-1/2}.$$

Equivalently, we have

$$E[e^{t(n-1)S^2/\sigma^2}]=(1-2t)^{-(n-1)/2}, \qquad t<\frac{1}{2}$$

This, of course, is the moment-generating function of a $\chi^2(n-1)$ variable and accordingly $(n-1)S^2/\sigma^2$ has this distribution.

3. Test Methods

Now I describe three statistical methods to test the lognormal assumption in the Black-Scholes Model, with confidence $1-\alpha$ for the first two methods with $\alpha=10^{-10}$.

3.1 First Method

The first test idea follows a suggestion from Dr. Ramsey. I use two separate estimates: the first an upper bound for σ and the second a lower bound for σ. If the lower bound is greater than the upper bound, we reject $H_0(\mu,\sigma^2)$ for all (μ,σ) with confidence $100(1-\alpha)\%$.

STEP 1. Estimate an upper bound for σ by using the standard Chi-Square distribution approach. That is, I use the fact that the distribution of $(n-1)S^2/\sigma^2$ is $\chi^2(n-1)$ by Theorem 4 to find a confidence interval for σ^2 with the confidence coefficient $1-\alpha/2$. So we want

$$P(a\le\frac{(n-1)S^2}{\sigma^2}\le b)=1-\frac{\alpha}{2}.$$

One way to do this is by selecting a and b so that $a=\chi^2_{1-\alpha/4}(n-1)$ and

$b = \chi^2_{\alpha/4}(n-1)$. Then, rearranging the inequalities, we have

$$1-\frac{\alpha}{2} = P(\frac{a}{(n-1)S^2} \le \frac{1}{\sigma^2} \le \frac{b}{(n-1)S^2})$$

$$= P(\frac{(n-1)S^2}{b} \le \sigma^2 \le \frac{(n-1)S^2}{a}).$$

Thus the probability that the random interval $[(n-1)S^2/b, (n-1)S^2/a]$ contains the unknown σ^2 is $1-\alpha/2$. It follows that a $100(1-\alpha/2)\%$ confidence interval for σ is given by $[\sqrt{\frac{n-1}{b}}S, \sqrt{\frac{n-1}{a}}S]$. Thus for $\tilde{\sigma} = \sqrt{\frac{n-1}{a}}S$, $P(\sigma \le \tilde{\sigma}) \ge 1-\alpha/2.$

STEP 2: For an independent random sample $\{X_i\}_{i=1}^n$ from a fixed normal distribution, let A be the number of i such that $|X_i - \mu| \ge 3\sigma$. Then A is a binomial random variable with parameters n and p, where

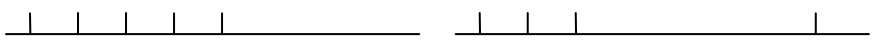
$$p = P(|\frac{X-\mu}{\sigma}| \ge 3) = P(|Z| \ge 3) \approx 2(1-0.99865) = 0.0027$$

since

$$P(Z \le 3) = \Phi(3) \approx 0.998650 \quad \text{by Mathematica.}$$

Determine the smallest integer M such that $P(A \ge M) < \alpha/2$. Since n, p, α are already known, we can solve for this minimum M.

STEP 3: Order the data points increasingly, $\{Y_i\}_{i=1}^n$.



$Y_i \quad Y_{i+1} \quad Y_{i+2} \quad \ldots\ldots \qquad\qquad Y_{n-M+i} \quad \ldots\ldots \qquad\qquad Y_n$

By construction, the probability is at least $1-\alpha/2$ that $[\mu-3\sigma, \mu+3\sigma]$ covers at least n-M+1 of the points $Y_1, Y_2, ..., Y_n$. Thus, with probability at least $1-\alpha/2$, there is some $i \in \{1,2,\ldots,M\}$ such that $[Y_i, Y_{n-M+i}] \subset [\mu-3\sigma, \mu+3\sigma]$. From this, it follows that if we define $\hat{\sigma} = \min_{1 \le i \le M} \frac{Y_{n-M+i} - Y_i}{6}$, then $P(\hat{\sigma} \le \sigma) \ge 1-\alpha/2.$

STEP 4: Let C be the event that $\sigma \le \tilde{\sigma}$; from step 1

$$P(C) = P(\sigma \le \tilde{\sigma}) \ge 1 - \alpha/2 .$$

Let D be the event that $\hat{\sigma} \le \sigma$; from step 3 $P(D) = P(\hat{\sigma} \le \sigma) \ge 1 - \alpha/2$.

So

$$P(\hat{\sigma} \le \tilde{\sigma}) \ge P(\hat{\sigma} \le \sigma \le \tilde{\sigma}) = P(C \cap D) = P(C) + P(D) - P(C \cup D)$$

$$\ge 1 - \alpha/2 + 1 - \alpha/2 - 1 = 1 - \alpha .$$

If $\hat{\sigma} > \tilde{\sigma}$, we reject $H_0(\mu, \sigma)$ for all (μ, σ) with the confidence $1 - \alpha$.

The following is the sample test chosen from one of my 100 sample companies with 34 closing prices with company symbol “QQQQ” (from my.yahoo.com):

Date	Close	Xi	Yi	Y(n-M+i)	[Y(n-M+i)-Yi]/6
6-Feb-06	40.81	0.002691792	-0.02202732	0.0076421	0.004944903
3-Feb-06	40.92	0.012386315	-0.01796739	0.0087337	0.004450178
2-Feb-06	41.43	0.017229428	-0.01338772	0.0115418	0.004154914
1-Feb-06	42.15	-0.00356507	-0.0103553	0.0123863	0.003790269
31-Jan-06	42	0.004513608	-0.00766643	0.0160003	0.003944462
30-Jan-06	42.19	-0.00189798	-0.00731739	0.0172294	0.004091137
27-Jan-06	42.11	-0.01338772	-0.00700401	0.0303234	0.006221242
26-Jan-06	41.55	-0.00700401	-0.0065431		min=0.003790269
25-Jan-06	41.26	0.00435309	-0.00438597		
24-Jan-06	41.44	-0.00435309	-0.00435309		
23-Jan-06	41.26	-0.0002424	-0.00430314		
20-Jan-06	41.25	0.030323441	-0.00397522		
19-Jan-06	42.52	-0.00731739	-0.00356507		
18-Jan-06	42.21	0.01154176	-0.00189798		
17-Jan-06	42.7	0.006535971	-0.00073287		
13-Jan-06	42.98	0.000465224	-0.00069987		
12-Jan-06	43	0.004871834	-0.0002424		
11-Jan-06	43.21	-0.00766643	0		
10-Jan-06	42.88	-0.00069987	0.000465224		
9-Jan-06	42.85	-0.00397522	0.001219066		
6-Jan-06	42.68	-0.01796739	0.002691792		
5-Jan-06	41.92	-0.00430314	0.00435309		
4-Jan-06	41.74	-0.0103553	0.004513608		
3-Jan-06	41.31	-0.02202732	0.004871834		
30-Dec-05	40.41	0.007642093	0.006535971		
29-Dec-05	40.72	0.006608762	0.006608762		
28-Dec-05	40.99	0.001219066	0.007642093		
27-Dec-05	41.04	0.00873368	0.00873368		
23-Dec-05	41.4	0	0.01154176		
22-Dec-05	41.4	-0.0065431	0.012386315		
21-Dec-05	41.13	-0.00438597	0.016000341		
20-Dec-05	40.95	-0.00073287	0.017229428		
19-Dec-05	40.92	0.016000341	0.030323441		
16-Dec-05	41.58				

In my statistics, the followings are some important variables by computation: $n = 33, \overline{X} = 0.000566, S^2 = 0.000107, a = 3.2538, b = 115.2258, M = 7$.

Using these, I computed the upper bound for σ in step 1 is $\tilde{\sigma} = 0.0325$ while the lower bound in step 3 is $\hat{\sigma} = 0.00379$ (minimum value in the last column). Since the lower bound is less than the upper bound, we do not reject the lognormal assumption for this company.

3.2 Second Method

The second idea is called the Chi-Square Goodness-Of-Fit test, which was first developed by Karl Pearson (see [DS], p.370). For the normal distribution the mechanics of this test consist of discretizing the hypothesized distribution into a multinomial distribution of k cells, counting the observed number of observations in each cell and contrasting these, via a Chi-Square statistic, with the expected number of observations for each cell.

STEP 1: Determine $a_1 < a_2 < a_3 < \cdots < a_k$ to put all the data points into k+1 bins: $(-\infty, a_1), [a_1, a_2), [a_2, a_3) ..., [a_{k-1}, a_k), [a_k, +\infty)$, and compute the number of data points n_i in each interval. Clearly $\sum_{i=1}^{k+1} n_i = n.$

STEP 2: Compute p_i for each interval where

$$p_1 = \Phi(\frac{a_1 - \overline{X}}{S}),$$

$$p_i = \Phi(\frac{a_i - \overline{X}}{S}) - \Phi(\frac{a_{i-1} - \overline{X}}{S}) \quad \text{for} \quad 2 \le i \le k \text{, and}$$

$$p_{k+1} = 1 - \Phi(\frac{a_n - \overline{X}}{S}).$$

STEP 3: For this test, I use the test statistic T= $\sum_{i=1}^{k+1} \frac{(n_i - np_i)^2}{np_i}$ and the critical value for this test comes from the Chi-Square distribution with degrees of freedom equal to the number of terms in the sum (k+1) minus 1, minus the number of unknown parameters, i.e. k+1-1-2=k-2. There are a number of rules that have been proposed for deciding when the test is reasonably accurate. They center around the values np_i. The most conservative rule states that each must be at least 5. Some authors claim that values as low as 1 are acceptable. All agree that the test works best when the values are about equal from term to term. If the data are grouped, there is little choice but to use the groups as given, although adjacent groups could be combined to increase np_i. For individual data, the data can be grouped for the purpose of performing the test.

STEP 4: If $\chi_\alpha^2(k-2) \le T$, then we reject $H_0(\overline{X}, S^2)$ with confidence $1-\alpha$.

According to this method, the sample test on the same company "QQQQ" mentioned in the first method is:

Yi	n=33	Ni	Pi	nPi	(Ni-nPi)^2/nPi
-0.0220273	<-0.005	8	0.295409596	9.74851665	0.313618019
-0.0179674	[-0.005, 0)	9	0.182775131	6.03157934	1.460897841
-0.0133877	[0, 0.007)	9	0.254648638	8.40340504	0.042354919
-0.0103553	>=0.007	7	0.267166635	8.81649897	0.374260634
-0.0076664		33			T=2.191131414
-0.0073174					
-0.007004					
-0.0065431					
-0.004386					
-0.0043531					
-0.0043031					
-0.0039752					
-0.0035651					
-0.001898					
-0.0007329					
-0.0006999					
-0.0002424					
0					
0.00046522					
0.00121907					
0.00269179					
0.00435309					
0.00451361					
0.00487183					
0.00653597					
0.00660876					
0.00764209					
0.00873368					
0.01154176					
0.01238632					
0.01600034					
0.01722943					
0.03032344					

In this case, the degree of freedom is 4-2-1=1. Thus we do not reject it as lognormal, since $\chi_{\alpha}^{2}(1)$ = 41.8215 (by Mathematica) > T=2.191131414.

I have to admit that I chose $a_1, a_2, a_3 ..., a_k$ subjectively, but my results are consistent with those Dr. Ramsey achieved by choosing equal probability bins with np_i at least 5.

4. Graphical Method

Another way to see how well the model and data match is to plot the respective density and distribution functions.

There are several graphical ways to describe the difference. Here I use a P-P plot, which is also called a probability plot. The plot is created by ordering the observations as $y_1 \le y_2 \le y_3 ... \le y_n$. A point is then plotted corresponding to each value. The point's coordinates to are $(F_n(y_j), F^*(y_j))$, where $F_n(y_j)$ is the empirical distribution function and $F^*(y_j)$ is the cumulative distribution function for a normal distribution with mean $\mu = \overline{X}$ and variance $\sigma^2 = S^2$. If the Black-Scholes model fits well, the plotted points will be near the $45°$ line running from (0,0) to (1,1). However, for this to be the case, a different definition of the empirical distribution function is needed. It can be shown that the expected value of $F_n(y_j)$ is j/(n+1). Therefore the empirical distribution should be that value and not the usual j/n. If two observations have the same value, either plot both points or plot a single value by averaging the two x-coordinates values. I used the first choice in my plot (Figure 2).

For the same data points from company "QQQQ", here is the P-P plot:

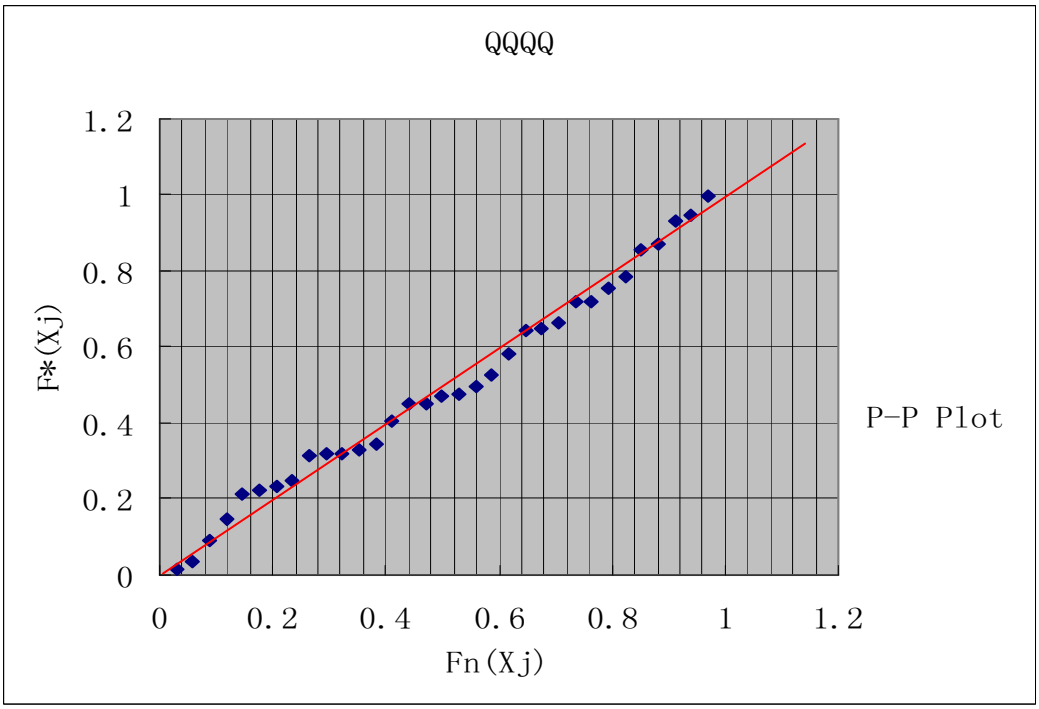


Figure 2: P-P Plot of "QQQQ"

From this plot, we can see the difference between the plotted dot points and line $y=x$ is not excessive, so we do not reject the null hypothesis with parameter $(\overline{X}, S^2)$.

Also, we can quantify this "eye-ball" test. Let

$$\text{D=max } \{|F_n(y_j) - F^*(y_j)|, 1 \le j \le n\},$$

be the test statistic, and use the critical value for this test $k/\sqrt{n}$, where the constant k can be found from the table in [DS], p.112. If $D \ge k/\sqrt{n}$, we reject $H_0(\overline{X}, S^2)$ by Kolmogorov-Smirnov test (see [KPW], p.428).

Now I use the P-P Plot to check the non-lognormality of the data points from the company with symbol“ AGP ” (from my.yahoo.com).

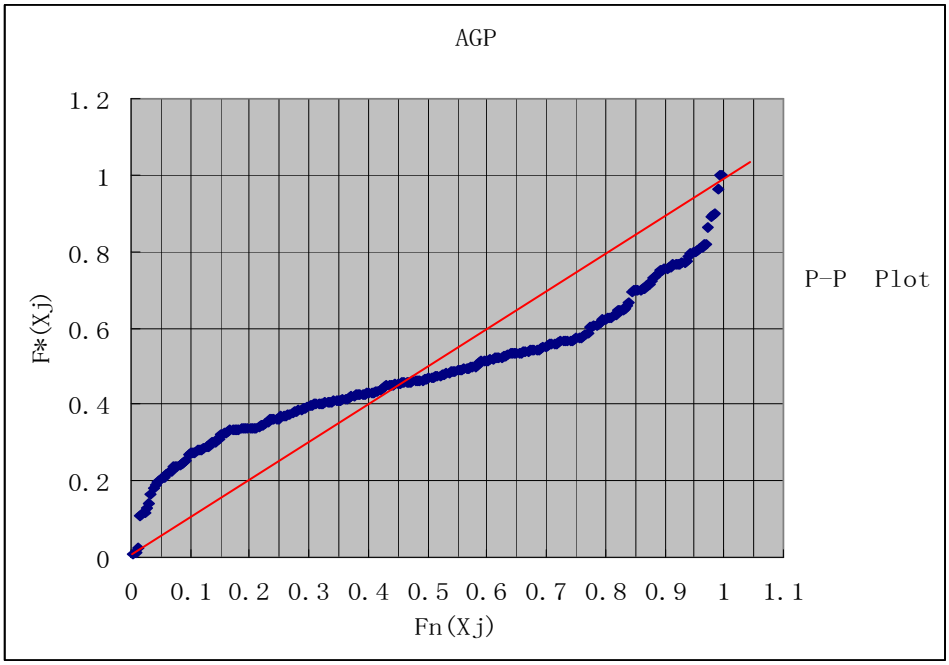


Figure 3: P-P Plot of “AGP

From the above plot, we can see the difference between the dot points and the main diagonal is much bigger than the first one. At the same time, the test statistic D is greater than the critical value. So we reject $H_0(\overline{X}, S^2)$.

The following P-P plot shows that “CRMT” fits the lognormal hypothesis with parameter $(\overline{X}, S^2)$ very well.

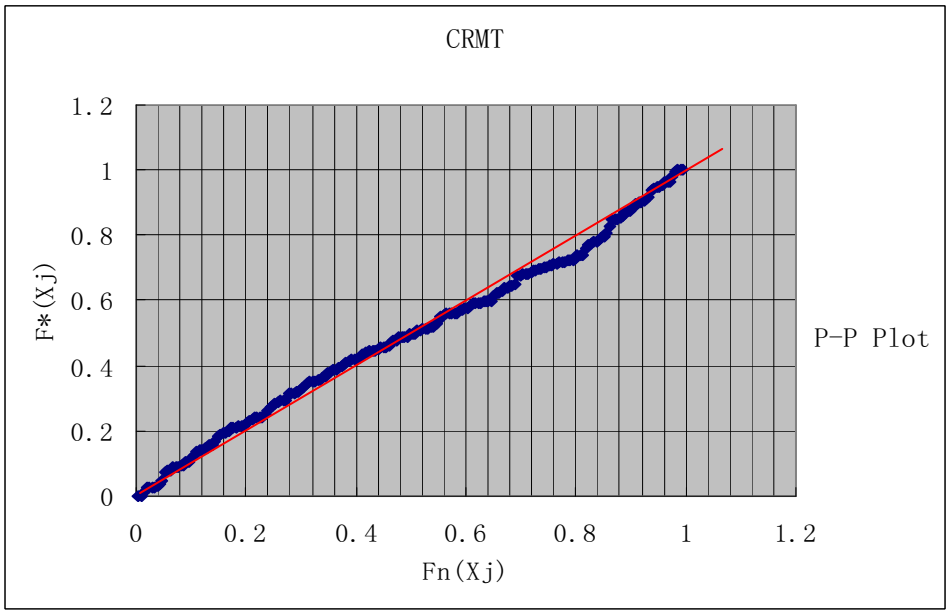


Figure 4: P-P Plot of “CRMT”

However, there exist some “bad” P-P plots which illustrate why we might reject the lognormal hypothesis with parameter $(\overline{X}, S^2)$ as in Figure 5 below:

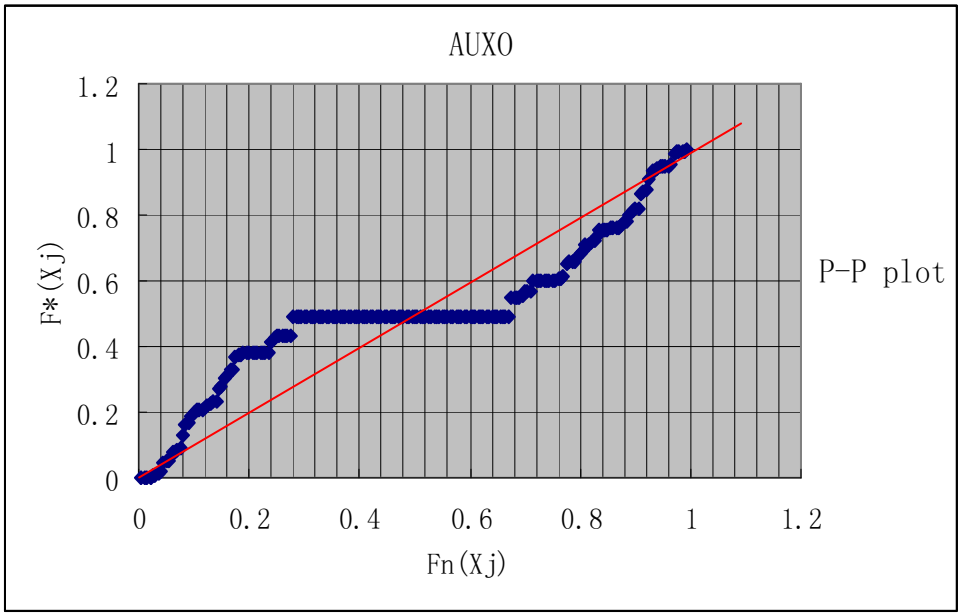


Figure 5: P-P Plot of “AUXO”

5. Test Results

Methods 1 and 2 produced the following results:

Compamy Quotes	# Of Data Points	First Method	Second Method
AAPL	237	Lognormal	Lognormal
ABGX	1402	Not Lognormal	Not Lognormal
ACAD	431	Not Lognormal	Not Lognormal
ACP	50	Lognormal	Lognormal
ACPW	1380	Not Lognormal	Not Lognormal
ADLS	130	Lognormal	Lognormal
AEA	54	Lognormal	Lognormal
AEGG	199	Lognormal	Lognormal
AES	1430	Not Lognormal	Not Lognormal
AGP	268	Not Lognormal	Not Lognormal
AGU	28	Lognormal	Lognormal
AM	23	Lognormal	Lognormal
AMAT	55	Lognormal	Lognormal
AMIC	754	Not Lognormal	Not Lognormal
AMZN	1608	Not Lognormal	Not Lognormal
AN	1407	Not Lognormal	Not Lognormal
APD	31	Lognormal	Lognormal
APN	2186	Not Lognormal	Not Lognormal
ARI	27	Lognormal	Lognormal
ASPR	199	Lognormal	Lognormal
ATLS	439	Lognormal	Lognormal
ATVI	71	Lognormal	Lognormal
AUXO	199	Not Lognormal	Not Lognormal
B	52	Lognormal	Lognormal
BP	64	Lognormal	Lognormal
BPG	37	Lognormal	Lognormal
BRCM	1503	Not Lognormal	Not Lognormal
BRNC	120	Lognormal	Lognormal
CECE	3120	Not Lognormal	Not Lognormal
CENT	3124	Not Lognormal	Not Lognormal
CHA	206	Lognormal	Lognormal
CHU	198	Lognormal	Lognormal
CIX	44	Lognormal	Lognormal
CLRI	199	Lognormal	Not Lognormal
CMCSA	1692	Not Lognormal	Not Lognormal
CNFL	2640	Not Lognormal	Not Lognormal
CNXT	657	Not Lognormal	Lognormal
CPRT	1021	Not Lognormal	Not Lognormal
CRMT	208	Lognormal	Lognormal
CSCO	1475	Not Lognormal	Not Lognormal

Compamy Quotes	# Of Data Points	First Method	Second Method
CUO	1592	Not Lognormal	Not Lognormal
DELL	1741	Not Lognormal	Not Lognormal
DF	159	Lognormal	Lognormal
DGIT	2506	Not Lognormal	Not Lognormal
DIS	39	Lognormal	Lognormal
EBAY	232	Not Lognormal	Lognormal
FMCN	148	Lognormal	Lognormal
GAI	1418	Not Lognormal	Not Lognormal
GE	29	Lognormal	Lognormal
GNI	27	Lognormal	Lognormal
GOOG	370	Lognormal	Lognormal
GPI	2076	Not Lognormal	Not Lognormal
GTW	1607	Not Lognormal	Not Lognormal
HCAR	199	Lognormal	Not Lognormal
HD	46	Lognormal	Lognormal
HON	60	Lognormal	Lognormal
HYTM	543	Lognormal	Lognormal
IAL	37	Lognormal	Lognormal
INTG	719	Not Lognormal	Not Lognormal
IPG	809	Not Lognormal	Not Lognormal
JDSU	1484	Not Lognormal	Not Lognormal
JNJ	52	Lognormal	Lognormal
JNPR	1417	Not Lognormal	Not Lognormal
KO	46	Lognormal	Lognormal
KTCC	3998	Not Lognormal	Not Lognormal
LECO	30	Lognormal	Lognormal
LFC	537	Lognormal	Not Lognormal
LTON	489	Not Lognormal	Not Lognormal
LU	931	Not Lognormal	Not Lognormal
LVLT	1877	Not Lognormal	Not Lognormal
LWAY	488	Not Lognormal	Not Lognormal
MAS	26	Lognormal	Lognormal
MET	63	Lognormal	Lognormal
MO	34	Lognormal	Lognormal
MSC	44	Lognormal	Lognormal
MXWL	2301	Not Lognormal	Not Lognormal
ORCL	1322	Not Lognormal	Not Lognormal
POIG	199	Lognormal	Not Lognormal
QCOM	43	Lognormal	Lognormal
QQQQ	33	Lognormal	Lognormal
RBC	30	Lognormal	Lognormal
RFMD	1363	Not Lognormal	Not Lognormal
RNDC	1385	Not Lognormal	Not Lognormal
SFP	1635	Not Lognormal	Not Lognormal

Compamy Quotes	# Of Data Points	First Method	Second Method
SINA	1454	Not Lognormal	Not Lognormal
SIRI	2870	Not Lognormal	Not Lognormal
SUNW	1296	Not Lognormal	Not Lognormal
SYMC	298	Not Lognormal	Lognormal
SYT	140	Lognormal	Lognormal
TATTF	79	Lognormal	Lognormal
TRCI	30	Lognormal	Lognormal
UNT	4553	Not Lognormal	Not Lognormal
VNK	585	Not Lognormal	Not Lognormal
VTSS	1574	Not Lognormal	Not Lognormal
WAL	24	Lognormal	Lognormal
WCG	407	Lognormal	Lognormal
WHR	64	Lognormal	Lognormal
WPC	27	Lognormal	Lognormal
XOM	60	Lognormal	Lognormal
YAHOO	426	Not Lognormal	Not Lognormal

The following summarizes the previous table.

	Do not reject Lognormal by Methods 1 and 2	Not Lognormal by Methods 1 and 2
$n \le 250$	45	1
$n > 250$	4	43
7 companies did not have identical results by Methods 1 and 2.		

Among those 7 companies that did not have identical results by the first two methods, I evaluated their lognormality by examing the P-P plot. For these "split decision" cases, method 3 rejected H_0 for 3 of them, which happens to coincide with having a sample size at least 250, while I did not reject the null hypothesis for the remaining 4 companies with sample sizes less than 250. For example, "EBAY" with n=232 did not have the identical results from the first two methods, but the P-P plot did not support rejecting $H_0(\overline{X}, S^2)$.

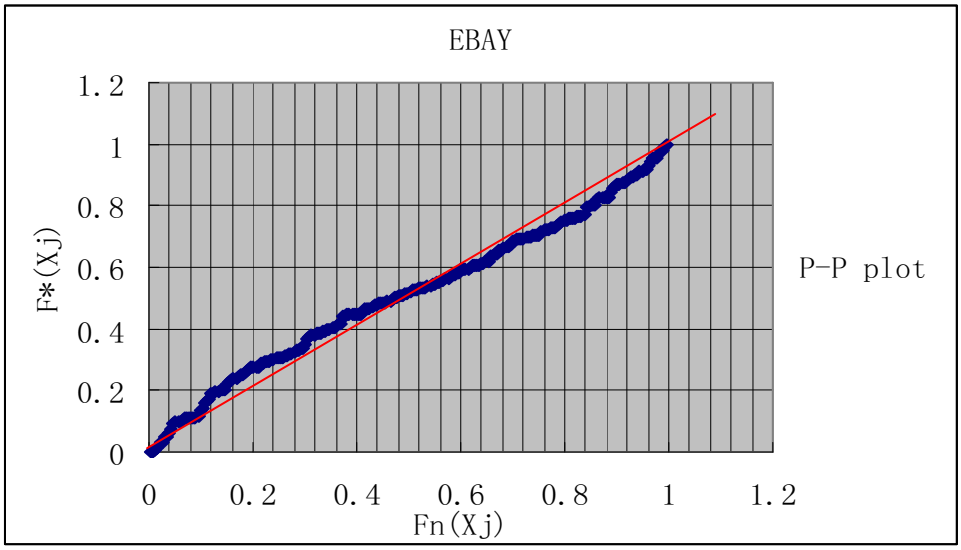


Figure 6: P-P Plot of "EBAY"

6. Tail Analysis

According to the test results above, it seems that almost all of the short sequences of daily stock prices satisfy the lognormal assumption in Geometric Brownian Motion. However, we should reject the lognormal assumption if we deal with long sequences of daily stock prices.

Furthermore, according to the theory of power-law distribution in financial market fluctuations, the probability that a return, r_i, where $r_i = (X_i - \overline{X})/S$, has an absolute value larger than x is found empirically to be P($|r_i|$>x)=O(x^{-3}) (see [GGPS], p.267).

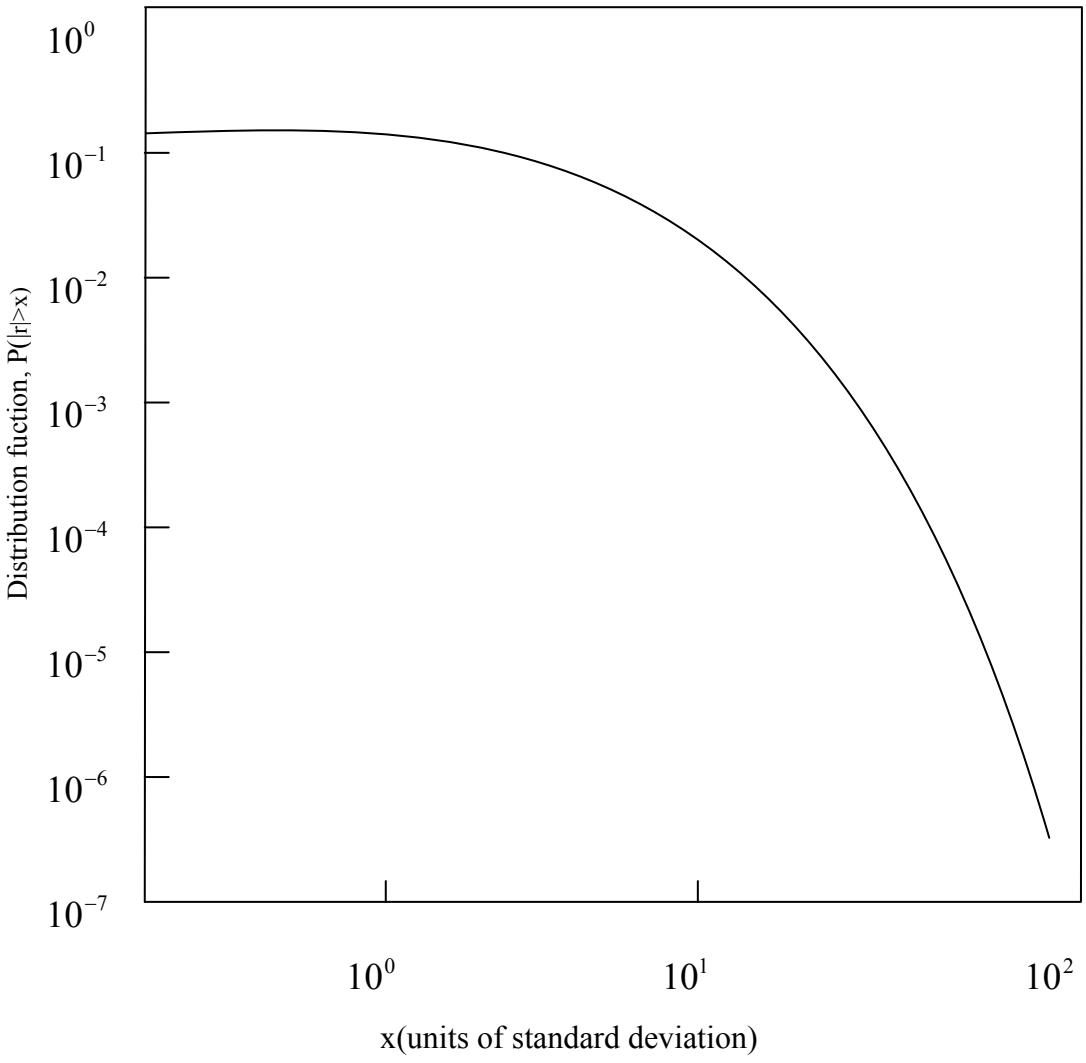


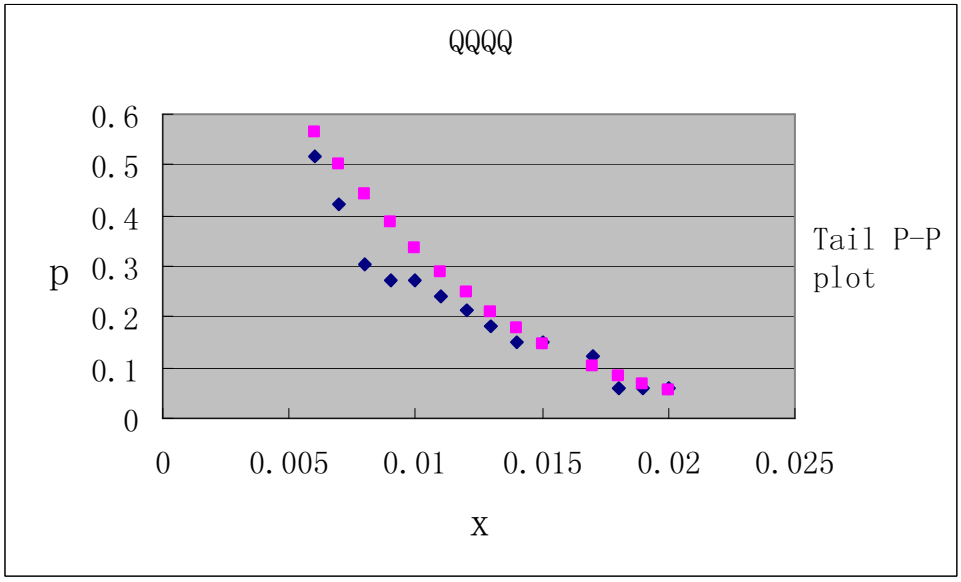
Figure7: Cumulative distributions of the normalized 15-min absolute return of the 1000 largest companies in the 'Trades and Quotes' database for the 2-yr period 1994-1995. We define the normalized return as $r_i = (X_i - \overline{X})/S$. We obtain P($|r_i| > x$)=O(x^{ζ}) with $\zeta = 3.1 \pm 0.1.$

For this reason, I double-checked the tail part of the curve with data points from company "QQQQ".

x	P(x)=P(|Xi|>x)	P*(x)
0.006	0.515151515	0.562818
0.007	0.424242424	0.499609
0.008	0.303030303	0.440382
0.009	0.272727273	0.385399
0.01	0.272727273	0.334828
0.011	0.242424242	0.288745
0.012	0.212121212	0.247141
0.013	0.181818182	0.209927
0.014	0.151515152	0.176949
0.015	0.151515152	0.147994
0.017	0.121212121	0.1011
0.018	0.060606061	0.082566
0.019	0.060606061	0.066887
0.02	0.060606061	0.053746

Here P*(x) represents the probability of model distribution, i.e. a normal distribution with parameter $(\bar{X}, S^2)$.

In the following graph, the darker points (diamonds) correspond to P(x) and lighter points (squares) correspond to P*(x):



From the two curves above, we can see the end of tail of P*(x) is a little heavier than P(x), which is identical with the result mentioned in [Jo].

References:

[DS] **Ralph B. D'Agostino and Michael A. Stephens,** Goodness-Of-Fit Techniques, 1986.

[HT] **Robert V. Hogg and Elliot A. Tanis**, Probability And Statistical Inference, *Sixth Edition*, 2001.

[GGPS] Xavier Gabaix, Parameswaran Gopikrishnan, Vasiliki Plerou and H.Eugene Stanley, A theory of power-law distribution in financial market fluctuations, *Letters to nature, vol423,* 2003.

[Jo] **John C. Hull,** Options, Functions, And Other Derivatives, *Sixth Edition*, 2006.

[KPW] **Stuart A. Klugman, Harry H. Panjer and Gordon E. Willmot,** Loss Models, *Second Edition*, 2004.

[Mo] **Morris H. DeGroot,** Probability And Statistics, 1975.

[Sh] **Sheldon M. Ross,** An Elementary Introduction To Mathematical Finance, *Second Edition*, 2003.

[Sw] **Stephen Wolfram,** The Mathematica book, *fifth Edition*, 2003.