Journal of Cost Analysis and Parametrics, 5:5–24, 2012
ISSN: 1941-658X print
DOI: 10.1080/1941658X.2012.682922

www.iceaaonline.com

The Fractal Nature of Cost Risk: The Portfolio Effect, Power Laws, and Risk and Uncertainty Properties of Lognormal Distributions

CHRISTIAN SMART

Missile Defense Agency, Redstone Arsenal, Alabama

Cost risk can be added to the list of the many phenomena in nature that follow a power-law probability distribution. Both the normal and lognormal, neither of which is a power-law distribution, underestimate the probability of extreme cost growth, as shown by comparison with empirical data. This situation puts the widely debated "portfolio effect" into further dispute. However, even though power laws are useful for modeling extreme events, budgets are not typically set at extreme percentiles, such as the 90th. Indeed, budgets are usually set at the 70th percentile or below. In addition, it is shown that the lognormal distribution is also problematic in that region and for percentile funding in general. To model cost risk for an individual program by setting budgets and/or reserves using percentile funding with a percentile chosen at or below the 70th percentile, it appears that the normal distribution may be the best option.

> One must do no violence to nature, nor model it in conformity to any blindly formed chimera. (Janos Bolyai, 1820)

Introduction

Janos Bolyai was a 19th century Hungarian mathematician noted for discovering non-Euclidean geometry. While researching the parallel postulate of Euclid as teenager, he became convinced that there could be a geometry independent of the parallel postulate, which inspired him to write to his father, also a mathematician, that "One must do no violence to nature, nor model it in conformity to any blindly formed chimera; that on the other hand, one must regard nature reasonably and naturally, as one would the truth, and be contented only with a representation of it which errs to the smallest possible extent," (Gray, 2004). It was later discovered by Einstein and others that physical space is, in fact, non-Euclidean. Thus, the mathematical theory developed by Bolyai and others provided a geometric tools needed for much of 20th century physics, including the development of general relativity (Kiss, 1999).

In the spirit of Bolyai, the contemporary mathematician Benoit Mandlebrot, building upon and synthesizing the work of many disparate predecessors, discovered that nature often does not act in accordance with the linear mathematics that were developed in the 18th and 19th centuries (Mandlebrot, 1983, 1997). Just as the Euclidean space model of nature is too simple, Mandlebrot discovered that numerous natural and man-made phenomena are subject to random behavior that does not follow the statistical distributions commonly

Address correspondence to Christian Smart, Missile Defense Agency, Redstone Arsenal, AL 35898. E-mail: christian.smart@mda.mil

taught in elementary statistics courses, such as the normal and lognormal. Examples include financial markets, energy of incoming cosmic rays, frequency of words, amount of damage a fire causes to a house, and distribution of incomes, to name only a few. Normal and lognormal distributions underestimate the frequency of extreme events for such phenomena, which are said to be "self-similar" and follow a power law. An event or object is self-similar if it is exactly like or similar to a part of itself; that is, the whole has the same shape as one or more of its parts. An example is the map of a coastline. As Mandelbrot pointed out in his book *The Fractal Geometry of Nature* (Mandlebrot, 1983), while the maps of coastlines rendered at different scales differ in specific details, they have the same types of features: "in a rough approximation, the small and large details of coastlines are geometrically identical except for scale" (Mandlebrot, 1983, p. 34).

In this article, it is shown that cost risk can be added to the list of the many phenomena to which Mandelbrot's fractal model applies.

Cost Growth and Power Laws

Cost growth is the amount by which a program exceeds its initial budget. It is typically expressed as a percentage. For example, for a program initially expected to cost \$100 million at the beginning of the program, but which actually costs \$150 million by the end of the program's development, the program is said to have experienced 50% cost growth. Cost growth has been shown to be an endemic and universal phenomenon. Studies by Shaffer (2003), the U.S. Government Accountability Office (GAO, 1992), and Smart (2002, 2007) have shown that on average, over three-quarters of all NASA programs experience cost growth, with an average cost growth ranging to 35% and higher, with many programs experiencing much higher growth, including 100% or more. Cost overruns are not limited to space and weapons systems development projects. A 2002 study found that 90% of construction projects experienced cost overruns (Flyvbjerg et al., 2002). Boston's Big Dig project experienced a \$11 billion overrun (Flyvbjerg, 2005). The Suez Canal cost 20 times as much as initially planned, and the Sydney Opera House cost 15 times as much as was originally projected. The Concorde supersonic airplane, a technology marvel in the late 1960s and 1970s that was able to travel from New York to Paris in only 3.5 hours, cost 12 times more than predicted (Flyvbjerg et al., 2002). Even smaller projects are not immune to large amounts of cost growth. The new Madison County Jail in north Alabama, which was supposed to cost \$29 million when the project began, eventually grew to \$79 million (Doyle, 2010).

Cost risk is the probability that an estimate will exceed a specified amount, such as \$100 million or \$150 million. Cost growth and cost risk are intrinsically related. Historical cost growth provides an excellent means for determining the overall level of risk for cost estimates. For example, if 95% of past programs have experienced less than 100% growth, it should be expected that the ratio of actual cost to the initial estimate should be less than 100% with approximately 95% confidence.

A power law is a polynomial relationship that exhibits scale invariance, which means that the relationship does not vary as the scale changes. Scale invariance is closely related to self-similarity and is a prime consideration in fractal geometry (Mandlebrot, 1983). A real-valued function $\boldsymbol{f}$ is defined as scale- invariant if and only if there exists a constant $\boldsymbol{k}$ such that

$$\boldsymbol{f(cx) = c^k f(x)}$$

for all real values x in the function's domain and for any real number $c \neq 0$. The power function

$$f(x) = ax^b$$

is scale-invariant since

$$f(cx) = a(cx)^b = ac^bx^b = c^bax^b = c^bf(x).$$

However, the affine function

$$f(x) = 2x + 1$$

is not scale-invariant. To see this, note that

$$f(cx) = 2(cx) + 1 = 2cx + 1.$$

Then there would have to exist a k such that, for all values of c and x,

$$2cx + 1 = f(cx) = c^k f(x) = c^k(2x + 1).$$

When $x = 0$, note that equality of these two terms requires that

$$2c \cdot 0 + 1 = c^k(2 \cdot 0 + 1),$$

which means that

$$1 = c^k,$$

namely $c = 1$ or $k = 0$ and $c \neq 0$. But then when $x = 1$,

$$2c \cdot 1 + 1 = f(c \cdot 1) = c^k f(1) = c^k(2 \cdot 1 + 1) = 2c^k + 1 = 3 \text{ (since } c^k = 1).$$

Therefore, it follows that

$$f(c) = 3,$$

which is true only when $c = 1$. Scale invariance requires that the invariance works for all $c \neq 0$, not only particular values of c, so the affine function is not scale-invariant.

A single-parameter Pareto distribution is a probability distribution that embodies this concept. If X is a random variable that follows a Pareto distribution, then the tail probability that X is greater than a number x is given by

$$\mathbf{Pr}\,(X \geq x) = \left(\frac{x}{x_m}\right)^{-k}$$

for all $x \geq x_m$, where x_m is the (necessarily positive) minimum possible value of X, and $k > 0$. The tail probability of the single-parameter Pareto distribution is scale-invariant, because it is a power function with $a = (x_m)^k$ and $b = -k$.

The probability density function of a single-parameter Pareto distribution is defined as

$$\mathbf{Pr}(x) = \frac{kx_m^k}{x^{k+1}},$$

and the cumulative distribution function is defined as

$$\mathbf{Pr}(X \leq x) = \mathbf{1} - \left(\frac{x_m}{x}\right)^k.$$

An example of a scaling phenomenon is the distribution of income. Consider, for example, income distributions in the United States. For incomes above \$70,000, incomes scale according to a power law; see Figure 1.

For any given income in Figure 1, the diamond-shaped markers on the solid line represent the empirical likelihood that an income exceeds that value. For example, slightly over 10% of U.S. households have incomes in excess of \$100,000. The graphs of the normal, lognormal, and Pareto show the predicted likelihood of exceeding a given income, based on a regression using the income data. This graph is designed to show the tails of the distribution and how likely an income is to be exceeded. It is the opposite of an S-curve plot that shows the cumulative frequency.

Note that in Figure 1, the power law fit is nearly perfect, with a Pearson $R^2 = 99.49\%$. Census Bureau data from 2006 (available at www.census.gov) indicate that 1.7% of U.S. households have incomes greater than \$250,000. The Pareto power law estimates this probability to be equal to 1.73%, a very close approximation. The lognormal distribution fit to the Census Bureau data estimates that the probability is equal to 0.53%, while a normal distribution estimates the probability to be equal to less than 3.5 in one million. The normal distribution has thin tails, which, in practical terms, means that the probability of extreme events is extremely small. The lognormal distribution provides a better estimate of the right tail, but is not as good as that of the Pareto distribution.

Cost growth data follow a similar pattern. Figure 2 shows a probability plot for the cost growth from the Schaffer study (2003). It is seen from Figure 2 that cost growth data follows a pattern similar to that of income distribution. That is, the Pareto power law better

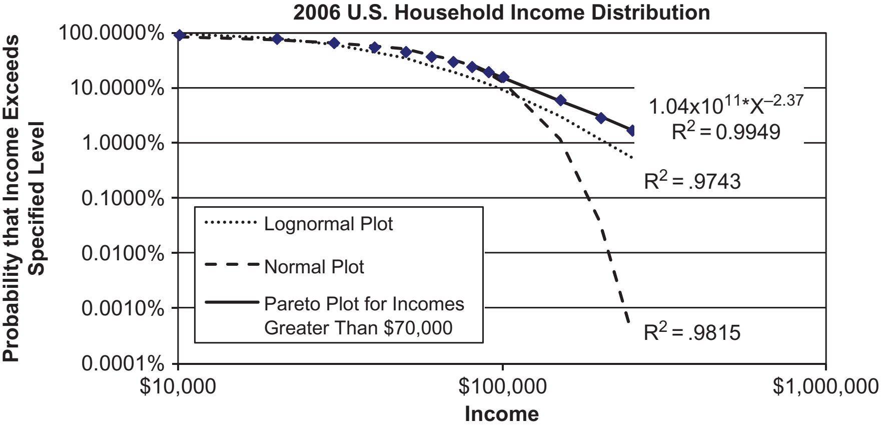

FIGURE 1 U.S. household incomes, Pareto distributions, normal and lognormal distributions for U.S. income (source: U.S. Census Bureau) (color figure available online).

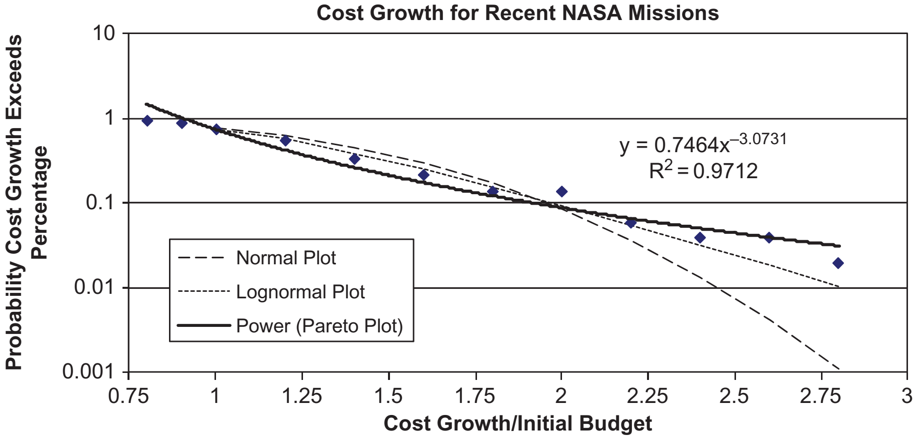

FIGURE 2 Cost growth probability plot for Schaffer study (color figure available online).

predicts the probability of extreme cost growth (in this case, greater than 100%) than either the normal or lognormal law. For example, the probability that cost growth exceeds 180% is actually 2%. The Pareto distribution estimates this probability to be equal to 3.1%, while the lognormal estimate is 1.0%, and the normal estimate is 0.1%. In this case, the lognormal and Pareto provide approximately equal accuracy in estimating the extreme right tail, but the lognormal underestimates the probability of this level of cost growth.

What is the impact of scale invariance on cost growth? One facet of cost risk is that budgets are not set in isolation. Rather, budgets are set in the context of multiple ongoing projects. Thus, in practice, a portfolio of projects is often considered. In this case, it has been suggested that due to diversification across a suite of missions, it is possible to achieve a high level of confidence in the overall budget while setting budgets for individual missions at a lower level (Taleb, 2007). This draws on ideas in economics, such as modern portfolio theory developed by Nobel laureate Harry Markowitz (1959); see Table 1 for an example (Anderson, 2004). In Table 1, there are ten mutually independent normal distributions. For independent normal variates, the sum of the variates is also normal,

TABLE 1 Example of the portfolio effect for ten mutually independent normal distributions

Project	μ	σ	61% Confidence level
Project 1	$1,696	$539	$1,846
Project 2	$1,481	$404	$1,594
Project 3	$1,395	$435	$1,516
Project 4	$874	$288	$954
Project 5	$840	$219	$901
Project 6	$1,449	$371	$1,552
Project 7	$1,638	$537	$1,788
Project 8	$1,031	$259	$1,103
Project 9	$1,271	$323	$1,361
Project 10	$1,937	$602	$2,105
Portfolio metrics	$13,612	$1,317	$14,720

with mean equal to the sum of the means of the individual random variable and standard deviation equal to the square root of the sum of squares of the individual standard deviations. Therefore, the portfolio level mean in Table 1 is the sum of individual project means, and the portfolio level standard deviation in Table 1 is the square root of the sum of the individual project variances. Because of this fact, in this particular case, it is possible to achieve 80% confidence for the full portfolio of ten projects while budgeting each individual project at the 61% confidence level. This is a significant savings. The portfolio level 80% confidence level in Table 1 is the sum of the 61% confidence levels for the individual projects.

However, when a Pareto distribution is applied to the same data (see Table 2), using the means and a common scale to represent the probability of cost growth that is consistent with the cost-growth historical data, a drastically different conclusion is reached.

The scale derived from the cost growth data in the Schaffer study is used for all ten projects with means equal to those defined in Table 1; that is, the empirical data summarized in Figure 2 are used to define the scale for each distribution, assuming that cost growth follows a power law. This distribution is consistent with the cost-growth data shown in Figure 2. Assuming these ten Pareto distributions to be mutually independent, a Monte Carlo simulation with 5000 trials was performed using @Risk in Microsoft Excel. For these ten distributions, it was found that the portfolio effect is minimal and that each individual project must be funded at the 77.5% confidence level to achieve an overall portfolio confidence level equal to 80%; see Table 2 for a summary of these data.

The primary consequence of scale invariance for cost growth and cost risk is that the portfolio effect is minimal, if it even exists at all. This is due to the fact that the normal, and, to a lesser extent, the lognormal, distributions underestimate the probability of extreme cost growth. When this probability is modeled more accurately in accordance with the empirical data, the portfolio effect vanishes, or at best is minimal. Thus, policy makers should be careful in assuming that such an effect will help diversify risk among missions. If policy makers want to achieve a high level of confidence for their overall budget, their focus should be on sufficiently funding each individual mission at a sufficient confidence level. To do otherwise is to place faith in the "blindly formed chimera" of the normal distribution.

TABLE 2 Example of the minimal portfolio effect when scale invariance is taken into account (Pareto data)

Project	μ	Scale	77.5% Confidence level
Project 1	$1,696	3.0731	$1,859
Project 2	$1,481	3.0731	$1,623
Project 3	$1,395	3.0731	$1,529
Project 4	$874	3.0731	$958
Project 5	$840	3.0731	$921
Project 6	$1,449	3.0731	$1,588
Project 7	$1,638	3.0731	$1,795
Project 8	$1,031	3.0731	$1,130
Project 9	$1,271	3.0731	$1,393
Project 10	$1,937	3.0731	$2,123
Portfolio metrics	$13,612		$14,919

The Lognormal Paradox: Considerations for Modeling Cost Risk When Budgeting Near the Median (50% to 70% Confidence Levels)

While the Pareto distribution better models cost risk at the right tails, such as at the 90th percentile, it does not do a good job of modeling the central portion of the distribution, which is where funding is often set; see Figure 3 for an example with a common mean.

In Figure 3, the normal and lognormal distributions have common means and standard deviations, while the one-parameter Pareto distribution is defined so that its mean is the same as those of the normal and lognormal, while the scaling parameter is based on the empirical analysis of cost growth data. Note that the Pareto is more conservative at the right tail (at or above the 90th percentile) but is much less conservative at lower levels. This is the tradeoff in fat-tailed distributions—they better reflect the reality of extreme events, but they do not always do a good job of representing lower confidence levels, such as those at which budgets are typically set. Even Mandlebrot admits that scale invariance does not do as good a job as the normal and lognormal at representing the bulk of the distribution (Mandlebrot, 1997). Also note that the Pareto distribution in this example largely represents risk, but it does not realistically represent opportunities, or the potential for cost savings, below the 50th percentile.

Program reserves are typically set at the 70th percentile or below. For example, NASA policy dictates that all NASA programs must be funded at a 70% confidence level, so the concern is not with getting the tails correct, but rather in providing realistic assessments of the bulk of the cost risk distribution.

The lognormal distribution seems to represent a compromise between the fat-tailed distributions and the normal distribution. As we have seen, the lognormal distribution has a fatter tail than the normal distribution, and fatter-tailed distributions, such as the Pareto, have even fatter tails that better represent the 90th and 95th percentiles. However, the lognormal distribution does a better job of representing the bulk of the distribution. When funding is near the central tendency, the lognormal or normal distribution may do better job of representing the reserves needed.

One of the properties of the normal distribution is that the mean and median are equal to one another. Also, for two normal distributions with common mean $\boldsymbol{\mu}$ and standard deviations $\boldsymbol{\sigma_1} < \boldsymbol{\sigma_2}$, the distributions intersect at the 50th percentile, but for all percentiles

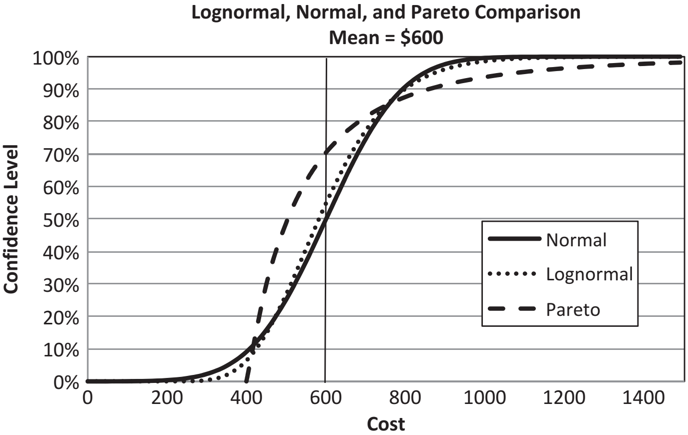

FIGURE 3 Comparison of lognormal, normal, and Pareto distributions with a common mean.

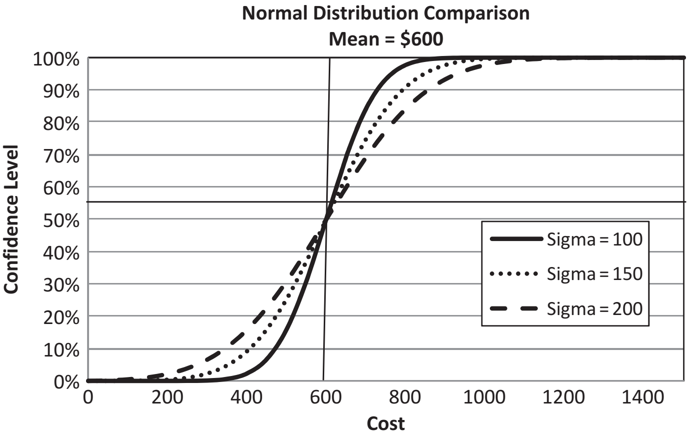

FIGURE 4 Comparison of normal distributions with common mean.

above the 50th percentile, more reserves are required for the distribution with the higher standard deviation. This is intuitive, since the standard deviation is a quantitative measure of uncertainty. Thus, a higher standard deviation means that there is more uncertainty. When uncertainty is higher, a higher budget is usually required to achieve any confidence level above the median. Figure 4 gives a graphical comparison of three normal distributions with common mean but different standard deviations.

For example, for the distribution with standard deviation equal to 100, the 55th percentile is $612, and when the standard deviation is equal to 200, the 55th percentile is $624.

Note that risk is not equivalent to standard deviation. Standard deviation is a measure of uncertainty, reflecting both the potential for cost to increase (risk) and the potential for cost to decrease (opportunity). In the case of normal and lognormal distributions, the distribution is completely specified by its mean and standard deviation. When the standard deviation increases, uncertainty increases, since the right tail also becomes fatter. Therefore, as the standard deviation increases, risk also increases. It follows that risk is positively associated with standard deviation, although it is not equivalent to it.

This common and intuitively appealing situation does not apply to the lognormal distribution. Indeed, for the lognormal distribution there is a paradox—when risk increases, less money may actually be required to achieve a higher amount of confidence! This seeming contradiction partly stems from a basic property of the lognormal distribution, which is that the mean is actually strictly greater than the median; consider the same means and standard deviations as earlier for a lognormal distribution used to represent cost risk (see Figure 5).

In this case, the 55th percentile, when the standard deviation is equal to 100, is $607, but when the standard deviation increases to 200, the 55th percentile drops to $598. This is not logical, since the risk has doubled, but less money is needed to achieve a relatively high level of confidence. These two distributions intersect at approximately the 60th percentile, so funding at a level equal to the 60th percentile would be needed just so that the two different risk profiles would require the same level of funding.

This is just an example illustrating this property that for two lognormals with a given mean, the one with the higher standard deviation, and hence the one with more risk, requires less funding at the 50th percentile and, at levels slightly above that, creates logical problems. This is because it implies that when budgeting to the 50th percentile, or slightly above

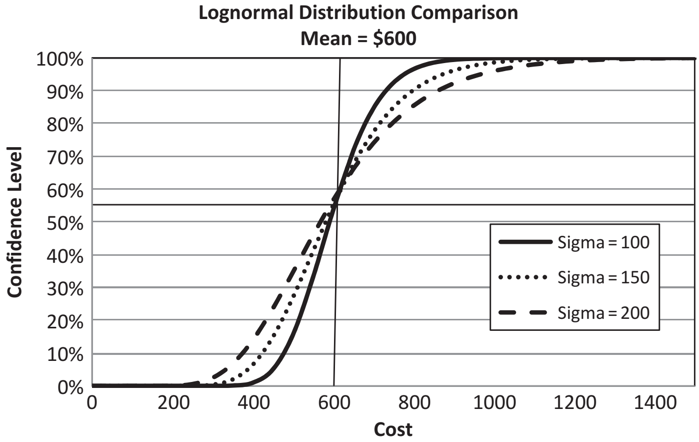

FIGURE 5 Comparison of lognormal distributions with a common mean.

that level, riskier events may require less funding than less risky events. Suppose an analyst develops a cost risk analysis for his manager that is modeled with a lognormal distribution. The analyst presents it to his manager, telling him that the 50th percentile is $100 million. After discussing the risks used in the analysis, the manager asks the analyst to go back and add an additional risk to the analysis and to update him with the new results. Suppose that the risk does not affect the mean, but it increases the standard deviation. Then the new analysis, which is riskier than the previous analysis, will require less than $100 million funding at the 50th percentile. Now the analyst goes back to his manager and reports to him that adding the additional risk to the analysis means that now only $98 million is needed to achieve the 50th percentile of the new risk distribution.While the analyst has done nothing wrong, he will have a difficult time explaining that a new analysis that considers additional risk will mean that less funding is required. This is formalized in the theorem that follows.

A lognormal distribution is the probability distribution of any random variable whose logarithm is normally distributed. Equivalently, a lognormal is the exponentiation of a normal distribution. The lognormal distribution can be represented by two parameters, such as the mean $\boldsymbol{\mu}$ and standard deviation $\boldsymbol{\sigma}$. It can also be represented by the mean $\boldsymbol{p}$ and standard deviation $\boldsymbol{q}$ of the log-transformed random variable that follows a normal distribution. Because $\boldsymbol{p}$ is also the median, as well as the mean, of the normal distribution, it follows that $\mathbf{exp}(\boldsymbol{p})$ is the median of the associated lognormal distribution, Note that the following additional relationships between $\boldsymbol{\mu}$, $\boldsymbol{\sigma}$, $\boldsymbol{p}$, and $\boldsymbol{q}$ apply:

$$\boldsymbol{\mu = e^{p+0.5q^2},}$$

$$\boldsymbol{\sigma^2 = e^{2p+q^2}\left(e^{q^2}-1\right),}$$

$$\boldsymbol{p = 0.5\ \ln\left(\frac{\mu^4}{\mu^2+\sigma^2}\right),}$$

$$\boldsymbol{q = \sqrt{\ln\left(1+\frac{\sigma^2}{\mu^2}\right)}.}$$

Theorem 1. *For two lognormal distributions with common mean μ and standard deviations $\sigma_1 > \sigma_2 > 0$, the 50th percentile of the lognormal distribution with mean μ and standard deviation σ_1 is less than the 50th percentile of the lognormal distribution with mean μ and standard deviation σ_2.*

Proof. Since $\sigma_1 > \sigma_2 > 0$, it holds that

$$\sigma_1^2 > \sigma_2^2$$

and

$$\mu^2 + \sigma_1^2 > \mu^2 + \sigma_2^2,$$

which means that

$$\frac{1}{\mu^2 + \sigma_1^2} < \frac{1}{\mu^2 + \sigma_2^2}$$

and

$$\frac{\mu^4}{\mu^2 + \sigma_1^2} < \frac{\mu^4}{\mu^2 + \sigma_2^2};$$

from this, it can be seen that

$$\sqrt{\frac{\mu^4}{\mu^2 + \sigma_1^2}} < \sqrt{\frac{\mu^4}{\mu^2 + \sigma_2^2}},$$

and so

$$\ln\left(\sqrt{\frac{\mu^4}{\mu^2 + \sigma_1^2}}\right) < \ln\left(\sqrt{\frac{\mu^4}{\mu^2 + \sigma_2^2}}\right),$$

which, in turn, implies that normal distribution means are such that $p_1 < p_2$. By the fact that the median of a lognormal is equal to $\exp(p)$, the proof is completed.

It is known from Theorem 1 that two lognormal distributions with a common mean intersect at some percentile greater than the 50th. The exact percentile at which the two intersect in terms of the parameters for a lognormal distribution can be determined. For a lognormal distribution the kth percentile can be represented as

$$e^{p+zq},$$

where z represents the z-score of the kth percentile for a standard normal distribution. This result, which will be shown as a complicated function, is of more than theoretical interest. It is important in showing that for a lognormal, as the risk increases but the mean remains constant, the maximum point at which the two distributions intersect increases without limit. This is significant, because it means that there is no limit at which the two lognormals may intersect. Even though the riskier lognormal, the one that has a fatter tail, will eventually intersect the less risky lognormal, this point of intersection has no upper bound. So simply increasing the percentile at which funding is set to the 60th, 70th, 80th, or even 90th percentiles will not overcome the logical problem that a less risky lognormal may require more funding than a riskier lognormal.

The next theorem involves the concept of the "coefficient of variation of a distribution," namely the ratio of the distribution's standard deviation to its mean.

Theorem 2. *For two lognormal distributions with a common mean $\boldsymbol{\mu}$ and standard deviations $\boldsymbol{\sigma_1} > \boldsymbol{\sigma_2} > 0$ with $\boldsymbol{\sigma_1} = \boldsymbol{\alpha} \cdot \boldsymbol{\sigma_2}$, where $\boldsymbol{\alpha} > \mathbf{1}$, the normal distribution z-score $\boldsymbol{z}$ at which the two intersect is the following function of the coefficients of variation of the two distributions:*

$$z = \mathbf{0.5}\left(\sqrt{\ln(1+\alpha^2 CV^2)} + \sqrt{\ln(1+CV^2)}\right),$$

where $\boldsymbol{CV} = \boldsymbol{\sigma_2}/\boldsymbol{\mu}$ is the coefficient of variation for the lognormal distribution with parameters $\boldsymbol{\mu}$ and $\boldsymbol{\sigma_2}$. Since $\boldsymbol{\sigma}_1 = \boldsymbol{\alpha} \cdot \boldsymbol{\sigma_2}$, the coefficient of variation for the lognormal distribution with parameters $\boldsymbol{\mu}$ and $\boldsymbol{\sigma_1}$ is $\boldsymbol{\alpha} \cdot \boldsymbol{CV}$. Also, the minimum value of z is $\boldsymbol{q_2}$, namely the standard deviation of the normal distribution associated with the lognormal distribution that has the smaller standard deviation $\boldsymbol{\sigma_2}$.

Proof. The $\boldsymbol{k}$th percentile of the lognormal distribution with mean $\boldsymbol{\mu}$ and standard deviation $\boldsymbol{\sigma}$ can be written as

$$e^{p+zq}.$$

Setting the two expressions for the $\boldsymbol{k^{th}}$ percentile where intersection occurs equal yields

$$e^{p_1+zq_1} = e^{p_2+zq_2},$$

and so

$$p_1 + zq_1 = p_2 + zq_2.$$

Solving for z yields

$$z = \frac{p_2 - p_1}{q_1 - q_2},$$

and substituting for $\boldsymbol{p}$ and $\boldsymbol{q}$ yields

$$z = \frac{\mathbf{0.5}\ln\left(\frac{\mu^4}{\mu^2+\sigma_2^2}\right) - \mathbf{0.5}\ln\left(\frac{\mu^4}{\mu^2+\sigma_1^2}\right)}{\sqrt{\ln\left(1+\frac{\sigma_1^2}{\mu^2}\right)} - \sqrt{\ln\left(1+\frac{\sigma_2^2}{\mu^2}\right)}},$$

which can be simplified, using properties of logarithms, to

$$z = \frac{0.5 \ln\left(\frac{\mu^4}{\mu^2+\sigma_2^2} \div \frac{\mu^4}{\mu^2+\sigma_1^2}\right)}{\sqrt{\ln\left(1+\frac{\sigma_1^2}{\mu^2}\right)} - \sqrt{\ln\left(1+\frac{\sigma_2^2}{\mu^2}\right)}} = \frac{0.5 \ln\left(\frac{\mu^2+\sigma_1^2}{\mu^2+\sigma_2^2}\right)}{\sqrt{\ln\left(1+\frac{\sigma_1^2}{\mu^2}\right)} - \sqrt{\ln\left(1+\frac{\sigma_2^2}{\mu^2}\right)}},$$

Substituting $\sigma_1 = \alpha \cdot \sigma_2$, it is found that

$$z = \frac{0.5 \ln\left(\frac{\mu^2+\alpha^2\sigma_2^2}{\mu^2+\sigma_2^2}\right)}{\sqrt{\ln\left(1+\frac{\alpha^2\sigma_2^2}{\mu^2}\right)} - \sqrt{\ln\left(1+\frac{\sigma_2^2}{\mu^2}\right)}}.$$

Recall that the coefficient of variation is the ratio of the standard deviation to the mean. The numerator for this expression can be rewritten in terms of the coefficients of variation, viz.,

$$\frac{\mu^2+\alpha^2\sigma_2^2}{\mu^2+\sigma_2^2} = \frac{\mu^2\left(1+\frac{\alpha^2\sigma_2^2}{\mu^2}\right)}{\mu^2\left(1+\frac{\sigma_2^2}{\mu^2}\right)} = \frac{1+\alpha^2 CV^2}{1+CV^2},$$

yielding

$$z = \frac{0.5\left(\ln\left(1+\alpha^2 CV^2\right) - \ln\left(1+CV^2\right)\right)}{\sqrt{\ln\left(1+\alpha^2 CV^2\right)} - \sqrt{\ln\left(1+CV^2\right)}} \quad .$$

Setting

$$x = \sqrt{\ln\left(1+\alpha^2 CV^2\right)}, \qquad y = \sqrt{\ln\left(1+CV^2\right)}\ ,$$

$$z = 0.5\left(\frac{x^2-y^2}{x-y}\right) = 0.5\left(\frac{(x-y)(x+y)}{x-y}\right) = 0.5\,(x+y)\ .$$

Therefore,

$$z = 0.5\left(\sqrt{\ln(1+\alpha^2 CV^2)} + \sqrt{\ln(1+CV^2)}\right),$$

where CV is the standard deviation (σ_2) of the lognormal distribution divided by the mean, and α is the ratio of σ_1 to σ_2.

Note that since $\sigma_1 > \sigma_2 > 0$, $\alpha > 1$. Because z is increasing as a function of α, notice that the greatest lower bound of z is attained when $\alpha = 1$. Therefore

$$z \geq 0.5\left(\sqrt{\ln(1+\alpha^2 CV^2)} + \sqrt{\ln(1+CV^2)}\right)$$

$$= \mathbf{0.5}\left(\sqrt{\mathbf{ln(1+}\boldsymbol{CV}^{\mathbf{2}}\mathbf{)}} + \sqrt{\mathbf{ln(1+}\boldsymbol{CV}^{\mathbf{2}}\mathbf{)}}\right)$$

$$= \mathbf{0.5}\cdot\mathbf{2}\left(\sqrt{\mathbf{ln}\left(\mathbf{1}+\boldsymbol{CV}^{\mathbf{2}}\right)}\right) = \sqrt{\mathbf{ln}\left(\mathbf{1}+\boldsymbol{CV}^{\mathbf{2}}\right)}$$

But by the expression for $\boldsymbol{q}_{\mathbf{2}}$,

$$\boldsymbol{q}_{\mathbf{2}} = \sqrt{\mathbf{ln}\left(\mathbf{1}+\boldsymbol{CV}^{\mathbf{2}}\right)},$$

so the minimum possible value of z is seen to be $\boldsymbol{q}_{\mathbf{2}}$, and this completes the proof and establishes the theorem.

As an example of the application of Theorem 2, suppose that two lognormal distributions have a common mean but different standard deviations. The points of intersection for numerous pairs of coefficients of variation for the two distributions are shown in Table 3.

Note that in the table, the percentile at which the two distributions intersect ranges from approximately the 55th to over the 68th percentile. Also, note that the greater the difference between the two standard deviations (since the means are the same, the ratio of coefficients of variation is the same as the ratio of the standard deviations), the higher the percentile at which the two distributions intersect. This is counter to intuition. For example, this situation could arise when applying correlation to a risk analysis. Correlation does not affect the mean, but it has a significant impact on the standard deviation (Book, 1999; Smart, 2002). A higher correlation value, on average, translates to higher risk, which should mean that more money should be required to achieve the 60th percentile. But if, for example, the coefficient of variation of the first distribution is 10% while the second is 55%, the second standard deviation is 5.5 times the first. However, the two distributions intersect at the 62nd percentile, so less funding is needed to achieve the 60th percentile for the riskier cost distribution. Figure 6 provides an illustration of this concept.

Theorem 2 established a lower bound on the intersection (the z-score must be at least $\boldsymbol{q}_{\mathbf{2}}$). A further question can be asked based on the information in Table 3: Is there any upper limit at which two lognormals can intersect? The next theorem (Theorem 3) shows that, given a lognormal distribution and any percentile greater than the lower bound, there is another lognormal distribution with the same mean that intersects the original lognormal at the specified percentile. The conclusion is that two lognormal distributions can intersect at any percentile between the 50th and the 100th. Theorem 4 will establish points of intersection between a normal and a lognormal.

Theorem 3. *Given a lognormal distribution and any percentile **k** between the lower bound (corresponding to the lower limit of z) that is established in Theorem 2 and the 100th percentile, there is another lognormal distribution with the same mean that intersects that distribution at the **k**th percentile.*

Proof. At one point near the end of the proof of Theorem 2, it was seen that

$$z = \mathbf{0.5}\left(\sqrt{\mathbf{ln(1+}\boldsymbol{\alpha}^{\mathbf{2}}\boldsymbol{CV}^{\mathbf{2}}\mathbf{)}} + \sqrt{\mathbf{ln(1+}\boldsymbol{CV}^{\mathbf{2}}\mathbf{)}}\right).$$

TABLE 3 Percentiles at which two lognormal distributions intersect

Coefficient of variation	Coefficient of variation of first lognormal								
of second lognormal	10.0%	15.0%	20.0%	25.0%	30.0%	35.0%	40.0%	45.0%	50.0%
15.0%	55.0%								
20.0%	55.9%	56.9%							
25.0%	56.9%	57.8%	58.8%						
30.0%	57.8%	58.8%	59.7%	60.6%					
35.0%	58.7%	59.7%	60.6%	61.5%	62.4%				
40.0%	59.6%	60.5%	61.5%	62.4%	63.3%	64.2%			
45.0%	60.4%	61.4%	62.3%	63.2%	64.1%	65.0%	65.8%		
50.0%	61.3%	62.2%	63.1%	64.0%	64.9%	65.8%	66.6%	67.4%	
55.0%	62.1%	63.0%	63.9%	64.8%	65.7%	66.5%	67.4%	68.1%	68.9%

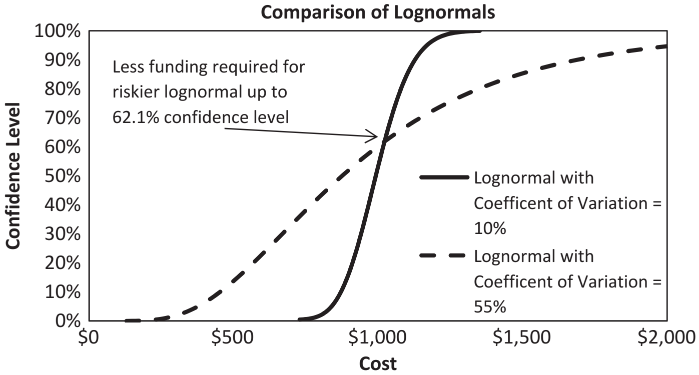

FIGURE 6 Comparison of funding levels for two lognormals.

Solving for $\boldsymbol{\alpha}$, it is found successively that

$$\sqrt{\ln(1+\alpha^2 CV^2)} = 2z - \sqrt{\ln(1+CV^2)},$$

$$\ln(1+\alpha^2 CV^2) = \left(2z - \sqrt{\ln(1+CV^2)}\right)^2,$$

$$1+\alpha^2 CV^2 = \exp\left(2z - \sqrt{\ln(1+CV^2)}\right),$$

from which it can be seen that

$$\alpha = \sqrt{\frac{e^{\left(2z-\sqrt{\ln(1+CV^2)}\right)^2} - 1}{CV^2}}.$$

Now the coefficient of variation must be set sufficiently high relative to the baseline lognormal (#2) to find another lognormal that intersects it at percentile $\boldsymbol{k}$. This is done by calculating $\boldsymbol{\alpha}$, which is the ratio of the coefficients of variation of other lognormals relative to the baseline lognormal for the z-score that corresponds to the percentile at which the two distributions will intersect. That value of $\boldsymbol{\alpha}$ defines lognormal (#1), which has the associated normal distribution z-score for percentile $\boldsymbol{k}$. Furthermore, as $\boldsymbol{\alpha}$ increases to ∞, the equation above shows that $\boldsymbol{z}$ also increases to ∞, and that, in turn, implies that $\boldsymbol{k}$ increases to 100%. This completes the proof and hence establishes the theorem.

Thus, given a lognormal distribution, there is another lognormal distribution with a greater standard deviation that intersects the first lognormal at the 80th percentile, another at the 90th percentile, another at the 95th, the 99th, the 99.99th, etc. That is, given two lognormals with the same mean, there is no upper bound for the intersection between them. For example, suppose that a lognormal distribution has coefficient of variation equal to 20% and the 75th percentile is selected. Then the z-score is the inverse of a standard normal distribution (calculated in Excel using the formula "=NORMSINV(0.75)"), which is approximately 0.67449. This z-score and the coefficient of variation of 20% are used as inputs in the formula derived in the proof of Theorem 3; that is,

$$\boldsymbol{\alpha} = \sqrt{\frac{\boldsymbol{e}^{\left(\mathbf{2(0.67449)}-\sqrt{\mathbf{ln}\left(\mathbf{1+(0.20)^2}\right)}\right)^2} - \mathbf{1}}{\mathbf{(0.20)^2}}},$$

which results in $\boldsymbol{\alpha} \approx \mathbf{4.369}$.

For the baseline lognormal, the coefficient of variation is fixed. Then from Theorem 3, as $\boldsymbol{\alpha}$ increases, the z-score increases. As the value of $\boldsymbol{\alpha}$ increases, so does the standard deviation of the other lognormal, and thus, the risk for the other lognormal increases. But $\boldsymbol{z}$ also increases, which means that the riskier the lognormal becomes, the higher the percentile at which the two distributions intersect. This process continues with no limit: the higher the risk, the higher the point of intersection.

By definition, the normal and lognormal distributions have the same mean. For a normal distribution, the mean and 50th percentiles are the same value. However, for a lognormal distribution, the 50th percentile is always less than or equal to the mean. Thus, at the 50th percentile, the lognormal requires less funding than a normal distribution with the same mean and standard deviation. Thus, when funding to the 50th percentile, for a given mean and standard deviation, measuring risk with a lognormal distribution requires fewer risk reserves than when measuring risk with a normal distribution, even though the lognormal tail is fatter and hence is riskier.

However, funding policies typically focus on percentiles greater than the 50th, such as the 70th or 80th percentiles. However, it turns out that another peculiar feature of the lognormal distribution is that for percentiles less than the 85th, the normal distribution percentiles exceed those of a lognormal with the same mean and standard deviation. This means that percentile funding policies at these levels consider a normal distribution to be riskier than a lognormal.

This issue does not impact estimating at the mean, which is not the same as percentile funding; see Figure 7 for an example.

Theorem 4. *For a lognormal distribution and a normal distribution with common mean and standard deviation, the maximum point at which the distributions intersect is at least the 84% confidence level.*

Proof. At the 50th percentile, the normal mean is equal to the normal 50th percentile, and it is given that this is equal to the lognormal mean. Because the lognormal's 50th percentile is less than its mean, as was discussed earlier, if funding is set at the 50th percentile, a lognormal risk measure will require less funding than the normal with the same mean. Since the lognormal has a fatter tail than the normal distribution, this means that the normal and lognormal distributions must intersect at some point above the 50th percentile of the normal distribution. From Theorem 3, there is no bound to the upper limit for the intersection, it is only known that it will be at some point less than the 100th percentile. Since lognormal and normal distributions are unbounded above, the theoretical 100th percentile is not a finite number.

The probability level of a normal distribution one standard deviation above the mean is approximately 84.1%, and here, the normal z-score is equal to 1. Let $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ denote the mean and standard deviation of the normal distribution, respectively. Let $\boldsymbol{p}$ and $\boldsymbol{q}$ denote the mean and standard deviation, respectively, of the normal distribution associated with the lognormal. Note that the normal distribution that is associated with the

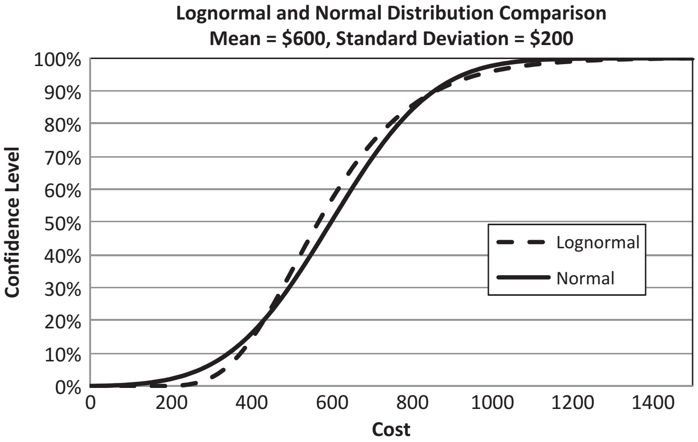

FIGURE 7 Comparison of lognormal and normal distribution percentiles with common mean = \$600 and common standard deviation = \$200.

lognormal is not the same distribution as the normal distribution with mean μ and standard deviation σ.

The mean of the lognormal distribution is therefore

$$e^{p+0.5q^2},$$

and the standard deviation is

$$e^{p+0.5q^2}\sqrt{e^{q^2}-1}\,.$$

So, at the 84.1th percentile of the normal distribution, the normal distribution percentile is $\mu+\sigma$, and the lognormal distribution percentile is e^{p+q}.

This is thus comparing $\mu+\sigma$ with e^{p+q}. Since the mean of the lognormal distribution is the same as the mean of the normal distribution, $e^{p+0.5q^2}$ can be substituted for μ and $e^{p+0.5q^2}\sqrt{e^{q^2}-1}$ for σ in the term $\mu+\sigma$, and this can be compared with e^{p+q}:

$$\mu+\sigma = e^{p+0.5q^2} + e^{p+0.5q^2}\sqrt{e^{q^2}-1} = e^{p+0.5q^2}\left(1+\sqrt{e^{q^2}-1}\right).$$

Setting this equal to e^{p+q} gives

$$e^{p+0.5q^2}\left(1+\sqrt{e^{q^2}-1}\right) = e^{p+q},$$

which simplifies to

$$e^{0.5q^2}\left(1+\sqrt{e^{q^2}-1}\right) = e^{q}.$$

Note that one solution to this transcendental equation is $q=0$, since $e^0=1$, although it is an extraneous solution because it forces $\sigma=0$, which is not possible for such distributions. In fact, $q>0$ follows from the fact that q is itself the standard deviation of a normal distribution.

Also note that

$$e^{q^2} = 1 - 1 + e^{q^2} \le 1 + 2\sqrt{e^{q^2} - 1} + \left(e^{q^2} - 1\right) = \left(1 + \sqrt{e^{\,q^2} - 1}\right)^2,$$

which implies that

$$1 + \sqrt{e^{\,q^2} - 1} \ge \sqrt{e^{\,q^2}} = e^{0.5q^2}.$$

When $q > 1$,

$$e^{0.5q^2}\left(1 + \sqrt{e^{q^2} - 1}\right) \ge e^{0.5q^2} e^{0.5q^2} = e^{q^2} > e^{q},$$

and when $q = 1$, this inequality holds also because $e^{q} = e = 2.718$ and $e^{0.5q^2}\left(1 + \sqrt{e^{q^2} - 1}\right) = \sqrt{e}\left(1 + \sqrt{e - 1}\right) = 3.810$. Thus equality is not attained when $q \geq 1$, and so the normal percentile one standard deviation greater than the mean exceeds the lognormal percentile (the same) one standard deviation greater than the (same) mean.

When $q < 1$, the inequality

$$e^{0.5q^2}\left(1 + \sqrt{e^{q^2} - 1}\right) > e^{q}$$

is difficult to verify algebraically or even by using calculus. Numerical routines such as Mathematica and Wolfram Alpha (http://www.wolframalpha.com) or direct calculations made in Excel can be used to verify that equality only holds at $q = 0$ and that the difference increases as q increases. The difference can be expressed as the continuous function $f(q) = e^{0.5q^2}(1 + \sqrt{e^{q2} - 1}) - e^{q}$, and so the fact that the inequality holds when $q \ge 1$ and the only root is at 0 imply that the inequality must also hold for $q \le 1$. Figure 8 shows a graph of $f(q)$, namely of the differences between the normal and lognormal distribution percentiles at one standard deviation above the mean.

Because of the paradox, care should be taken in applying lognormal distributions when reserve levels are set at the 80th percentile or below. In these cases, a normal distribution will provide more logically consistent results and will better aid decision makers in making logical choices in establishing budgets and reserve levels. Also, when reserve levels and

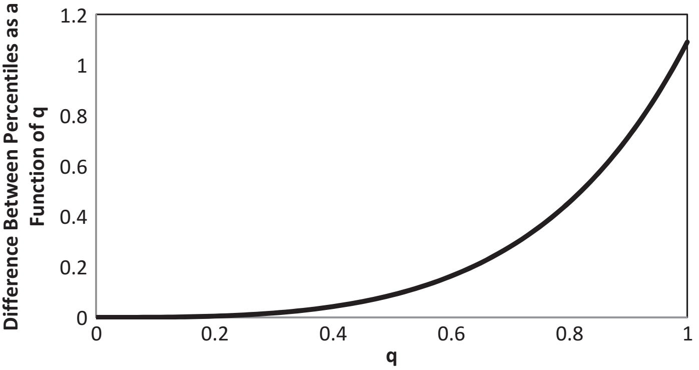

FIGURE 8 Graphical depiction of differences between normal and lognormal percentiles.

budgets are set below the 85th percentile, the normal distribution is more conservative than the lognormal distribution.

Summary

Numerous natural and economic phenomena have been shown to have the property of scale invariance. Cost growth and cost risk can be added to this extensive list. Also, normal distributions were shown to significantly underestimate the likelihood of extreme cost growth. While the portfolio effect for cost risk is intuitively appealing, the data indicate that this phenomenon is at best minimal and should not be relied upon by policy makers to reduce risk. Instead, the focus of confidence levels should be placed on individual missions.

When considering individual missions, funding levels are often set at confidence levels near the central portion of the distribution, such as the 60th or 70th percentiles. In this case, fat-tailed distributions like the Pareto are not the best choice for modeling cost risk. When funding levels are set at such levels, the lognormal also proves to be problematic. Oddly enough, in such cases, the normal distribution proves to be the most logically consistent and conservative choice. However, the triumph of the normal distribution in such a case indicates that by setting funding levels at the 60th or 70th percentile may be a fool’s game, since this seemingly contradicts the empirical evidence provided in the first part of the article that the normal distribution does not do a good job of modeling the higher percentiles and that there is little, if any, portfolio effect. Thus, setting funding levels at the 60th or even the 70th percentile for individual missions is not likely to provide sufficient confidence levels at the agency level to provide a comfortable margin for future cost growth.

References

Anderson, T. P. (2004, June 22–24). The Trouble With Budgeting to the 80th Percentile. In *72nd Military Operations Research Society Symposium*, Monterey, CA.

Book, S. A. (1999). Why Correlation Matters in Cost Estimating. In Advanced Training Session, *32nd Annual DOD Cost Analysis Symposium*, Williamsburg, VA.

Doyle, S. (2010, January 13). Final bill for Madison County jail is in: $79,811,693.31. *The Huntsville Times*. Retrieved from http://blog.al.com/breaking/2010/01/final_bill_for_madison_county.html

Flyvbjerg, B., Holm, M. S., & Buhl, S. (2002). Underestimating Costs in Public Works Projects: Error or lie? *Journal of the American Planning Association*, 68(3), 279–295.

Flyvbjerg, B. (2005). *Policy and Planning for Large Infrastructure Projects: Problems, Causes, Cures* (pp. 4–5). Washington, D.C.: World Bank Publications.

Gray, J. (2004). *Janos Bolyai, Non-Euclidean Geometry, and the Nature of Space*. Cambridge, MA: Burndy Library Publications.

Gray, J. (2004). *Janos Bolyai, Non-Euclidean Geometry, and the Nature of Space.* Cambridge, MA: Burndy Library Publications.

Kiss, E. (1999). *Mathematical Gems from the Bolyai Chests: Janos Bolyai’s Discoveries in Number Theory and Algebra*. Budapest: TypoTEXLtd. Electronic Publishing Co.

Mandlebrot, B. B. (1983). *The Fractal Geometry of Nature*. New York: W.H. Freeman and Company.

Mandlebrot, B. B. (1997). *Fractals and Scaling in Finance: Discontinuity, Concentration, Risk*. New York: Springer Verlag.

Markowitz, H. (1959). *Portfolio Selection: Efficient Diversification of Investments*, Yale, CT: Wiley, Yale University Press.

Schaffer, M. (2003). NASA Cost Growth: A Look at Recent Performance. Unpublished Powerpoint presentation, NASA HQ, Washington, D.C.

Smart, C. (2002). Predicting Differences between Estimated and Actual Cost and Schedule. Unpublished white paper.

Smart, C. (2007, July 17–19). Cost and Schedule Interrelationships. Presented at the *2007 NASA Cost Symposium*, Denver, CO.

Taleb, N. N. (2007). *The Black Swan: The Impact of the Highly Improbable*. New York: Random House.

United States Government Accountability Office. (1992). Space Missions Require Substantially More Funding than Initially Estimated (GAO-NSIAS-93-97). Washington, DC: U.S. GAO.

About the Author

Christian Smart, Ph.D., is the Director for Cost Estimating at the Missile Defense Agency. In this capacity, he is responsible for overseeing all cost estimating activities developed and produced by the agency, and directs the work of a 100-person team. In 2010, he received an Exceptional Public Service Medal from NASA for his contributions to the Ares I Joint Cost Schedule Confidence Level Analysis and his support for the Human Space Flight Review Panel led by Norm Augustine. In 2009, he was awarded the Parametrician of the Year award by the International Society of Parametric Analysts. He has won several best paper awards at the annual International Society of Parametric Analysts and the Society of Cost Estimating and Analysis annual conferences. One of these conference papers was the basis for this journal article.