

1. Introduction

Part A of the "Superior Semiconductors" case introduced a simple decision problem with one unknown quantity that had finitely many possible values. In Part B, the decision analyst is confronted a broader scope of uncertainty, involving several unknown quantities, some of which could have infinitely many possible values. In this chapter, we will develop methods for analyzing such complex situations.

In this example, we learn that the cost of developing a new product, which had been roughly estimated at \$26 million, might be significantly more or less than that amount. When we recognize the full scope of uncertainty about this dollar cost, we must admit that the actual development cost could be virtually any positive number. To this huge set of possible values, we need to assign probabilities that describe our information and beliefs about this unknown quantity.

The way to make this probability-assessment task tractable is to assume that our beliefs about the unknown quantity can be reasonably approximated by some probability distribution in one of the mathematical families of probability distributions that have been studied by statisticians. Each of these families includes a variety of different probability distributions that are characterized by a few parameters. So our task is, first to pick an appropriate family of probability distributions, and then to assess values for the corresponding parameters that can fit our actual beliefs about the unknown quantity. The most commonly used families of probability distributions typically have only two or three parameters to assess. So this parameter-assessment problem may actually be easier than the problem of assessing a discrete probability distribution that we saw in Part A of the "Superior Semiconductors" case (where the decision analyst had to generate five probability numbers that sum to one).

This chapter introduces some of the best-known families of probability distributions, each of which is considered particularly appropriate for certain special kinds of unknown quantities. But in general practice, the Normal and Lognormal distributions are widely recognized as applicable in the broadest range of situations. So in this chapter, we will give most careful consideration to these two families, as well as the Generalized-Lognormal family which includes the Normal and Lognormal families as special cases.

2. Normal distributions

The central limit theorem (which was introduced in the previous chapter) tells us that, if we make a large number of independent samples from virtually any probability distribution, our sample average will have a probability distribution that is approximately Normal. This remarkable theorem assures us that there will be many unknowns in the real world that are well described by Normal distributions. In fact, a Normal probability distribution will be a good fit for any unknown quantity that can be expressed as the sum of many small contributions that are independently drawn from some fixed probability distribution. For example, if our uncertainty about each day's revenue is the same as any other day and is independent of all other days' revenues, and each day's revenue will be small relative to the whole year, then total annual revenue should be Normal random variable. (Here the phrase Normal random variable is used to denote any unknown quantity that has a Normal probability distribution.)

The Normal probability distributions are a two-parameter family that are indexed by the mean and standard deviation. That is, for any number μ and any positive number σ , there is a Normal probability distribution that has mean μ and standard deviation σ .

A linear transformation of any Normal random variable is also Normal. That is, if a random variable \mathbf{X} has a Normal probability distribution with mean μ and standard deviation σ , and if c and d are any two nonrandom numbers, then the random variable $c*\mathbf{X}+d$ also has a Normal probability distribution, with mean $c*\mu+d$ and standard deviation $|c|*\sigma$.

Every number (positive or negative) is a possible value of a Normal random variable. Of course there are infinitely many numbers, and so no single number gets positive probability under a Normal distribution. Instead, we can only talk meaningfully about intervals of numbers having positive probability. In the language of probability theory, such distributions that do not put positive probability on any single value are called continuous.

Excel provides several functions for dealing with Normal distributions. Here we will use two: NORMINV and NORMSDIST. Of these, the function NORMINV will be much more important in this course, but I should explain both of them now.

	A	B	C	D	E	F	G	H
1	Mean	Stdev	Normal random variable			19.9	20.1	
2	26	4.5		22.93948		P(F1 < D2 < G1)		
3						0.00729		
4	NORMAL DISTRIBUTION				FORMULAS			
5	CumvProby	Value	ProbyDensity		D2. =NORMINV(RAND(),A2,B2)			
6	0.001	12.0939		F3. =NORMSDIST((G1-A2)/B2)-NORMSDIST((F1-A2)/B2)				
7	0.01	15.5315	0.00407		B6. =NORMINV(A6,\$A\$2,\$B\$2)			
8	0.02	16.7581	0.00998		B6 copied to B6:B106			
9	0.03	17.5364	0.01467		C7. =(A8-A6)/(B8-B6)			
10	0.04	18.1219	0.01884		C7 copied to C7:C105			
11	0.05	18.5982	0.02269		A6 contains 0.001			
12	0.06	19.0035	0.02629		A7:A105 is filled by a series from .01 to .99, step .01			
13	0.07	19.3589	0.02969		A106 contains 0.999			
14	0.08	19.6772	0.03291					
15	0.09	19.9666	0.03598	Inverse Cumulative Chart plots (A6:A106,B6:B106)				
16	0.1	20.233	0.03891					
17	0.11	20.4806	0.04171					
18	0.12	20.7126	0.04438					
19	0.13	20.9312	0.04695					
20	0.14	21.1386	0.04941					
21	0.15	21.3361	0.05176					
22	0.16	21.5249	0.05402					
23	0.17	21.7063	0.05619					
24	0.18	21.8809	0.05827					
25	0.19	22.0495	0.06027					
26	0.2	22.2127	0.06218	Probability Density Chart plots (B7:B105,C7:C105)				
27	0.21	22.3711	0.06401					
28	0.22	22.5251	0.06577					
29	0.23	22.6752	0.06745					
30	0.24	22.8216	0.06906					
31	0.25	22.9648	0.07059					
32	0.26	23.1049	0.07206					
33	0.27	23.2423	0.07346					
34	0.28	23.3772	0.07479					
35	0.29	23.5098	0.07605					
36	0.3	23.6402	0.07725					
37	0.31	23.7687	0.07838					
38	0.32	23.8954	0.07946					
39	0.33	24.0204	0.08046					
40	0.34	24.1439	0.08141					
41	0.35	24.2661	0.0823					
42	0.36	24.3869	0.08313					
43	0.37	24.5067	0.08389					
44	0.38	24.6253	0.0846					
45	0.39	24.7431	0.08525					
46	0.4	24.8599	0.08584					
47	0.41	24.976	0.08638					

Figure 1. The Normal distribution.

The cumulative probability below any value in a Normal distribution depends on how many standard deviations this value is above or below the mean, according to the Excel function NORMSDIST. That is, if \mathbf{X} is any Normal random variable with mean μ and standard deviation σ , and w is any number, then

$$P(\mathbf{X} < w) = \text{NORMSDIST}((w - \mu) / \sigma)$$

(Notice, there are two S's in this function's name.) If we want to know the probability that \mathbf{X} , a Normal random variable with mean μ and standard deviation σ , is between some pair of numbers w and y , where $w < y$, then we can use the formula

$$\begin{aligned} P(w < \mathbf{X} < y) &= P(\mathbf{X} < y) - P(\mathbf{X} < w) \\ &= \text{NORMSDIST}((y - \mu) / \sigma) - \text{NORMSDIST}((w - \mu) / \sigma) \end{aligned}$$

as illustrated by cell F3 in Figure 1.

The Normal distribution has a symmetry around its mean so that, for any number w ,

$$\text{NORMSDIST}(-w) = 1 - \text{NORMSDIST}(w)$$

That is, a Normal random variable with mean μ and standard deviation σ is as likely to be less than $\mu - w * \sigma$ as it is to be greater than $\mu + w * \sigma$.

Some special values of the NORMSDIST function worth noting are

$\text{NORMSDIST}(0) = 0.5$, $\text{NORMSDIST}(0.675) = 0.75$, $\text{NORMSDIST}(1) = 0.841$,
 $\text{NORMSDIST}(1.96) = 0.975$, $\text{NORMSDIST}(3) = 0.99865$. These cumulative probabilities imply that, when \mathbf{X} is any Normal random variable with mean μ and standard deviation σ ,

$$P(\mathbf{X} < \mu) = P(\mathbf{X} > \mu) = 0.5$$

$$P(\mathbf{X} < \mu - 0.675 * \sigma) = 0.25$$

$$P(\mathbf{X} < \mu + 0.675 * \sigma) = 0.75$$

$$P(\mu - \sigma < \mathbf{X} < \mu + \sigma) = 0.683$$

$$P(\mu - 1.96 * \sigma < \mathbf{X} < \mu + 1.96 * \sigma) = 0.95$$

$$P(\mu - 3 * \sigma < \mathbf{X} < \mu + 3 * \sigma) = 0.9973$$

The Excel function NORMINV is the inverse cumulative-probability function for Normal probability distributions. That is, if \mathbf{X} is any Normal random variable with mean μ and standard deviation σ , then

$$P(\mathbf{X} < \text{NORMINV}(q, \mu, \sigma)) = q$$

Here the NORMINV parameters q , μ , and σ can be any three numbers such that $\sigma > 0$ and $0 < q < 1$. Using NORMINV, we can make an inverse cumulative distribution chart by plotting probability numbers from 0 to 1 against the corresponding NORMINV(probability, μ , σ) values, as shown in Figure 1 (in rows 16-30) for the values $\mu=26$ and $\sigma=4.5$.

Recall from chapter 2 that we can simulate a random variable with any probability distribution by applying a RAND() value as input into the inverse cumulative-probability function. Thus, we can simulate a Normal random variable with mean μ and standard deviation σ by the formula

$$\text{NORMINV}(\text{RAND}(), \mu, \sigma)$$

as shown in cell D2 of Figure 1 (where the mean μ is in cell A2 and the standard deviation σ is in cell B2).

I have tried to argue that the inverse cumulative probability curve shows us everything that we could want to know about the probability distribution of a random variable. But people often prefer to look at a different way of representing probability distributions graphically, called the probability density curve. We will not need to use probability density functions for any calculations in this course, but I have to admit that they are handy for drawing pictures of probability distributions that many people like to look at. The probability density function provides a useful way to visualize a random variable's probability distribution because, for any interval of possible values, the probability of the interval is the area under the probability density curve over this interval. So in case you want to look at these pictures too, I will explain how to make them and what they mean. Excel provides a rather complicated formula for computing Normal probability-density functions [NORMDIST(\bullet , μ , σ , FALSE)]. But to emphasize the importance of the inverse-cumulative function NORMINV, and to help you better understand what probability densities are, I will instead show you here how to estimate such probability densities with NORMINV.

A probability density function for a random variable at any possible value is actually defined mathematically to be the slope of the cumulative probability function at that value. So when a mathematician says that a random variable \mathbf{X} has a probability density $f(y)$ at some value

y, this statement means that, for any small interval around y, the probability of **X** being in that interval is approximately equal to the length of the interval multiplied by the density f(y).

According to Figure 1, for example, a Normal random variable with mean 26 and standard deviation 4.5 has, at the value 20, a probability density of approximately 0.036. So we may infer that the probability of such a random variable being between 19.9 and 20.1 (an interval of length 0.2 that includes the value 20) is approximately $0.2 \times 0.036 = 0.0072$.

To make a chart of probability densities for a given distribution, let us begin by filling a range of cells in column A with probability values that go from 0 to 1. For the spreadsheet of Figure 1, such probability values from 0 to 1 with step size 0.01 are filled into the range A6:A106, except that the values 0 and 1 at the top and bottom of this range are changed to 0.001 and 0.999, because NORMINV and many other inverse cumulative functions are not defined at the probabilities 0 and 1. Next, in column B, let us compute the corresponding inverse-cumulative values. So for the spreadsheet in Figure 1, with the mean in cell B2 and the standard deviation in cell C2, we enter in cell B6 the formula

$$=NORMINV(A6, \$A\$2, \$B\$2)$$

and we copy B6 to cells B6:B106. Then the probability density can be estimated in column C by entering into cell C7 the formula

$$=(A8 - A6) / (B8 - B6)$$

and then copying cell C7 to the range C7:C105. (This formula in the C column is not applied in the top and bottom rows 6 and 106, because it needs to refer to A and B values in the rows above and below the current row.) In this C7 formula, the denominator B8-B6 is the length of an interval from a value (12.09 in B6) that is slightly less than B7 to a value (16.76 in B8) that is slightly greater than B7 (which is 15.53). The numerator A8-A6 is the probability of the random variable being in this short interval, because A8 (0.02) is the probability of the random variable being less than B8 (16.76), while A6 (0.001) is the probability of the random variable being less than B6 (12.09). So $(A8 - A6) / (B8 - B6)$ is the probability in a short interval around B7 divided by the length of this interval, and so it can be used to approximate the probability density of this random variable at B7.

The probability density chart shown in Figure 1 (in rows 33 to 47) is plotted from the

values in B7:B105 and the estimated densities in C7:C105. The characteristic shape of this Normal density chart is often called a bell-shaped curve. The probability density function provides a useful way to visualize a random variable's probability distribution because, for any interval of possible values, the probability of the interval is the area under the probability density curve over this interval.

Normal random variables can take positive or negative values, and their probability of going negative can be non-negligible when the mean is less than twice the standard deviation. But for there are many unknown quantities, such as prices, where a negative value would be nonsensical. For such unknowns, it is more appropriate to use a probability distribution in which only the positive numbers are possible values. The statistical literature features several families of such distributions, of which the most widely used in finance and decision analysis is the family of Lognormal probability distributions. But before introducing Lognormal random variables, we briefly turn aside to introduce two important mathematical functions: EXP and LN.

3. EXP and LN

The function EXP is called the exponential function. If you enter the formula =EXP(1) into any cell, you will get a magic number 2.71828... which mathematicians often denote by the symbol "e". For any other number r, the formula EXP(r) returns the value of e raised to the power r, that is EXP(r) equals 2.71828^r . (In Figure 2, see cells B6, D6, and E6.)

The function LN is called the natural logarithm function, and it is the inverse of EXP. That is, for any positive number x, the formula LN(x) returns the number r such that EXP(r) = x. (In Figure 2, see cells B3, B6, and B7.)

You may have studied these functions EXP and LN in mathematics classes, but let me introduce them here with an basic financial application that may help you to see what is so special about the powers of this magic number e. Suppose that you put money into a bank account that offers an annual interest rate of some nominal interest rate r that will be compounded k times during the year (where k is some integer). Then k times per year, the bank will pay you r/k dollars for every dollar in your account (including past interest payments as well as your original investment). That is, k times during the year, the value of your account will be multiplied by

$1+r/k$, and so every dollar that you initially invested will yield at the end of the year $(1+r/k)^k$ dollars. For example, Figure 2 shows that, when the nominal interest rate r is 0.12 (in cell B3) and the times compounded is $k = 4$ (in cell B4), then each dollar initially invested will grow to $(1+r/k)^k = 1.1255$ dollars (in cell B5) at the end of the year. This 1.1255 ratio of the end-of-year value divided by the beginning-of-year value for the account may be called its annual growth ratio.

	A	B	C	D	E	F
1	THE EXP FUNCTION FROM COMPOUND INTEREST:					
2	Suppose we start with \$1 in a bank account.					
3	Nominal interest rate	0.12				
4	Times compounded/year	4	<i>(try 1,2,4,12,365)</i>			
5	\$Value at end of year	1.1255088	The number e:	e^r		
6	EXponential formula	1.1274969	2.7182818	1.1274969		
7	LN, the inverse of EXP	0.12				
8						
9	So EXP(r) is the actual yield, per \$1 initial investment, after a year					
10	when nominal interest rate r is compounded continuously.					
11	LN(x) is the nominal interest rate that would yield \$ x after a year,					
12	per \$1 initial investment, with continuous compounding.					
13						
14	AN EXAMPLE TO SHOW THE ADVANTAGE OF LOGARITHMIC GROWTH RATES					
15		Values:	Growth rates:			
16	Year 0	100	Percent	Logarithmic		
17	Year 1	80	-20.00%	-22.31%		
18	Year 2	100	25.00%	22.31%		
19	Average growth rate:		2.50%	0.00%		
20						
21	FORMULAS FROM RANGE A1:D19					
22	B5. = $(1+B3/B4)^{B4}$	C17. = $(B17-B16)/B16$				
23	B6. =EXP(B3)	C18. = $(B18-B17)/B17$				
24	B7. =LN(B6)	D17. =LN(B17/B16)				
25	D6. =EXP(1)	D18. =LN(B18/B17)				
26	E6. = $D6^{B3}$	C19. =AVERAGE(C17:C18)				
27		D19. =AVERAGE(D17:D18)				

Figure 2. Compound interest and growth rate calculations with EXP and LN.

Now consider what happens when the same nominal interest rate r is compounded more times per year. (Try $k=12$ or $k=365$ in cell B4 of Figure 2, to represent monthly or daily compounding.) The result is that the annual growth ratio will increase towards the limiting value of EXP(r) as k becomes large. That is, EXP(r) is the annual growth ratio (the end-of-year value

divided by the beginning-of-year value) for an account that pays the nominal interest rate r but is compounded a large number of times during the year.

Conversely, for any annual growth ratio x , the formula $\text{LN}(x)$ is the nominal annual interest rate which, if compounded many times during the year, would yield x dollars at the end of the year for every dollar invested at the beginning of the year. We may call $\text{LN}(x)$ the logarithmic growth rate that corresponds to the growth ratio x .

If an investment that was worth w dollars at the beginning of a year becomes worth y dollars at the end of the year, then people would traditionally say that its growth rate during the year has been $(y-w)/w$. But its logarithmic growth rate is the natural logarithm of the growth ratio, that is $\text{LN}(y/w)$. When the growth ratio y/w is close to 1, these logarithmic growth rates are close to the traditional growth rates.

Let me give you an example to show the advantage of using logarithmic growth rates instead of traditional growth rates in sophisticated quantitative analysis. Consider a two-year investment that was worth \$100 initially, was worth \$80 at the end of the first year, and was worth \$100 again at the end of the second year. In traditional terms, the growth rate in the first year was $(80-100)/100 = -0.20$, and the growth rate in the second year was $(100-80)/80 = 0.25$. So the average growth rate for these two years seems to be $(-0.20 + 0.25)/2 = 0.025$, which gives us a positive average growth rate during a period when the value did not grow at all! The logarithmic growth rates in this example are $\text{LN}(80/100) = -0.2231$ in the first year and $\text{LN}(100/80) = 0.2231$ in the second year, which correctly yields an average growth rate of $(-0.2231+0.2231)/2 = 0$ over the two-year period.

There are two other formulas that you should know about EXP and LN:

$$\text{EXP}(x+y) = \text{EXP}(x) * \text{EXP}(y)$$

$$\text{LN}(r*s) = \text{LN}(r) + \text{LN}(s)$$

(Here x and y can be any two numbers, and r and s can be any two positive numbers.) The first of these formulas tells us that EXP converts addition into multiplication, while the second tells us that LN converts multiplication into addition. These formulas are illustrated in Figure 3 by the equality between cells K24 and L24, and by the equality between cells K26 and L26.

	H	I	J	K	L
16	Other facts about EXP and LN:				
17	EXP(r1+r2) = EXP(r1)*EXP(r2)				
18	LN(x1*x2) = LN(x1)+LN(x2)				
19					
20	Example:				
21	r (random)	s (random)			
22	-0.8441	1.1442			
23	x = EXP(r)	y = EXP(s)		EXP(r+s)	EXP(r)*EXP(s)
24	0.4299	3.1401		1.3500	1.3500
25	LN(x)	LN(y)		LN(x*y)	LN(x)+LN(y)
26	-0.8441	1.1442		0.3001	0.3001
27					
28	FORMULAS FROM RANGE H21:L26				
29	H22. =NORMINV(RAND(),0,1)				
30	I22. =NORMINV(RAND(),0,1)				
31	H24. =EXP(H22)		H26. =LN(H24)		
32	I24. =EXP(I22)		I26. =LN(I24)		
33	K24. =EXP(H22+I22)		K26. =LN(H24*I24)		
34	L24. =H24*I24		L26. =H26+I26		

Figure 3. The multiplication and addition properties of EXP and LN.

4. Lognormal distributions

By definition, a random variable Y is a Lognormal random variable if its natural logarithm $LN(Y)$ is a Normal random variable. Because the EXP function is the inverse of LN, a Lognormal random variable can be equivalently defined as any random variable that can be computed by applying the EXP function to a Normal random variable. Thus, for any number m and any positive number s , the formula

$$=EXP(NORMINV(RAND(),m,s))$$

is a Lognormal random variable. The parameter m and s are called the log-mean and log-standard-deviation of the random variable. The log-mean and the log-standard-deviation are not the mean and standard deviation of this Lognormal random variable, instead they are the mean and standard deviation of its natural logarithm. In general, the Lognormal random variables have a family of probability distributions that can be parameterized by their log-mean and log-standard-deviation. If you multiply a Lognormal random variable with log-mean m and

log-standard-deviation by a (nonrandom) positive constant c , the result is another Lognormal random variable that has log-mean $m+\text{LN}(c)$ and the same log-standard-deviation s .

Lognormal random variables are often used in finance to describe the growth ratio of an asset's value over some period of time. To see why, let \mathbf{Y} denote the growth ratio for the value of some financial asset over the coming year. That is, \mathbf{Y} is the unknown value of this asset a year from today divided by the asset's value today. For each integer i from 1 to 52, let \mathbf{X}_i denote the growth ratio of this asset during the i 'th week of the coming year. So

$$\mathbf{Y} = \mathbf{X}_1 * \mathbf{X}_2 * \dots * \mathbf{X}_{52}.$$

Taking the natural logarithm of both sides, we get

$$\text{LN}(\mathbf{Y}) = \text{LN}(\mathbf{X}_1 * \mathbf{X}_2 * \dots * \mathbf{X}_{52}) = \text{LN}(\mathbf{X}_1) + \text{LN}(\mathbf{X}_2) + \dots \text{LN}(\mathbf{X}_{52})$$

because LN converts multiplication to addition. Now suppose that the growth ratio each week is drawn independently from the same probability distribution. Suppose also that there cannot be a large change in any one week, that is, the growth ratio \mathbf{X}_i in any one week must be close to 1. Then $\text{LN}(\mathbf{X}_1), \dots, \text{LN}(\mathbf{X}_{52})$ will be independent random variables, each drawn from the same distribution on values close to $\text{LN}(1) = 0$. But recall from the beginning of Section 2 that a Normal probability distribution will be a good fit for any unknown quantity that can be expressed as the sum of many small (near 0) contributions that are independently drawn from a fixed probability distribution. So $\text{LN}(\mathbf{Y}) = \text{LN}(\mathbf{X}_1) + \text{LN}(\mathbf{X}_2) + \dots \text{LN}(\mathbf{X}_{52})$ must be an approximately Normal random variable, which implies that the growth ratio \mathbf{Y} itself must be an approximately Lognormal random variable. Furthermore, multiplying this growth ratio \mathbf{Y} by the known current value of this asset, we find that this asset's value a year from today must also be an approximately Lognormal random variable.

	A	B	C	D	E	F	G
1	FACT: The multiplicative product of many independent identically						
2	distributed near-1 random variables is approximately Lognormal.						
3	EXAMPLE: Weekly growth ratios are uniform between 0.95 and 1.05						
4		Value					
5	Weekly ratio	1		Growth ratio for 52 weeks			
6	0.9733	0.9733		1.2329			
7	1.0240	0.9967					
8	1.0077	1.0044	(rand)	Lognormal random variables			
9	1.0292	1.0337	0.2804	EXP(NORMINV)	LNORMINV	GENLINV	
10	0.9998	1.0335		0.8669	0.8668	0.8668	
11	1.0386	1.0734	FORMULAS				
12	0.9925	1.0654	A6. =0.95+0.1*RAND()				
13	0.9896	1.0544	B6. =B5*A6				
14	1.0016	1.0560	A6:B6 copied to A6:A57				
15	1.0366	1.0947	D6. =B57				
16	1.0286	1.1260	C9. =RAND()				
17	1.0102	1.1375	E10. =EXP(NORMINV(C9,-0.02168,0.2084))				
18	0.9717	1.1054	F10. =LNORMINV(C9,1,0.2107)				
19	0.9789	1.0820	G10. =GENLINV(C9,0.8502,0.9786,1.1262)				
20	1.0498	1.1360					
21	0.9699	1.1017	Line chart plots B5:B57.				
22	0.9738	1.0729					
23	0.9609	1.0309					
24	0.9963	1.0272					
25	1.0255	1.0533					
26	0.9583	1.0094					
27	1.0217	1.0313					
28	1.0131	1.0448					
29	1.0014	1.0463					
30	0.9816	1.0270					
31	1.0372	1.0652					
32	0.9876	1.0519					
33	0.9609	1.0108					
34	1.0301	1.0412					
35	1.0169	1.0589					
36	0.9907	1.0491					

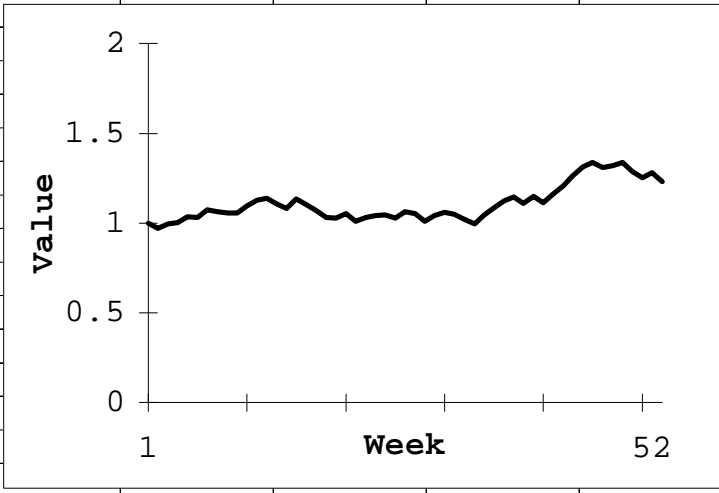


Figure 4. Multiplying random variables to get approximate Lognormal growth.

This result is illustrated in Figure 4, which illustrates a financial asset whose value can change by up to 5% on in any given week. To be specific, this model assumes that the growth ratio of the assets value is drawn from a Uniform probability distribution over the interval from 0.95 to 1.05. These growth ratios for the 52 weeks of the year are simulated by entering the

formula $=0.95+0.10*\text{RAND}()$ into each cell in the range A6:A57. Supposing that the initial value of this asset is 1, cell B5 is given the value 1, the formula $=B5*A6$ is entered into cell B6, and then cell B6 is copied to B6:B57. The result is that cells B6:B57 simulate the value of this asset at the end of each of the 52 weeks of the year. The growth ratio for the whole 52-week year (which is also the value at the end of the year if the initial value was 1) is echoed in cell D6 by the formula $=B57$. Because this annual growth ratio is the multiplicative product of many independent random ratios, each of which is close to 1, we know that it must be an approximately Lognormal random variable.

Knowing that the annual growth ratio in cell D6 of Figure 4 is a Lognormal random variable, we may now ask which Lognormal distribution is the probability distribution for this random variable. Recall that Lognormal probability distributions can be parameterized by their log-mean and log-standard-deviation, which are respectively the mean and standard deviation of the natural logarithm of the random variable. To estimate these parameters, we can compute the sample mean and standard deviation of many independent recalculations of LN(D6), by entering the formula $=\text{LN}(D6)$ into a cell that serves as the model output at the top of a large simulation table (not shown in the figure). In this way (using a few thousand simulations), one can show that the log-mean is approximately -0.02 and the log-standard-deviation is approximately 0.21 . That is, the formula

$$=\text{EXP}(\text{NORMINV}(\text{RAND}(), -0.02, 0.21))$$

should yield a random variable that has approximately the same probability distribution as the random annual growth ratio in cell D6. Such a formula has been entered into cell E10 of Figure 4.

Of course the realized value of cells D6 and E10 are very different in Figure 4, because they are independent random variables. To show that they have the same probability distribution, we should make a simulation table that records several hundred recalculations of D6 and E10 (not shown in the figure). Then these simulated values of D6 and E10 should be sorted separately and plotted against the percentile index, to yield a chart that estimates the inverse cumulative probability distributions of these two random variables. These estimated inverse cumulative curves for D6 and E10 should be very close to each other.

In some situations, you may prefer to parameterize a Lognormal random variable by its own expected value and standard deviation, instead of by the expected value and standard deviation of its natural logarithm. Unfortunately, the formulas for computing the mean μ and the standard deviation σ of a Lognormal random variable from its log-mean m and log-standard-deviation s are rather complicated:

$$\mu = \text{EXP}(m + 0.5*s^2), \quad \sigma = ((\text{EXP}(s^2) - 1)^{0.5})*\mu$$

So to let you avoid such formulas, Simtools provides a function called LNORMINV which is the inverse cumulative-probability function for Lognormal random variables parameterized by the mean and standard deviation. That is, when the mean μ and the standard deviation σ are derived from the log-mean m and the log-standard-deviation s according to the (complicated) two formulas shown above, then for any cumulative probability q between 0 and 1, the formula LNORMINV(q, μ, σ) will yield the same value as EXP(NORMINV(q, m, s)). Thus a Lognormal random variable that has expected value μ and standard deviation σ can be simulated in Excel with Simtools by the formula

$$=\text{LNORMINV}(\text{RAND}(), \mu, \sigma)$$

Such a formula has been entered into cell F10 in Figure 4, using the same RAND() as the EXP(NORMINV) formula in cell E10. (The slight difference between the values of cells E10 and F10 is due to roundoff error in computing μ and σ from m and s .)

We have argued in this section that the growth ratio of a financial investment over some period of time might be naturally assumed to be a Lognormal random variable. We have noted that this assumption then implies that the value of the investment as any particular date in the future would also be a Lognormal random variable (because the future value is the known current value multiplied by the Lognormal growth ratio between now and that future date). But this assumption also implies that logarithmic rates of return over any interval of time are Normal random variables, because the natural logarithm of a Lognormal random variable is a Normal random variable. To remember this distinction, notice that growth rates can be positive or negative, and Normal random variables can be positive or negative. But Lognormal random variables are never negative, and a financial asset's value at any point in time cannot be negative, and its growth ratio over any interval of time also cannot be negative.

5. Generalized Lognormal distributions

Now I want to introduce a three-parameter family of continuous probability distributions that may be called the Generalized-Lognormal family. These Generalized-Lognormal distributions are defined to include the probability distributions of all random variables of the form $c*\mathbf{X}+d$ where \mathbf{X} is a Normal or Lognormal random variable and c and d are nonrandom constants. So these Generalized-Lognormal distributions include the Normals and Lognormals as special cases.

Any Generalized-Lognormal probability distribution can be characterized by specifying its value at the 0.25, 0.50, and 0.75 cumulative-probability levels, which we may call the quartile boundary points for the distribution, denoted by Q_1 , Q_2 , and Q_3 respectively. That is, we may say that Q_1 , Q_2 , and Q_3 are the quartile boundary points for an unknown quantity \mathbf{X} if

$$P(\mathbf{X} < Q_1) = 0.25, P(\mathbf{X} < Q_2) = 0.50, \text{ and } P(\mathbf{X} < Q_3) = 0.75.$$

These three boundary points divide the number line into four separate quartiles which are equally likely to include the actual value of \mathbf{X} . The first quartile is the set of all numbers less than Q_1 , the second quartile is interval between Q_1 and Q_2 , the third quartile is the interval between Q_2 and Q_3 , and the fourth quartile is the set of numbers greater than Q_3 . Assuming that \mathbf{X} is a continuous random variable, each of these quartiles has the same probability 1/4 of including the actual value of \mathbf{X} . The second quartile boundary point Q_2 is also called the median of the distribution of \mathbf{X} .

Simtools provides a function called GENLINV which is the inverse cumulative probability function for a Generalized-Lognormal probability distribution parameterized by its quartile boundary points. That is, if Q_1 , Q_2 , and Q_3 are any three numbers such that $Q_1 < Q_2 < Q_3$, then the formula

$$=\text{GENLINV}(\text{RAND}(),Q_1,Q_2,Q_3)$$

returns a Generalized-Lognormal random variable that has quartile boundary points Q_1 , Q_2 , and Q_3 .

In the special case where the quartile boundary points satisfy the following equal-difference condition

$$Q_3 - Q_2 = Q_2 - Q_1,$$

this Generalized-Lognormal random variable is a Normal random variable with mean $\mu = Q_2$ and

standard deviation $\sigma = (Q_3 - Q_1)/1.349$. On the other hand, in the special case where the quartile boundary points satisfy the following equal-ratio condition

$$Q_3/Q_2 = Q_2/Q_1,$$

the Generalized-Lognormal random variable is a Lognormal random variable with log-mean $m = \text{LN}(Q_2)$ and log-standard-deviation $s = (\text{LN}(Q_3) - \text{LN}(Q_1))/1.349$. Thus the GENLINV function gives us another way of constructing Normal or Lognormal random variables, based on their quartiles. An example is shown in cell G10 of Figure 4, with quartiles that satisfy the equal-ratio condition and match the log-mean and log-standard deviation in cell E10.

For people who more technical details about GENLINV, I should explain how one could equivalently compute these Generalized-Lognormal random variables without Simtools.xla. First, given the quartile boundary points Q_1 , Q_2 , and Q_3 , compute the ratio of quartile widths

$$\beta = (Q_3 - Q_2)/(Q_2 - Q_1).$$

If β is different from 1, then $\text{GENLINV}(\text{RAND}(), Q_1, Q_2, Q_3)$ is equivalent to the formula

$$Q_2 + (Q_3 - Q_2) * (\beta^{\text{NORMINV}(\text{RAND}(), 0, 1/0.67449) - 1}) / (\beta - 1)$$

If β is equal to one, which happens when $Q_3 - Q_2 = Q_2 - Q_1$, then $\text{GENLINV}(\text{RAND}(), Q_1, Q_2, Q_3)$ is instead equivalent to the formula

$$\text{NORMINV}(\text{RAND}(), Q_2, (Q_3 - Q_2)/0.67449)$$

Beyond the cumulative-probability parameter and the three quartile boundary points, the GENLINV function also optionally can be given a fifth and a sixth parameter. These parameters impose bounds on the possible values of GENLINV. If you want to make sure that the value of GENLINV never goes below some lowest value L, then you can include this number L as the optional fifth parameter. The value of the formula $\text{GENLINV}(p, Q_1, Q_2, Q_3, L)$ is the same as $\text{GENLINV}(p, Q_1, Q_2, Q_3)$ unless it would be less than L, in which case the value is rounded up to L. For example, $\text{GENLINV}(\text{RAND}(), 2, 5, 9)$ is negative with probability 0.103, while $\text{GENLINV}(\text{RAND}(), 2, 5, 9, 0)$ is never negative but equals 0 with probability 0.103. Similarly, if you want to make sure that the value of GENLINV never goes above some highest value H, then you can include this number H as the optional sixth parameter. The value of the formula $\text{GENLINV}(p, Q_1, Q_2, Q_3, L, H)$ is the same as $\text{GENLINV}(p, Q_1, Q_2, Q_3, L)$ unless it would be greater than H, in which case the value is rounded down to H.

6. Subjective probability assessment

There are some unknown quantities for which finding comparable data may be relatively straightforward. For example, if we want to predict the rate of return next year for a stock of an old established company, we might naturally assume that next year's annual rate of return will be another independent draw out of the same probability distribution that generated the past 30 years' annual rates of return for this stock. In such cases, we can collect data about past realizations of this quantity, and we can use our sample mean and standard deviation to estimate the mean and standard deviation (or log-mean and log-standard-deviation) of an appropriate Normal or Lognormal distribution for our future unknown quantity.

But we may also be confronted by unknown quantities that have a more unique quality, for which there may not be any obviously comparable data of past realizations. For example, if we want to estimate the potential demand for a particular new product like the one in the "Superior Semiconductors" case, we might be very uncomfortable with a naive assumption that this new product will look like a random draw from the previous 30 new products that we have introduced. Such an assumption would ignore all the unique technical characteristics of this new product and its market which should be shaping our beliefs in this area.

So let us now consider the problem of assessing a probability distribution for an unknown quantity where analysis of data about similar quantities seems to be out of the question. Even where there is no data, if we have uncertainty about an unknown quantity, then we should be able to describe our uncertainty by a probability distribution. To make this point forcefully, let me take up a specific example to which you might have thought that probability modeling could not even be applied: the question of when was Napoleon Bonaparte born.

So let N denote the year when Napoleon Bonaparte was born. We can make this a continuous random variable by allowing fractions, so that $N = 1066.5$ would mean that Napoleon was born in the middle of the year 1066 AD. If we do not know when Napoleon was born, then N is certainly an unknown quantity. But how can we define its probability distribution?

A probability distribution is a description of uncertainty and beliefs, and different people may have different beliefs about an unknown quantity such as N . A French professor of history might feel sure about the year in which Napoleon was born, while an American professor of

management might be unsure about the century in which he was born. To describe the different beliefs of different people, we must admit that they may subjectively assess different probability distributions for the same unknown. So we should not simply speak about "the" probability distribution of N . Instead, we should speak of a particular individual's subjective probability distribution for the unknown quantity N , as assessed at a particular point in time with reference to the information available to this individual at this point in time. This subjective probability distribution is thus defined to be a quantitative description of what this individual believed about this unknown quantity at this point in time.

When we set out to measure your subjective probability distribution for N , given your current information, you might feel a sense of reluctance if you feel disqualified by relative ignorance about French history. In some circumstances, an individual should be reluctant to express his or her beliefs where others have better information. In the "Superior Semiconductors," the decision analyst and the executive decision-maker rightly defer to the business-marketing manager's beliefs about the potential demand for the new T-regulator device. So you might say that your beliefs about this unknown quantity should be the beliefs of the best expert that you can consult. And if the best expert that you can consult is an encyclopedia that has the exact date of Napoleon's birth, then of course you should check this encyclopedia and assign subject probability 1 to the date that it gives you. But (to prepare ourselves for practical problems like demand forecasting where the encyclopedia cannot help us) let us suppose that you have no access to an encyclopedia or French history professor. Instead let us suppose that your best available expert is me with the information and beliefs that I had at the time that I first sat to write this essay on subjective-probability assessment. In fact, I have taken this exercise seriously (resisting the temptation to open that forbidden encyclopedia) and have assessed my subjective probability distribution for this unknown quantity. As an illustration of subjective probability in practice, let me now describe how I assessed this subjective probability distribution.

The first thing that I began to do was to recall some of the things that I knew about Napoleon from high school. I knew that the French revolution began in 1789, but I did not know what Napoleon was doing at that time. I knew that Napoleon invaded Russia in 1812, which I could remember because of the name of the famous 1812 Overture (by some famous composer).

I believed that Napoleon died in the 1820s or 1830s, but it might have been a relatively young death from something like cancer. I felt quite sure that he had risen to be the most powerful man in France by 1799 (that date stuck in my mind for some reason), and I felt pretty sure that he had been remarkably young for such a position of power at that time. I recalled someone saying that revolutions have often been times of great opportunity for ambitious young men like Napoleon. So I knew that he had been born sometime in the 1700s, and I drew a time line for this century with intervals marked to represent the decade years 1710, 1720, through 1790. Then I started thinking about each of these years as possible birth-years for Napoleon, starting in the middle, at 1750. I soon found myself calculating, for each of these years, how old Napoleon would have been in late 1799. It just seemed easier to think about these ages-in-1799 than to think about the birth-years themselves.

You might have supposed that the first questions that I would undertake would be to assess an expected value and a standard deviation for my probability distribution, where the mean is supposed to be a measure of the center of my subjective probability distribution, and the standard deviation is supposed to be a measure of the amount of uncertainty around this center. But we have seen that expected values and standard deviations are rather sophisticated mathematical concepts that are not so easy to understand. Even I, a professional statistics teacher, would have trouble judging whether my subjective standard deviation of N should be (say) 4 years or 7 years. It is simpler (and thus better) to think first about assessing the probabilities of events.

I might try, for example, to assess my subjective probability of N being less than 1750 (that is, of Napoleon being older than 49 at the end of 1799). My subjective probability of the event " $N < 1750$ " could be defined as the number p such that I would be indifferent between the following two alternatives, if they were offered to me uninformatively:

- (1) a lottery ticket that pays \$1000 if $N < 1750$, and pays \$0 otherwise,
- (2) a lottery ticket that pays \$1000 if a particular `RAND()` in a spreadsheet is less than p after its next recalculation, and pays \$0 otherwise.

By being "offered uninformatively," I mean that these lotteries are not supposed to be available only because someone who knows more than me wants to see me lose (or win), so that getting

such a choice should not change my beliefs about N . I could say that I would prefer alternative (1) above when p is $1/2$ or $1/10$, but I would prefer alternative (2) when p is $1/100$, and so my subjectively assessed $P(N < 1750)$ must be somewhere between $1/100$ and $1/10$. As I think about p decreasing below $1/10$, alternative (2) becomes worse until somewhere it ceases to be better than alternative (1) for me. But to assess that point of indifference that defines $P(N < 1750)$, I may need to consider some probability numbers for which I have relatively little intuitive feeling. Is $p=0.05437$ a number that makes lottery (2) just as good as lottery (1), or might it be $p=0.06819$? Even a trained statistician might have little or no confidence in his ability to answer such questions subjectively.

So it is better to think instead about events that have simple probabilities like $1/2$ or $1/4$. Thus, subjective probability assessment in practice often begins with an assessment of the three quartile boundary points that (for an unknown quantity with a continuous distribution) divide the number line into four equally likely intervals.

Thus, the first question that I actually asked myself was: "For what number Q_2 would I be indifferent between (1a) a bet that pays me \$1000 if $N < Q_2$ and pays me \$0 otherwise, and (2a) a bet that pays me \$1000 if $N \geq Q_2$ and pays me \$0 otherwise?" After trying several numbers, I settled on $Q_2=1764$, because I felt that Napoleon was equally likely to have been more or less than 36 years old in late 1799. (Actually, with $Q_2 = 1764$, if I imagined I already had either of these two bets and that I was approached by someone with an offer to exchange it for the other bet, then I found that I always wanted to stay with the bet that I already had, but the intensity of my desire to reject the "new" offer was equally intense, whichever side I imagined myself having first. So I decided that I really was indifferent.) Thus my subjective probabilities should satisfy the equations,

$$P(N < 1764) = P(N \geq 1764) = 1/2.$$

Next, I asked myself, for what number Q_1 would I be indifferent between (1b) a bet that pays me \$1000 if $N < Q_1$, and (2b) a bet that pays me \$1000 if $Q_1 \leq N < 1764$? After going back and forth for a while, I finally decided that I would be indifferent for $x = 1758$. That is, I figured that the events " $N < 1758$ " (aged at least 42 in late 1799) and " $1758 \leq N < 1764$ " (aged at least 36 but not yet 42 in late 1799) were equally likely, in my opinion. Since the union of

these two equally likely and mutually exclusive events was " $N < 1764$ " which I had just assigned subjective probability $1/2$, I could conclude

$$P(N < 1758) = 1/4.$$

Next, I asked myself, for what number Q_3 would I be indifferent between (1c) betting on the event " $1764 \leq N < Q_3$ " and (2c) betting on the event " $N \geq Q_3$ "? After more introspection, I decided that I would be indifferent for $Q_3 = 1767$. That is, I figured that the events " $1764 \leq N < 1767$ " (aged at least 33 but not yet 36 in late 1799) and " $N \geq 1767$ " (younger than 33 in late 1799) were equally likely, in my opinion. So I now had two more mutually exclusive and equally likely events whose union (" $N \geq 1764$ ") had probability $1/2$, and so I could conclude that $P(N \geq 1767) = P(1764 \leq N < 1767) = 1/4$. Thus, I wrote that

$$P(N < 1767) = 1 - P(N \geq 1767) = 3/4.$$

To check my judgments, I asked myself whether I would be indifferent between (1d) betting on the event " $1758 \leq N < 1767$ " (aged at least 33 but not yet 42 in late 1799) and (2d) betting on the event " $N > 1767$ or $N \leq 1758$ " (younger than 33 or more than 42 years old in late 1799)? This was a hard question, but I decided that the answer was probably yes, so I felt comfortable with my assessments. If I had felt that " $1758 \leq N < 1767$ " was less likely than " $N > 1767$ or $N \leq 1758$ ", then I would have reconsidered my earlier answers, to see whether I might have underestimated Q_3 or overestimated Q_1 in the preceding two paragraphs.

This process of assessing equally likely intervals had thus yielded three points on my (inverse) cumulative probability curve for this unknown N :

$$P(N < 1758) = 0.25, \quad P(N < 1764) = 0.50, \quad P(N < 1767) = 0.75$$

I could have continued splitting intervals to get values with cumulative probability 0.125, 0.375, etc., but instead I tried a simpler tactic. I just assumed that my beliefs about this unknown quantity can be represented by a Generalized-Lognormal distribution with my subjective assessed quartile boundary points $Q_1 = 1758$, $Q_2 = 1764$, and $Q_3 = 1767$, and I used the GENLINV function with these parameters to see what the rest of the inverse-cumulative distribution would look like. But this assumption needed to be checked by comparing some of the other probabilities that the distribution would predict with my subjective beliefs.

The shape of a Generalized-Lognormal may be symmetric around its median or it may be

skewed, depending on the ratio of the widths of the two inner quartiles $Q_3 - Q_2$ and $Q_2 - Q_1$. In general, the possible values of a Generalized-Lognormal distribution are bounded on the side of the shorter inner quartile, while the possible values extend indefinitely on the other side where the wider inner quartile expressed more uncertainty. With my assessed quartile boundaries, the ratio of quartile widths is $(1767 - 1764) / (1764 - 1758) = 0.5$, which expresses much more uncertainty on the low side than on the high side. So the Generalized Lognormal distribution is very skewed distribution, with the region of substantial probability extending much farther on the low side than on the high side. In this case the extreme values that have a probability 1/100 below or above them are

$$\text{GENLINV}(0.01, 1758, 1764, 1767) = 1704,$$

$$\text{GENLINV}(0.99, 1758, 1764, 1767) = 1769.$$

Notice that the distance from low value with cumulative probability 0.01 (1704) to the lower quartile boundary point (1758) is much farther than the distance from the higher quartile boundary point (1769) to the high value with cumulative probability 0.99 (1769).

Looking at such low and high values in the probability distribution can be very important for deciding whether the estimated Generalized-Lognormal distribution is a good fit for the actual subjective probability distribution that we are trying to assess. So I asked myself, did I really believe that the probability of Napoleon being born before 1704 (aged 96 or more in late 1799) was 1/100? And did I really believe that the probability of Napoleon being born in 1769 or later (younger than 31 in late 1799) was also 1/100? An event that has probability 1/100 should be considered extremely unlikely, but not completely out of the range of reasonable possibilities. By these standards, I felt strongly that " $N < 1704$ " was not even in the range of reasonable possibilities, and so $P(N < 1704)$ should be substantially less than 1/100. I also felt that, in my subjective probability distribution, $P(N \geq 1769)$ was substantially larger than 1/100. So one possible conclusion was that my true beliefs about N could not be fit into the Generalized-Lognormal family of distributions. But rather than give up on such a nice functional form, it was worthwhile ask instead whether I might have been somewhat mistaken about my subjective quartile boundary points. After all, it was not so easy to decide whether those intervals were equally likely.

So I thought next about what values with subjective probability 1/100 below or above would have seemed reasonable for $\text{GENLINV}(0.01, Q_1, Q_2, Q_3)$ and $\text{GENLINV}(0.99, Q_1, Q_2, Q_3)$. After some introspection about the low end of the likely values of \mathbf{N} , I estimated that the probability of \mathbf{N} being less than 1745 (55 or older in late 1799) should be about 1/100. At the high end of the distribution, I estimated that the probability of \mathbf{N} being greater than 1776 (not yet 24 in late 1799) should also be about 1/100.

Next I thought about small adjustments to my assessed quartiles. I felt more confident about the median value Q_2 that I had assessed, and so I focused on changing Q_1 and Q_3 . The problem seemed to be that my first estimated distribution was much too skewed, and so I tried increasing Q_1 and Q_3 each by one, which brought the ratio of inner quartile widths up to

$$(Q_3 - Q_2) / (Q_2 - Q_1) = (1768 - 1764) / (1764 - 1759) = 0.8$$

The 0.01 and 0.99 cumulative-probability values with these parameters became

$$\text{GENLINV}(0.01, 1759, 1764, 1768) = 1741,$$

$$\text{GENLINV}(0.99, 1759, 1764, 1768) = 1775,$$

which seemed compatible with my beliefs about these low and high ends of the distribution.

Revisiting the questions that I originally used to assess the quartiles, I felt that these new estimates were not clearly less plausible than my original estimates, so I was willing to write

$$P(\mathbf{N} < 1759) = P(1759 \leq \mathbf{N} < 1764) = P(1764 < \mathbf{N} \leq 1768) = P(1768 < \mathbf{N}) = 1/4$$

Thus, I concluded that the Generalized-Lognormal distribution with quartile boundaries at 1759, 1764, and 1768, as shown in Figure 5, was a reasonable description of my beliefs about Napoleon's birth year, given the information that I then had available. That is, I might have said that Napoleon's birth year was as likely to be in any given interval as the random variable $\text{GENLINV}(\text{RAND}(), 1759, 1764, 1768)$.

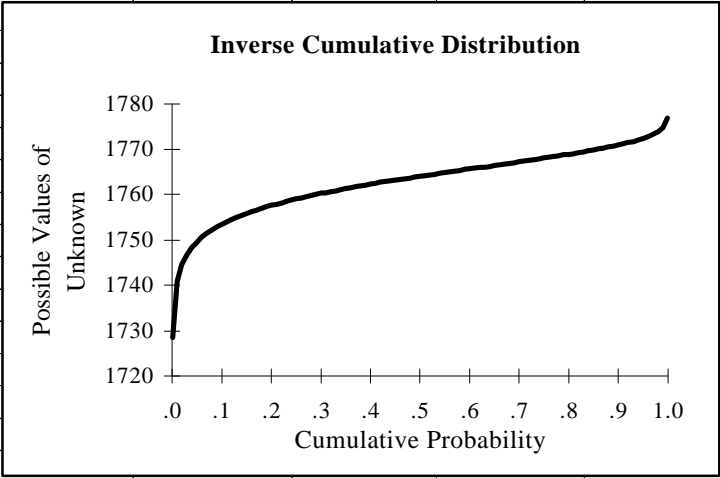
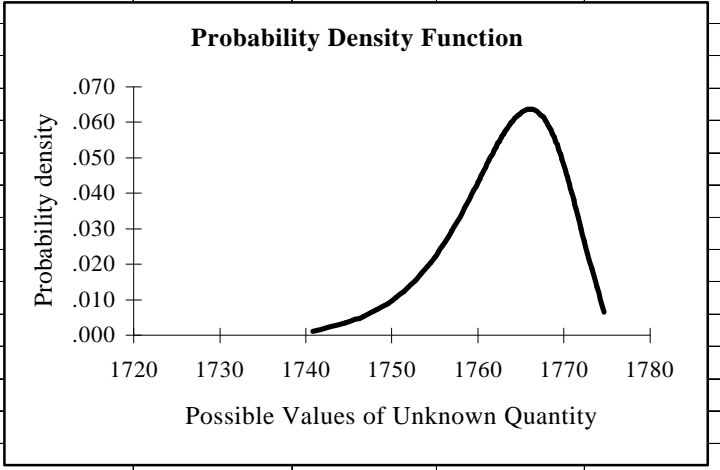
	A	B	C	D	E	F	G	H
1		Quartiles of Unknown Quantity				Ratio of quartile widths		
2		1759	1764	1768		0.8		
3	1%-point	25%point	50%point	75%point	99%-point		Simulated value	
4	1740.82				1774.74		1770.60	
5								
6	GENERALIZED-LOGNORMAL DISTRIBUTION							
7	CumvProby	Value	ProbyDensity		FORMULAS			
8	0.001	1728.41			A4. =GENLINV(0.01,\$B\$2,\$C\$2,\$D\$2)			
9	0.01	1740.82	0.0012		E4. =GENLINV(0.99,\$B\$2,\$C\$2,\$D\$2)			
10	0.02	1744.54	0.0034		G4. =GENLINV(RAND(),\$B\$2,\$C\$2,\$D\$2)			
11	0.03	1746.74	0.0053		F2. =(D2-C2)/(C2-B2)			
12	0.04	1748.31	0.0071		B8. =GENLINV(A8,\$B\$2,\$C\$2,\$D\$2)			
13	0.05	1749.54	0.0089		B8 copied to B8:B108			
14	0.06	1750.55	0.0107		C9. =(A10-A8)/(B10-B8)			
15	0.07	1751.41	0.0124		C9 copied to C9:C107			
16	0.08	1752.16	0.0140					
17	0.09	1752.83	0.0157	Inverse Cumulative Chart plots (A8:A108,B8:B108)				
18	0.1	1753.44	0.0173					
19	0.11	1753.99	0.0189					
20	0.12	1754.50	0.0205					
21	0.13	1754.97	0.0220					
22	0.14	1755.41	0.0235					
23	0.15	1755.82	0.0250					
24	0.16	1756.21	0.0264					
25	0.17	1756.58	0.0279					
26	0.18	1756.93	0.0293					
27	0.19	1757.26	0.0307					
28	0.2	1757.58	0.0320					
29	0.21	1757.88	0.0333					
30	0.22	1758.18	0.0346					
31	0.23	1758.46	0.0359					
32	0.24	1758.74	0.0372					
33	0.25	1759.00	0.0384					
34	0.26	1759.26	0.0396	Probability Density Chart plots (B9:B107,C9:C107)				
35	0.27	1759.51	0.0408					
36	0.28	1759.75	0.0419					
37	0.29	1759.98	0.0431					
38	0.3	1760.21	0.0442					
39	0.31	1760.43	0.0452					
40	0.32	1760.65	0.0463					
41	0.33	1760.87	0.0473					
42	0.34	1761.08	0.0483					
43	0.35	1761.28	0.0493					
44	0.36	1761.48	0.0502					
45	0.37	1761.68	0.0511					
46	0.38	1761.87	0.0520					
47	0.39	1762.06	0.0529					
48	0.4	1762.25	0.0537					
49	0.41	1762.44	0.0545					

Figure 5. Subjectively assessed quartiles and a Generalized-Lognormal distribution.

7. A decision problem with discrete and continuous unknowns

In part B of the "Superior Semiconductor" case, the decision analyst (Eastmann) needed to assess quantitatively the uncertainty that her colleagues had about the development cost and potential market value for the proposed new product. For each of these unknown quantities, we may understand that she asked the best available expert some version of the quartile-assessment questions that were described in the previous section. (In summary, she would asking the expert to specify three point that would divide the number line into four regions such that the expert could not say that any of these regions was more likely to contain the unknown quantity than any other.) Then the Generalized-Lognormal distribution would be a convenient way to extrapolate these assessed quartiles to a full probability distribution. (To check the fit of this distribution, she might have asked the experts to think about all-but-1/100-probability lower and upper bounds for the unknown quantity, and she would compare these bounds to the 0.01 and 0.99 cumulative-probability values returned by the GENLINV function.)

The quartile boundary points for development cost assessed by the chief production engineer are 23, 26, and 29 \$million, which happen to satisfy the equal-difference condition that characterizes the Normal distributions among the Generalized-Lognormals ($29 - 26 = 26 - 23$). For a Normal distribution, the quartile boundary points Q_1 , Q_2 , Q_3 depend on the mean μ and standard deviation σ by the formulas

$$Q_1 = \mu - 0.675 * \sigma, \quad Q_2 = \mu, \quad Q_3 = \mu + 0.675 * \sigma.$$

So with these assessed quartiles, the mean and standard deviation of the development cost must be $\mu = Q_2 = 26$ and $\sigma = (Q_3 - Q_2) / 0.675 = (29 - 26) / 0.675 = 4.444$. Thus, the development cost in this case could be simulated equally well by the formula $\text{NORMINV}(\text{RAND}(), 26, 4.444)$ or the formula $\text{GENLINV}(\text{RAND}(), 23, 26, 29)$.

The quartile boundary points for total market value assessed by the business-marketing manager are 80, 100, and 125 \$million, which happen to satisfy the equal-ratio condition that characterizes the Lognormal distributions among the Generalized-Lognormals ($125/100 = 100/80$). For a Lognormal distribution with log-mean m and log-standard-deviation s , the quartile boundary points Q_1 , Q_2 , Q_3 depend on the log-mean and log-standard-deviation by the

formulas

$$\text{LN}(Q_1) = m - 0.6745*s, \text{LN}(Q_2) = m, \text{LN}(Q_3) = m + 0.6745*s.$$

So with these assessed quartiles, the log-mean and log-standard-deviation of the total market value must be $m = \text{LN}(Q_2) = \text{LN}(100) = 4.605$ and $s = (\text{LN}(Q_3) - \text{LN}(Q_2)) / 0.6745 = (\text{LN}(125) - \text{LN}(100)) / 0.6745 = 0.3308$. (These log-mean and log-standard-deviation numbers are very hard to interpret!) Thus, the total market value in this case could be simulated equally well by the formula $\text{EXP}(\text{NORMINV}(\text{RAND}(), 4.605, 0.3308))$ or the formula $\text{GENLINV}(\text{RAND}(), 80, 100, 125)$.

Assuming that these unknown quantities are independent, the resulting simulation model is shown in Figure 6. The subjectively assessed probabilities and quartile points are listed in the top 13 rows of the spreadsheet. Then the simulation model is developed in rows 15-20. The development cost is simulated in cell A16 by the formula

$$=\text{GENLINV}(\text{RAND}(), \text{A3}, \text{B3}, \text{C3})$$

using the assessed quartile points for the development cost that are listed in cells A3:C3. The success or failure of the product development is simulated in cell A17 by the formula

$$=\text{IF}(\text{RAND}() < \text{F3}, 1, 0)$$

where the value 1 denotes success, the value 0 denotes failure, and cell F3 contains the assessed probability of success (0.95). The total value of the market for this new product is simulated in cell A18 by the formula

$$=\text{A17} * \text{GENLINV}(\text{RAND}(), \text{A6}, \text{B6}, \text{C6})$$

using the assessed quartile points for the market value that are listed in cells A6:C6. In cell A18, the Generalized-Lognormal random variable is multiplied by A17, because the market value will actually be zero if the new product is not developed. (If A18 is 1 then multiplying by it makes no difference, but if A18 is 0 then multiplying by it makes the total market value 0.) The number of competitors is simulated in cell A19 by the formula

$$=\text{DISCRINV}(\text{RAND}(), \text{A9:A13}, \text{B9:B13})$$

using the possible values listed in A9:A13 with corresponding probabilities in B9:B13. Finally, the bottom line for Superior Semiconductors is their profit, computed in cell B20 by the formula

$$=\text{A18} / (1 + \text{A19}) - \text{A16}$$

That is, profit equals total market value divided by the number of firms in the market, minus development costs.

The use of simulation to estimate expected profit may have seemed unnecessary when we analyzed part A of this case in Chapter 2, because we then had only one discrete random variable, and we could easily compute the expected value directly. But now part B, we have four different random variables, some of which have infinitely many possible values. In examples like this, we really need simulation to estimate the expected profit, because other direct methods of calculation may be difficult or impossible.

So the simulated profit was echoed in cell B27 (by the formula =A20), and a simulation table contains the results of 801 independent recalculations of this simulated profit below in the range B28:B828. The sample mean and standard deviation from this data is used in B23 and B24 to estimate the expected value and standard deviation of profit. These calculations suggest that we could still recommend this new product under the expected value criterion, because the estimated expected profit is a positive number (1.454 \$million). Analysis of more simulation data might give a different and more accurate estimate of the expected profit, but the 95% confidence interval computed in cells E25:F25 makes us quite confident that the expected profit is bigger than zero (because the lower bound in E25 is positive). A previous analysis using a sample size of 401 yielded a 95% confidence interval that included negative values, and so the sample size was increased to the 801 shown here, to be able to confidently decide this question.

So in the expected-value analysis for part B seem quite similar to part A in Chapter 2, where uncertainties about development cost, development success, and market value were all ignored. But the measures of risk for part B look very different from part A. To begin with, the standard deviation of profit is reported now to be about 16.4 \$million, instead of the 9.3 \$million which was computed for part A.

	A	B	C	D	E	F	G	H
1	"SUPERIOR SEMICONDUCTOR" CASE (B)							
2	Development-cost quartile points:					P(Successful development)		
3	23	26	29			0.95		
4								
5	Market-value quartile points (if successful):							
6	80	100	125					
7								
8	k	P(#competitors=k)						
9	1	0.10			FORMULAS FROM RANGE A1:F27			
10	2	0.25			A16. =GENLINV(RAND(),A3,B3,C3)			
11	3	0.30			A17. =IF(RAND(<F3,1,0)			
12	4	0.25			A18. =A17*GENLINV(RAND(),A6,B6,C6)			
13	5	0.10			A19. =DISCRINV(RAND(),A9:A13,B9:B13)			
14					A20. =A18/(1+A19)-A16			
15	Simulation model					B27. =A20		
16	28.287	Development cost				B23. =AVERAGE(B28:B828)		
17	1	Successful development?				B24. =STDEV(B28:B828)		
18	83.165	Total market value				E23. =COUNT(B28:B828)		
19	3	Number of competitors				E25. =B23-1.96*B24/(E23^0.5)		
20	-7.496	Profit				F25. =B23+1.96*B24/(E23^0.5)		
21								
22	Analysis of simulation data				Sample size			
23		1.454	E(Profit)		801			
24		16.433	Stdev(Profit)		95% confidence interval for E(Profit)			
25					0.316	2.592		
26		Profit (\$millions)						
27	SimTable	-7.496						
28	0	-35.1034						
29	0.00125	-33.0313						
30	0.0025	-32.2422						
31	0.00375	-32.1554						
32	0.005	-31.6148						
33	0.00625	-31.5656						
34	0.0075	-31.3032						
35	0.00875	-30.1216						
36	0.01	-30.0159						
37	0.01125	-29.527						
38	0.0125	-29.2109						
39	0.01375	-29.0051						
40	0.015	-28.6758						
41	0.01625	-27.623						
42	0.0175	-27.2858						
43	0.01875	-26.347						
44	0.02	-25.9593	Simulation data in B28:B828 is sorted.					
45	0.02125	-25.8212	Cumulative risk profile plots (A28:A828,B28:B828).					

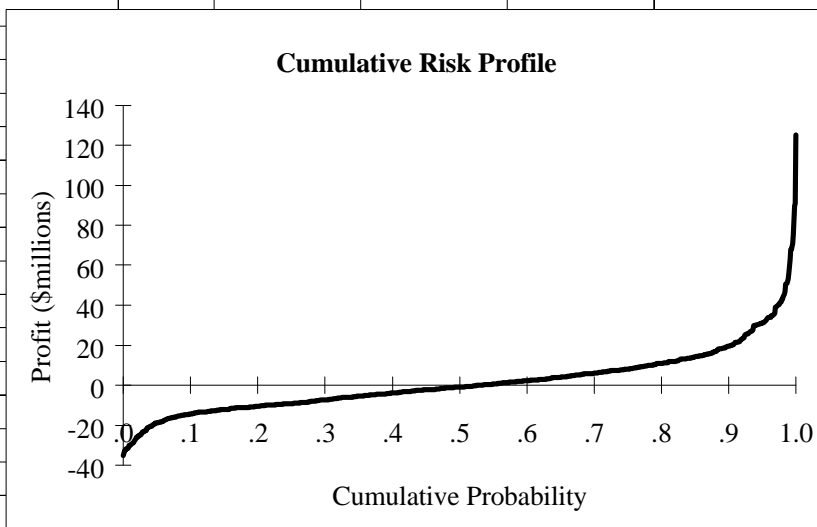


Figure 6. Simulation analysis of the Superior Semiconductor (B) case.

The cumulative risk profile in Figure 5 was made by sorting the simulated profit data in B28:B828 and plotting it in an XY-chart against the percentile index in A28:A828. (Because of this sorting, Figure 5 shows only the worst profits that were sorted to the top of the data range.) Reading from this chart at the 0.05 and 0.10 cumulative-probability levels we find values at risk of about -19 and -14 \$million respectively, which are substantially worse than the 5% value at risk found in our analysis of part A (-9.33 \$million). However, the total probability of losing money our analysis of part B (about 0.5) is lower than in part A (where it was 0.65).

8. Other probability distributions

Uniform random variables. A continuous random variable has a "Uniform distribution over the interval from A to B" if the random variable is sure to be between the numbers A and B and the value of this random variable (rounded to any given number of decimal places) is equally likely to be any possible number between A and B. The probability density associated with this Uniform distribution is a flat constant ($1/(B-A)$) over the whole interval from the lower bound A and the upper bound B, and the probability density is zero outside this interval.

The Excel function RAND() returns a random variable with a Uniform distribution over the interval from 0 to 1, which we use (with the appropriate inverse cumulative function) to create random variables with any other distribution. To make Uniform random variable over some other interval, from the lower bound A to the upper bound B, we use the linear formula

$$=A+(B-A)*RAND()$$

For example, $95+(105-95)*RAND()$ gives us a Uniform random variable over the interval from 95 to 105.

If X is a random variable with a Uniform distribution over the interval from A to B, then X has quartile boundary points $Q_1 = (3*A+B)/4$, $Q_2 = (A+B)/2$, and $Q_3 = (A+3*B)/4$. Its mean is $\mu = (A+B)/2$, and its standard deviation is approximately $\sigma = 0.2887*(B-A)$.

A Uniform random variable over the interval from 95 to 105 is (rounded to one decimal place) just as likely to be 104.9 as 97.2 or any other number in the interval, and yet 105.1 (just 0.2 more than 104.9) is completely impossible. Such a sudden break from "equally likely" to

"impossible" seems unrealistic in most application. Thus, there is rarely good reason to believe that any unknown quantity is well represented by a simple Uniform random variable.

Triangular random variables. Like Uniform random variables (and Beta random variables which are discussed below), Triangular random variables are random variables that are guaranteed to stay within some bounded interval. A Triangular probability distribution is parameterized by its lower bound, its mode (or "most likely" value), and by its upper bound. With Simtools, a random variable with a Triangular distribution, with lower bound b_1 , mode b_2 , and upper bound b_3 can be constructed by the formula

$$=TRIANINV(RAND(),b_1,b_2,b_3)$$

The probability density for such a triangular random variable has a graph that looks like a triangle over the interval from b_1 to b_3 , with its peak at b_2 . Statisticians consider such triangular random variables to be rather crude objects, but triangular random variables are popular for simulating random variables where the decision-maker is comfortable assessing a lower bound, a most-likely value, and an upper bound.

Beta random variables. Beta random variables are also bounded within a given interval, like Triangular and Uniform random variables, but Beta random variables are considered more sophisticated and more respectable to the refined tastes of statisticians.

For example, suppose that we are wondering about some unknown fraction, such as, what is fraction of Kellogg students prefer to drink "Lite" beer. This unknown quantity must be between 0 and 1 (that is, between 0% and 100%). Thus, to describe our beliefs about this unknown fraction, we do not want to use a distribution that allows negative values or values that are greater than 1. The Beta distributions are such a family of distributions, yielding values that are always between 0 and 1. Beta distributions are often used by statisticians for to describe beliefs about such unknown fractions.

The Beta distributions form a two-parameter family. Excel provides a function BETAINV that takes two parameters called α ("alpha") and β ("beta"). (To avoid confusion between the family name and the parameter name, notice that I am capitalizing the distributional family's name. Some books use a somewhat different notation in which the parameters of the Beta distribution are called r and n , where $r = \alpha$, and $n = \alpha + \beta$.) The Excel formula

$$=BETAINV(RAND(),\alpha,\beta)$$

gives us a Beta random variable with parameters α and β , taking values between 0 and 1. The meaning of these parameters α and β can be difficult to explain, so Simtools.xla provides an alternative function called BETINV that is parameterized by the mean μ and standard deviation σ . That is,

$$=BETINV(RAND(),\mu,\sigma)$$

is a Beta random variable with mean μ and standard deviation σ , taking values between 0 and 1. The names of these two functions differ by just one letter, but you can remember that the Simtools function drops the letter "A" because it does not require you to understand a strange "Alpha" parameter. But an even better way to avoid confusion between these functions is to always use the Insert-Function dialogue box to select either of them, because the dialogue box will show you which function has "alpha" and "beta" as parameters, and which has "mean" and "stdevn" as parameters. The relationship between the unintuitive α and β parameters for this Beta distribution and the better-understood μ and σ is given by the formulas

$$\mu = \alpha / (\alpha + \beta)$$

$$\sigma = (\alpha * \beta)^{.5} / ((\alpha + \beta) * (\alpha + \beta + 1)^{.5})$$

The Uniform distribution over the interval from 0 to 1 is actually a special case of a Beta distribution, using the special values $\alpha = 1$ and $\beta = 1$ (or $\mu = 0.5$ and $\sigma = 0.2887$).

The Beta distribution can apply to any example in which we are wondering what fraction of a large population belongs to some subpopulation, such as the fraction of Kellogg students who drink Lite beer. Suppose that we began by thinking that this unknown fraction is drawn from a Uniform distribution over the interval from 0 to 1, but we then sampled m members of this large population and we found that s of these members belong to the subpopulation. Then after observing the sample, our new probability beliefs about this unknown fraction must be a Beta distribution with parameters $\alpha = s + 1$, and $\beta = (m - s) + 1$.

More generally, suppose that you thought that the unknown fraction of MM students who prefer Lite beer has a Beta distributions with some parameters α and β . Given these prior beliefs, you now learn that, among m randomly sampled MM students, s prefer to drink Lite beer. Then you should now assess a Beta distribution with parameters $\alpha' = \alpha + s$ and $\beta' = \beta + (m - s)$ for this

unknown fraction.

You can specify lower and upper bounds other than 0 or 1 by optional fourth and fifth parameters of both the BETINV and BETAINV functions. Thus, for example, BETINV(RAND(),24,3,20,30) returns a Beta random variable that takes values between 20 and 30, has mean 24 and standard deviation 3.

Exponential random variables. Exponential random variables are often used to represent the length of time that we will have to wait for something to happen. Like the Lognormal random variables, an Exponential random variable is always a positive number. Unlike Normal and Lognormal random variables, which have the highest probability density at or near the mean, an Exponential random variable always has its highest probability density at zero.

If \mathbf{X} is a random variable with an Exponential distribution then, for any two nonnegative numbers m and n ,

$$P(\mathbf{X} > m + n \mid \mathbf{X} > m) = P(\mathbf{X} > n).$$

To interpret this equation, suppose that \mathbf{X} represents the number of minutes that we will have to wait for our first customer to arrive. This equation says that the conditional probability of having to wait more than n additional minutes, given that we have already waited any number of minutes (m), is independent of the number of minutes that we have already waited. This no-memory property is often quite natural to assume for waiting times. So the Exponential distribution is often used for random waiting times.

The Exponential distributions have one parameter: the mean. To construct an exponential random variable with mean μ , you can use the Excel formula `= -μ*LN(1 - RAND())`. Equivalently, with Simtools, you can use the formula `=EXPOINV(RAND(),μ)`. This random variable has mean μ , and its standard deviation is also equal to μ . Its quartile points are $0.288*\mu$, $0.693*\mu$, and $1.386*\mu$.

Gamma random variables. The Gamma distribution is often used to represent beliefs about an unknown quantity that is a nonnegative number. The Gamma distribution has two parameters, often called α (alpha) and β (beta), which must also be nonnegative numbers. A Gamma random variable can be constructed in Excel by the formula

$$=\text{GAMMAINV}(\text{RAND}(),\alpha,\beta).$$

Such a Gamma random variable has expected value (mean)

$$\mu = \alpha * \beta,$$

and its standard deviation is

$$\sigma = (\alpha^{.5}) * \beta.$$

Conversely, if we know the mean μ and standard deviation σ of a Gamma random variable, then its parameters α and β must be

$$\alpha = (\mu/\sigma)^2 \text{ and } \beta = (\sigma^2)/\mu.$$

Thus, a Gamma random variable with mean μ and standard deviation σ can be constructed in Excel by the formula

$$=\text{GAMMAINV}(\text{RAND}(),(\mu/\sigma)^2,(\sigma^2)/\mu)$$

Simtools condenses the above formula into one function called GAMINV. That is, the formula

$$=\text{GAMINV}(\text{RAND}(),\mu,\sigma)$$

returns a random variable that has a Gamma distribution with mean μ and standard deviation σ .

To avoid confusion between these similarly-named functions, you should always use the

Insert-Function dialogue box when you want to use either of them, because the dialogue box will show you which function has "alpha" and "beta" as parameters, and which has "mean" and "stdevn" as parameters.

There are at least three common applications of Gamma distributions. The first is in statistics. Suppose that we have a range of K cells, in each of which the value is an independent Normal random variable with some mean μ and some standard deviation σ . Let \mathbf{S} denote the value returned by Excel's STDEV function applied to this range. Then \mathbf{S}^2 (the sample variance) is a Gamma random variable with mean σ^2 and standard deviation $(\sigma^2)*(2/(K-1))^{0.5}$. The sample standard deviation \mathbf{S} itself is not a Gamma random variable but, when K is large, its mean is close to σ , and its standard deviation is close to $\sigma/(2*(K-1))^{0.5}$. So we may use $\mathbf{S}/(2*(K-1))^{0.5}$ as an estimate of the standard error of our sample standard deviation statistic \mathbf{S} .

The second major application of Gamma distributions is in duration times. Recall that Exponential distributions may be applied to waiting times for the next customer's arrival. If $\mathbf{X}_1, \dots, \mathbf{X}_K$ are K independent random variables, each drawn from an Exponential distribution with mean τ , then the sum $\mathbf{X}_1 + \dots + \mathbf{X}_K$ has a Gamma distribution with parameters $\alpha = K$ and $\beta = \tau$.

So the time until the arrival of the K 'th customer may be a Gamma random variable.

A third application of Gamma distributions arises when we have watched customers arriving into a store for a limited period of time, and we are uncertain about the long-run mean arrival rate. Under some technical assumptions, it turns out to be natural to assume that our beliefs about this mean rate of arrival should be described by a Gamma distribution.

Gamma random variables are like a Lognormal random variables in that they can only take values greater than zero. But the Gamma distribution can have a positive probability density at zero, while the probability density of the Lognormal distribution cannot. In effect, Gamma random variables can come closer to zero than Lognormals. Thus, shape of the Gamma distribution often looks more similar to the Normal distribution with the same mean and standard deviation than does the corresponding Lognormal distribution. See for example Figure 7, which shows the probability densities for Gamma, Normal, and Lognormal distributions with mean 4 and standard deviation 3. When the mean is more than 5 times the standard deviation, the Gamma, Normal, and Lognormal distributions become very similar.

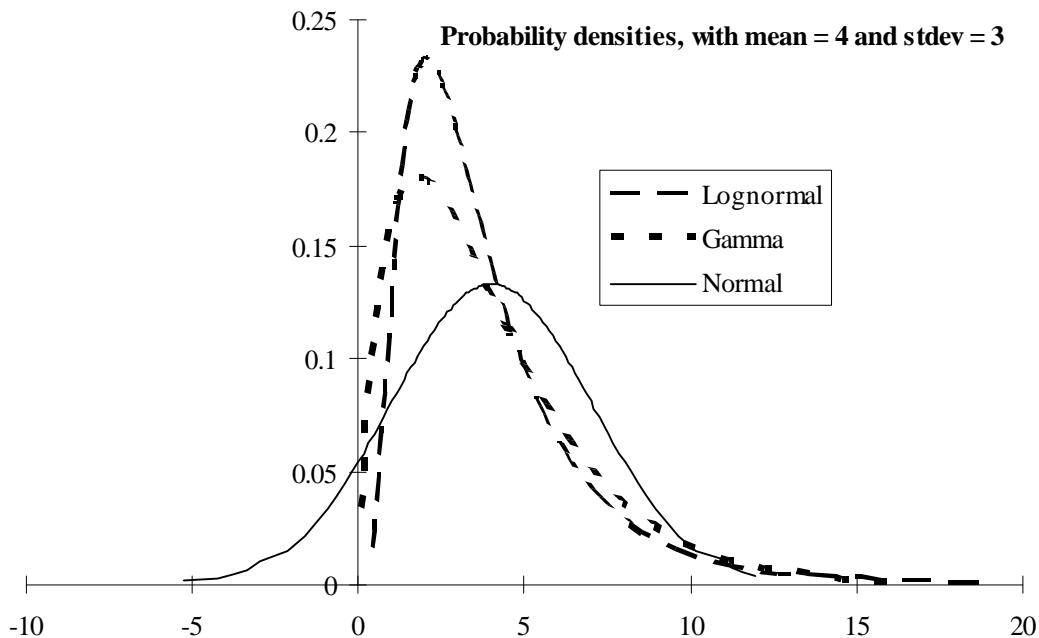


Figure 7. Comparison of Lognormal, Gamma, and Normal distributions.

Extreme-value random variables The Extreme-value (or Gumbel) distribution arises when we look at the maximum of a large number of independent random variables that are drawn from any of a wide family of probability distributions, including the Normals. Simtools.xls provides a function XTREMINV for making Extreme-value random variable parameterized by mean and standard deviation. With this function the formula

$$\text{XTREMINV}(\text{RAND}(),\mu,\sigma)$$

returns an Extreme-value random variable with mean μ and standard deviation σ . Without Simtools, such a random variable can also be made in Excel by the formula

$$\mu - \sigma * (0.45 + 0.78 * \text{LN}(-\text{LN}(\text{RAND}())))$$

Extreme-value random variables can be positive or negative. But unlike the Normals, the Extreme-value distributions are not symmetric about their means, but instead are skewed with a longer tail of likely values on the high side.

XTREMINV allows you to reverse the direction of skewness by entering a negative value as the standard deviation parameter. Such reverse-skewed distributions can arise when we look at the minimum of a large number of independent random variables from the same distribution.

A Weibull random variable with log-mean m and log-standard-deviation s is returned by the formula $\text{EXP}(\text{XTREMINV}(\text{RAND}(),m,-s))$. Notice the negative sign on s to reverse skewness. Weibull random variables are often used as models of a device's operating lifetime.

Binomial random variables. Suppose that we have an experiment whose outcome may be either a "success" or a "failure.". Suppose we perform this experiment n times independently, and let p denote the probability of a success each time we perform this experiment. Then the number of successes is a Binomial random variable with parameters n and p .

A Binomial random variable with parameters n and p can be simulated in a spreadsheet by taking the sum of n cells, into each of which we have entered the formula $=\text{IF}(\text{RAND}()<p,1,0)$. Using Simtools, we can also simulate a Binomial random variable with parameters n and p by the formula

$$=\text{BINOMINV}(\text{RAND}(),n,p).$$

A Binomial random variable with parameters n and p has a mean equal to $n*p$, and it has a standard deviation equal to $(n*p*(1-p))^{0.5}$. If this standard deviation is larger than 3, then the

Binomial random variable can also be well approximated by a Normal random variable with the same mean and standard deviation. If you want make sure that such a Normal approximation to a Binomial is integer-valued, then you can use the formula

$$=ROUND(NORMINV(RAND(),n*p,(n*p*(1-p))^0.5),0)$$

Poisson random variables. The Poisson distribution is commonly used to simulate the number of times that something will happen when there is no clear upper bound on how many times it might happen. For example, the number of customers who will make purchases in a particular shop tomorrow could be described by a random variable with a Poisson distribution.

Suppose that the number of minutes from one customer's arrival until the next customer arrives is a random variable drawn from an Exponential distribution with mean τ . Then the number of customers who arrive in any period of L minutes is a Poisson random variable, and its mean is L/τ . Thus, Poisson distributions are often used to represent our beliefs about the number of arrivals that will occur in a given period.

Poisson random variables always take values that are nonnegative integers. The Poisson distribution has one parameter: the mean. With Simtools, a Poisson random variable with mean μ can be simulated by the formula

$$=POISINV(RAND(),\mu).$$

A Poisson random variable with mean μ has a standard deviation that is equal to $\mu^{.5}$. If μ is larger than 20, then a Poisson random variable with mean μ can also be well approximated by a normal random variable with mean μ and standard deviation $\mu^{.5}$. That is, the formula

$$ROUND(NORMINV(RAND(),\mu,\mu^{.5}),0)$$

gives values that are very similar to $POISINV(RAND(),\mu)$, when $\mu > 20$.

A Poisson random variable with mean μ is also a good approximation for a Binomial random variable with parameters n and p if $n*p = \mu$ and p is very small (say $p < 0.1$). This fact gives you another way to simulate Poisson random variables, if you do not have the Simtools add-in. Given the mean μ of a Poisson random variable, pick a number n such that $n > 10*\mu$, and let $p = \mu/n$; then a Poisson random variable with mean μ can be well approximated by a Binomial random variable with these parameters n and p . (This Binomial random variable can be constructed in your spreadsheet by taking the sum of n cells that each contain the formula

=IF(RAND()<p,1,0.)

9. Summary

In this chapter we studied families of continuous probability distributions that are commonly used to describe uncertainty about unknown quantities which can take infinitely many possible values.

The Normal probability distributions are generally parameterized by their mean and standard deviation, and can be applied to unknowns (like sample means) that are the sum of many small terms drawn independently from some distribution. We learned to use Excel's NORMSDIST function to compute the probability of any given interval for a Normal random variable with any mean and standard deviation. The Excel function NORMINV was used to simulate Normal random variables, and plot their inverse cumulative curves. After introducing the concept of a probability density curve for a continuous random variable, we showed how to estimate it from an inverse cumulative function like NORMINV.

Before introducing the Lognormals, we reviewed the exponential functions EXP and the natural-logarithm function LN, showing how they arise when interest rates are compounded frequently. The advantages of using logarithmic rates of return were shown. We also reviewed how EXP converts addition to multiplication, while the inverse function LN converts multiplication to addition.

A Lognormal random variable is one whose natural logarithm is a Normal random variable. Lognormal random variables can be parameterized by their log-mean m and log-standard-deviation s , or by their true mean μ and true standard deviation σ . With the first parameterization, they can be simulated by the Excel formula $\text{EXP}(\text{NORMINV}(\text{RAND}(),m,s))$. With the second parameterization they can be simulated by the Simtools formula $\text{LNORMINV}(\text{RAND}(),\mu,\sigma)$. Lognormal distributions can be applied to unknown quantities (like growth ratios over some period of time) that are the product of many factors, each close to 1, that are drawn independently from some distribution.

The Generalized-Lognormals were introduced as a family of probability distributions which can be parameterized by the three quartile boundary points, and which includes as special

cases both the Normals (when the quartiles points have equal differences) and Lognormals (when the quartile points have equal ratios). The Simtools function GENLINV was introduced to work with Generalized-Lognormal distributions. Practical techniques were illustrated for subjectively assessing quartiles and fitting a Generalized-Lognormal distribution that can describe an individual's beliefs about a real unknown quantity for which there is no comparable statistical data. We then applied these techniques with simulation analysis to analyze a decision problem that involves several different unknown quantities, some continuous and some discrete.

Exercises

Case: Lepton Inc.

Lepton Inc. is a bioengineering firm, founded 5 years ago by a team of biologists. A venture capitalist has funded Lepton's research during most of this time, in exchange for a controlling interest in the firm. Lepton's research during this period has been focused on the problem of developing a bacterium that can synthesize HP-273, a protein with great pharmacological and medical potential. This year, using new methods of genetic engineering, Lepton has at last succeeded in creating a bacterium that produces synthetic HP-273 protein.

Lepton can now get a patent for this bacterium, but the patented bacterium will be worthless unless the FDA approves the synthetic protein that it produces for pharmacological use. The process of getting FDA approval has two stages. The first stage involves setting up a pilot plant and testing the synthetic protein on animal subjects, to get a preliminary certification of safety which is required before any human testing can be done. The second stage involves an extensive medical study for safety and effectiveness, using human subjects.

Lepton's engineers believe that the cost of the first stage could be is equally likely to be above or below \$9 million, has probability .25 of being below \$7 million, and has probability .25 of being above \$11 million. The probability of successfully getting a preliminary certification of safety at the end of this first stage is 0.3.

The cost of taking the synthetic protein through the second-stage extensive medical study has a net present value that is equally likely to be above or below \$20 million, has probability .25 of being below \$16 million, and has probability .25 of being above \$24 million. Given first-stage certification, if Lepton takes the synthetic HP-273 protein through this second stage then the conditional probability of getting full FDA approval for its pharmacological use is 0.6.

If Lepton gets full FDA approval for pharmacological use of the synthetic HP-273 protein, then Lepton will sell its patent for producing the synthetic HP-273 protein to a major pharmaceutical company. Given full FDA approval, the returns to Lepton from selling its patent would have a present value that is equally likely to be above or below \$120 million, has probability .25 of being below \$90 million, and has probability .25 of being above \$160 million.

Lepton's management has been operating on the assumption that they would be solely responsible for the development of this HP-273 synthetic protein throughout this process, which we may call Plan A. But there are two other alternatives that the owner wants to consider:

Plan B: Lepton has received an offer to sell its HP-273 patent for \$4 million now, before undertaking the any costs in the first stage of the approval process.

Plan C: If Lepton succeeds in getting preliminary FDA certification at the end of the first stage, then it could to sell a 50% share of the HP-273 project to other investors for \$25 million.

In your analysis of this situation, you may assume that each of the three unknown quantities described above has a probability distribution in the Generalized-Lognormal family.

1. Estimate the expected value and standard deviation of Lepton's profits under Plan A and Plan C. Then make a cumulative risk profile for each plan. If the owner of Lepton wants to maximize the expected value of profits, then what plan should we recommend, and what is the lowest price at which Lepton should consider selling its HP-273 patent now?

Base your answers on simulation data that includes enough so that the radius of your

95% confidence interval for the expected value of profit is less than \$2 million for both Plan A and Plan C. (That is, you should be able to generate 95% confidence intervals of the form $x \pm y$ where y is less than \$2 million.) How many simulations did you use? What is your 95% confidence interval for expected profit under each plan?

2. (a) Among the three unknown quantities described in the Lepton case, which have Normal distributions? Compute the mean and the standard deviation for each Normal random variable.
(b) Among the three unknown quantities described in the Lepton case, which have Lognormal distributions? Compute the log-mean and the log-standard-deviation for each of Lognormal random variable.
(c) Compute the probability that the stage 1 cost will be between \$8 and \$12 million. Also compute the probability that it will be more than \$12 million.

3. Consider the following five unknown quantities:

- (a) The number on the last numbered page of Besanko-Dranove-Shanley Ec's of Strategy.
- (b) Age of Columbus when he died.
- (c) Distance from Chicago to Seattle in miles.
- (d) Melting point of Aluminum (specify °F or °C).
- (e) Population of Italy on January 1, 1973.

For each of these unknown quantities, assess subjective quartile boundary points Q_1, Q_2, Q_3 , such that $Q_1 < Q_2 < Q_3$ and you would be indifferent between the following four lotteries:

Lottery 1: You get \$5000 if the unknown is less than Q_1 , but get \$0 otherwise.

Lottery 2: You get \$5000 if the unknown is between Q_1 and Q_2 , but get \$0 otherwise.

Lottery 3: You get \$5000 if the unknown is between Q_2 and Q_3 , but get \$0 otherwise.

Lottery 4: You get \$5000 if the unknown is greater than Q_3 , but get \$0 otherwise.

Then compute the points with 1% and 99% cumulative probability in the corresponding Generalized-Lognormal distribution by the formulas $\text{GENLINV}(0.01, Q_1, Q_2, Q_3)$ and $\text{GENLINV}(0.99, Q_1, Q_2, Q_3)$.

Check whether this Generalized-Lognormal distribution is a good fit to your actual beliefs by asking whether you think that the unknown has a 1/100 probability of being less than $\text{GENLINV}(0.01, Q_1, Q_2, Q_3)$, and whether you think that the unknown has a 1/100 probability of being greater than $\text{GENLINV}(0.99, Q_1, Q_2, Q_3)$. If not, then try to reassess the three quartile points until you also get a reasonable 1%-point and 99%-point from these GENLINV formulas.

Suggestions:

If $\text{GENLINV}(0.01, Q_1, Q_2, Q_3)$ seems too low, then try increasing Q_1 or decreasing Q_2 .

If $\text{GENLINV}(0.01, Q_1, Q_2, Q_3)$ seems too high, then try decreasing Q_1 or increasing Q_2 .

If $\text{GENLINV}(0.99, Q_1, Q_2, Q_3)$ seems too low, then try increasing Q_3 or decreasing Q_2 .

If $\text{GENLINV}(0.99, Q_1, Q_2, Q_3)$ seems too high, then try decreasing Q_3 or increasing Q_2 .