

Some Simple Applications of the Normal and Log Normal Probability Distributions

1. Introduction

As indicated by the title, this paper presents a brief and simple commentary on five applications of the normal and log normal probability distributions.

The motivation for writing this paper came from a realisation that of all the actuarial tools and techniques that I use in my professional life, there is only one where I find I have to dig out my original B3 (for such was the name of the General Insurance course at the time) study notes, to refresh my memory of the relevant formulae. In all other areas the relevant formula or technique can be derived quite easily from first principles.

This is one area where an initially quite complicated-looking formula is not the front for some highly theoretical and possibly impractical line of thinking: it is the key to some quite useful and valid applications.

This paper is therefore aimed at those actuaries who may initially have been deterred from looking further into the normal and log normal probability distributions, and who may therefore have missed out on quite a useful tool.

2. The Formulae:

I have not reproduced the formula for the normal probability distribution itself, since it is of little practical value. These days one does not even have to have a table of the cumulative normal probability distribution available, since the Excel function NORMDIST (used in some of the applications below) does just as well, if not better.

The log normal distribution is really an application of the normal distribution, rather than a separate distribution itself: it is used for applications to claim distributions that are “skewed”, where the natural logarithms of the claims are normally distributed.

Calculating the mean and standard deviation of a log normal probability distribution is quite straightforward:

Sample data

mean = a

standard deviation = b

Equivalent log normal distribution

$$\text{mean} = x = \log_e \left[a * \left(1 + \frac{b^2}{a^2} \right)^{\frac{1}{2}} \right]$$

$$\text{standard deviation} = y = \sqrt{\left[\log_e \left(1 + \frac{b^2}{a^2} \right) \right]}$$

It will be noted that if the standard deviation is calculated first, then the mean x can be calculated as:

$$x = \log_e a - \frac{y^2}{2}$$

And going the other way, from the log normal mean and standard deviation to the distribution of the sample data:

$$a = e^{\frac{x+y^2}{2}}$$

$$b = \sqrt{\left[e^{2x+y^2} * (e^{y^2} - 1) \right]}$$

Claim size weighted by probability (very clever formula)

- If:
- (i) A sample of claims follows a log normal distribution
 - (ii) Average claim is: a
Standard deviation is: b

Then the expected cost of claims above an excess point $E =$

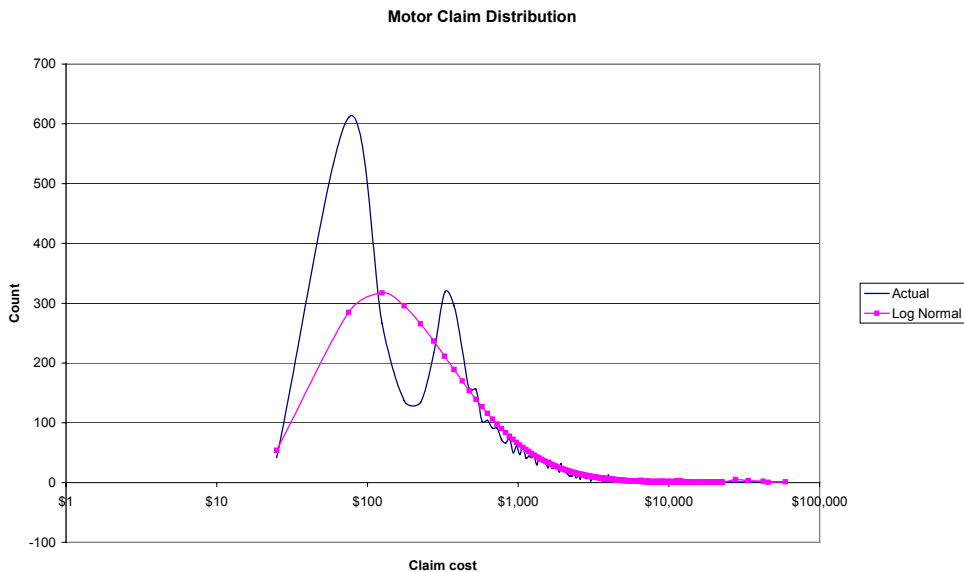
$a * [\text{probability of a claim exceeding } E],$ using a log normal distribution with mean $x + \frac{y^2}{2}$, standard deviation y (x and y as derived as above from a and b).

3. General and Health Insurance Claim Distributions

Perhaps the main reason why the log normal and normal probability distributions are so useful in motor and health insurance analysis, is that they actually do work.

This is particularly so in New Zealand where, because of the legislative environment, we have very little exposure to large bodily injury claims. In other environments, and in the case of some property damage risks, the normal and log-normal distributions tend to under-state the cost of “long-tailed” claims. This was referred to in Neil Christie’s paper to the Society in 1998: “An introduction to developing a claim size distribution for a reinsurance programme”, in which he noted that the Pareto distribution provided a more appropriate solution for the particular reinsurance issue covered by that paper.

The graph below looks at the distribution of the natural logarithm of a portfolio of motor insurance claims, with a sum insured band of between \$10,000 and \$20,000. The data comes from a random selection of claims from a number of different pools, and has been “shifted” away from the actual data.

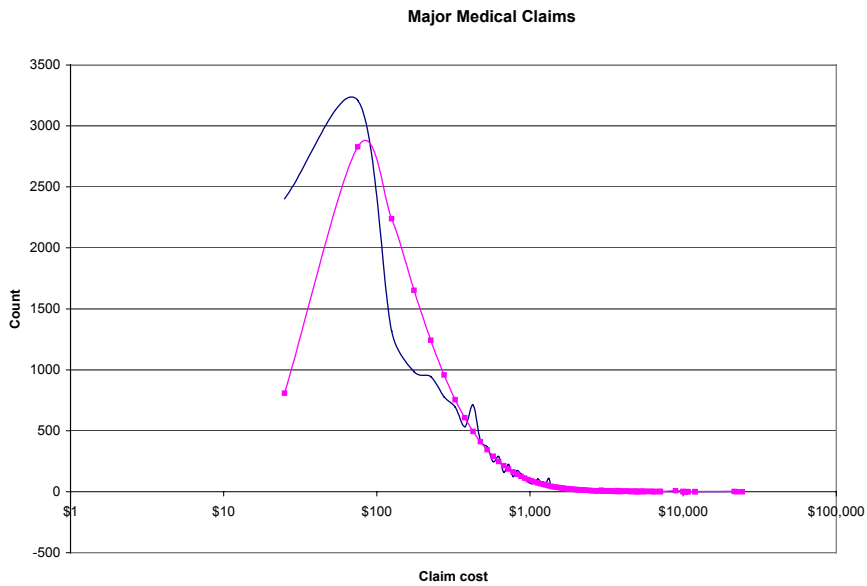


The first peak relates to assessors’ fees, and claims for windscreen cracks, which can easily be filtered out to obtain a smoother curve. I decided not to filter these small claims out in this graph (thereby presenting a “tidier” picture), to illustrate some of the issues that need to be dealt with at the smaller end of the scale. Even with this distortion present, the log normal distribution curve follows the actual claim pattern from about \$400 upwards reasonably closely.

At the upper reaches of the distribution, the proportion of claims above each limit is as follows:

Excess	Actual proportion of claims above this excess	Proportion above excess per log normal
1,000	30.82%	33.14%
5,000	6.76%	5.01%
10,000	2.03%	1.53%
20,000	0.23%	0.37%
30,000	0.08%	0.15%

For major medical claims illustrated in the graph below, again we have some distortion below about the \$400 level, but above that level the log normal curve produces a close fit to the sample data:



The position for larger claims is as follows:

Excess	Actual proportion of claims above this excess	Proportion above excess per log normal
1,000	7.8%	7.8%
5,000	0.2%	0.3%

Having access to a probability distribution that is quite accurate has a number of applications, including:

- a) Assessing the impact of different levels of excess.
- b) Assessing the impact of a change in no-claim bonus structure, if the new structure is expected to lead to behavioural change that varies by claim size.
- c) Assessment of reinsurance programmes.
- d) Producing confidence intervals for future financial forecasts.

Personally I do relatively few statistical exercises using deterministic probability calculations, and use stochastic modelling for most exercises involving variable claim sizes. For complicated applications, (e.g. work on no-claim discount structures) I find that the programming code for simulation models is much simpler to write, and more accurate, than the code for deterministic calculations.

For illustration purposes I have set out the results of a calculation below, using both methods. In the first case, I have set out the results and the formulae of an Excel spreadsheet that calculates the probability of a motor claim, for vehicles in the sum insured bracket \$10,000 to \$20,000 exceeding a certain amount. In the second instance I have shown the Visual Basic code that produces the identical answer using a simulation technique.

To model the normal or log normal distribution in Visual Basic, I used to call up a text file of data which, when choosing a random number between 1 and 1,000, would send back a number of standard deviations that were normally distributed. Subsequently however I am indebted to Ian Perera who showed me a simple formula which is effectively a *sine* curve that has been tweaked a bit to replicate the normal distribution. The Visual Basic code set out below includes this technique.

In both calculations, if the mean claim size is \$1,425 and the standard deviation \$2,400, then there is a 19.16% probability of a claim being larger than \$2,000, and the expected cost of claims in excess of \$2,000 is \$873.69. The claim cost net of excess is \$873.69 less \$2,000 x .1916 = 490.54.

	A	B	C	D	E
1	Motor Insurance Example				
2	Sum insured band \$10,000 - \$20,000				
3		Sample data	Log normal	Going the other way	For claim cost weighted by probability
4	Mean	1,425.0	6.5896384	1,425.00	7.93
5	Standard deviation	2,400.0	1.1595592	2,400.00	1.1596
6					
7					
8	Excess	2,000			
9	Probability of claim	19.157%			
10	Expected claim cost,	873.69			
11	Cost above excess	490.54			
12	Excess discount	65.58%			
13					
14					
15					

	A	B	C	D	E
1	Motor Insurance Example				
2	Sum insured band \$10,000 - \$20,000				
3		Sample data	Log normal	Going the other way	For claim cost weighted by probability
4	Mean	1425	=LN(B4)-0.5*C5^2	=EXP(C4+0.5*C5^2)	=+C4+C5^2
5	Standard deviation	2400	=LN(1+(B5/B4)^2)^0.5	=D4*((EXP(C5^2)-1)^0.5)	=+C5
6					
7					
8	Excess	2000			
9	Probability of claim above excess	=1-NORMDIST(LN(B8),C4,C5,TRUE)			
10	Expected claim cost, claims above excess	=+B4*(1-NORMDIST(LN(B8),E4,E5,TRUE)			
11	Cost above excess	=+B10-B8*B9			
12	Excess discount	=1-B11/B4			
13					
14					
15					

Visual Basic code that simulates the same result:

```

Sub carindiv()
carmean = 6.5896384
carsdev = 1.1595592
excess_limit = 2000
For j = 1 To 10000000
    y1 = Rnd
    y2 = Rnd
    devno = (-2 * Log(y1)) ^ 0.5 * Sin(2 * 3.14159265358979 * y2)
    claimsize = Exp(carmean + carsdev * devno)
    If claimsize >= excess_limit Then
        excess_count = excess_count + 1
        excess_cost = excess_cost + claimsize - excess_limit
    End If
Next j
Debug.Print "Frequency:", excess_count / 10000000
Debug.Print "Cost", excess_cost / 10000000
End Sub

```

It is worth making a brief comment to note that Excel and Visual Basic have a minor inconsistency – the Log function in Visual Basic produces the natural logarithm, whereas in Excel the equivalent function is LN. LOG in Excel returns the logarithm to base 10 by default.

4. **General Insurance Solvency Standards:**

Insurance jurisdictions around the world have varying levels of sophistication regarding their required level of solvency, and the degree to which this is prescribed by statute. New Zealand has very little by way of prescribed solvency margins, the only requirement is that members of the Insurance Council are simply required to maintain a solvency margin equal to 20% of net premiums, although most operate at 40% or more.

Australia has moved to a system whereby claim provisions are required to have a 75% or more probability of sufficiency. In the absence of any guidance this would be a potentially highly subjective criterion, but fortunately the Institute of Actuaries of Australia has commissioned a paper: “Research and Data Analysis Relevant to the Development of Standards and Guidelines on Liability Valuations for General Insurance” by Robyn Bateup and Ian Reed that sets out some quite simple parameters for arriving at a suitable margin over the central estimate.

By way of illustration, it is suggested that for a book of liability business, the coefficient of variation (being the standard deviation divided by the mean) should be as follows:

$$\text{Coefficient} = \sqrt{(.028 + \frac{1.6}{L})}$$

L in this formula equals the size of the central claims estimate in millions of (Australian) dollars. The factors .028 and 1.6 come from tables in the paper. So a portfolio with a central estimate of \$50m would have a coefficient of variation of 24.5%.

I understand that actuaries in Australia are generally using either the normal or the log normal distribution for their sufficiency calculations, depending on the line of business, and the size of the portfolio (the larger the business, the more “normal” the distribution). The log normal distributions have the fatter “tail” and could therefore be considered generally more prudent. However for a company with several diverse lines of business, the log normal distribution is not additive, and this can create difficulties with diversification discounts (diversification discounts are also dealt with in the Bateup & Reed paper).

Assuming a log normal probability distribution for our liability example, this translates to a prudential margin using a goal seek as follows:

	A	B	C
1	Australian Prudential Margin Log Normal Example		
2			
3		Claim distribution	Log normal
4			
5	Mean	1.00000	-0.02915
6	Coefficient of variation	0.24500	0.24144
7			
8			Sufficiency
9	Loading on central estimate	1.1430514	75.0000%
10		Calculate with goal seek	

	A	B	C
1	Australian Prudential Margin Log Normal Example		
2			
3		Claim distribution	Log normal
4			
5	Mean	1	=LN(B5)-0.5*C6^2
6	Coefficient of variation	0.245	=(LN(1+(B6/B5)^2))^0.5
7			
8			Sufficiency
9	Loading on central estimate	=1.1430514	=NORMDIST(LN(B9),C5,C6,TRUE)
10		Calculate with goal seek	

In this example therefore, the central estimate needs to be loaded by 14.3% to provide a 75% probability of sufficiency.

5. Asset Models

Stochastic investment models form the basis of a significant amount of actuarial advice that is offered these days. The problem with these models though is they tend to be quite complex, and without understanding their inner workings, the basis of this advice has the potential for being flawed.

For my own part I am indebted to the writers of a paper presented to the Staple Inn Actuarial Society in November 1999: "The TY Model". The authors Y.H. Yakoubov, M.H. Teeger and D.B. Duval, have produced a very readable document which sets out the building blocks of a stochastic investment model in a practical format. The model can be replicated without too much difficulty with a day or two of programming and testing.

In the bowels of stochastic investment models one will inevitably find normal distributions, and the TY Model requires the use of a normal distribution in each of its building blocks.

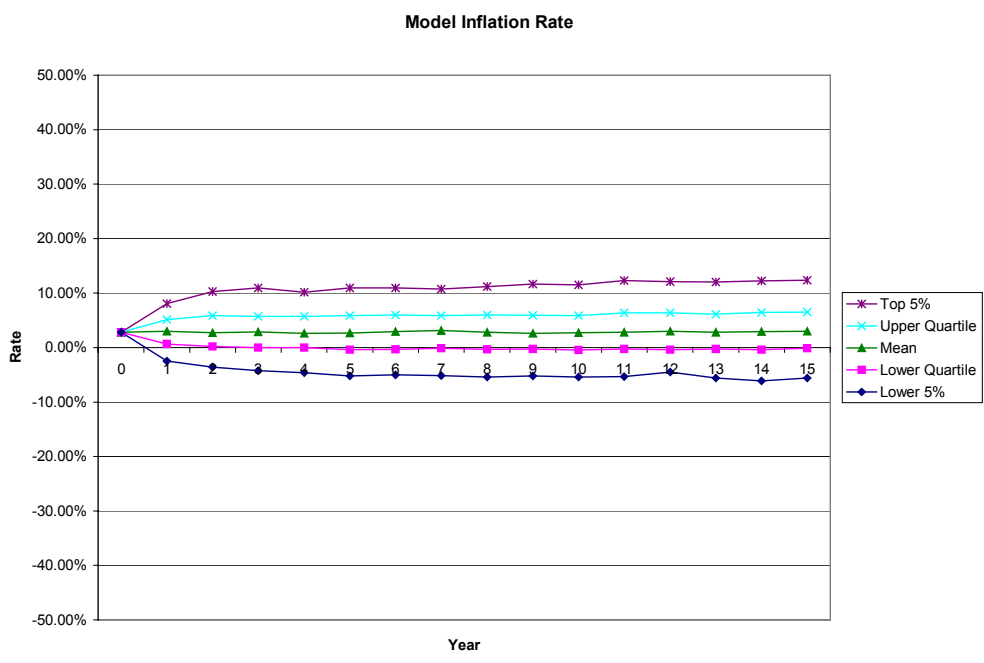
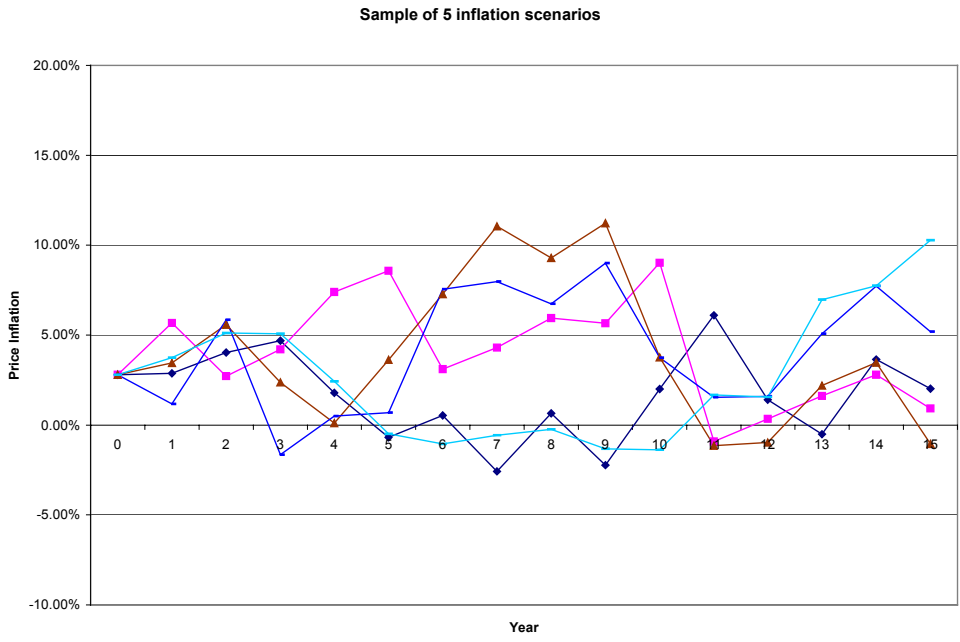
The starting point of this particular model is the generation of retail price inflation, with the inflation rate in any period being a function of:

- a) The expected mean rate of inflation.
- b) A correlation factor applied to the inflation rate in the previous period.
- c) A normally distributed random factor with a mean of 0 and a standard deviation related to the random factor for the previous period.

I have found that the following code reproduces the retail price inflation tables in the paper quite closely:

```
Sub main()
Dim inf_rate(1000, 15)
Open "d:\client\assets\results.prn" For Output As #2
Randomize
For iterate = 1 To 1000
  cpimean = 0.028
  inf_rate(iterate, 0) = cpimean
  infalpha1 = 0.6
  Rem paper says infbeta1 should be "9.38 (= 3.06 ^2)". However only
  Rem .0306 ^ 2 gives suitable results
  infbeta1 = 0.000938
  Rem paper says infbeta2 should be 0.72 but the results are more volatile
  Rem than results in the charts later in the paper
  infbeta2 = 0.5
  cpivariation0 = 0
  cpiforce0 = cpimean
  For t = 1 To 15
    cpisdev = (infbeta1 + infbeta2 * cpivariation0 ^ 2) ^ 0.5
    REM these next three lines are the formula sourced via Ian Perera
    REM for simulating the normal distribution
    Y1 = Rnd
    Y2 = Rnd
    devno = (-2 * Log(Y1)) ^ 0.5 * Sin(2 * 3.14159265358979 * Y2)
    cpivariation1 = cpisdev * devno
    cpiforce1 = cpimean + infalpha1 * (cpiforce0 - cpimean) + cpivariation1
    inf_rate(iterate, t) = Exp(cpiforce1) - 1
    cpivariation0 = cpivariation1
    cpiforce0 = cpiforce1
  Next t
Next iterate
Write #2, "Iteration", 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15
For i = 1 To 1000
  Write #2, i, inf_rate(i, 0), inf_rate(i, 1), inf_rate(i, 2), inf_rate(i, 3), _
  inf_rate(i, 4), inf_rate(i, 5), inf_rate(i, 6), inf_rate(i, 7), inf_rate(i, 8), _
  inf_rate(i, 9), inf_rate(i, 10), inf_rate(i, 11), inf_rate(i, 12), inf_rate(i, 13), _
  inf_rate(i, 14), inf_rate(i, 15)
Next i
End Sub
```

A chart of 5 results, and a chart with the combined effect of 1,000 simulations with a 75% and 95% distribution band, are set out below.



The other elements of the model described in the paper can be derived in turn with a similar approach.

6. Claim Frequencies: Binomial Distribution

Quite often in life insurance, health insurance, or general insurance applications questions arise as to the variability of the number of claims.

For example, if the expected number of death claims in a portfolio is 30 in any one year, what is the likelihood that the number could be 25 or less, or 35 or more?

With this type of exercise I would normally just simulate the experience, and see how many iterations came out below or above these two limits. However the normal approximation to the binomial distribution does provide reasonable answers, and to repeat the basic formulae:

$$\text{mean} = \text{number of lives insured (n)} \times \text{average mortality rate (q)}$$

$$\text{standard deviation} = \sqrt{[n * q * (1-q)]}$$

The normal approximation to the binomial distribution tends to be quite accurate for variations close to the mean, although it becomes proportionately less accurate as one moves further away from the mean. The following Excel extract compares the results of a normal approximation to the binomial distribution, with the true results from a simulation run. It is not possible to obtain the “true” results from a binomial distribution calculation, since readers will recall that the binomial distribution becomes unmanageable with large populations.

I find that once the number of expected events reaches 5 or more then the normal approximation works reasonably well, although I understand that at least one actuarial textbook suggests a level of 10 rather.

	A	B	C	D
1	Life Insurance Claim Variability			
2				
3	Lives	100,000		
4	Average qx	0.001		
5	Limit	125		
6	Binomial mean	100.00000		
7	Std deviation	9.99500		
8				
9	Probability of claims in the year exceeding 125:			
10	Normal - excess of		124.5	
11	Using normal approximation to binomial:			0.71%
12				
13	Using simulation model:			1.05%

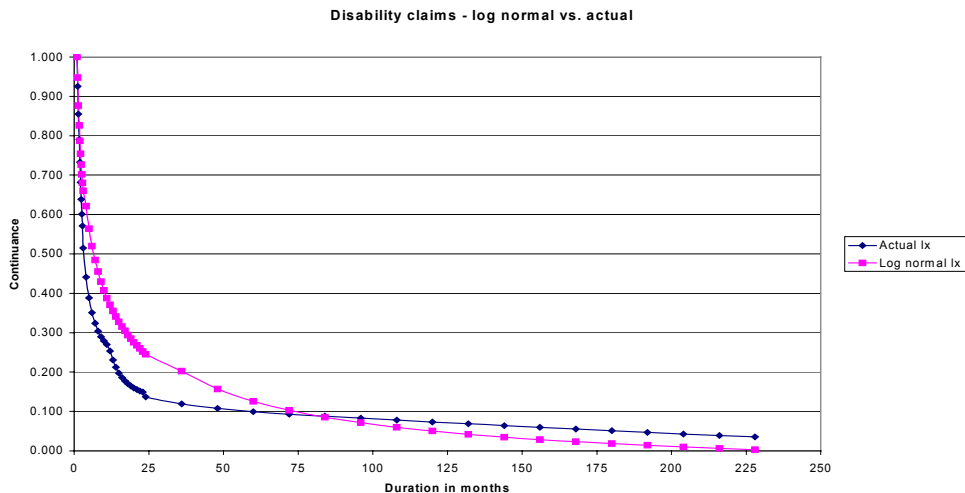
	A	B	C	D
1	Life Insurance Claim Variability			
2				
3	Lives	100000		
4	Average qx	0.001		
5	Limit	=+B3*B4*1.25		
6	Binomial mean	=+B3*B4		
7	Std deviation	=+(B3*B4*(1-B4))^0.5		
8				
9	Probability of claims in the year exceeding 125:			
10	Normal - excess of		124.5	
11	Using normal approximation to binomial:			=1-NORMDIST(C10,B6,B7,TRUE)
12				
13	Using simulation model:			0.010499999858439

7. **Disability Income Claims:**

It is possible to use a log normal distribution for some applications involving disability income claims, that produces a reasonably good result. This approach however tends to over-estimate the number of short-term claims, and under-estimate the number of long-term claims.

I quite often use the log normal distribution when setting up and testing a projection programme, as the processing time is quicker than running along the claim continuance curve with every claim, and terminating the claim randomly at the end of each week/month/year.

The chart below sets out how the continuance pattern of claims whose duration is log normally distributed will compare against the actual claim continuance pattern. It will be noted that the termination rates of log normally distributed claims are too low up until about month 24, and are then too rapid. We end up with significantly fewer long term claims using the log normal distribution than is actually the case. Overall the undiscounted claim cost for each of these two continuance curves is identical.



This graph has been obtained by:

- a) Simulating a large number of disability income claims with a realistic set of termination assumptions, terminating randomly about the mean in each week/month/year.
- b) Calculating the mean and standard deviation of the natural logarithm of the resulting sample data of claim durations.
- c) Creating a distribution of claim durations that follow a log normal distribution with the mean and standard deviation from (b).
- d) Setting up a claim continuance curve from the distribution of claim durations in (c).

In summary, the aim of this paper was to describe some applications of the normal and log normal distributions, and I hope that some actuaries may find it of value in applying them to practical professional issues.

Peter Davies B.Bus.Sc., FIA, FNZSA, AIAA, ASA
September 2002