PROBABILITY DISTRIBUTION OF FIRE LOSSES

A.M. HASOFER and I.R.THOMAS
CESARE, Victoria University of Technology
P.O.Box 14428 MMC Melbourne, Victoria 3000 Australia

ABSTRACT

Data on fire losses in Hotels and Motels collected by the National Fire Incident Reporting System in the USA between 1983-1995 are statistically analysed. It is shown, using as illustration the year 1988 data, that the non-zero fire losses closely follow a lognormal distribution. For the years 1983-1995 the deflated mean and coefficient of variation, as well as the percentage of zero-loss fires, have remained practically constant.

A new technique for estimating high quantiles of the distribution, based on recent work in the Theory of Extreme Value Distributions, is presented and used to obtain the 99% quantile of the fire loss for the year 1988. The result matches closely the estimate obtained by linear interpolation, but an estimate based on the lognormal distribution parameters derived from the global data seriously underestimates the percentile. A table of the coefficients required to calculate high quantiles for the whole period (1983-1995) is given.

KEY WORDS: Fire losses, hotels and motels, NFIRS data base, lognormal distribution, high quantile estimation, fire loss time variation.

INTRODUCTION

In 1977 Rogers published a paper [1] in which he studied the probability distribution of fire losses in the United Kingdom for various industrial occupancies. The purpose of the paper was to evaluate the effect of sprinkler protection on fire losses. The data he analysed were yearly values over the period 1966-1972. Only individual fires for which the total damage to structure and content was £10,000 or more were included. The methodology used was due to Ramachandran [2].

The purpose of the present paper is to revisit Rogers' findings about the statistical properties of the data, using the far more extensive data available from the USA, as well as some more refined statistical techniques that have been recently developed.

Rogers assumed that the fire loss was lognormally distributed. Although he claimed that, based on previous work, this was a "reasonable" assumption, he was not able to justify it on the basis of the data presented, because the data available to him represented "only a small percentage of the number of fires", namely the upper tail, with the bulk of the distribution absent. Moreover, because of this restriction, one important estimate obtained by Rogers, namely the expected loss, is of particularly questionable accuracy. The data available to the authors and presented in this paper do cover the full range of losses (subject to some uncertainties, as discussed later in the paper).

Copyright © International Association for Fire Safety Science

Since the purpose of this paper is mainly to illustrate the proposed methodology, the analysis will concentrate on just one type of occupancy: Hotels and Motels.

According to Rogers, there are two main parameters that have practical application in fire engineering: the average loss and the maximum property loss.

Average losses can be used to estimate the probable reduction in loss per fire due to the installation of fire fighting equipment such as sprinklers. Indeed, such installations can be justified on economic grounds only if the total reduction in fire losses exceeds the total cost of installing and maintaining these systems. Thus, for various purposes, including broad planning of fire cover, it is necessary to estimate the total loss (or, equivalently, the average loss) for different occupancies and a variety of fire fighting equipment. An optimum fire cover would be one which would minimize the sum of total fire loss and fire fighting costs, subject to certain constraints such as risk to life.

On the other hand, there is usually a maximum property loss that an individual owner (whether a private person or a corporation) and/or their insurers would be prepared to put up with. This allowable maximum loss depends on the probable consequential losses as well as on the assets of the owner (or even on the size of the insurance company in the case of a large building). However, the yearly maximum of the raw data is a highly variable measure. For example, for Hotels and Motels in USA over 1983-1995, the maximum of the maximum yearly loss was $5 million while the minimum of the maximum yearly loss was $1.5 million. In risk analysis, it is customary to specify the loss that is exceeded with a given fixed probability, e.g.1%. It is known as the 99% quantile of the distribution of losses and is far more stable than the maximum itself. Thus, for the same data, the maximum 99% quantile was $329,000 and the minimum $210,000.

DESCRIPTION OF THE DATA

The data used in this study is obtained from the National Fire Incident Reporting System of the USA (NFIRS). The NFIRS database is maintained by the United States Fire Administration, Federal Emergency Management Agency and consists of systematic reports of fires under a uniform system contributed by many Fire Departments throughout the USA. Each incident reported is coded systematically in maintaining the database.

ANALYSIS OF A TYPICAL DATA SET

The method of analysis will be illustrated on one particular data set, namely Hotels and Motels for the years 1983-1995. The analysis is carried separately for each year. A full analysis will be given for the year 1988, which should be typical of other years.

The first task is to test whether the data are consistent with a lognormal distribution. We have from the outset a problem, since the probability of the value zero in a lognormal distribution is zero. In the data, however, there is a comparatively large proportion of zeroes. For 1988, the total number of fires listed was 3377 of which 943, or 27.9% had a zero fire loss.

There were clearly two ways to deal with the problem. The first (and most obvious one) was to imagine that the zero values were actually distributed lognormally in the positive neighbourhood of the origin. The second one was to just disregard the zero values and fit a lognormal distribution to the loss values greater than zero. The fitting was carried out using a quantile-quantile plot [3] of the logarithm to the base 10 of the non-zero fire loss against the quantiles of the standard normal distribution.

The *p*-th quantile of a random variable X is a number η such that the probability that X is less than η is equal to *p*. A normal quantile-quantile plot consists of a plot of the ordered values of the data versus the corresponding quantiles of a standard normal distribution. If the quantile-quantile plot is fairly linear, the data are reasonably normal. Linearily is conveniently measured by the correlation coefficient between the ordered data and the corresponding normal quantiles.

Figure 1 shows the quantile-quantile plot when the zeroes are taken into account, while Fig. 2 shows the plot when the zeroes are ignored. It is quite clear that the second plot is a better fit. Numerically, this can be confirmed by calculating the correlation between the quantiles. For Fig. 1 the correlation is 0.984, while for Fig. 2 it is 0.994.

It might have been thought that it is quite possible that there were many more fires with negligible losses than shown in the data, but they were not reported. This would suggest that the fit would be improved by increasing the proportion of zeroes in the data. This, however, is not the case. Increasing the number of zeroes actually decreases the correlation. For example, if the proportion of zeroes is increased to 50%, the correlation decreases to 0.976. From here on, the analysis will ignore the zero loss fires, and we will be content to quote the percentage of zero fires.

The next task is to find the mean and standard deviation of the logarithm of loss. This can be obtained in two ways:

1. By direct calculation of the mean and standard deviation of the logarithmic data. This yields a mean of 2.85 and a standard deviation of 1.022.
2. By calculating the coefficients of regression of the data quantiles on the quantiles of the standard normal distribution. A standard least-square regression without weights was performed. This yielded a mean of 2.86 and a standard deviation 1.014, a result not significantly different from the first calculation.

As far as the mean of the non-zero fire losses is concerned, it can, of course, be easily calculated from the raw data. However, as pointed out in [4], this is an unreliable method that suffers from great variability. It is much better to calculate the mean μ_U and standard deviation σ_U of the logarithm U (to the base 10) of the fire loss X, and then use the formula [4]

$$\mu_X = 10^{\mu_U + \frac{1}{2}\log_e(10)\sigma_U^2} \qquad (1)$$

where $\log_e(10) = 2.303$.

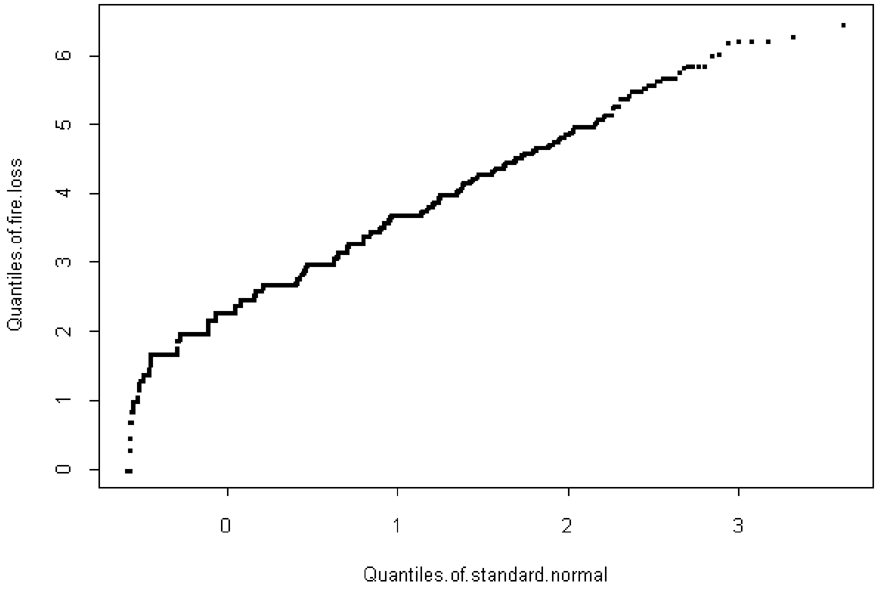

Fig. 1 - Lognormality of fire loss, including zeroes.

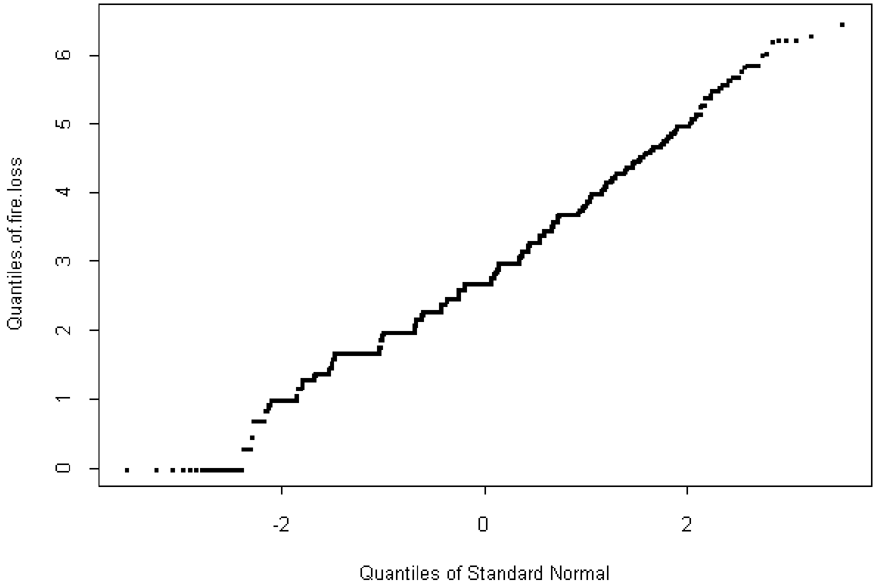

Fig. 2 - Lognormality of fire loss, after deleting zeroes

VARIATION OF THE FIRE LOSS OVER TIME

Table 1 illustrates the variation of the mean and standard deviation of the non-zero fire losses over the years 1983-1995. One clear feature of the table is that the number of fires has been steadily decreasing.

However, in absolute terms, the mean fire loss is steadily increasing. But following the lead of Rogers [1] the means were deflated, using the Consumer Price Index published by the U.S. Department of Labor. The particular index used was for all urban consumers, U.S. city average, based on all items. The values used were for January of each year.

Table 1: Fire loss variation with time

Year	No of fires	Percentage of zero-loss fires	Mean of log	Deflated mean of log	Standard deviation of log	Deflated mean dollar loss
1983	3005	25.32	2.78	2.78	1.01	9149
1984	3336	25.87	2.77	2.76	1.00	7863
1985	3505	26.90	2.81	2.77	1.00	8588
1986	3361	25.83	2.85	2.80	1.01	9312
1987	3317	27.40	2.87	2.82	1.00	9376
1988	3377	27.92	2.85	2.78	1.02	9627
1989	3251	29.44	2.86	2.77	1.02	9222
1990	3200	27.75	2.89	2.77	1.01	8956
1991	2982	27.77	2.90	2.77	1.01	8645
1992	2894	25.74	2.95	2.80	1.05	11850
1993	2721	26.39	2.91	2.75	1.02	8665
1994	2563	25.83	2.97	2.80	0.99	8485
1995	2362	23.71	3.03	2.84	0.99	9390

The last column of Table 1 gives the deflated mean loss in (US) dollars calculated by using Eq. 1. There is no longer a discernible trend, either upwards or downwards.

It is interesting to note that overall there is very little variation and no clearly discernible trend over time in all the measures reported in Table 1 during the thirteen year period covered, apart from the number of fires. Another point of interest is that the standard deviation of the logarithm of the loss (which is approximately equal to the coefficient of variation of the losses) remained surprisingly constant throughout the period under consideration.

ESTIMATION OF HIGH QUANTILES

While the mean of non-zero fire losses is of great interest in estimating the social damage caused by fires, it is even more important for insurance purposes to be able to determine the high quantiles of the distribution of fire losses, e.g. the fire loss that will be exceeded with a probability of 1%. (See the discussion in Rogers' paper [1] p.16 and in Ramachandran's book [5] pp.186-188).

One might have thought that once a probability distribution has been globally fitted to the data set, it is possible to estimate the high quantiles from that distribution. This is unfortunately not the case. The reason is that the fitting of the distribution involves two types of error:

1. a modelling error,
2. a sampling error, due to random variations in the data.

Both errors tend to be different in the centre of the distribution and the tails. This can clearly be seen in Fig. 2, where the fitting in the centre of the distribution is much better than in the tails. The modelling error is particularly worrying when estimating high (or low) quantiles, because it is quite possible that the physical mechanisms that generate very high (or low) values of the variable are different from those that generate the central values.

To avoid this problem, it is customary to estimate high quantiles from the upper tail of the data alone.

Rogers' approach, using the work of Ramachandran [2], was a weighted least squares estimation with correlated errors based on extreme value theory. Since then, simpler methods, based on the pioneering work of Weissman [6] and Hasofer and Wang [7], have been evolved. A comprehensive statement, specially written for engineers, is given in Hasofer [8]. In this paper, an outline of the methodology, sufficient for the practising engineer, is given, dispensing with theoretical proofs.

Suppose the given sample has n elements. The first step in the estimation is to put the sample values in descending order: $X_1 \geq X_2 ... \geq X_n$. We then choose the k largest values $X_1,...,X_k$ and base the quantile estimation on them. When there is a large sample available, as in the case of fire losses, a simple rule of thumb is to choose $k = 1.5\sqrt{n}$. (There are more sophisticated methods to choose k. See [8] p.208 and the references therein).

The next step is to determine the domain of attraction of the underlying distribution. The idea of a domain of attraction is based on the fact that the distribution of the maximum of a random sample, when suitably scaled, tends, with increasing size of sample, to one of three so-called “Extreme Value Distributions”. They are:

1. Type I, also known as the Gumbel distribution,
2. Type II, also known as the Cauchy distribution,
3. Type III, whose mirror image is known as the Weibull distribution.

Weissman has shown that the domain of attraction of a distribution determines the shape of its upper tail. Using that property, it is possible to obtain simple, but highly efficient formulae for the high quantiles of the distribution, based on the top k values of the sample.

Determination of the domain of attraction of the distribution is accomplished by calculating the test statistic

$$W(X_1,...,X_k) = \frac{k(\overline{X} - X_k)^2}{(k-1)\sum_{i=1}^{k}(X_i - \overline{X})^2} \qquad (2)$$

where

$$\overline{X} = \left(\sum_{i=1}^{k} X_i\right) / k \qquad (3)$$

The value of W is then tested against the upper and lower limits W_U and W_L given in Table 3 in the Appendix for values of k between 13 and 500. The table is an extract from [8] p 204. The same table is also available in Hasofer and Wang [7]. Values of W_U and W_L for intermediate values of k can be calculated by interpolation. For values of k larger than 500 we can take $W_L = 1/k - 2.56/k^{3/2}$ and $W_U = 1/k + 2.56/k^{3/2}$. If W lies between the two limits, it is concluded that the underlying distribution belongs to the Type I ("Gumbel") domain of attraction. In that case, the estimator of the quantile q that is exceeded with probability ε is given by

$$q = a\ln(k/n\varepsilon) + X_k \qquad (4)$$

where $a = \overline{X} - X_k$ and $\ln$ is the natural logarithm (to the base e).

If, on the other hand, $W < W_L$, we conclude that the underlying distribution is Type II ("Cauchy"). In that case, we carry out the tranformation

$$Y_i = \log_{10}(X_i - \omega) \qquad (5)$$

for $i = 1,...,k$ and an appropriate value of ω (see below). The resulting transformed sample now belongs to the domain of attraction of the Type I distribution and the required quantile can be evaluated by using Eq. 4.

(In the papers referred to, natural logarithms are used, but for the transformation step this makes no difference, and using logarithms to the base 10 makes it easier to compare the results with those based on the analysis above (Analysis of a typical data set).)

The appropriate value of ω is obtained by solving the equation

$$W(Y_1,...,Y_k) = \frac{1}{k} \qquad (6)$$

It is shown in Hasofer and Li [9] that when $W < W_L$ this equation always has a unique solution. As stated above, the required quantile q_x of the original distribution can now be evaluated by calculating the corresponding quantile q_y of the tranformed sample, using equation

$$q_y = a\ln(k / n\varepsilon) + Y_k \tag{7}$$

where $a = \overline{Y} - Y_k$ and then obtaining the original quantile q_x by carrying out the inverse transformation to Eq. 5, namely

$$q_x = \omega + 10^{q_y}\,. \tag{8}$$

The case $W > W_U$ corresponds to the case when the underlying distribution is in the domain of attraction of Type III. The Type III distribution has a finite upper bound and does not normally show up in situations where there no fixed upper bound within the range of interest of the variable being studied. This is clearly the case for fire losses. The mirror image of the Type III distribution, on the other hand, is extensively used to model load resistances, which are by their very nature non-negative and thus have a finite lower bound. As mentioned above, it is associated with the name of the celebrated Swedish engineer W.Weibull.

APPLICATION TO THE TYPICAL DATA SET

The method just described was applied to the data set discussed above in 'Analysis of a typical data set', namely Hotels and Motels for the year 1988. We use the undeflated figures.

The analysis dealt with fires with non-zero losses, of which there were 2434. The value of k chosen was $1.5\sqrt{2434} \approx 75$. The value of ω calculated from the raw data was 5.31×10^{-3}. The corresponding values of W_L and W_U (obtained by interpolation from Table 3) were 10.64×10^{-3} and 18.45×10^{-3}. The conclusion is that the sample belongs to the Type II domain of attraction.

Solving Eq. 6) we find ω = 142,911.6. Using that value, we find that for the transformed sample $Y_1,\ldots,Y_k$ the value of Y_k is 5.348 and the value of a is 0.2799. The value of the 99% quantile for the transformed sample $Y_1,\ldots,Y_k$ is 5.663. Reverting to the X values, we find that the 99% quantile is $317,519.

This value can be compared with the value obtained directly by linear interpolation from the sample, namely $350,000, which is about 10% higher than the calculated value. However, we must remember that with this type of lognormal distribution the 99% quantile is expected to have a coefficient of variation of about 23%, so that the agreement is quite satisfactory. (The coefficient of variation was obtained from a Monte Carlo simulation of the lognormal distribution.)

On the other hand, if we calculate the 99% quantile from the mean and standard deviation of the lognormal distribution, namely 2.85 and 1.02, we find $170,408, which is far below the value derived from the sample.

The calculations just carried out confirm the above statement that estimating high quantiles from a hypothesised global distribution, with parameters derived from the whole sample, can be severely misleading.

On the other hand, the proposed asymptotic method is probably more reliable than the linear interpolation, as far as prediction of future extreme values are concerned. For a full discussion of this topic, see Boos [10].

Table 2 gives the values of ω, a , Y_k and the 99% quantile for the period 1983 to 1995. The values are based on raw loss figures. Other quantiles can be calculated from the first three values by using Eq. 7 and 8.

Table 2: Values of ω, a , Y_k and the 99% quantile for 1983-1995.

Year	ω	a	Y_k	99% quantile ($)
1983	-17,700	0.435	4.85	219,000
1984	-158,000	0.196	5.35	210,000
1985	-62,000	0.275	5.18	238,000
1986	9,270	0.339	4.93	212,000
1987	-61,300	0.330	5.13	262,000
1988	-142,000	0.280	5.35	318,000
1989	-121,000	0.249	5.28	256,000
1990	-86,900	0.273	5.21	253,000
1991	-157,000	0.230	5.35	273,000
1992	-305,000	0.150	5.61	329,000
1993	-66,700	0.282	5.14	256,000
1994	-130,000	0.216	5.30	265,000
1995	-33,000	0.336	4.99	269,000

REFERENCES

[1] Rogers, F.E. (1977) *Fire losses and the effect of sprinkler protection of buildings in a variety of industries and trades*. Building Research Establishment. Fire Research Station. Borehamwood, Hertfordshire, UK.

[2] Ramachandran, G. (1974) Extreme value theory and large fire losses. *Astin Bull.* **VII**, 3, 293-310.

[3] Hoaglin, D. C., Mosteller, F. and Tukey, J. W., editors (1983). *Understanding Robust and Exploratory Data Analysis.* Wiley, New York.

[4] Johnson, N. L. and Kotz, S. (1970) *Distributions in Statistics: continuous univariate distributions*. Wiley, New York.

[5] Ramachandran, G. (1998)*The Economics of Fire Protection.* E & FN Spon, London.

[6] Weissman, I. (1978) Estimation of Parameters and Large Quantiles Based on the *k* Largest Observations. *J.Am. Stat. Assoc.*, **73**, 364, 812-815.

[7] Hasofer, A.M. and Wang, J.Z. (1992) A Test for Extreme Value Domain of Attraction. *J. Am. Stat. Assoc.*, **87***,* 417, 171-177.

[8] Hasofer, A.M. (1996) Non-parametric estimation of failure probabilities. Chapter 4 of *Mathematical Models for Structural Reliability Analysis* (F.Casciati and J.B.Roberts eds.) CRC Mathematical Modelling Series, 195-226.

[9] Hasofer, A.M. and Li, S. (1999) Estimation for Type II domain of attraction. *Aust. & N.Zealand J. of Statistics*, **41**, 2, 223-232.

[10] Boos, D.D.(1984) Using Extreme Value Theory to Estimate Large Percentiles. *Technometrics*, **26**,1, 33-39.

APPENDIX

Values of W_L and W_u for k =13,…, 500.

Table 3 - Values of of W_L and W_U for given k.

k	$W_L \times 10^3$	$W_U \times 10^3$
13	53.84	159.95
14	48.96	140.63
15	45.65	127.56
16	43.15	118.33
17	40.77	109.50
18	38.65	101.71
19	36.85	95.04
20	35.07	88.88
21	33.36	83.60
22	32.10	78.73
25	28.54	67.21
30	24.17	53.55
40	18.64	37.71
50	15.18	28.95
60	12.88	23.29
80	9.89	16.84
100	8.02	12.96
200	4.19	5.94
500	1.80	2.26