# SOME NOTES ON THE STATISTICAL THEORY OF EXTREME VALUES

R. E. BEARD
London, England

In preparing the notes on the subjects for discussion at the 4th ASTIN Colloquium at Trieste [1] I used some of the material which formed the basis of a talk given to the Scandinavian Actuarial Societies in September 1962. The papers presented in Trieste have established on a firm mathematical footing the formula for the excess loss premium derived in my note but the discussions also showed that some of the other 1962 material would be of interest.

It will be appreciated that the question originally proposed, i.e. to calculate an excess of loss reinsurance premium when the only information available is the largest claim experienced in each of a succession of periods, was deliberately phrased in this form as being the most troublesome case likely to arise in practice. Clearly if other information is available it would not be rejected in arriving at a premium, but this immediately extends the problem to one of finding the best methods of combining information of different kinds. For example if, say, the largest 5 claims in each period are known, what technique will make the maximum use of the data? Such extensions of the problem are not discussed in this note. Furthermore, I am not unmindful of the valuable comment by Jung "that there is a natural law which states that you can never get more out of a mincing machine than what you have put into it" [2].

From essentially heuristic reasoning I derived the following formula for the calculation of an excess of loss premium

$$P = \frac{1}{\alpha} \exp - \{\alpha (x-u)\}$$

where $x$ is the level from which the excess operates and $u$ and $\alpha$ are values derived from the mean $(\bar{x})$ and variance $(s)$ of the sequence of observed extreme values, i.e. $u = \bar{x} - y(\bar{N})/\alpha$, $\alpha = \frac{\sigma(\bar{N})}{s}$ where $y(\bar{N})$ and $\sigma(\bar{N})$ are tabulated.

This formula was derived from the exponential class of functions which Gumbel has called Type I [3]. d'Hooge [4] has established this formula on a rigorous basis and has also given the results for Gumbel functions types II and III. Using a different approach which brings in the very general theorems of Franckx [5], Jung [2] has derived the same formula but with the extension that equality of numbers of claims in the successive intervals is not a necessary condition. It thus appears that provided the observed claims are properly adjusted for changes in the value of money the technique is on a sound theoretical footing although its practical utility has still to be established.

However, the vexed question arises as to which of the three types I, II, or III should be adopted as the assumed asymptotic distribution. In my original approach I adopted type I (exponential) as this included the log-normal distribution which well represents many of the non-life claims distributions and is only slightly less dangerous (in the Benktander sense [6]) than the Pareto distribution. On a Gumbel plot, i.e. the ordered claims plotted against their probability, the graph should be approximately linear. A departure from linearity could indicate that the type I distribution was unsuitable as a limiting distribution and a warning thus given that special care was needed.

In order to examine the nature of the estimate of the excess loss premium some experiments were made on the assumption that the underlying claim distribution was a normal curve of errors. This was deliberately chosen because the approach to the asymptotic form is slow (the formula is, of course, exact if the claim distribution is exponential).

As a first test the formula $P = \frac{1}{\alpha} e^{-y}$ can be examined against a known distribution; thus for examples of 10 from a $N(0,1)$ distribution we find $u$ from $\{1 - \Phi(u)\} = .1$, i.e. $u = 1.2816$ and $\varphi(u) = .17550$; this gives $x$ $(= n\varphi(u)) = 1.755$. We can then compare $\frac{1}{\alpha} e^{-\alpha(x-u)}$ with $10\,[\varphi(x) - x\{1 - \Phi(x)\}]$. The following table sets out a comparison of approximate values and the true values for various sample sizes and different excess limits.

The "excess loss premiums" are seen to be in all cases greater

| $x$ | $n = 10$ Approx. | $n = 10$ True | $n = 50$ Approx. | $n = 50$ True | $n = 100$ Approx. | $n = 100$ True | $n = 1{,}000$ Approx. | $n = 1{,}000$ True |
|---|---|---|---|---|---|---|---|---|
| 1.0 | .935 | .833 | 5.30 | 4.16 | 12.83 | 8.33 | 340.0 | 83.3 |
| 1.5 | .389 | .293 | 1.57 | 1.46 | 3.39 | 2.93 | 63.2 | 29.3 |
| 2.0 | .161 | .085 | .47 | .42 | .89 | .85 | 11.7 | 8.5 |
| 2.5 | .067 | .020 | .14 | .10 | .24 | .20 | 2.2 | 2.0 |
| 3.0 | .027 | .0038 | .042 | .019 | .062 | .038 | .40 | .38 |
| 3.5 | .010 | .0006 | .012 | .003 | .016 | .006 | .07 | .06 |
| u | 1.282 | | 2.054 | | 2.326 | | 3.090 | |

$$\text{Approx.} = \frac{1}{\alpha} e^{-\alpha(x-u)} \qquad \text{True} = n\,[\varphi(x) - x\{1\text{-}\Phi(x)\}]$$

than the true values. A reasonably close approximation is given if $x$ is not too far removed from $u$, particularly if $n$ is large. This suggests that provided the excess limit is within the range of the extreme values, i.e. that some claims will fall on the reinsurer, the method may give reasonable practical results.

Now the normal curve converges more rapidly than the exponential and the inference from the foregoing is that the behaviour for a curve which converges more slowly than the exponential will be such that the approximate values might be less than the true. A few test calculations on a log-normal distribution suggested this inference was justified, a warning that care will be needed for the practical distributions of non-life insurance.

A next experiment was made using the table of 25000 random normal deviates in Tracts for Computers No. XXV, these being based on the random numbers prepared by Kendall and Babington Smith, the object being to test an "excess premium" calculated from these values with the known true value. The maximum (positive) value for each of the 500 sets of 50 values was first found; various aggregations were also made, finishing with the 5 maximum values from 5 sets of 5,000 values. In studying these figures it was found that as the sample size increased (and the number of extremes decreased) the approximations deviated further from the theoretical value. This was finally traced to the fact that the number of deviates in the table for $x \geqslant 3.00$ was substantially greater than the expected number. The figures are given below as being of possible value in

assessing the use of this table for calculations in which the tail could be significant.

| $x$ | *Expected No.* | *Actual No.* |
|---|---|---|
| 3.00 | 33.7 | 41 |
| 3.10 | 24.2 | 31 |
| 3.20 | 17.2 | 25 |
| 3.30 | 12.1 | 23 |
| 3.40 | 8.4 | 17 |
| 3.50 | 5.8 | 9 |
| 3.60 | 4.0 | 6 |
| 3.70 | 2.7 | 1 |

The excess deviates were sufficient to produce a bias in the calculations. For the 5 samples of 5,000 the extreme values were 3.63, 3.48, 3.63, 3.68 and 3.91. The mean and standard deviations of these are 3.666 and .1392 respectively and the values of $y(\bar{N})$ and $\sigma(\bar{N})$ .459 and .802 respectively. These give $y = \frac{1}{.174}(x - 3.586)$ which leads to the following "excess loss premiums" for the levels mentioned for a total of 25,000 values:—

| *Excess level* | $\frac{1}{\alpha}e^{-y}$ | *Theoretical* | *Actual value in random numbers table* |
|---|---|---|---|
| 3.00 | 25.55 | 9.55 | 13.25 |
| 3.10 | 14.30 | 6.65 | 9.75 |
| 3.20 | 8.01 | 4.65 | 6.94 |
| 3.30 | 4.49 | 3.22 | 4.63 |
| 3.40 | 2.54 | 2.15 | 2.69 |
| 3.50 | 1.30 | 1.42 | 1.31 |
| 3.60 | .80 | 1.00 | .51 |
| 3.70 | .45 | .62 | .21 |

For values of $x$ greater than 3.20 the approximate values are reasonable approximations to the values derived from the actual observations in this region; below this value the approximate values become significantly greater than the actual values. The figures show that the approximate method is not a very satisfactory estimation of the true underlying values, but having regard to the bias in the random numbers this cannot be regarded as surprising.

If the foregoing experiment is interpreted as a practical case, the 5 yearly premium of 8.01 can be looked at as based on the 5 maximum claims in 5 successive years while the value of 6.94 is derived from the value of the 25 claims greater or equal to 3.20 occurring during this period.

To further test the suitability of non-life data for the technique a search was made for suitable fire statistics. The Property Insurance Fact Book 1960 [7], gives the larger fires occurring in the U.S.A. for the years 1949-1959 inclusive. The largest fire in each year was taken and a "Gumbel plot" showed a departure from linearity, suggesting that the data departed from the type I (exponential). Adjusting the amounts of the fires for the rise in building costs over the period produced a rather better result, although the plot was still non-linear, even allowing for some distortion arising from the "Livonia" fire. A further plot was made using the logarithm of the amount of fire, which would give a linear plot if the Gumbel type II limiting distribution applied. However, this overcorrected and the plot became concave instead of convex.

A similar experiment was tried on a series of figures relating to claims arising from a motor portfolio for the period 1940-1961. Near linearity of the Gumbel plot was obtained when the basic figures were corrected for cost of living changes (measured by retail price index) over the period. The results for the shorter period 1952-1961 did not lead to a linear graph and no improvement could be obtained by adjusting the basic figures for price index changes or by plotting the logarithm.

The correct interpretation of these preliminary experiments is uncertain, but the results of the fire data suggest that the proper distribution lies between the exponential and the Cauchy types. In other words the underlying claim distribution approaches the limit more rapidly than the exponential but less rapidly than the Pareto. The inference from the motor figures is that there is some lack of homogeneity, but otherwise the exponential is a reasonable limiting distribution.

As a further experiment some figures have been derived from the actual claims experienced from a portfolio of miscellaneous accident business over a period of five years. These derived figures are: —

| Year | *Claims in excess of 2000* | | | |
|---|---|---|---|---|
| 0 | 2656, | 3296 | | |
| 1 | 2299, | 4078, | 11418 | |
| 2 | 3076, | 3654, | 3840 | |
| 3 | 3412, | 5188, | 6664, | 6921 |
| 4 | 2249 | | | |

The net estimated premiums for excess loss cover at certain limits have been calculated from the 5 extreme claims by the formula $P = \frac{1}{\alpha} e^{-\alpha(x-u)}$ and the actual net cost found from the actual claims over the limits with the following results: —

| | Excess limit | | |
|---|---|---|---|
| | 2000 | 4000 | 6000 |
| Approx. annual net premium | 4050 | 2510 | 1550 |
| Net cost from 5 year claims | 6550 | 2850 | 2334 |

The values by the approximate formula are seen to be in all cases less than those derived from all the relevant claims. In the region of the average claim (4500) or of the average extreme claim (5500) the values are not widely dissimilar, but it is apparent that some care must be used if the approximate method is used (in default of other information). This result supports the suggestion made earlier in these notes that the approximate premiums may be underestimations if the underlying distribution converges more slowly than the exponential, which is suggested by the Gumbel plots referred to in the previous section. It would be of great interest to experiment on the basis of the Gumbel type II limiting distribution and see what sort of answers are derived. These could well be overestimations and provided a reasonably simple method could be found for the type II calculations much greater confidence could be placed on the use of extreme values in this particular problem as well as opening the way for other applications.

[1] BEARD, R. E.: ASTIN Bulletin, Vol. II, Pt. III, page 313, 1963.
[2] JUNG, J.: ASTIN Bulletin, Vol. III, Pt. II. *)
[3] GUMBEL, E. J.: Statistics of Extremes, Columbia University Press, New York, 1958.
[4] D'HOOGE, L.: ASTIN Bulletin, Vol. III, Pt. II. *)
[5] FRANCKX, E.: ASTIN Bulletin, Vol. II, Pt. III, page 415, 1963.
[6] BENKTANDER, G.: ASTIN Bulletin, Vol. II, Pt. III, page 387, 1963.
[7] Property Insurance Fact Book 1960, Insurance information Institute, New York.

*) These papers will appear in 1964.