Journal of Econometrics 23 (1983) 91–102. North-Holland

ON THE ESTIMATION OF LONG TAILED SKEWED DISTRIBUTIONS WITH ACTUARIAL APPLICATIONS

Robert V. HOGG and Stuart A. KLUGMAN

University of Iowa, Iowa City, IA 52242, USA

Very long tailed skewed distributions often arise in practice. In particular, we find that size-of-loss distributions in casualty insurance are mainly of this type. Compounding explains why many of these losses have approximate Pareto, generalized Pareto, Burr, and log-*t* distributions. An adaptation of the empirical mean residual life function helps the statistician select the correct model to fit in these cases. It is then discovered that minimum distance estimates, in particular that of Cramér–von Mises, and minimum chi-square estimates are extremely valuable and easy to use in the case of grouped data. Two substantial examples are given, one involving hurricane losses and the other dealing with malpractice claims.

1. Introduction

A problem of great interest to actuaries is the modelling of loss distributions. In order to set premiums, evaluate the effects of deductibles and limits and determine the impact of inflation, it is necessary to have information about the process producing the losses. The probability distribution of the loss variable is sufficient for answering these questions. In this paper we seek appropriate models for these distributions and methods of fitting these models. The overriding characteristic of these distributions is a heavy right tail (we assume all our models give probability zero to negative losses, so the left tail is of lesser importance). In section 2 we develop several heavy tailed models and in section 3 we give methods of estimation. Finally, in section 4, these methods are illustrated with two examples. These examples, and specific questions we hope to answer with our models, are detailed in the remainder of this section.

The first example concerns losses from hurricanes occurring from 1949 to 1980 as provided by the American Insurance Association. All values are in 1981 dollars and only those greater than 5,000,000 have been included. Table 1 gives the losses (ordered) along with values of $e_n(x)$, as defined in section 3. With a probability model we could estimate the frequency of losses in excess of a specified amount, or we could estimate the expected number of years between hurricanes causing a specified amount of damage. While these losses are not relevant to any one company, the example provides a good illustration of the loss process. The second data set includes malpractice

Table 1

Indexed hurricane losses, 1949–1980
(000 omitted).

Year	Loss	$e_n(x)$
1964	6,766	203,962
1968	7,123	209,775
1971	10,562	212,784
1956	14,474	215,610
1961	15,351	221,890
1966	16,983	227,853
1955	18,383	234,541
1958	19,030	242,557
1974	25,304	245,371
1959	29,112	251,225
1971	30,146	260,616
1976	33,727	268,210
1964	40,596	273,220
1949	41,409	285,379
1959	47,905	292,827
1950	49,397	306,669
1954	52,600	320,325
1973	59,917	331,420
1980	63,123	348,727
1964	77,809	356,311
1955	102,942	354,833
1967	103,217	381,832
1957	123,680	391,483
1979	140,136	409,121
1975	192,013	392,968
1972	198,446	429,483
1964	227,338	450,665
1960	329,511	398,277
1961	361,200	427,686
1969	421,680	440,647
1954	513,586	435,926
1954	545,778	538,312
1970	750,389	500,552
1979	863,881	774,119
1965	1,638,000	—

claims paid for insured hospitals in 1975. This was taken from *Medical Malpractice Closed Claims* 1975–1978, Vol. 2, No. 2, NAIC. Table 2 gives, for each range the number of claims, the average claim amount and the total losses. Once again, values of $e_n(x)$ are presented. Claims for which nó indemnity was paid have been ignored. The fitted model may be used to evaluate the effects of a policy limit or to price a layer of coverage. In addition, we will look at a similar set of data for 1978 to illustrate the effects of inflation on the model.

Table 2

Empirical distribution of amounts of indemnity paid by hospitals in 1975.

Amount of indemnity	Frequency	Cumulative frequency	Average	e_n[a]
1–999	465	1739	401	12,563
1000–1999	281	1274	1276	16,002
2000–2999	202	993	2348	19,453
3000–3999	102	791	3241	23,331
4000–4999	64	689	4195	25,750
5000–5999	78	625	5107	27,361
6000–6999	65	547	6213	30,254
7000–7999	59	482	7452	33,305
8000–8999	26	423	8213	36,887
9000–9999	20	397	9253	38,289
10,000–19,999	164	377	13137	39,307
20,000–29,999	67	213	22721	57,156
30,000–39,999	24	146	32409	72,137
40,000–49,999	19	122	42603	75,854
50,000–59,999	15	103	52539	79,366
60,000–69,999	11	88	61032	82,461
70,000–79,999	13	77	73115	85,094
80,000–89,999	11	64	83091	90,543
90,000–99,999	2	53	90000	98,693
100,000–199,999	31	51	120107	92,563
200,000–299,999	13	20	234500	104,870
300,000–399,999	5	7	354564	135,557
500,000–999,999	2	2	638040	138,040

[a] e_n is computed at the lower limit of the range.

2. Development of models

With data like that given in section 1, many statisticians might first try to fit the exponential distribution, with p.d.f.

$$f(x\mid\theta)=\theta e^{-\theta x}, \qquad 0<x<\infty.$$

This would be a reasonable first guess, but it actually does not provide a very good fit because the tails are not thick enough.

However, suppose we can assume that each loss actually has an exponential distribution, but that the values of the parameter θ change as we go from loss to loss, possibly due to the fact that different parts of the country, seasons, or, in example 2, hospitals, have different 'safety' parameters. Moreover, let us say that the distribution of the various θ values can be described by the p.d.f. $g(\theta)$, which we assume is continuous on $\theta>0$.

Thus, in fact, the p.d.f. of the loss X is the mixture of exponential p.d.f's, namely,

$$h(x)=\int_0^\infty [\theta\,\mathrm{e}^{-\theta x}]g(\theta)\,\mathrm{d}\theta.$$

By taking $g(\theta)$ to be a gamma p.d.f., one which provides a great deal of flexibility, we obtain

$$h(x)=\int_0^\infty [\theta\,\mathrm{e}^{-\theta x}]\left[\frac{\lambda^\alpha\theta^{\alpha-1}\,\mathrm{e}^{-\lambda\theta}}{\Gamma(\alpha)}\right]\mathrm{d}\theta=\frac{\alpha\lambda^\alpha}{(\lambda+x)^{\alpha+1}},\qquad 0<x.$$

This is one form of the *Pareto* p.d.f. and it has distribution function

$$H(x)=1-\frac{\lambda^\alpha}{(\lambda+x)^\alpha},\qquad 0<x,$$

and mean $\lambda/(\alpha-1)$. It should be noted that the Pareto p.d.f. has a thicker right tail than does the original exponential p.d.f.

An easy generalization of this follows by replacing the original exponential distribution with a gamma one with p.d.f.

$$f(x\,|\,\theta)=\frac{\theta^k x^{k-1}\,\mathrm{e}^{-\theta x}}{\Gamma(k)},\qquad 0<x.$$

If $g(\theta)$ is gamma, then the mixture of $f(x\,|\,\theta)$ results in

$$h(x)=\int_0^\infty\left[\frac{\theta^k x^{k-1}\,\mathrm{e}^{-\theta x}}{\Gamma(k)}\right]\left[\frac{\lambda^\alpha\theta^{\alpha-1}\,\mathrm{e}^{-\lambda\theta}}{\Gamma(\alpha)}\right]\mathrm{d}\theta$$

$$=\frac{\Gamma(k+\alpha)\lambda^\alpha x^{k-1}}{\Gamma(k)\Gamma(\alpha)(\lambda+x)^{k+\alpha}},\qquad 0<x.$$

We call this a generalized Pareto distribution and note that it is a member of the Pearson Type VI family. As a special case it includes the p.d.f. of the well-known F-distribution, but unfortunately the distribution function cannot be written in closed form and that makes it less useful to us.

Possibly a more realistic generalization would be one that starts with a Weibull fit to the data rather than an exponential or gamma. That is, we first assume the loss X has a Weibull distribution with p.d.f.

$$f(x\,|\,\theta)=\theta\tau x^{\tau-1}\,\mathrm{e}^{-\theta x^\tau},\qquad 0<x.$$

Of course, this is exponential when $\tau=1$; but, with the flexibility of the parameter τ, it provides a reasonable fit to many data sets. Mixing (or compounding) as before gives the *Burr* p.d.f. [Burr (1942), Burr and Cislak (1968)], namely,

$$h(x)=\int_0^\infty [\theta\tau x^{\tau-1}\,\mathrm{e}^{-\theta x^\tau}]\left[\frac{\lambda^\alpha\theta^{\alpha-1}\,\mathrm{e}^{-\lambda\theta}}{\Gamma(\alpha)}\right]\mathrm{d}\theta$$

$$=\frac{\alpha\tau\lambda^\alpha x^{\tau-1}}{(\lambda+x^\tau)^{\alpha+1}},\qquad 0<x.$$

This has distribution function

$$H(x)=1-\frac{\lambda^\alpha}{(\lambda+x^\tau)^\alpha},\qquad 0<x.$$

This distribution may also be derived through the transformation, $Y=X^{1/\tau}$, of a Pareto random variable, X.

One other family of distributions, along with its compounded generalization, that should be mentioned is the *lognormal* distribution with p.d.f.

$$f(x\,|\,\theta)=\frac{1}{x}\sqrt{\frac{\theta}{2\pi}}\,\mathrm{e}^{-\theta(\ln x-\mu)^2/2},\qquad 0<x.$$

That is, we assume that $\ln X$ has a normal distribution with mean μ and variance $\sigma^2=1/\theta$. If this is compounded with a gamma p.d.f., we obtain

$$h(x)=\int_0^\infty\left[\frac{1}{x}\sqrt{\frac{\theta}{2\pi}}\,\mathrm{e}^{-\theta(\ln x-\mu)^2/2}\right]\left[\frac{\lambda^\alpha\theta^{\alpha-1}\,\mathrm{e}^{-\lambda\theta}}{\Gamma(\alpha)}\right]\mathrm{d}\theta$$

$$=\frac{\lambda^\alpha\Gamma(\alpha+\tfrac{1}{2})}{\sqrt{2\pi}\,\Gamma(\alpha)x[\lambda+(\ln x-\mu)^2]^{\alpha+1/2}},\qquad 0<x.$$

That is, $\ln X$ is distributed much like a Student's t and is called the *log-t*. The distribution function of this cannot be found in closed form either.

Of course, the *gamma* distribution with parameters α and λ and p.d.f.

$$\frac{\lambda^\alpha x^{\alpha-1}\,\mathrm{e}^{-\lambda x}}{\Gamma(\alpha)},\qquad 0<x<\infty,\quad \alpha>0,\quad \lambda>0,$$

fits many data sets quite well. In addition, if $\ln X$ does not have a symmetric distribution, then neither the lognormal nor the log-t will fit very well, but the *log-gamma* distribution might be very satisfactory. That is, the distribution of X might be such that $\ln X$ has a gamma distribution with parameters α and λ.

In the next section, we make suggestions about selecting one or more of these distributions as a model.

3. Selection of model

For years, demographers have found the 'mean residual life' to be a valuable function. That is, if X is the length of life with p.d.f. $k(x)$, then

$$e(x)=E(X-x\,|\,X\geqq x)=\int_x^\infty (w-x)\frac{k(w)}{\int_x^\infty k(w)\,\mathrm{d}w}\,\mathrm{d}w$$

is the mean residual life. In actuarial literature this is $\mathring{e}_x$, the complete expectation of life. However there is no reason why X cannot be a loss and $k(x)$ the p.d.f. of a loss distribution. Hall and Wellner (1981) made some observations about the shape of $e(x)$ for some of the distributions in section 2. For our purposes, a few of these might be best described in fig. 1.

This suggests that we do the following in selecting a model or two. Using the data, compute the empirical $e_n(x)$, where n is the sample size; this is

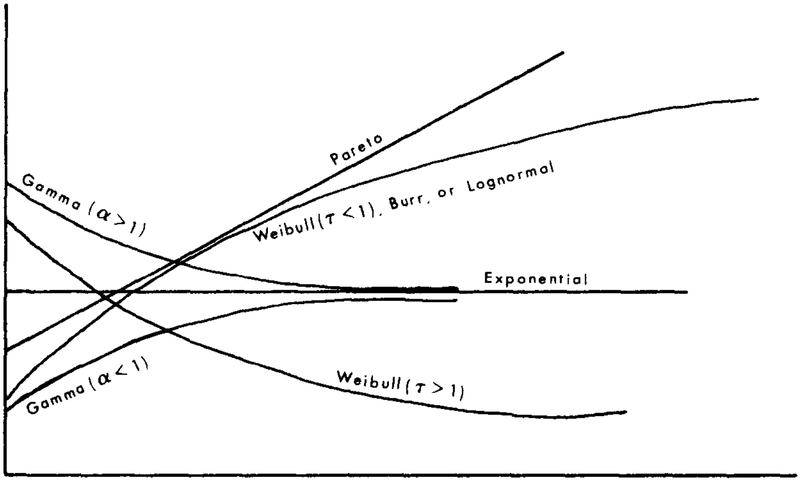

Fig. 1. Mean residual life functions.

rather easy as it is just an averaging process. For illustration, take the data in table 1. We obtain, with $n=35$,

$$e_{35}(863{,}881)=1{,}638{,}000-863{,}881=774{,}119.$$

$$e_{35}(750{,}389)=(1{,}638{,}000+863{,}881)/2-750{,}389=500{,}882,$$

and so on. The values of $e_{35}(x)$ at the data points are given in table 1 and are plotted in fig. 2 with a smooth curve drawn through them.

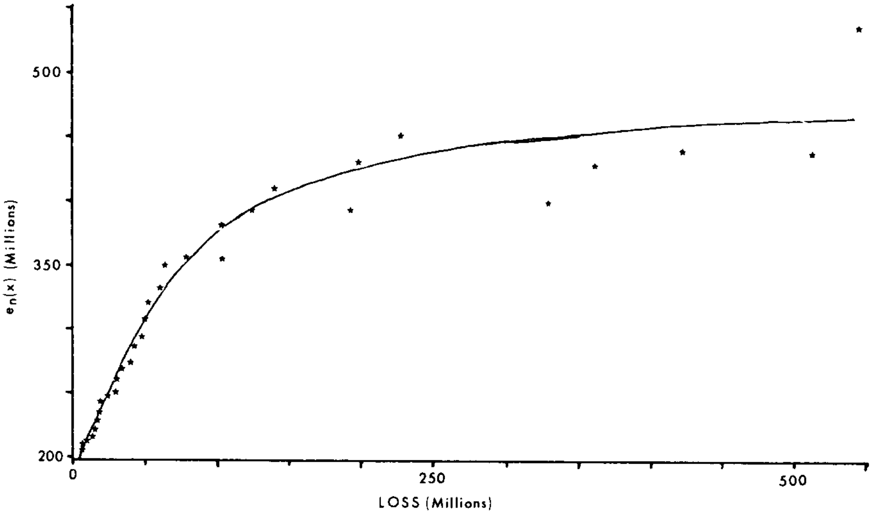

Fig. 2. Empirical mean residual life hurricane data.

From this point it appears that the lognormal or Weibull ($\tau<1$) or possibly a gamma ($\alpha<1$) may provide good models. We may also consider the Burr distribution as a generalization of the Weibull if none of the two parameter families are satisfactory.

The malpractice data in table 2 show a similar pattern leading us to consider the same distributions as in the first example. We now turn to three methods of parameter estimation.

The first is maximum likelihood estimation. When the individual data points are available, the quantity to maximize is

$$L=\prod_{i=1}^{n} h(x_i).$$

When the data are grouped (as in the second example) we cannot compute the exact moments nor the regular likelihood function. We certainly prefer maximum likelihood estimation, if at all possible, as it is well known that those estimators are better in the sense of BAN (best asymptotic normal) than other estimators. However, the best that we can provide is a likelihood function based upon the probabilities of the k cells, namely,

$$L=\prod_{i=0}^{k-1}[H(x_{i+1})-H(x_i)]^{f_{i+1}},$$

where f_{i+1} is the frequency of the cell with boundaries $(x_i, x_{i+1}]$, $i=0,1,2,\ldots,k-1$. It is now clear why it is desirable to have the distribution function H in closed form. Of course, H and thus L are functions of the parameters associated with the distribution, and we maximize L with respect to the parameters.

A computationally more efficient procedure is based on the concept of minimum distance, which is reviewed nicely in a bibliography by Parr (1981). It is based on the minimization of a measure of the distance between the empirical distribution function, $H_n(x)$, and the distribution function $H(x)$. An advantage over maximum likelihood is that this approach is unchanged when the data are grouped, we merely use $H_n(x)$ only at the group boundaries.

One popular measure is that of Cramér–von Mises which, in the grouped data case, is

$$K=\frac{1}{n}\sum_{i=0}^{k}[H_n(x_i)-H(x_i)]^2 w(x_i),$$

where $w(x_i)$ is the weight at point x_i. Popular weight functions are the uniform one, $w(x_i)=1$, and that of Anderson–Darling, $w(x_i)=1/[H(x_i)(1-H(x_i))]$. However, a particular situation might suggest other weights, giving large weights to the most important points. Since $H(x)$ is a nonlinear function of one or more parameters, the minimization of K is simply a nonlinear weighted least squares problem as $H_n(x_i)$ is the known cumulative frequency at x_i.

Finally, there is a third procedure that is extremely good for grouped data, namely minimum chi-square estimation, which is explained very well in an article by Moore (1978). If the frequency of the ith cell is

$$f_i=n\{H_n(x_i)-H_n(x_{i-1})\},$$

then the chi-square goodness-of-fit statistic is

$$Q=\sum_{i=1}^{k}\frac{[f_i-n\{H(x_i)-H(x_{i-1})\}]^2}{n\{H(x_i)-H(x_{i-1})\}}.$$

These minimum chi-square estimators are found by selecting those values of the parameters in the distribution function H that minimize Q. Note that the parameters appear in the denominator of Q too; and this complicates the problem, the resolution of which can be handled one of two ways:

(1) Initial estimates of the parameters can be inserted in the denominator and then it becomes a problem in nonlinear weighted least squares, iterating for a number of steps.

(2) The frequency f_i can be used to approximate the denominator, again resulting in a nonlinear weighted least squares problem. With this modification of Q, these modified estimators have the same asymptotic properties as the original minimum chi-square estimators.

For a detailed illustration of an application of this method to mortality data using suggestion (2), see Cramér and Wold (1935).

We now turn to the analysis of our two data sets.

4. Two examples

We are now prepared to fit models to our two data sets. The hurricane losses were fit by both maximum likelihood and minimum distance with Anderson–Darling weights. With only 35 observations, it did not seem wise to group in order to fit by the chi-square method. The results for the distributions suggested by the mean residual life function are given in table 3. Due to the non-reporting of losses below 5,000,000, we must use a conditional modification of the probability function. For m.l.e., we use $h^*(x)=h(x)/(1-H(t))$, and for m.d. we use $H^*(x)=(H(x)-H(t))/(1-H(t))$ where $t=5{,}000$ and all functions are for the random variable divided by 1000. The results in table 3 are parameters for the loss variable without truncation and after dividing by 1000.

With the lognormal distribution being best by m.d. and the Weibull best by m.l.e., it is difficult to choose between them. We can, however, eliminate the Burr distribution as it provides little improvement over the Wiebull. In

Table 3

Parameter estimates for hurricane losses.

Distribution	Method		Parameters		K	$-\ln L$
Lognormal	m.l.e.		$\mu=11.0456$	$\sigma=1.6028$	7.1338	454.18
	m.d.		$\mu=10.887$	$\sigma=1.7277$	4.6301	454.30
Weibull	m.l.e.		$\theta=0.0027037$	$\tau=0.51907$	7.0914	454.11
	m.d.		$\theta=0.0040616$	$\tau=0.49145$	5.0328	454.17
Burr	m.l.e.	$\alpha=3.7697$	$\lambda=6400.3$	$\tau=0.65994$	7.0133	454.27
	m.d.	$\alpha=3.6595$	$\lambda=3501.2$	$\tau=0.61914$	4.9279	454.36

table 4 we show the empirical and fitted distributions for the two ‘winners’. Our desire to have outstanding fit at the large values makes the lognormal distribution as estimated by m.d. a reasonable choice.

Table 4

Empirical and fitted distributions — hurricane data.

x	Empirical	Lognormal-m.d.	Weibull-m.l.e.
6,766	0.02857	0.03349	0.03751
7,123	0.05714	0.03995	0.04434
10,562	0.08571	0.09707	0.10117
14,474	0.11429	0.15258	0.15260
15,351	0.14296	0.16387	0.16280
16,983	0.17143	0.18393	0.18075
18,383	0.20000	0.20021	0.19521
19,030	0.22857	0.20747	0.20163
25,304	0.25714	0.27039	0.25690
29,112	0.28571	0.30314	0.28560
30,146	0.31429	0.31145	0.29289
33,727	0.34286	0.33854	0.31672
40,496	0.37143	0.38427	0.35673
41,409	0.40000	0.38921	0.36169
47,905	0.42857	0.42578	0.39461
49,397	0.45714	0.43351	0.40163
52,600	0.48571	0.44937	0.41610
59,917	0.51429	0.48222	0.44644
63,123	0.54286	0.49534	0.45870
77,809	0.57143	0.54747	0.50834
102,942	0.60000	0.61499	0.57521
103,217	0.62857	0.61561	0.57584
123,680	0.65714	0.65717	0.61871
140,136	0.68571	0.68469	0.64791
192,013	0.71429	0.74905	0.71894
198,446	0.74286	0.75533	0.72609
227,338	0.71143	0.78027	0.75490
329,511	0.80000	0.84019	0.82668
361,200	0.82857	0.85314	0.84263
421,680	0.85714	0.87330	0.86770
513,586	0.88571	0.89602	0.89616
545,778	0.91429	0.90238	0.90413
750,389	0.94286	0.93104	0.93958
863,881	0.97143	0.94141	0.95199
1,638,000	1.00000	0.97401	0.98671

The second data set is amenable to all three estimation methods. The results, for five distributional models are given in table 5. The clear winner is the generalized Pareto distribution, but it is not significantly better than the Pareto model. Parsimony suggests use of the Pareto.

Analysis of the 1978 malpractice data produced similar results. Again, the three parameter generalizations were not significantly superior to the Pareto.

Table 5

Parameter estimates for malpractice losses.

Distribution	Method	Parameters			Value[a]
Weibull	m.d.	$\theta = 0.022467$		$\tau = 0.44673$	0.13955
	m.c.	$\theta = 0.013230$		$\tau = 0.49363$	258.58
	m.l.e.	$\theta = 0.011688$		$\tau = 0.51199$	4315.8
Lognormal	m.d.	$\mu = 7.8398$		$\sigma = 1.8532$	0.025847
	m.c.	$\mu = 7.9513$		$\sigma = 1.7537$	65.680
	m.l.e.	$\mu = 7.9215$		$\sigma = 1.7412$	4232.9
Pareto	m.d.	$\alpha = 0.96114$		$\lambda = 2519.6$	0.013123
	m.c.	$\alpha = 0.95931$		$\lambda = 2593.0$	54.758
	m.l.e.	$\alpha = 0.99679$		$\lambda = 2669.2$	4229.5
Burr	m.d.	$\alpha = 0.97094$	$\lambda = 2435.5$	$\tau = 0.99395$	0.013115
	m.c.	$\alpha = 0.88675$	$\lambda = 3192.7$	$\tau = 1.0413$	54.331
	m.l.e.	$\alpha = 0.96517$	$\lambda = 2851.0$	$\tau = 1.0145$	5229.5
Generalized Pareto	m.d.	$\alpha = 0.95694$	$\lambda = 2432.5$	$k = 1.0216$	0.013103
	m.c.	$\alpha = 0.92684$	$\lambda = 2116.1$	$k = 1.1126$	53.967
	m.l.e.	$\alpha = 0.96562$	$\lambda = 2219.0$	$k = 1.1015$	4229.4

[a] K, Q or $-\ln L$ as appropriate.

A comparison of the parameters, as fitted by m.l.e. is:

	1975	1978
α	0.99679	1.0123
λ	2669.2	4030.0

It is interesting to note that the only major difference is in λ. We note that if a Pareto random variable is multiplied by a constant, say r, the result is a new Pareto variable with $\lambda_{\text{new}} = r\lambda_{\text{old}}$. This suggests that in the malpractice data, the 1978 experience is from a random variable which is 4030.0/2669.2 = 1.51 times that which was observed in 1975. This corresponds to an annual rate of increase of 14.6%. We note that this increase is uniform throughout the distribution and not just with respect to the 'average' claim.

We close by observing that for $\alpha < 1$ the mean does not exist. This is not a problem as malpractice insurance most always has an upper limit on the amount of loss paid. Of course, the expected amount paid by the insurance will always exist in this case.

5. Conclusions

We are very encouraged by our initial studies of long tailed skewed distributions, particularly in applications to actuarial work. We can find suitable distributions through compounding, and the empirical mean residual life then suggests certain of these models to use in various cases. Extremely good fits can be made using certain minimum distance estimators, in particular, that of Cramér–von Mises, and minimum chi-square estimators, as well as modified maximum likelihood estimators. Our experience has been that minimum distance estimates are the easiest to obtain (although numerical methods are still required) while maximum likelihood estimates are seemingly the most difficult to compute. We recommend beginning with m.d. then proceeding to m.c. and finally m.l.e., using the results of one method as starting values for the next.

Of course, the estimators of the parameters could be used to estimate functions of the parameters, like a tail-end probability. An *approximate* error structure of these estimators can be found by using the approximate error structure of the estimators of the parameters, which is available in the minimum chi-square case; see Moore (1978). Also we know the approximate variances and covariances of the maximum likelihood estimators; either of these could also be used with minimum distance estimators since they are so similar. Granted that there is much in the way of approximating in these suggestions, but it would at least provide an indication of how worthwhile each estimate is.

Another question that frequently arises in these skewed cases concerns the use of robust methods, primarily in the treatment of outliers. Of course, if the goal is the estimation of the total cost (or loss), as it might be in hospitalization insurance, we definitely need an estimate of the mean and thus outliers cannot be down weighted. On the other hand, we might use robust methods in getting estimates of other characteristics, such as percentiles. Then too, if we really believe that we truly have contamination (a mixture of a basic distribution and at least one other contaminating distribution), we could estimate the basic distribution using robust methods. But, these and other ideas must be investigated in other studies.

References

Burr, I.S., 1942, Cumulative frequency functions, Annals of Mathematical Statistics, 215–232.

Burr, I.S. and P. Cislak, 1968, On a general system of distributions, I: Its curve-shape characteristics, II: The sample median, Journal of the American Statistical Association, 627–635.

Cramér, H. and H. Wold, 1935, Mortality variations in Sweden: A study in graduation and forecasting, Skandinavisk Aktuarie Tidskrift, 161–241.

Hall, W.J. and J.A. Wellner, 1981, Mean residual life in: M. Csörgö, ed., Statistics and related topics (North-Holland, Amsterdam) 169–184.

Parr, W.C., 1981, Minimum distance estimation: A bibliography, Communications in Statistics A, 1205–1224.

Moore, D.S., 1978, Chi-square tests, in: Robert V. Hogg, ed., Studies in statistics (Mathematical Association of America, Washington, DC) 66–106.