DEALING WITH UNCERTAINTY IN TAILS

presents:

Comparison of estimation on distribution assumptions

APRIL 28, 2017

M&A SOLUTIONS
Hong Kong

Contents

Section 1 - Introduction

Current implementation of RBC framework in HK leaves out catastrophe modelling because of absence of resources and experience to implement it. In order to help local companies to bridge that gap, we offer an insight into potential ways to estimate worst case outcomes for unlimited exposures, and thereby try to get a sense of what a catastrophic claim may look like.

With increasing amount of data we are now in a world where risks can be split into increasingly minute categories. Whether it's pleasure craft portfolio, or commercial reinsurance, with few data points there are increased uncertainties with regards to the claim outcomes. But even without that, in our work actuaries are regularly faced with censored distributions - from deductibles, to limits, from capping to cupping.

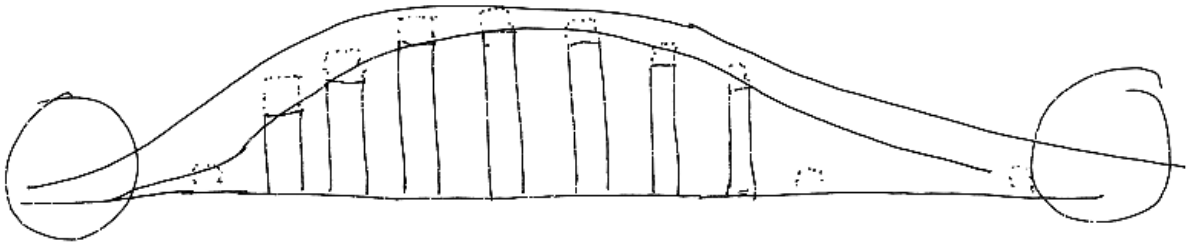
Recent research into the subject has yielded exploration of robotic reserving [**Robotic Reserving – Are We There Yet**], individual stochastic model creation [CR Larsen, 2017]. However, none of these papers addressed the inherent censoring that is present when so few data points exist.

Maximum likelihood principle is ideally suited to operate in those situations, unlike method of moments, which does not generate the most efficient estimators of the unknown parameters. This is particularly true when the distributions have non-neutral skew and kurtosis.

To simplify the idea of the expectation maximisation (EM) algorithm beyond permissible measure is to say that "the algorithm takes the change from one probability column of the histogram to another and tries to find parameters of a distribution that would fit these changes well". This explanation is riddled with problems but is clear enough for a practitioner. The original paper introducing the algorithm comes from Dempster et al. (1977) [Marginal pdf to fit to distribution]

One of the attractive features of ML estimators in trying to assess the underlying distribution of a particular values is its lack of conservatism, it is true that for large uncertainty around the parameters ML estimator has twice the variance of other estimators [example 18.5 Kendall 2A p57]

Our approach also helps to combat anchoring bias by providing practitioners a set of tools to reassess anchoring bias (p5) from the previous results (where the estimate was made based on existing data, and future estimates are made to insure consistency with the previous estimates, instead of assessing the emerging distribution of outcomes). In more simplistic terms, when you are facing more extreme outcomes that fall within the same range, it typically implies that the underlying distribution is probably more volatile than originally assessed (standard and trending outside of you insurance window). [Trends changes can be unclear like increasing variance isn't identical to trending mean]



Although in ideal world we would be able to test the estimations we make on new data, or at least set aside some portion of the data to test the fit, we typically lack sufficient data to split the observations into test and sample set. Hence, we put forward an approach to think about the uncertainty surrounding the models by fitting several models to the data.

The paper is structured as follows: in section 2 basic methodology and EM algorithm, Pearson's distributions, and Extreme value estimation are outlined. In section 3 We describe its implementation and adjustments that we make to account for categorical variables, and variables with limited boundary values. In section 4 we show how the method can be applied to 3 datasets. In section 5 we analyse the results. In section 6 we summarise our work and explore future developments.

Section 2 - Basic Methodology

When we talk about censoring it is important to know that it's not just deductibles and excesses we speak off, but also lack of availability of the whole information set limited by the newness of portfolio or claim type.

We've attempted to adopt a comprehensive approach to the analysis by breaking down the problems with the current approach(es) and trying to address them.

2.1 Why Pearsons family of distributions?

Most of the back-of-the envelope approaches to tail estimation center around distributions where skew and kurtosis cannot be set by the user. By moving away from more commonly accepted ones towards Pearsons family, we've attempted to address the issue of being able to explicitly adjust the skew and kurtosis of the severity distributions.

The reasons Pearson's offer lucrative set of distributions is because they include fat tailed distributions such as Levy distribution (type V), Cauchy, t-distributions (both two-tailed type IV distributions) and their shapes are far more varied than generalised Pareto which is extensively used for tail modelling but limited to skewed distribution. The reasons why Pearson's sometimes doesn't fare much worse than Generalised Extreme Value, as we will show, is because it does not tie the numbers to any particular shape of the distribution, and attempts to estimate them outright. Although one may argue that this makes for an unappealing feature of the method, we do not

believe this to be the case - since this leaves the flexibility to create any pdf shape that the data fits best.

Ideally, we would want to use an EM algorithm to estimate the key parameters of the distribution. However, as of now the approach is not available, and we tried to shy away from building new tools.

The method used to estimate the key parameters for the Pearson's distributions was a maximum likelihood method which addresses some of the concerns of the missing data that we've tried to capture for the two other methods. Further discussion of the method can be found in section 3.1.

To quote directly from R documentation of the creator: "First, the empirical moments of the input vector are calculated. In the second step, the moments are altered, such that the moment restrictions for the current sub-class are fulfilled (if necessary), and the method of moments estimator is calculated to obtain starting values for the optimizer. In the last step, the starting values are adjusted (if necessary) in order to assure that the whole sample lies in the support of the distribution"

2.2 Why EM algorithm on mixed distributions?

EM algorithm is brilliant it works on incomplete datasets, by working with the relationships of the existing pdf values. There are a number of papers explaining it's elegance, and we strongly recommend to anybody with similar inclinations to read them (please see Appendix 1). We will test several distributions with EM algorithm and try to determine the best one to use.

The EM algorithm made up of two steps (E for expectation, M for maximisation) acts as follows:

In step 1: We estimate the missing data, on the currently available data: assume that $Q(\emptyset'|\emptyset_k) = \boldsymbol{E}\{\boldsymbol{log}\ \boldsymbol{L}(x|\emptyset')|\ y,\emptyset_k\}$ exists for all **x** and $\emptyset$, where L is the likelihood to be maximised. Let $\emptyset_k$ denote the estimate of $\emptyset$ obtained on the kth iteration. Then we progress as follows :

Evaluate $\boldsymbol{Q}(\emptyset'|\emptyset)$, which is the conditional expectation taken over the unknown (missing) elements of **x**.

In step 2 we maximise the likelihood under assumption that the missing data is known: We determine$\emptyset_{(k+1)}$ such that the value of $\emptyset$maximises $Q(\emptyset'|\emptyset_k)$

Inserting theta into the probabilities allows us to optimise how much of the distribution is "missing".

We are faced with a situation where only part of the density function is known (the lower observations are censored because of deductible), higher may have not been observed yet. In this situation we neither know the parameters of the distribution, nor its parameters. One of the main methods advantages is that it requires only the gradient.

The tail distributions carries little correlation to the body of the distribution, so if this principle is extended beyond simple separation into body and tail, and into 'normal' tail and 'extreme' tail mixture distributions with multiple means start to make sense. For this analysis (and partially due to low data availability) we've limited the number of distributions to 2. In larger datasets, there is no reason to set this number at 2. The idea is to place reliance on the ability of EM algorithm to discern between two underlying datasets.

We've performed this analysis for both normal and lognormal distributions for severity. As expected, even with allowance for mixture distributions normal distributions proved to a poor proxy for the tail of severity, however lognormal fittings created mostly far more reasonable values.

2.3 Why Extreme Value Theory?

Not every large claim in the portfolio will fall in the storyline of extreme value, and hence to some degree Generalised Extreme value approach (GEV) serves as an indicator of a conservative approach to estimation of severity. We have used the GEV as the approach allowed us to forgo making assumptions about the data. Generalized Pareto can be used instead, by practitioners with more insight into the process.

Section 3 - Implementation

In working with the data we have made extensive use of several excellent packages: "PearsonsDS", "extRemes", and "mixtools", which allows users to customised distribution based on a few inputs.

3.1 PEARSONS Family of Distributions

The packages were selected according to their capability to fit distributions to long tail of parameters. Pearsons was chosen because of the ability to estimate an entire family of distributions based on few inputs, extRemes, because of its Extreme Value capability, and mixtools for the gaussian mixture distributions, which also extended to lognormal models.

Additionally, we've used mixtools to estimate mix gaussian models - as Rekik and co. paper demonstrates, they offer an excellent fit to distributions which have two or more "underlying drivers" so to speak.

In the three examples we have used, all of them had specific challenges tied with the data. The first had few data points, the second had extreme data points, and the third estimate had to be initiated using moments of distribution derived from industry data. More sophisticated approach would have looked at the prior and posterior distribution of the data (I've outlined the results in Appendix 3)

To bypass moments estimation, I've used Maximum Likelihood (ML) estimation for Pearson's (note that implementation of that algorithm essentially precludes from certain distributions (0, II, III, V, VII)) from being selected (REF - because those distribution represent a fixed relationship between parameters, vs a range of relationships), however this has not been the result for this analysis, as we've generated both III and V type of distribution depending on different random seed.

3.2 EM and mixture distributions

For Gaussian Mixture estimates, we have used EM algorithm, which, as ML estimation process reduces reliance on absent data. More information on estimating cumulants under ML can be found in Appendix 2.

Note, that for lognormal model the EM algorithm could not be initiated without providing the method with a guesstimate of the probably means. We have used the moments derived from aggregated data (See Appendix 4).

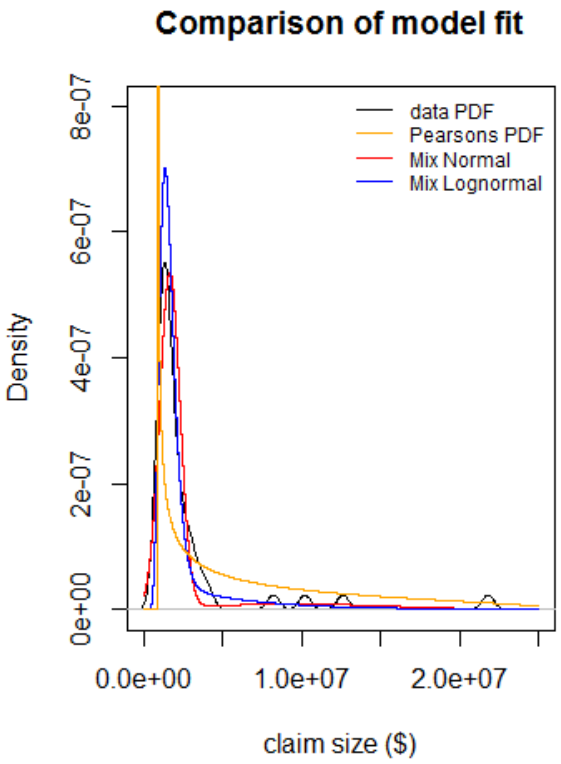
3.3 Extreme Value distributions

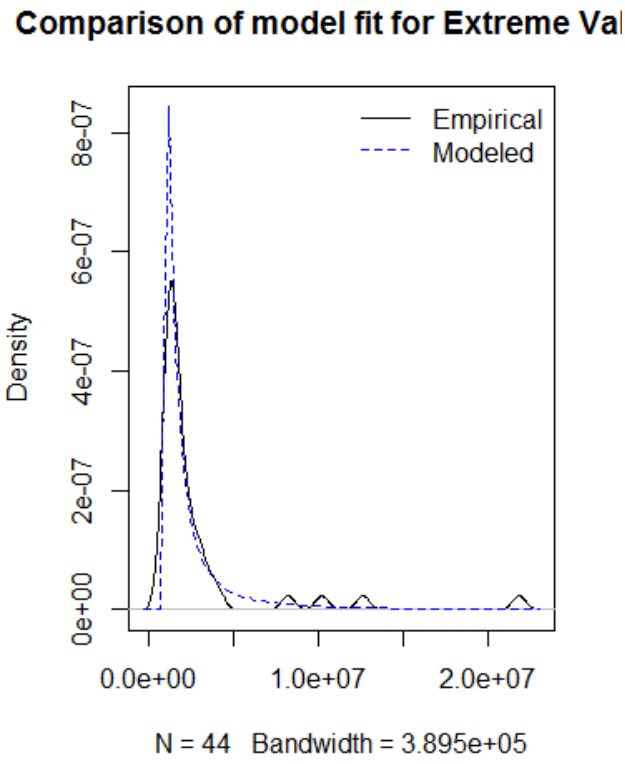
For Extreme Value estimation I've used the embedded default values of the algorithm, which fit a Generalised Extreme Value distribution to the data. Default methodology is similar to Pearson's approach in that it involves ML estimation procedure, and then decides on the distribution (Gumbel, Frechet, or Weibull) based on the parameters generated. Here my assumption was that all of the data points are large claims, which is not unreasonable given the description of the circumstances for each dataset.

Section 4 - Comparison of approaches

There are a number of different ways companies can account for the severity of their claims (large or otherwise). As the analysis above demonstrates, depending on the approach the tail percentile shifts significantly. This uncertainty is something that actuaries dealing with the subject should reflect upon, and draw conclusions. Not every dataset will generate reasonable tail values for lognormal data, or even more exotic pearson's distributions. We use the graphs illustrated below to summarise our observations by dataset:

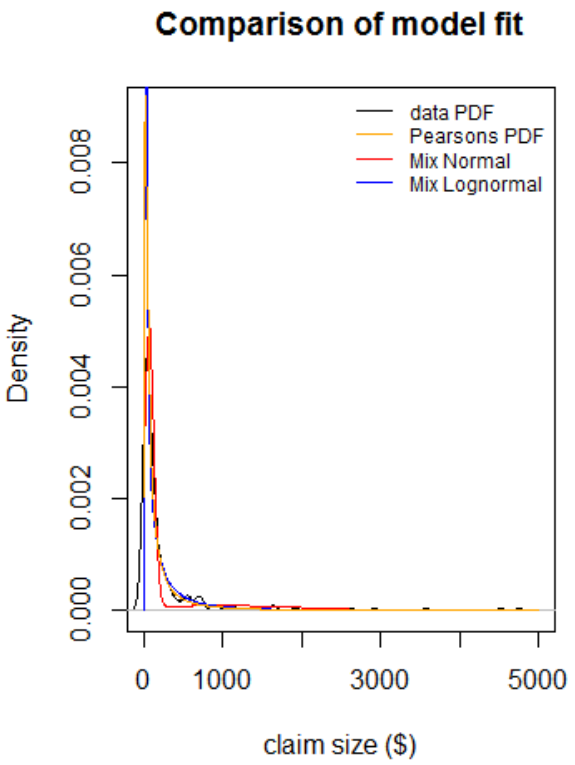
4.1 Dataset 1: 44 values, representative of "other category" of large claims

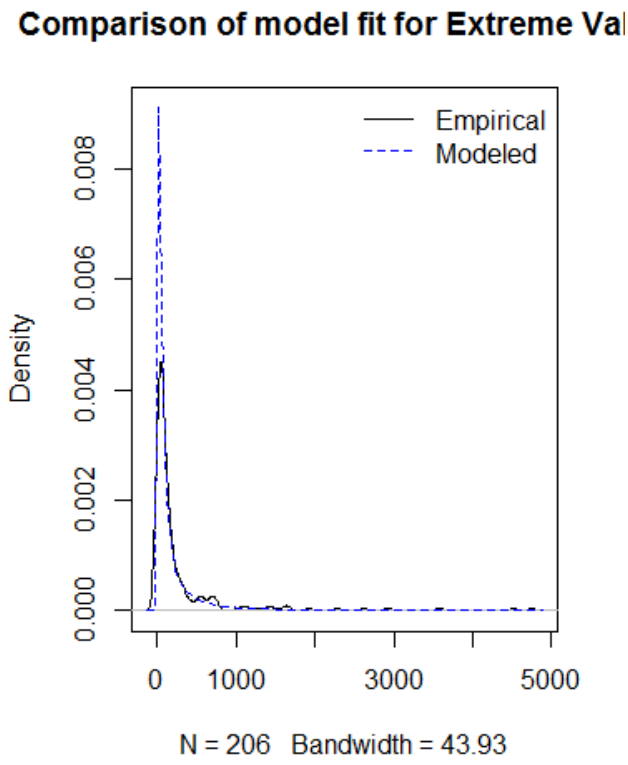




As can be observed that the dataset is scarce, and contains a significant outlier. The outlier will be common feature between first two datasets, but the absence of significant data to calibrate the tail presents a significant challenge to Pearson's distribution, although the it is able to cope with the far better than the remaining options. Lognormal distribution and the EVT outcomes have somewhat similar shapes, however Lognormal does worse on the tail outcomes. Mixed normal distribution performs second worst after Pearson's.

4.2 Dataset 2: 206 values, representative of large losses to an entire portfolio over a period spanning 20 years



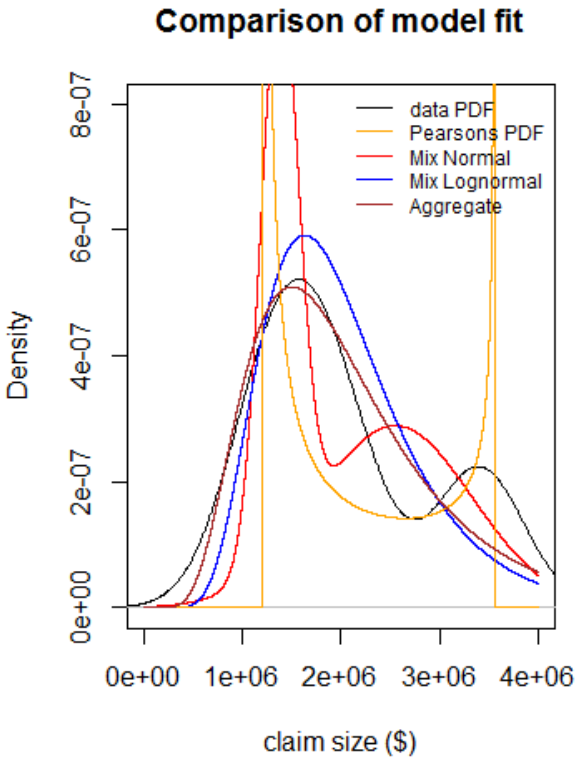


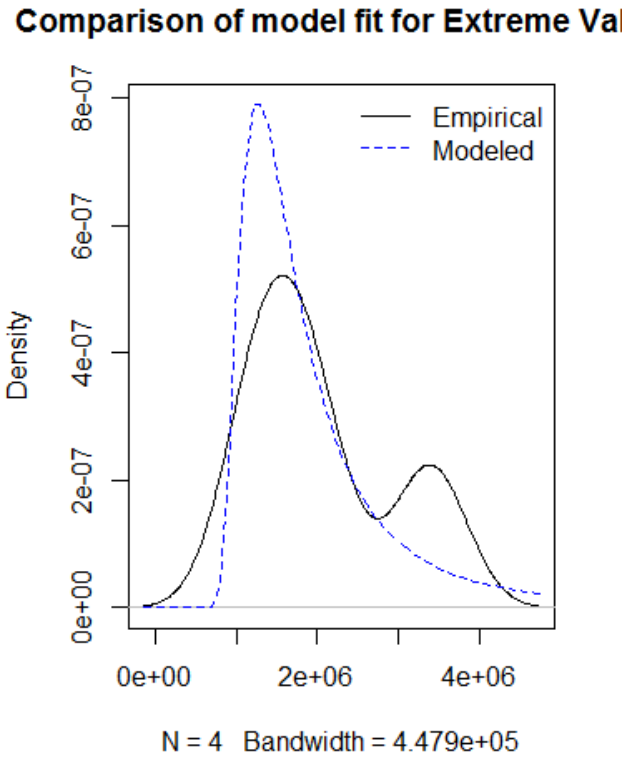
The presence of a number of datapoint helps Pearsons distribution to perform better here - it's fitting is comparable to the EV outcomes. Both Normal and Lognormal distribution struggle to accommodate both the long tail and the high kurtosis, and end up underfitting. In fact, mixed normal performs worse than the empirical data.

4.3 Dataset 3: 4 values, with additional aggregated data from industry wide statistics

In addition to analysing the four above approaches, we compare the outcomes against the distribution of sample means that are generated from the aggregate data statistics. Although not entirely comparable to other distributions they provide a useful tail cut-off and sense check. For Normal distribution these values are almost identical (the estimated tails for the four points, and smoothed industry statistics) suggesting that mixed normal distribution is an outright bad fit for the data.

Despite the fact that the graph suggests that aggregate lognormal distribution has a longer tail than the one calibrated on the four points with some guidance for the industry statistics, this isn't true. Because the mean of aggregate distribution is lower (as can be observed from higher probability of lower outcome values). This implies that despite the longer tail of the aggregate distribution, 99.5% is actually smaller than the estimate based on the 4 points of data.





Pearson's distribution throws an error, which results in two peaks of the outcomes. Considering that the data is very scarce (4 data points) this is not unexpected. Normal approximation does the best job of accentuating the probable two underlying distributions. However lognormal seems a more sensible approach as it ignores the two humps and smoothes over them.

Section 5 - Analysis

Below we present a summary table of 99.5% of outcomes for the fitted values of the distribution.

Models give the ratio of the number that a particular fit produces to empirical 99.5% percentile. For most distributions EVT gives the most conservative number, however, Pearsons looks to be able to give comparable number for datasets with larger number of points.

Example	Number of data points	Skew	kurtosis	Normal at 1 in 200	Lognormal at 1 in 200	Pearsons	EVT	Professional estimate
1	44	3.64	17.03	1.4	2.8	1.3	3.6	[please fill]
2	206	4.37	24.9	0.8	1.3	4.4	4.3	[please fill]
3a (small estimated from large)	4 (110)	0.45	1.87	1.8 (1.8)	2.4 (1.9)	1.1	3.6	[please fill]

The above approaches present a range of outcomes that the underlying distribution will take. Although expert software will go a long way to helping you estimate the underlying uncertainty of the individual risk severity

Section 6 - Conclusion

The body of the claim severity is rarely correlated to the tail. We ignored the considerations of cut-off and exploration of the implications of this cut off. This is not meant to imply that this analysis is not significant, but rather that we are not able to incorporate all the individual considerations that go into these assumptions systematically.

Our goal was to provide a useful tool in terms of how risk appetite statement really sits with the current insurance gaps. Although the original intention was to include modelling of correlation between frequency and severity into the picture, we hope this paper provides the necessary push to get capital actuaries and risk managers to think about the real tails of their severity distributions.

Appendix 1 - References by topic

http://imaging.mrc-cbu.cam.ac.uk/methods/BayesianStuff?action=AttachFile&do=get&target=bilmes-em-algorithm.pdf <-understandable, contains HMM

https://www.cs.utah.edu/~piyush/teaching/EM_algorithm.pdf p4 EM algorithm

http://cs229.stanford.edu/notes/cs229-notes8.pdf similarly

Several methods: http://math.usask.ca/~longhai/teaching/stat812-1409/rdemo/EM_examples.pdf

https://en.wikibooks.org/wiki/Data_Mining_Algorithms_In_R/Clustering/Expectation_Maximization_(EM)

http://rstudio-pubs-static.s3.amazonaws.com/1001_3177e85f5e4840be840c84452780db52.html

https://pdfs.semanticscholar.org/d770/534dd5cf35883c365a258c77770fbb4bdc78.pdf <-excellent!!!!

Extremes Primer http://grotjahn.ucdavis.edu/EWEs/extremes_primer_v9_22_15.pdf

Appendix 2 - Pearsons ML cumulants derivation example

reducing to $\sigma^2/(2n)$ as above if the distribution is normal with $\mu_1' = 0$, when $\mu_4' = 3$

more robust estimator of var $\hat{\sigma}$ is

$$\frac{m_4' - m_2'^2}{4nm_2'} = \frac{m_2'}{4n}(\hat{\gamma}_2 + 2),$$

where $\hat{\gamma}_2 = m_4'/(m_2')^2 - 3$, and is less efficient than $\hat{\sigma}^2/(2n) = m_2'/(2n)$ in the

because the latter is based on the sufficient statistic and is not affected by sampling va

18.19 Even when the regularity conditions for its efficiency do not hold, the ML-

have MV properties conferred upon it by the sufficient statistic(s) of which it is alway

Thus, for example, in Example 18.1 $\hat{\theta} = x_{(n)}$ is a multiple $n/(n+1)$ of the unbiased

minimum variance (cf. Exercise 17.24), and consequently $\hat{\theta}$ will be asymptotically unbias

same asymptotic variance as t. However, we saw in Example 11.4 that here

$$\text{var}\, x_{(n)} = \frac{n\theta^2}{(n+1)^2(n+2)}.$$

Thus when we consider the m.s.e. of $x_{(n)}$ as in **17.30**, the square of its bias, $\theta^2/(n+1)$

same order of magnitude as its variance. If we apply the result of Example 17.14, we find

$$\frac{n(n+2)}{(n+1)^2}t = \left(\frac{n+2}{n+1}\right)x_{(n)}$$

has m.s.e. $\theta^2/(n+1)^2$, against $\theta^2/\{n(n+2)\}$ for the unbiased t and $2\theta^2/\{(n+1)(n+2)\}$

Thus $\hat{\theta}$ in this case has asymptotically, and indeed almost exactly, twice as large an m.s.e. as

multiple has.

The cumulants of an ML estimator

18.20 Haldane and Smith (1956) obtained expressions for the first four cumulants of

estimator. Suppose that the distribution sampled, $f(x|\theta)$, is grouped into k classes, and

probability of an observation falling into the rth class is π_r $(r = 1, 2, \ldots, k)$. We

any distribution to a multinomial distribution (**7.7**), and if the range of the original

is independent of the unknown parameter θ, we seek solutions of the likelihood equation

Since the probabilities π_r are functions of θ, we write, from **7.15**,

$$L(x|\theta) \propto \prod_r \pi_r^{n_r},$$

where n_r is the number of observations in the rth class and $\sum_r n_r = n$, the sample

likelihood equation is

$$\frac{\partial \log L}{\partial \theta} = \sum_r n_r \frac{\pi_r'}{\pi_r} = 0,$$

where a prime denotes differentiation with respect to θ. Of course, the ML estimator $\hat{\theta}$

differ from what it would be for the ungrouped observations, when more information is

:s 18.24–25 discuss the difference, which is important in some contexts – cf. **25.15–19**.

r, as $k \to \infty$ the difference disappears.

Now, using Taylor's theorem, we expand π_r and π_r' about the true value θ_0, and obtain

$$\begin{aligned}\pi_r(\hat{\theta}) &= \pi_r(\theta_0) + (\hat{\theta} - \theta_0)\pi_r'(\theta_0) + \tfrac{1}{2}(\hat{\theta} - \theta_0)^2\pi_r''(\theta_0) + \cdots \\ \pi_r'(\hat{\theta}) &= \pi_r'(\theta_0) + (\hat{\theta} - \theta_0)\pi_r''(\theta_0) + \tfrac{1}{2}(\hat{\theta} - \theta_0)^2\pi_r'''(\theta_0) + \cdots.\end{aligned} \tag{18.38}$$

If we insert (18.38) into the likelihood equation, expand binomially, and sum the series, we have, writing

$$\begin{aligned}A_i &= \sum_r \{\pi_r'(\theta_0)\}^{i+1}/\{\pi_r(\theta_0)\}^i, \\ B_i &= \sum_r \{\pi_r'(\theta_0)\}^i\pi_r''(\theta_0)/\{\pi_r(\theta_0)\}^i, \\ C_i &= \sum_r \{\pi_r'(\theta_0)\}^{i-1}\{\pi_r''(\theta_0)\}^2/\{\pi_r(\theta_0)\}^i, \\ D_i &= \sum_r \{\pi_r'(\theta_0)\}^i\pi_r'''(\theta_0)/\{\pi_r(\theta_0)\}^i, \\ \alpha_i &= \sum_r \{\pi_r'(\theta_0)\}^i\left\{\frac{n_r}{n} - \pi_r(\theta_0)\right\}/\{\pi_r(\theta_0)\}^i, \\ \beta_i &= \sum_r \{\pi_r'(\theta_0)\}^{i-1}\pi_r''(\theta_0)\left\{\frac{n_r}{n} - \pi_r(\theta_0)\right\}\Big/\{\pi_r(\theta_0)\}^i, \\ \delta_i &= \sum_r \{\pi_r'(\theta_0)\}^{i-1}\pi_r'''(\theta_0)\left\{\frac{n_r}{n} - \pi_r(\theta_0)\right\}\Big/\{\pi_r(\theta_0)\}^i,\end{aligned}$$

the expansion

$$\begin{aligned}&\alpha_1 - (A_1 + \alpha_2 - \beta_1)(\hat{\theta} - \theta_0) + \tfrac{1}{2}(2A_2 - 3B_1 + 2\alpha_3 - 3\beta_2 + \delta_1)(\hat{\theta} - \theta_0)^2 \\ &- \frac{1}{6}(6A_3 - 12B_2 + 3C_1 + 4D_1)(\hat{\theta} - \theta_0)^3 + \cdots = 0.\end{aligned} \tag{18.39}$$

For large n, (18.39) may be inverted by Lagrange's theorem to give

$$\begin{aligned}\hat{\theta} - \theta_0 = {} & A_1^{-1}\alpha_1 + A_1^{-3}\alpha_1[(A_2 - \tfrac{3}{2}B_1)\alpha_1 - A_1(\alpha_2 - \beta_1)] \\ & + A_1^{-5}\alpha_1[\{2(A_2 - \tfrac{3}{2}B_1)^2 - A_1(A_3 - 2B_2 + \tfrac{1}{2}C_1 + \frac{2}{3}D_1)\}\alpha_1^2 \\ & - 3A_1(A_2 - \tfrac{3}{2}B_1)\alpha_1(\alpha_2 - \beta_1) + \tfrac{1}{2}A_1^2\alpha_1(2\alpha_3 - 3\beta_2 + \delta_1) \\ & + A_1^2(\alpha_2 - \beta_1)^2] + O(n^{-3}).\end{aligned} \tag{18.40}$$

Expression (18.40) enables us to obtain the moments of $\hat{\theta}$ as series in powers of n^{-1}.

Consider the sampling distribution of the sum

$$W = \sum_r h_r\left\{\frac{n_r}{n} - \pi_r(\theta_0)\right\},$$

where the h_r are any constant weights. From the moments of the multinomial distribution (cf. 7.2) we obtain for the moments of W, writing $S_i = \sum_r h_r^i \pi_r(\theta_0)$,

$$\begin{aligned}
&\mu_1'(W) = 0,\\
&\mu_2(W) = n^{-1}(S_2 - S_1^2),\\
&\mu_3(W) = n^{-2}(S_3 - S_1S_2 + 2S_1^3),\\
&\mu_4(W) = 3n^{-2}(S_2 - S_1^2)^2 + n^{-3}(S_4 - 4S_1S_3 - 3S_2^2 + 12S_1^2S_2 - 6S_1^4),\\
&\mu_5(W) = 10n^{-3}(S_2 - S_1^2)(S_3 - 3S_1S_2 + 2S_1^3) + O(n^{-4}),\\
&\mu_6(W) = 15n^{-3}(S_2 - S_1^2)^3 + O(n^{-4}).
\end{aligned} \tag{18.4}$$

From (18.41) we can derive the moments and product moments of the random variables α_i, β and δ_i appearing in (18.40), for all of these are functions of the form W. Finally, we substitute these moments into the powers of (18.40) to obtain the moments of $\hat{\theta}$. Expressed as cumulants these are

$$\begin{aligned}
\kappa_1 &= \theta_0 - \tfrac{1}{2}n^{-1}A_1^{-2}B_1 + O(n^{-2}),\\
\kappa_2 &= n^{-1}A_1^{-1} + n^{-2}A_1^{-4}[-A_2^2 + \tfrac{7}{2}B_1^2 + A_1(A_3 - B_2 - D_1) - A_1^3] + O(n^{-3}),\\
\kappa_3 &= n^{-2}A_1^{-3}(A_2 - 3B_1) + O(n^{-3}),\\
\kappa_4 &= n^{-3}A_1^{-5}[-12B_1(A_2 - 2B_1) + A_1(A_3 - 4D_1) - 3A_1^3] + O(n^{-4}),
\end{aligned} \tag{18.42}$$

whence

$$\begin{aligned}
\gamma_1 &= \kappa_3/\kappa_2^{3/2} = n^{-1/2}A_1^{-3/2}(A_2 - 3B_1) + o(n^{-1/2}),\\
\gamma_2 &= \kappa_4/\kappa_2^2 = n^{-1}A_1^{-3}[-12B_1(A_2 - 2B_1) + A_1(A_3 - 4D_1) - 3A_1^3] + o(n^{-1}).
\end{aligned} \tag{18.43}$$

The first cumulant in (18.42) shows that the bias in $\hat{\theta}$ is of order of magnitude n^{-1} unless $B_1 = 0$, when it is of order n^{-2}, as may be confirmed by calculating a further term in the first cumulant. If $k \to \infty$, $A_i \to E\{(\partial \log f/\partial\theta)^{i+1}\}$ and so on, and thus the leading term in the second cumulant corresponds to the asymptotic variance previously established in **18.16**. Expressions (18.43) illustrate the tendency to normality, established in **18.16** for ungrouped observations.

If the terms in (18.42) were all evaluated, and unbiased estimates made of each of the first four moments of $\hat{\theta}$, a Pearson distribution or Edgeworth expansion could be fitted and an estimate of the small-sample distribution of $\hat{\theta}$ obtained which would provide a better approximation than the ultimate normal approximation of **18.16**. Pfanzagl (1973) gives an Edgeworth expansion to improve the normal approximation in the continuous case, which is equivalent to using the cumulants (18.42).

Winterbottom (1979) shows how to use (18.42) in a Cornish–Fisher expansion to find approximate confidence intervals for θ.

The next higher-order terms in (18.42)–(18.43) are derived by Shenton and Bowman (1963).

Peers (1978) obtained the asymptotic joint cumulants of

$$n^{1/2}\hat{\theta} \quad \text{and} \quad t_r = n^{-r/2}(\partial^r \log L/\partial\theta^r)_{\theta=\hat{\theta}}, \quad r = 2, 3, 4,$$

in terms of $n^{-s/2}\partial^s \log L/\partial\theta^s$, $s = 1, 2, 3, 4$. The set of statistics $(\hat{\theta}, t_2, t_3, t_4)$ is 'second-order sufficient' for θ, since the Taylor expansion of $\log L$ about $\hat{\theta}$ contains only these statistics to the second order of approximation in n.

Successive approximation to ML estimators

18.21 In most of the examples we have considered, the ML estimator has been obtained in explicit form. An exception was in Example 18.3, where we were left with a cubic equation to solve for the estimator, and this can be done without much trouble when the sample is given. Sometimes, however, the likelihood equation can only be solved numerically, starting from some trial value t.

As at (18.31), we expand $\partial \log L/\partial\theta$ in a Taylor series, but this time about its value at t, obtaining

$$0 = \left(\frac{\partial \log L}{\partial\theta}\right)_{\hat{\theta}} = \left(\frac{\partial \log L}{\partial\theta}\right)_t + (\hat{\theta} - t)\left(\frac{\partial^2 \log L}{\partial\theta^2}\right)_{\theta^*},$$

where θ^* lies between $\hat{\theta}$ and t. Thus

$$\hat{\theta} = t - \left(\frac{\partial \log L}{\partial\theta}\right)_t \Big/ \left(\frac{\partial^2 \log L}{\partial\theta^2}\right)_{\theta^*}. \tag{18.44}$$

If we can choose t so that it is likely to be in the neighbourhood of $\hat{\theta}$, we can replace θ^* in (18.44) by t and obtain

$$\hat{\theta} = t - \left(\frac{\partial \log L}{\partial\theta}\right)_t \Big/ \left(\frac{\partial^2 \log L}{\partial\theta^2}\right)_t, \tag{18.45}$$

which will give a closer approximation to $\hat{\theta}$. The process can be repeated until no further correction is achieved to the desired degree of accuracy.

The most common method for choice of t is to take it as the value of some (preferably simply calculated) consistent estimator of θ, ideally one with high efficiency, so that, by (17.61), it will be highly correlated with the efficient $\hat{\theta}$. Then, as $n \to \infty$, we shall have the two consistent estimators t and $\hat{\theta}$ converging to θ_0, and θ^* consequently also doing so. The three random variables $(\partial^2 \log L/\partial\theta^2)_{\theta^*}$, $(\partial^2 \log L/\partial\theta^2)_t$ and $[E(\partial^2 \log L/\partial\theta^2)]_t$ will all converge to $[E(\partial^2 \log L/\partial\theta^2)]_{\theta_0}$. Use of the second of these variables, instead of the first, in (18.44) gives (18.45) above: use of the third instead of the first gives the alternative iterative procedure

$$\hat{\theta} = t - \left(\frac{\partial \log L}{\partial\theta}\right)_t \Big/ \left[E\left(\frac{\partial^2 \log L}{\partial\theta^2}\right)\right]_t = t + \left(\frac{\partial \log L}{\partial\theta}\right)_t (\operatorname{var}\hat{\theta})_t, \tag{18.46}$$

$\operatorname{var}\hat{\theta}$ being the asymptotic variance obtained in **18.16**. Expression (18.45) is known as the Newton–Raphson iterative process; (18.46) is sometimes known as 'the *method of scoring* for parameters', due to Fisher (1925), because $(\partial \log L/\partial\theta)_{\hat{\theta}}$ is called the *score* function. Kale (1961) shows that although (18.45) ultimately converges faster, (18.46) will often give better results for the first few iterations when n is large. It is usually less laborious.

Both (18.45) and (18.46) may fail to converge in particular cases. Even when they do converge, if the likelihood equation has multiple roots there is no guarantee that they will converge to the root corresponding to the absolute maximum of the LF; this should be verified by examining the changes in sign of $\partial \log L/\partial\theta$ from positive to negative and searching the intervals in which these changes occur to locate, evaluate and compare the maxima. Barnett (1966) discusses a systematic method of doing this. As with all such methods, the location of the maximum should be verified by repeating the analysis with several different starting-points.

Appendix 3 - [TBC] Derivation of posterior mixed distribution for Dataset 3

Prior example assumption of lognormal distribution for the

Data

Posterior

Estimates

Appendix 4 - Dealing with aggregated dataset

Estimating moments from large claims dataset.

Given aggregated numbers for frequency and aggregate severity of large claims in the industry, we:

1. Made assumption about frequency being distributed as poisson process
2. That allowed us to assume that aggregate severity was a compound poisson distribution
3. Which in turn allowed us to derive individual mean and standard deviation of the distribution
4. To compute the possible input values into the two sample distribution we calculated the natural logarithm of the average claim for each of the past years, and computed the standard deviation of them. Then we applied roughly 1 standard deviation mean on either side of the average claim mean, to get potential initiation values for lognormal distribution

When computing the aggregate lognormal distribution we used the parameters estimated in step 2. This (almost certainly) provides an underestimation of overall aggregate distribution, but the tail serves as a useful lower boundary check for the fitted lognormal distribution, (which looks more or less reasonable).

Further information on the procedure can be found in Excel file.

Appendix 5 - R code

```
## Generate sample

library(extRemes)

library('PearsonDS')## find Pearson distribution with these parameters


library(mixtools)

# gives the distribution number and it's parameters

## compare with method of moments estimator




# result mean - 0.95, var = 1.64 skew  = 0.9, kurtosis = 1.4??


Book1 <- read.csv("C:/Users/ANNBAB/Desktop/papers/Book1.csv", header=FALSE, stringsAsFactors=FALSE)

Book2 <- read.csv("C:/Users/ANNBAB/Desktop/papers/Book2.csv", header=FALSE, stringsAsFactors=FALSE)

Book3<-c(3400000,1560000,2000000,1200000)# data set 3

data<-as.numeric(unlist(Book3))#change for dataset

hist(data, breaks=15)

#pearsons

ppar<-pearsonFitML(as.vector(data))

print(unlist(ppar))

vector1<-c(ppar$a,ppar$b,ppar$location,ppar$scale)# for Book2 use ppar$a,ppar$b, ppar$shape,ppar$location,ppar$scale)

#EVT

EVT_fit<-fevd(data)

p <- EVT_fit$results$par

EVT_r<-revd(1000,loc = p[ 1 ], scale = p[ 2 ], shape = p[ 3 ] )

qevd( 0.995, loc = p[ 1 ], scale = p[ 2 ], shape = p[ 3 ] )
```

```
#lognorm

log_data<-log(data, base = exp(1))

set.seed(1234)

gm<-normalmixEM(log_data,k=2, mu = c(14,15), sigma = 0.5) #use this for Book3->#, mu = c(14,15), sigma = 0.5)

l_l<-gm$lambda

l_mu<- gm$mu  #

l_s<-gm$sigma #

#norm

gm<-normalmixEM(data,k=2)

l<-gm$lambda

mu<- gm$mu  #

s<-gm$sigma #

# compare



max(data)

x = seq(0,4000000,10)

DATA1<-rpearsonI(10000,params=vector1)#Book1, Book3

p_truth<-dpearsonI(x,params=vector1)#Book1, Book3

DATA2<-rpearsonV(10000,params=vector1)#Book2

p_truth<-dpearsonV(x,params=vector1)#Book2

l_truth = l_l[1]*dlnorm(x,l_mu[1],l_s[1]) + l_l[2]*dlnorm(x,l_mu[2],l_s[2])

l_truth1 = rlnorm(l_l[1]*10000,l_mu[1],l_s[1]) + rlnorm(l_l[2]*10000,l_mu[2],l_s[2])

l_truth0 = dlnorm(x,14.44,0.468) # aggregate data for Book3

l_truth01=rlnorm(20000,14.44,0.468) #aggregate data for Book3

truth = l[1]*dnorm(x,mu[1],s[1]) + l[2]*dnorm(x,mu[2],s[2])

truth1 = rnorm(l[2]*10000,mu[1],s[1]) + rnorm(l[2]*10000,mu[2],s[2])

plot(density(data),lwd=1, xlab="claim size ($)",xlim=c(0, max(x)),ylim=c(0,0.0000008), main = "Comparison of model fit")# for book3 use ylim ylim=c(0,0.0000008)

lines(x,truth,col="red",lwd=1)

lines(x,l_truth,col="blue",lwd=1)
```

```
lines(x,p_truth,col="orange",lwd=1)

lines(x,l_truth0,col="brown",lwd=1)

#legend("topright",c("data PDF", "Pearsons PDF","Mix Normal", "Mix Lognormal"), lty=1, col=c('black', 'orange', 'red','blue'), bty='n', cex=.75)

#use for book 3 #

legend("topright",c("data PDF", "Pearsons PDF","Mix Normal", "Mix Lognormal", "Aggregate"), lty=1, col=c('black', 'orange', 'red','blue','brown'), bty='n', cex=.75)

plot(EVT_fit,"density",main = "Comparison of model fit for Extreme Value")

quantile(DATA1, 0.995)/quantile(data,0.995)

qevd( 0.995, loc = p[ 1 ], scale = p[ 2 ], shape = p[ 3 ] )/quantile(data,0.995)

par(mfrow=c(1,2))
```