Minimax Estimation of Insurance Loss Extreme Percentiles

Andrew D Smith[1]

[1]University College Dublin, ansmith@ucd.ie

Probability Seminar, 20 February 2019

How Insurance Works (roughly)

- Suppose you own a ship worth €10, 000, 000
- You buy insurance against loss at sea, for an annual premium of 0.5% of the sum assured, that is €50 000. The insurer writes many other similar policies.
- A regulator requires the insurer to hold more than the premium, so claims can still be settled even if losses are bigger than the premiums.
- The insurer has to ask shareholders for a capital injection which they can't hand back until the insurance contracts expire.
- If all goes as expected, a year later the shareholders get back the capital they injected, plus the expected profit margin built into the premiums.

From the Insurer's Perspective

		€
	Expected loss	40, 000
+	Profit margin	10, 000
=	Insurance premium	50, 000
+	New subscribed capital	8, 000
=	Available financial resources	58, 000

Capital Requirement Calculation

The insurer had to demonstrate €58, 000 of available resources, even though the expected loss is only €40,000.

- This requirement is based on a probability distribution with mean of €40,000 whose 99.5%-ile is €58,000.
  - that is, there is a 99.5% probability that the insurance loss is $\leq$ €58,000.
- The ratio of 45% $= \frac{18{,}000}{40{,}000}$ is the *capital requirement factor* or CRF.

Generally, shareholders will argue for a lower CRF so they have to put less cash at risk, while regulators argue for higher CRF to protect policyholders from insurer default.

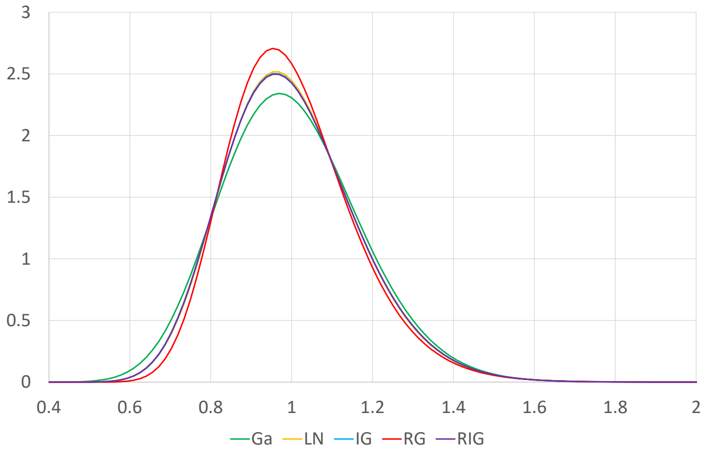
Loss Distribution Fitting

Capital Requirement Factors are usually estimated based on probability distributions fitted to historic claims data. One of the following families is customarily chosen (in decreasing order of popularity):

- The Gamma distributions (Ga).
- The Lognormal distributions (LN).
- The Inverse Gaussian (Wald) distributions (IG).
- Reciprocal Gamma distributions (RG).
- Reciprocal Inverse Gauss distributions (RIG).

These are all examples of scale-invariant 2-parameter exponential families.

Probability Densities: Mean = 1.0 and CRF = 0.5



The Need for Estimation

We have:

- A family of probability density functions $f(x;\theta) \geq 0$ with unknown parameter $\theta \in \Theta \subset \mathbb{R}^2$ to be estimated.
- All are normalised such that $\int_0^\infty f(x;\theta)dx = 1$
- Historic data $x_1, x_2, \ldots x_n$, which we will assume are a random sample of (independent) observations from an unknown probability distribution.

How to pick θ given data $x_1, x_2, \ldots x_n$? We denote the selected value by $\hat{\theta}$; the hat indicates a parameter estimate.

Method of Moments (MoM)

Solve for $\hat{\theta} \in \Theta$ that satisfies:

$$\int_0^\infty x f(x;\hat{\theta})dx = \frac{1}{n}\sum_{j=1}^{n} x_j$$

$$\int_0^\infty x^2 f(x;\hat{\theta})dx = \frac{1}{n}\sum_{j=1}^{n} x_j^2$$

More generally, if $\Theta \subset \mathbb{R}^d$ then we match means of $x, x^2, \ldots x^d$. Equivalently, make $f(x;\hat{\theta}) - \frac{1}{n}\sum_{j=1}^{n} \delta(x - x_j)$ orthogonal to all polynomials in x of order $\leq d$.

Maximum Likelihood (MLE)

As an alternative to the method of moments, the method of *maximum likelihood* means choosing the value of $\theta \in \Theta$ to maximise the product of the densities.

$$\hat{\theta} = \arg\max_{\theta} \prod_{j=1}^{n} f(x_j;\theta) = \arg\max_{\theta} \sum_{j=1}^{n} \ln f(x_j;\theta)$$

Distribution Origins

There are classical contexts for the distributions we fit:

- Gamma = sum of exponential waiting times.
- Lognormal = product of many independent variables.
- Inverse Gauss = first time a Brownian motion hits a linear boundary.

None of these are obvious generating mechanisms for insurance claims. So it is not clear which family we should prefer. Unstated reasons for choice include analytical tractability, availability in popular spreadsheets and statistical tools, use by peers or in the literature, a belief that ‘it works in practice’.

Classifying Right Tail Behaviour

Distribution	Asymptotic Behaviour	Tail Fatness
Ga	$-\ln f(x) = O(x)$	Thin
LN	$-\ln f(x) = O(\ln x)^2$	Intermediate
IG	$-\ln f(x) = O(x)$	Thin
RG	$-\ln f(x) = O(\ln x)$	Fat
RIG	$-\ln f(x) = O(x)$	Thin

We might guess that the RG produces highest CRFs (true), while Ga, IG and RIG produce lowest CRFs (partly true). This gives a clue about who will be arguing for each distribution, but does not tell us which distribution to use.

Futility of Hypothesis Tests

- The five families under consideration share many properties: bell-shaped, positive, right-skewed and the number of data points is small, often fewer than 20.
- Goodness of fit tests on small data sets measure wildness of data, eg if empirical data looks bimodal, even if due to chance, then GoF test will fail.
- Generally, GoF tests pass for all dists (if data not too wild) or fail for all dists (wild data). Very unusual that test pasts for just one distribution.
- The different families are similar enough to be statistically indistinguishable, but different enough to make a material difference to CRFs.

What else can go Wrong?

Sources of error:

- Parameter error: fitted parameters differ from the distribution that generated the data.
- Mis-Specification error
  - Data is drawn from a different distribution family
  - Data points not independent
  - Data points from multiple distributions
- Data may be incorrect or unrepresentative
- Someone is trying to mislead you

Isolating Mis-Specification Error

Treating the different forms of error:

- The first type of error - parameter error - has been widely explored. We will summarise the results next.
- Regulators and auditors address data issues and human biases with data quality standards.
- This leaves mis-specification error, which is our topic.

All sources of error may be important, but in order to isolate mis-specification error, we take a theoretical limit of large data size where sampling variability can be ignored.

Use of Data

- We debate the *best* methodology without looking at actual loss data, beyond knowing the range of CRFs that are of practical interest.
- Practitioners find this approach frustrating: commercially they prefer to look at the data and then decide what method they like (hint: industry wants low CRF, regulators want high CRF).
- On the other hand consistency is also valued (as most firms will estimate hundreds of these distributions) and cherry-picking is frowned on.

Conclusion: there is a role for methodologies justified by statistical principles rather than a specific data set.

Classical Properties of Parameter Estimators

- **Consistency**. The estimate converges (with probability 1) to the true parameter as the sample size increases to infinity. Generally a low bar - most sensible methods should pass this test.
- **Unbiasedness**. The mean of the estimate is the true parameter. Generally hard to achieve, but easier for exponential families.
- **Efficiency**. The estimate is of minimum variability among unbiased estimators. MLE scores well by this metric, and generally has lower variance of parameter estimates than MoM, and indeed than any other method. MLE attains the Cramér-Rao bound for exponential families.

Stated Justifications for Method of Moments

Method of moments is still widely used, because:

- It is easier to implement than MLE (especially for Gamma distributions).
- No optimisation required.
- Easy to audit.
- Allows comparisons across alternative distribution families.

There is a general feeling among statisticians that these are not ‘good’ theoretical reasons.

Large Sample Limit

- Suppose the data is drawn from a reference distribution with density $f_{Ref}(x)$, not necessarily from the fitted family. Consider the (almost sure) limit of $\hat{\theta}$.
- For MoM we solve for $\hat{\theta}$

$$\int_0^\infty x^k f(x;\hat{\theta})dx = \int_0^\infty x^k f_{Ref}(x)dx; k = 1, 2$$

- For MLE we solve for $\hat{\theta}$:

$$\hat{\theta} = \arg\max_{\theta} \int_0^\infty \ln f(x;\theta) f_{Ref}(x)dx$$

Defining CRF Error

For a given:

- Reference family and CRF
- Fitted family
- Confidence level

We can determine the fitted CRF. This is an almost-sure limit, ie a number, not a distribution. Measure badness as the absolute difference between fitted and reference CRF's. Where the reference family is unknown, we take the worst case.

Geometric Analogy: Imagine distribution families as points in space and CRF errors as distances between those points. We want to fit from a family near the centre of the smallest sphere containing all families, as this minimises the maximum distance.

Results: Reference CRF = 0.50, 99.5%-ile

Table of fitted CRF:

Fitted	Ref Model Method	Ga	LN	IG	RG	RIG
Ga	MLE	0.5000	0.4669	0.4687	0.4395	0.4695
Ga	MoM	0.5000	0.4715	0.4733	0.4473	0.4740
LN	MLE	0.5413	0.5000	0.5020	0.4665	0.5030
LN	MoM	0.5318	0.5000	0.5020	0.4731	0.5028
IG	MLE	0.5389	0.4980	0.5000	0.4649	0.5008
IG	MoM	0.5295	0.4980	0.5000	0.4714	0.5008
RG	MLE	0.5945	0.5416	0.5439	0.5000	0.5328
RG	MoM	0.5651	0.5298	0.5320	0.5000	0.5328
RIG	MLE	0.5379	0.4973	0.4992	0.4643	0.5000
RIG	MoM	0.5286	0.4973	0.4992	0.4708	0.5000

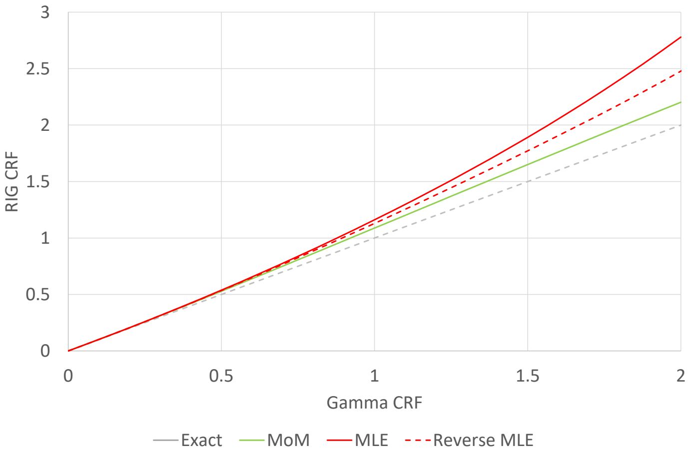
Worst Case Absolute Error for each Methodology

Fitted	Method	CRF = 0.25	CRF = 0.50	CRF = 0.75	Rank
Ga	MLE	0.0164	0.0605	0.1264	8
Ga	MoM	0.0151	0.0527	0.1046	7
LN	MLE	0.0096	0.0413	0.0992	6
LN	MoM	0.0084	0.0318	0.0673	3
IG	MLE	0.0093	0.0389	0.0921	5
IG	MoM	0.0081	0.0295	0.0605	2
RG	MLE	0.0205	0.0945	0.2455	10
RG	MoM	0.0171	0.0651	0.1372	9
RIG	MLE	0.0092	0.0379	0.0877	4
RIG	MoM	0.0080	0.0292	0.0603	1

Which Methods come out Best?

- Minimax ranking of methods, best first (but all close): RIG (MoM), IG (MoM), LN (MoM), RIG (MLE), IG (MLE), LN (MLE)
- The next four are substantially worse: Ga (MoM), Ga (MLE), RG (MoM), RG (MLE).
- For the best six methods, the worst reference models are Ga (overestimate CRF) and RG (underestimate CRF) in some order; these two are close.
- Geometrically, Ga and RG are the extreme points and other families are bunched around the midpoint.
- The same ranking of methods holds for a range of reference CRF's (provided they are small enough) and %-ile levels (provided they are high enough).

Fitting RIG to Gamma Data: Minimax Extreme Case



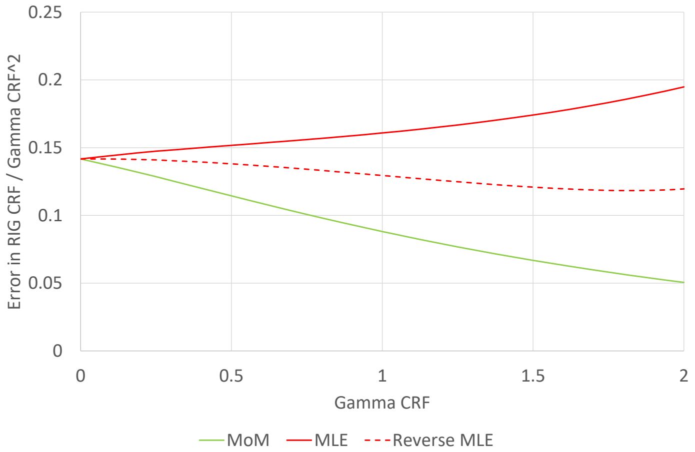
Error Behaviour when CRF is Small

We deduce that, for a given distribution pair:

- The error is zero if the model is well-specified (this is the consistency property).
  - So there is no saddle point; if Mother Nature wants to frustrate your estimates, she has to pick families at random.
- The error is $O(CRF^2)$, unless the reference and fitted families are both elements of $\{LN, IG, RIG\}$ in which case the error is $O(CRF^3)$.
- The difference between MLE and MoM is $O(CRF^3)$

These are the leading terms of power series whose higher coefficients can be derived using Cornish-Fisher and Saddle-Point expansions.

CRF Error divided by Squared CRF



Why did MoM Perform So Well?

- The method of moments involves raising observations to powers. This implies that a single outlier can have a large effect on parameter estimates.
- The sensitivity to outliers is a reason why MoM scores badly by traditional efficiency measures.
- MLE makes better use of the whole distribution, using behaviour of the smallest losses to infer the shape of the right hand tail.
- That right-from-left trick only works if you know the distribution family; can go badly wrong if the model family is mis-specified.
- MoM uses the right tail data to estimate the right tail. Small observations have little impact. So MoM is more robust than MLE to mis-specification.

Posing the Right Problem

- We have found the least bad of ten methodologies in a minimax sense. All the alternatives we considered were ‘sensible’ approaches.
- Just measuring an empirical quantile is not a sensible approach, as it cannot be implemented with ≤ 200 data points. However, empirical percentiles have zero error according to our criterion because there is no model error in the limit of large samples.
- We are still looking for the best mathematical way to define ’sensible’ so as to exclude such trivial cases.

Picking a Robust Methodology: Lessons Learned

- If you do not know what distribution your data comes from, estimate within a family in the middle of the possibilities.
- In classical statistics, the method of maximum likelihood is generally theoretically preferred to other methods (such as MoM), on the grounds of efficiency.
- MoM is sometimes used; people feel guilty about taking an easy but theoretically inferior option.
- We have shown that when models are mis-specified, MoM can be more robust than MLE, so practitioner guilt may be misplaced.

Limitations

We have **not** shown robustness of MoM in a general sense, but only in the very specific situation where:

- We are interested in accurate estimation of CRFs
- These are set at extreme percentiles
- The relevant distributions are in a neighbourhood of normal limits where asymptotic expansions are valid.
- We can list five distribution families, from one of which the data has been drawn.

MoM might be a terrible approach to other equally interesting problems. In the context of the statistical literature, it is remarkable to find even one problem where MoM beats MLE.

Probability and Effective Regulation

- Most advanced economies operate regimes where insurers have to hold enough capital to be reasonably sure of being able to pay claims, but short of absolute certainty.
- European Parliament thinks this balance is specified by
  - Enshrining a 99.5% confidence level in regulation.
  - Setting data quality standards.
- This leaves unspecified how best to allow for model and parameter uncertainties.
- Important role for probabilists to define estimation methodologies and make sense of this ambiguity.