1 Modelling claim size

1.1 Introduction

Models describing variation in claim size lack the theoretical underpinning provided by the Poisson point process in Chapter 8. The traditional approach is to impose a family of probability distributions and estimate their parameters from historical claims $z_1, \ldots, z_n$ (corrected for inflation if necessary). Even the family itself is often determined from experience. An alternative with considerable merit in the computer age is to throw all prior matematical conditions overboard and rely solely on the historical data. This is known as a **non-parametric** approach. Much of this chapter is on the use of historical data.

How we go about is partly dictated by the size of the record, and here the variablity from one case to another is enormous. With automobile insurance the number of observations n may be huge, providing a good basis for deducing the probability distribution of the claim size Z. By contrast, major incidents in industry (like the collapse of an oil rig!) are rare, making historical material scarce. This span of variation is reflected in the presentation. Basic issues are parametric versus non-parametric methods and (above all) the extreme right tail of distributions. Lack of historical data in the region that matters most financially suggests that this problem deserves special attention. The mathematical framework is **mixing** (Section 9.6) in combination with **Pickands'** theorem from the theory of extremes (Section 9.4). A final, general issue is the uitilization of incomplete or **censored** information.

1.2 Parametric and non-parametric modelling

Introduction

Claim size modelling can be **parametric** through families of distributions such as the Gamma, log-normal or Pareto with parameters tuned to historical data or **non-parametric** where each claim z_i of the past is assigned a probability $1/n$ of re-appearing in the future. A new claim is then envisaged as a random variable $\hat{Z}$ for which

$$\Pr(\hat{Z} = z_i) = \frac{1}{n}, \qquad i = 1, \ldots, n. \tag{1.1}$$

As model for Z this is an entirely proper distribution (since its sum over all i is one). If it appears peculiar, there are actually several points in its favour (one in its disfavour too); see below. Note the notation $\hat{Z}$ which is the familiar way of emphasizing that estimation has been involved. The model is known as the **empirical distribution function** and will in Section 9.5 employed as a brick in an edifice that also involves the Pareto distribution. The purpose of this section is to review parametric and non-parametric modelling on a general level.

Scale families of distributions

All sensible parametric models for claim size are of the form

$$Z = \beta Z_0, \tag{1.2}$$

where $\beta > 0$ is a parameter, and Z_0 is a standardized random variable corresponding to $\beta = 1$. This proportionality is inherited by expectations, standard deviations and percentiles; i.e. if ξ_0, σ_0

and $q_{0\epsilon}$ are expectation, standard deviation and ϵ-percentile for Z_0, then the same quantities for Z are

$$\xi = \beta\xi_0, \qquad \sigma = \beta\sigma_0 \qquad \text{and} \qquad q_\epsilon = \beta q_{0\epsilon}. \tag{1.3}$$

To see what β stands for, suppose currency is changed as a part of some international transaction. With c as the exchange rate the claim quoted in foreign currency becomes cZ, and from (1.2) $cZ = (c\beta)Z_0$. The effect of passing from one currency to another is simply that $c\beta$ replaces β, the shape of the density function remaining what it was. Surely anything else makes little sense. It would, for example, be contrived to take a view on risk that differed in terms of US$ from that in British £ or euros, and the same point applies to inflation (Exercise 9.2.1).

In statistics β is known as a **parameter of scale** and parametric models for claim size should always include them. An example worth commenting is the log-normal distribution used in earlier chapters. If it is on the form $Z = \exp(\theta + \tau\varepsilon)$ where ε is $N(0,1)$, we may also write it

$$Z = \xi Z_0 \qquad \text{where} \qquad Z_0 = \exp(-\frac{1}{2}\tau^2 + \tau\varepsilon) \quad \text{and} \quad \xi = \exp(\theta + \frac{\tau^2}{2}).$$

Here $E(Z_0) = 1$, and ξ serves as both expectation and scale parameter. The mean is often the most important of all quantities associated with a distribution, and it makes sense to make it visible as the scale parameter. Such tactics has in this book been followed whenever practical.

Fitting a scale family
Models for scale families satisfy the relationship

$$\Pr(Z \le z) = \Pr(Z_0 \le z/\beta) \qquad \text{or} \qquad F(z|\beta) = F_0\left(\frac{z}{\beta}\right).$$

where $F_0(z)$ is the distribution function of Z_0. Differentiating with respect to z yields the family of density functions

$$f(z|\beta) = \frac{1}{\beta} f_0(\frac{z}{\beta}), \quad z > 0 \qquad \text{where} \qquad f_0(z) = F_0'(z). \tag{1.4}$$

Additional parameters describing the shape of the distributions are hiding in $f_0(z)$. All scale families have density functions on this form.

The standard way of fitting such models is through likelihood estimation. If $z_1, \ldots, z_n$ are the historical claims, the criterion becomes

$$\mathcal{L}(\beta, f_0) = -n\log(\beta) + \sum_{i=1}^{n} \log\{f_0(z_i/\beta)\}, \tag{1.5}$$

which is to be maximized with respect to β and other parameters. Numerical methods are usually required. A useful extension covers situations with **censoring**. Typical examples are claims only registered as above or below certain limits, known as censoring **to the right** and **left** respectively. Consider n_r of cases the former where $b_1, \cdots, b_{n_r}$ are lower bounds on the losses that actually occurred. Event i among those has probability is $1 - F_0(b_i/\beta)$, and its *logarithm* is added to the

log likelihood (1.5) of the fully observed claims $z_1, \ldots, z_n$. In other words, likelihood estimation is now undertaken by maximizing

$$\mathcal{L}(\beta, f_0) = -n\log(\beta) + \underbrace{\sum_{i=1}^{n} \log\{f_0(z_i/\beta)\}}_{\textit{complete information}} + \underbrace{\sum_{i=1}^{n_r} \log\{1 - F_0(b_i/\beta)\}}_{\textit{censoring to the right}}, \tag{1.6}$$

Censoring to the left is similar and discussed in Exercise 9.2.2. Details will be developed for the Pareto family in Section 9.4.

Skewness as simple description of shape

A major issue in claim size modelling is the degree of asymmetry towards the right tail of the distribution. A useful, *simple* summary is the **coefficient of skewness** defined as

$$\gamma = \text{skew}(Z) = \frac{\mu_3}{\sigma^3} \qquad \text{where} \qquad \mu_3 = E(Z - \xi)^3. \tag{1.7}$$

The numerator is the **third order moment**. Skewness should *not* depend on the currency being used and doesn't since

$$\text{skew}(Z) = \frac{E(Z-\xi)^3}{\sigma^3} = \frac{E(\beta Z_0 - \beta\xi_0)^3}{(\beta\sigma_0)^3} = \frac{E(Z_0 - \xi_0)^3}{\sigma_0^3} = \text{skew}(Z_0)$$

after inserting (1.2) and (1.3). Neither is the coefficient changed when Z is shifted by a fixed amount; i.e. $\text{skew}(Z+b) = \text{skew}(Z)$ through the same type of reasoning. These properties confirm skewness as a (simplified) representation of the shape of a distribution.

The standard estimate of the skewness coefficient γ from observations $z_1, \ldots, z_n$ is

$$\hat{\gamma} = \frac{\hat{\mu}_3}{s^3} \qquad \text{where} \qquad \hat{\mu}_3 = \frac{1}{n - 3 + 2/n} \sum_{i=1}^{n} (z_i - \bar{z})^3. \tag{1.8}$$

Here $\hat{\mu}_3$ is the natural estimate of the third order moment[1] and s the sample standard deviation. The estimate is for low n and heavy-tailed distributions typically severely biased downwards. *Under*-estimation of skewness, and by implication the risk of large losses, is a recurrent theme with claim size modelling in general and is common even when parametric families are used. Several of the exercises are devoted to the issue.

Shifted distributions

Sometimes the distribution of a claim starts at some some threshold b instead of at the orgin. Obvious examples are deductibles and contracts in re-insurance. Models can be constructed by adding b to variables Z starting at the origin; i.e.

$$Z_{>b} = b + Z = b + \beta Z_0,$$

where Z starts at the orgin. Hence

$$\Pr(Z_{>b} \le z) = \Pr(b + \beta Z_0 \le z) = \Pr\left(Z_0 \le \frac{z-b}{\beta}\right),$$

[1] Division on $n - 3 + 2/n$ makes it unbiased.

and differentiating with respect to z yields

$$f_{>b}(z|\beta) = \frac{1}{\beta} f_0\left(\frac{z-b}{\beta}\right), \qquad z > b, \tag{1.9}$$

as density function for variables starting ar b.

Sometimes historical claims $z_1, \ldots, z_n$ are known to exceed some *unknown* threshold b. Their *minimum* provides an estimate, precisely

$$\hat{b} = \min(z_1, \ldots, z_n) - C, \qquad \text{for unbiasedness: } C = \beta \int_0^\infty \{1 - F_0(z)\}^n \, dz; \tag{1.10}$$

see Exercise 9.2.4 for the unbiasedness correction. It is rarely worth the trouble to take that too seriously, and accuracy is typically high even when it isn't done[2]. The estimate is known to be **super-efficient**, which means that its standard deviation for large sample sizes is proportional to $1/n$ rather than the usual $1/\sqrt{n}$; see Lehmann and Casella (1998). Other parameters can be fitted by applying the methods below to the sample $z_1 - \hat{b}, \ldots, z_n - \hat{b}$.

Non-parametric estimation

The random variable $\hat{Z}$ that attaches probabilities $1/n$ to all claims z_i of the past is a possible model for *future* claims. Its definition in (1.1) as a discrete set of probabilities may seem at odds with the underlying distribution being continuous, but experience in statistics (see Efron and Tibshriani, 1994) suggests that this matters little. As with other distributions there are an expectation, a standard deviation, a skewness coefficient and also percentiles. All those are close related to the ordinary sample versions. For example, the mean and standard deviation of $\hat{Z}$ are by definition

$$E(\hat{Z}) = \sum_{i=1}^n \frac{1}{n} z_i = \bar{z}, \qquad \text{and} \qquad \mathrm{sd}(\hat{Z}) = \left(\sum_{i=1}^n \frac{1}{n} (z_i - \bar{z})^2 \right)^{1/2} \doteq s. \tag{1.11}$$

Percentiles are approximately the historical claims in descending order; i.e.

$$\hat{q}_\varepsilon = z_{(\varepsilon n)} \qquad \text{where} \qquad z_{(1)} \geq \ldots \geq z_{(n)}.$$

The skewness coefficient is also similar; see Exercise 9.2.4.

The empirical distribution function can only be visualized as **dot plot** where the observations $z_1, \ldots, z_n$ are recorded on a straight line to make their tightness indicate the underlying distribution. If you want a density function, turn to the kernel estimate in Section 2.2, which is related to $\hat{Z}$ in the following way. Let ε be a random variable with mean 0 and standard deviation 1, and define

$$\hat{Z}_h = \hat{Z} + hs\varepsilon, \qquad \text{where} \qquad h \geq 0. \tag{1.12}$$

The distribution of $\hat{Z}_h$ coincides with the estimate (**??**); see Exercise 9.2.5. Note that

$$\mathrm{var}(\hat{Z}_h) = s^2 + (hs)^2 \qquad \text{so that} \qquad \mathrm{sd}(\hat{Z}_h) = s\sqrt{1+h^2},$$

[2]The adjustment requires C to be *estimated*. It is in any case sensible to subtract some *small* number $C > 0$ from the minimum to make $z_i - \hat{b}$ strictly positive. Software may crash otherwise.

a slight inflation in uncertainty over that found in the historical data. With the usual choices of h that can be ignored. Sampling is still easy (Exercise 9.2.6), but usually there is not much point in using a positive h for other things than visualization.

In finance the empirical distribution function is often referred to as **historical simulation**. It is ultra-rapid to set up and to simulate (use Algorithm 4.1), and there is no worry as to whether a parametric family fits or not. On the other hand, *no simulated claim can be larger than what has been seen observed in the past.* How serious that drawback is depends on the situation. It may not matter too much when there is extensive experience to build on. In the big consumer branches of motor and housing we have presumably seen much of the worst. The empirical distribution function *can* also be used with big claims when the responsibility per event is strongly limited, but if it is not, the method can go seriously astray and under-estimate risk substantially. Even then is it possible to combine the method with specific techniques for tail estimation as in Section 9.5; see also some of the exercises.

1.3 The Gamma family

Introduction
Gamma distributions were introduced in Section 2.6. The family is a two-parameter one for which

$$Z = \xi Z_0 \qquad \text{where} \qquad Z_0 \sim \text{Gamma}(\alpha). \tag{1.13}$$

Here Z_0, with mean one, will be called a **standard** Gamma. Its density function is

$$f_0(z) = \frac{\alpha^\alpha}{\Gamma(\alpha)} z^{\alpha-1} \exp(-\alpha z), \quad z > 0; \tag{1.14}$$

see (**??**). Good operational qualities and flexible shape makes the Gamma model useful in many contexts. It is one of the most common distributions for claim size in property insurance, and has also been used with stochastic volatility and with stochastic claim intensity; see Sections 2.3. and 8.5.

Properties
Mean, standard deviation and skewness of a Gamma distributed Z are

$$E(Z) = \xi, \qquad \text{sd}(Z) = \xi/\sqrt{\alpha} \qquad \text{and} \qquad \text{skew}(Z) = 2/\sqrt{\alpha}, \tag{1.15}$$

and the model possesses a so-called **convolution** property. Let $Z_{01}, \ldots, Z_{0n}$ be an independent sample from Gamma(α). Then

$$\bar{Z}_0 \sim \text{Gamma } (n\alpha) \qquad \text{where} \qquad \bar{Z}_0 = (Z_{01} + \ldots + Z_{0n})/n;$$

see Appendix A. In other words, the average is another standard Gamma variable, the shape now being $n\alpha$. By the central limit theorem $\bar{Z}_0$ also tends to the normal as $n \to \infty$, and this proves that Gamma variables become normal as $\alpha \to \infty$. This is visible in Figure 9.1 left where Gamma percentiles are Q-Q plotted aganst normal ones. The line is much straightened out as $\alpha = 10$ is replaced by $\alpha = 100$. A similar tendency is seen among the density functions in Figure 9.1 right where two of the shapes were used with stochastic intensies in Section 8.5. More general versions of the convolution property are given among the exercises.

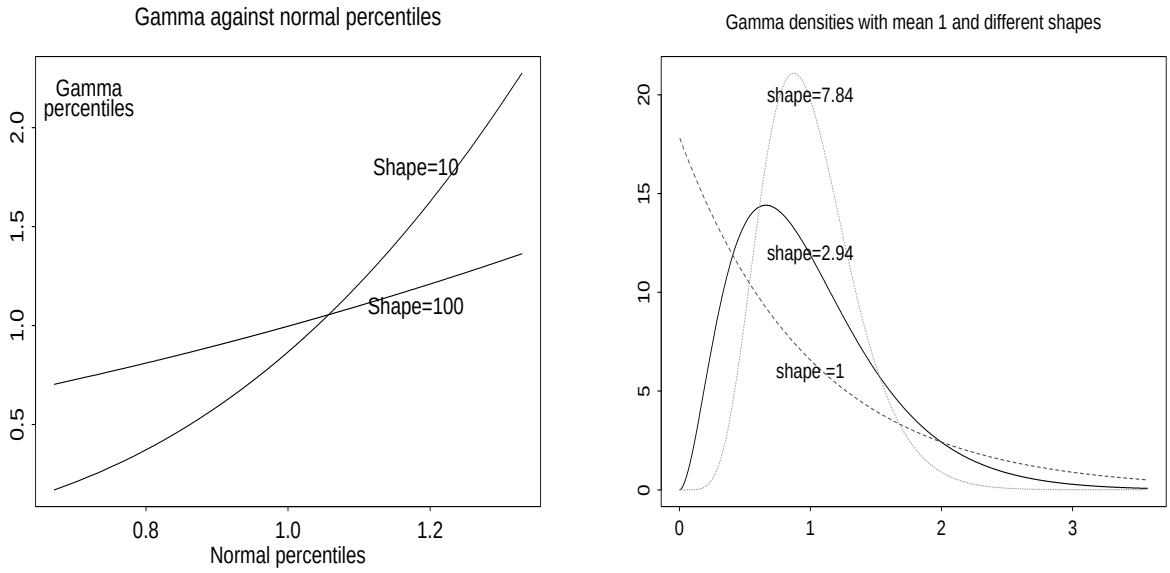


Figure 9.1 **Left***: Q-Q plot of standard Gamma percentiles against the normal.* **Right***: Standard Gxamma density functions.*

Fitting the Gamma familiy

The method of moments (Section 7.3) is the simplest way to determine Gamma parameters ξ and α from a set of historical data $z_1, \ldots, z_n$. If the theoretical expressions are matched sample mean and standard deviation $\bar{z}$ and s, we obtain

$$\bar{z} = \hat{\xi}, \quad s = \hat{\xi}/\sqrt{\alpha} \qquad \text{with solution} \qquad \hat{\xi} = \bar{z}, \quad \hat{\alpha} = (\bar{z}/s)^2.$$

Likelihood estimation is slightly more accurate, and is available in commercial software, but it is not difficult to implement on your own. The logarithm of the density function of the standard Gamma is

$$\log\{f_0(z)\} = \alpha \log(\alpha) - \log\{\Gamma(\alpha)\} + (\alpha - 1)\log(z) - \alpha z$$

which can be inserted into (1.5). After some simple manipulations this yields the log likelihood function

$$\mathcal{L}(\xi, \alpha) = n\alpha \log(\alpha/\xi) - n \log \Gamma(\alpha) + (\alpha - 1) \sum_{j=1}^{n} \log(z_j) - \frac{\alpha}{\xi} \sum_{j=1}^{n} z_j. \tag{1.16}$$

Note that

$$\frac{\partial \mathcal{L}}{\partial \xi} = -\frac{n\alpha}{\xi} + \frac{\alpha}{\xi^2} \sum_{i=1}^{n} z_i, \qquad \textbf{zero} \text{ when} \qquad \xi = (z_1 + \ldots + z_n)/n = \bar{z}.$$

It follows that $\hat{\xi} = \bar{z}$ is the likelihood estimate and $\mathcal{L}(\bar{z}, \alpha)$ can be tracked under variation of α for the maximizing value $\hat{\alpha}$; see also the bisection method in Appendix B.

Gamma and related regressions
Sometimes you may want to examine whether losses tend to be systematically higher with certain customers than with others. To the author's experience the issue is not so important as it was with claim frequency, but we should know how it's done. Basis are historical data similar to those in Section 8.4, now of the form

$$\begin{array}{cc} z_1 & x_{11} \cdots x_{1v} \\ z_2 & x_{21} \cdots x_{2v} \\ \cdot & \cdot \ \cdots \ \cdot \\ \cdot & \cdot \ \cdots \ \cdot \\ z_n & x_{n1} \cdots x_{nv}, \\ \textit{losses} & \textit{covariates} \end{array}$$

and the question is how we use them to understand how a future, reported loss Z are connected to explanatory variables $x_1, \ldots, x_v$. The standard approach is through

$$Z = \xi Z_0 \qquad \text{where} \qquad \log(\xi) = b_0 + b_1 x_1 + \ldots + b_v x_v,$$

and $E(Z_0) = 1$. As the explanatory variables fluctuate, so does the mean loss ξ.

Frequently applied models for Z_0 are log-normal and Gamma. The former simply boils down to ordinary linear regression. The logarithm of the claim in then used as as dependent variable and the explanatory variables fitted through least squares; see Exercise ??. Gamma regression is available in commercial software and implemented as likelihood fitting through an extension of (1.16), see Exercise 9.4.?, but you may be unlikely to implement this on your own. For an example, see Section 10.3.

Simulation algorithm
Gamma sampling is not so easy. A simple rejection/acceptance method was developped in Section 2.6 (Algorithm 2.11), but much faster procedures are available; see Devroye (1986). The following procedure is a more a fined-tuned version of rejection/acceptance. It is efficient and reasonably simple to implement:

Algorithm 9.1a Sampling Gamma (**Requirement**: $\alpha \geq 1$)
0 Input: ξ, α and $b = \alpha - 1$, $c = 3\alpha - 0.75$.
1 Repeat
2 Sample $U^* \sim$ uniform
3 $W^* \leftarrow U^*(1 - U^*)$, $Y^* \leftarrow \sqrt{c/W^*}(U^* - 0.5)$, $Z^* \leftarrow b + Y^*$
6 If $Z^* > 0$ then
7 Sample $V^* \sim$ uniform$(0, 1)$
8 $X^* \leftarrow 64(W^*)^3(V^*)^2$
9 If $X^* \leq 1 - 2(Y^*)^2/Z^*$ **or** if $\log(X^*) \leq 2\{b(\log(Z^*/b) - Y^*)\}$
then **stop** and return $Z^* \leftarrow \xi Z^*/\alpha$.

The loop is repeated until the stopping criterion is satisfied. Note that algorithm only works for when $\alpha \geq 1$. In the opposite case, we may invoke **Stuart's** theorem; i.e.

$$Z = \xi Y U^{1/\alpha} \sim \text{Gamma}(\alpha, \xi) \qquad \text{if} \qquad Y \sim \text{Gamma}(1 + \alpha),\ U \sim \text{Uniform};$$

see Devroye (1986). In other words, the case $\alpha < 1$ is referred back to the shape parameter $1 + \alpha$, which is solved by Algorithm 9.1a. In detail:

Algorithm 9.1b Sampling Gamma (**Requirement**: $\alpha < 1$)
0 Input: ξ, α
1 Sample $Z^* \sim$ Gamma$(1 + \alpha)$ *%Standard Gamma, Algorithm 9.1a*
2 Sample $U^* \sim$ uniform
3 Return $Z^* \leftarrow \xi Z^* (U^*)^{1/\alpha}$

Together the two sub-algorithms offers efficient sampling from the *general* Gamma distribution. The average number of trials in Algorithm 9.1a is about ??.

1.4 The Pareto family and extremes

Introduction
The Pareto distributions, introduced in Section 2.6, are among the most heavy-tailed of all models in practical use and potentially a conservative choice when evaluating risk in property insurance. Density and distribution functions are

$$f(z) = \frac{\alpha/\beta}{(1 + z/\beta)^{1+\alpha}} \quad \text{and} \quad F(z) = 1 - \frac{1}{(1 + z/\beta)^{\alpha}}, \qquad z > 0.$$

Simulation was easy (Algorithm 2.8), and the model was used for illustration in several of the earlier chapters. But Pareto distributions also play a special role in the mathematical description of the extreme right tail. There are, perhaps surprisingly, *general* results in that direction. That is the main topic of this section.

Properties
Pareto models are so-heavy-tailed that even the mean may fail to exist (that's why another parameter β represents scale). Formulae for expectation, standard deviation and skewness are

$$\xi = E(Z) = \frac{\beta}{\alpha - 1}, \quad \text{sd}(Z) = \xi \left(\frac{\alpha}{\alpha - 2} \right)^{1/2}, \quad \text{skew}(Z) = 2 \left(\frac{\alpha}{\alpha - 2} \right)^{1/2} \frac{\alpha + 1}{\alpha - 3}, \tag{1.17}$$

valid for $\alpha > 1$, $\alpha > 2$ and $\alpha > 3$ respectively. It is to the author's experience rare in practice that the mean doesn't exist, but infinite variances with values of α between 1 and 2 are not unfrequent in big claims situations. We saw in Section 2.6 that the exponential distribution appears in the limit when the ratio β/α is kept fixed and their common value raised to infinity. The Pareto and the Gamma families intesect in this sense, the exponential being the most *light*-tailed among the former and one of the *heavy*-tailed ones among the latter.

One of the most important properties of the Pareto family is its behaviour at the extreme right tail. The issue is defined by the **over-threshold** model which is the distribution of $Z_b = Z - b$ given $Z > b$. Its density function (derived in Section 6.2) is

$$f_b(z) = \frac{f(b + z)}{1 - F(b)};$$

see (**??**). It becomes particularly simple with Pareto models. Inserting the expressions for $f(z)$ and $F(z)$ yields

$$f_b(z) = \frac{(1+b/\beta)^\alpha \alpha/\beta}{(1+(z+b)/\beta)^{1+\alpha}} = \underbrace{\frac{\alpha/(\beta+b)}{\{1+z/(\beta+b)\}^{1+\alpha}}}_{\textit{Pareto density function}}$$

after some simple manipulations. This is again a Pareto density. The shape α is the same as before, whereas the parameter of scale has become $\beta_b = \beta + b$. In other words, over threshold models for Pareto variables remain Pareto with shape unaltered. The mean (if it exists) is known as the **mean excess function**, and becomes

$$E(Z_b|Z>b) = \frac{\beta_b}{\alpha-1} = \frac{\beta+b}{\alpha-1} = \xi + \frac{b}{\alpha-1} \qquad \text{(requires } \alpha > 1\text{)}. \tag{1.18}$$

It is larger than the original ξ and increases linearly with b.

Pickands' theorem

The tail property of Pareto models has a *general* extension. In 1975 Pickand discovered that only very limited classes of distribution can appear as over-threshold models when b becomes infinite. In fact, if the parent distribution for Z is continuous with no upper limit, the over-threshold model is bound to become Pareto as $b \to \infty$. That occurs *whatever distribution we started with*; see Pickands (1975). There is also a theory when Z is bounded by some given maximum, but such models are less frequent and perhaps less natural to employ; see Embrects, Klüppelberg and Mikosch (1997) for that extension.

The mathematical formulation requires some additional notation. Let $F(z|\alpha,\beta)$ be the distribution function of Pareto(α,β) and define

$$G_b(z) = \Pr(Z_b \le z|Z>b) = \Pr(Z \le b+z|Z>b)$$

as the over threshold distribution function of an *arbitrary* random variable Z. Suppose Z is continuous without fixed upper limit. Then *there exists a positive parameter α (possibly infinite) such that for all thresholds b there are parameters β_b (depending on b) that makes*

$$\max_{z\ge 0} |G_b(z) - F(z|\alpha,\beta_b)| \to 0, \quad \text{as} \quad b \to \infty.$$

This tells us that discrepancies between the two distribution functions vanish as the threshold grows. At the end they have become equal, and the over-threshold distribution a member of the Pareto family. The result applies for *finite* b (with $\beta_b = \beta + b$), if the original model was Pareto. Note that α can be infinite. The limit distribution is then the exponential which we may regard as a limiting member of the Pareto family.

Fitting under right censoring

Estimation of Pareto parameters was treated in Section 7.3, but we shall now add **censored** observations. That issue was introduced in Section 9.2. Only censoring to the right is worked out. A claim is then known to have exceeded a certain threshold, but not by how much. Observations are now in two groups, the ordinary, fully observed claims $z_1,\ldots,z_n$ and the n_r censored ones above

thresholds $b_1, \ldots, b_{n_r}$. The likelihood method offers a suitable way to draw on both sources of information. For the first group the log likelihood function is as in Section 7.3; i.e.

$$\mathcal{L}_1(\alpha, \beta) = n \log(\alpha/\beta) - (1+\alpha) \sum_{i=1}^{n} \log(1 + \frac{z_i}{\beta}).$$

Then there is the censored part where it is only known that $Z_i > b_i$. The probability of this happening is

$$\Pr(Z_i > b_i) = \frac{1}{(1 + b_i/\beta)^\alpha} \qquad \text{or} \qquad \log\{\Pr(Z_i > b_i)\} = -\alpha \log(1 + \frac{b_i}{\beta}),$$

and when all those are summed over i and added $\mathcal{L}_1(\alpha, \beta)$, we obtain the full log likelihood

$$\mathcal{L}(\alpha, \beta) = \underbrace{n \log(\alpha/\beta) - (1+\alpha) \sum_{i=1}^{n} \log(1 + \frac{z_i}{\beta})}_{\textit{complete information}} \quad - \quad \underbrace{\alpha \sum_{i=1}^{n_r} \log(1 + \frac{b_i}{\beta})}_{\textit{Censoring to the right}}.$$

which is to be maximized.

As a numerical problem this is about the same as in Section 7.3. An elementary procedure you can program on your own is the following. For given β ordinary differentiation proves that $\mathcal{L}(\alpha, \beta)$ is maximized by

$$\hat{\alpha}_\beta = n \left(\sum_{i=1}^{n} \log(1 + \frac{z_i}{\beta}) + \sum_{i=1}^{n_r} \log(1 + \frac{b_i}{\beta}) \right)^{-1},$$

and the optimum is found by tracking the function $\mathcal{L}(\hat{\alpha}_\beta, \beta)$ over β, as explained in Section 7.3. The best way of solving such one-dimensional problems is through the bisection method; see Appendix C.

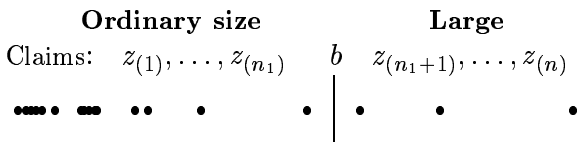
1.5 Large claim situations

Introduction

The big claims play a special role because of their importance financially. It is also hard to assess their distribution. They (luckily!) do not occur very often, and historical experience is therefore limited. Indeed, insurance companies may give cover to claims *larger* than have been seen earlier. What should our approach be in these situations? The simplest would be to fit a parametric family and extrapolate beyond past experience, but that may not be a very good idea. A Gamma distribution may fit well in the central regions without being reliable at all at the extreme right tail. Indeed, such a procedure may easily underestimate big claims risk severely; see Section 9.6. A Pareto model would be more conservative, and then there is the result due to Pickands that points to this distribution as a general description above all large thresholds. There is an idea here, and the purpose of the present section is to develope it.

An approach through mixtures

Historical claims look schematically like the following:



There are many values in the small and medium range to the left of the vertical bar and just a few (or none!) large ones to the right of it. What is actually meant by 'large' is not clear-cut, but let us say that we have selected a threshold b defining 'large' claims as those exceeding it. The original claims $z_1, \ldots, z_n$ have been ranked in ascending order as

$$z_{(1)} \leq z_{(2)} \cdots \leq z_{(n)}$$

so that observations from $z_{(n_1)}$ and smaller are below the threshold and those from $z_{(n_1+1)}$ and larger are above. How the threshold b is chosen in practice is discussed below; see also the numerical illustrations in Section 9.6.

Our strategy is to divide modelling into separate parts defined by the threshold. A random variable (or claim) Z may always be written

$$Z = (1 - I_b) Z_{\leq b} + I_b Z_{>b} \tag{1.19}$$

where

$$\begin{array}{lll} Z_{\leq b} = Z | Z \leq b, & Z_{>b} = Z | Z > b \quad \text{and} & I_b = 0 \quad \text{if } Z \leq b \\ \textit{central region} & \textit{extreme right tail} & = 1 \quad \text{if } Z > b. \end{array} \tag{1.20}$$

The random variable $Z_{\leq b}$ is Z confined to the region to the left of b, and $Z_{>b}$ is similar to the right. It is easy to check that two sides of (1.19) are equal, but at first sight this merely looks complicated. Why on earth can it help us? The point is that we have created a framework reaching out to two different sources of information. To the left of the threshold there is the historical data with which we may identify a model. On the right the result due to Pickands suggests a Pareto distribution. This defines a modelling strategy which will now be developped.

The empirical distribution mixed with Pareto

The preceding argument lead to a two-component approach which can be implemented in many ways. For example, to the left of b we could fit a parametric model. It would extend beyond b, but that may not matter too much; see Exercise ??. Another idea is to use non-parametric modelling, and this is the method that will be developed in detail with the threshold selected as one of the observations. *Choose* some small probability p and let $n_1 = n(1-p)$ and $b = z_{(n_1)}$. Then take

$$Z_{\leq b} = \hat{Z} \qquad \text{and} \qquad Z_{>b} = z_{(n_1)} + \text{Pareto}(\alpha, \beta), \tag{1.21}$$

where $\hat{Z}$ follows the empirical distribution function over $z_{(1)}, \ldots, z_{(n_1)}$; i.e.

$$\Pr(\hat{Z} = z_{(i)}) = \frac{1}{n_1}, \quad i = 1, \ldots, n_1. \tag{1.22}$$

The remaining part (the delicate one!) are the parameters are α and β and the choice of p. Plenty of historical data would deal with everything. Under such circumstances p can be determined low

enough (and hence b high enough) for the Pareto approximation to be a good one, and historical data to the right of b provides estimates $\hat{\alpha}$ and $\hat{\beta}$. There are even sophisticated, automated techniques for the selection of p, see ? and ?. In practice you might do just as well with trial and error. An example of this kind is discussed in the next section.

With more limited experience (as is common) is is hard to avoid a subjective element. One of the advantages of dividing modelling into two components is that it clarifies the domain where personal judgment enters. If you take the view that a degree of conservatism is in order when there is insufficient information for accuracy, that can be achieved by placing b low and using Pareto modelling to the right of it. Numerical experiments that supports such a strategy are carried out in the next section. Much material on modelling extremes can be found in Embrects, Klüppelberg and Mikosch (1997).

Sampling mixture models
As usual a sampling algorithm is also a summary of how the model is constructed. With the empirical distribution used for the central region it runs as follows:

Algorithm 9.2 Claims by mixtures
0 Input: Sorted claims $z_{(1)} \leq \ldots \leq z_{(n)}$, p, $n_1 = n(1-p)$, α and β.
1 Draw uniforms U_1^*, U_2^*
2 If $U_1^* > p$ then
3 $i^* \leftarrow 1 + [n_1 U_2^*]$ and $Z^* \leftarrow z_{(i^*)}$ *% The empirical distribution, Algorithm 4.1*
else
4 $Z^* \leftarrow b + \beta\{(U_2^*)^{-1/\alpha} - 1\}$ *%Pareto, Algorithm 2.8*
5 Return Z^*

The algorithm operates by testing whether the claim comes from the central part of the distribution or from the extreme, right tail over b. Other distributions could have been used on Line 3. The present version is extremely quick to implement.

1.6 Searching for the model

Introduction
A final model for claim size is the result of different deliberations. Historical data have typically been utilized through a non-parametric approach or with parametric families. We may also have changed the variable. The idea is then that standard families of distributions may fit a **transformed** variable better than the original one, and with re-transformation afterwards the model again applies to ordinary claims. One of our worries should be model error. Does the distribution selected reflect the uncertainty of real life? If there are small amounts of data to go on, the discrepancy could be huge. Should that lean us towards concervative choices? If accurate mathematical descriptions are beyond reach anyway, it could be an argument in favour of heavy-tailed distributions like Pareto.

The purpose of this section is to indicate how these themes enter by means of two very different examples. We have already met the Norwegian fund for natural disasters in chapter 7 where there were just $n = 21$ historical incidents to rely on. By contrast the so-called Danish fire claims will serve our needs for a 'large' data set. Many authors on actuarial science have used it as a

test case; see Embrechts, Klüppelberg and Mikosch(1997) where more on their orgin is given. The historical record comprises $n = 2167$ industrial fires. Damages start at one million Danish kroner (DKK)[3] with 263 as a maximum and with average $\bar{z} = 3.39$, standard deviation $s = 8.51$ and skewness coefficient $\gamma = 18.7$. The latter indicates very heavy tails and strong skewness towards the right. This also emerges clearly from the plots in Figure 9.2 and 9.3 below.

Working with transformations
A useful tool for modelling is to change data by means of a transformation, say $H(z)$. The situation is then as follows:

$$\underset{\textit{original data}}{z_1, \ldots, z_n} \qquad\qquad \underset{\textit{new data}}{y_1 = H(z_1), \ldots, y_n = H(z_n).}$$

Modelling is then attacked through $y_1, \ldots, y_n$ and $Y = H(Z)$ instead of the original Z. The idea is to make one of the simple models fit better than could be achieved with Z itself. At the end we re-transform back through $Z = H^{-1}(Y)$ with $Z^* = H^{-1}(Y^*)$ for the Monte Carlo. The log-normal is a familiar example. Then $H(z) = \log(z)$ and $H^{-1}(y) = \exp(y)$ with Y normal. The logarithm is the most commonly used transformation of all. Frequently applied alternatives are powers $Y = Z^\eta$ where $\eta \neq 0$ is a some given index. The choice of transformations (typically made by trial and error) is a *second* feature that adds flexibility to the usual families of distributions.

Variations on this theme are indeed many. With logaritms we might take

$$\underset{Y \textit{ positive}}{Y = \log(1 + Z)} \qquad\qquad \underset{Y \textit{ over the entire real line}}{Y = \log(Z),}$$

and entirely different families of distributions would be used for Y. As an example consider the Danish fire claims where we must take into account that they run from 1 and upwards (in milllion DKK). That makes $Y = \log(Z)$ positive, and one possibility could be the log-normal through

$$Z = e^Y, \quad Y = \xi_y e^{-\tau^2/2 + \tau\varepsilon} \qquad \text{with estimates} \qquad \hat{\xi_y} = 1.19,\ \hat{\tau} = 1.36.$$

Here ε is $N(0, 1)$. An alternative is the Gamma familiy. Let $Y_{0\alpha}$ be Gamma distributed with mean one and shape α and consider

$$Z = e^Y, \quad Y = \xi_y Y_{0\alpha} \qquad \text{with estimates} \qquad \hat{\xi_y} = 0.79,\ \hat{\alpha} = 1.16.$$

Both pairs of estimates are likelihood ones.

What is immediately clear from the huge discrepancy in the estimated means ξ_y is that both models can't fit. Indeed, the log-normal doesn't work. Its estimated density function (Figure 9.2 left, horizontal axis on *logarithmic* scale) matches the kernel density estimate poorly, but (as usual) Q-Q plotting (Figure 9.2 right) provides a better view. The right tail of the log-normal is too heavy and exaggerates the risk of extreme claims grossly[4]. By contrast the Gamma fit as displayed in Figure 9.3 is much better. Perhaps the extreme right tail is slightly too light, but the fit isn't an

[3]There are about eight Danish kroner in one euro.

[4]Note that the 167 largest observations have been left out to make the resolution in other parts of the plot better

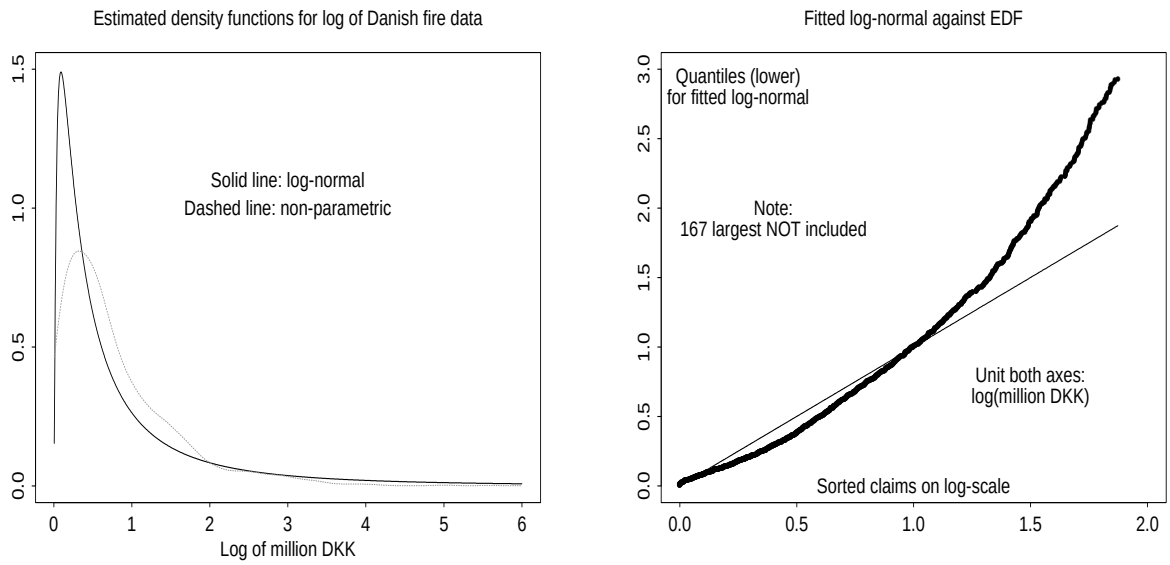


Figure 9.2 The **log-normal** *model fitted the Danish fire data on* **log***-scale. Density function with kernel density estimate (left) and Q-Q plot (right).*

end in itself, and consequences for the evaluation of the reserve is not necessarily serious. That will be examined in Section 10.3; see also Exercise 9.6.2 where a slight modification will improve the fit.

Pareto and Pareto mixing

The Pareto model is so heavy-tailed on its own that it could be tried on the raw Danish fire data directly (without log-transform). It is also a strong candidate for the extreme right tail (Section 9.4). Indeed, with such an extensive data record it is tempting to forget all about parametric families and use the strategy advocated in Section 9.5 using the empirical distribution function for the central part and Pareto on the right. Table 9.2 shows the results of fitting Pareto distributions (through maximum likelihood) over various thresholds b. As b is being raised, the situation should become more and more Pareto-like (Pickand's theorem). Under a strict Pareto regime, the shape parameter α is the same for all b whereas the scale parameter depends on b through $\beta_b = b - 1 + \beta/(\alpha - 1)$; see Exercise ?. Stretching the imagination a bit there are reminiscences of this in Table 9.1 where α is more stable than β; see Exercise 9.4.1 for detailed calculations.

But it would be a gross exaggeration to proclaim the data to be Pareto distributed. Q-Q plots for two of the over threshold distributions is shown in Figure 9.4. There is a reasonable fit on the right (above 5%), but not on the left (above 50%) where the Pareto distribution fitted has heavier tails that the empirical counterpart. Table 9.1 tell us why. The two shape parameters estimated (1.42 and 2.05) deliver quite unequal extreme uncertainty.

Tiny historical records

How should we confront a situation like the one in Table 7.1 (the Norwegian natural disasters) where there were no more than $n = 21$ claims in total, and where the phenomenon itself surely is

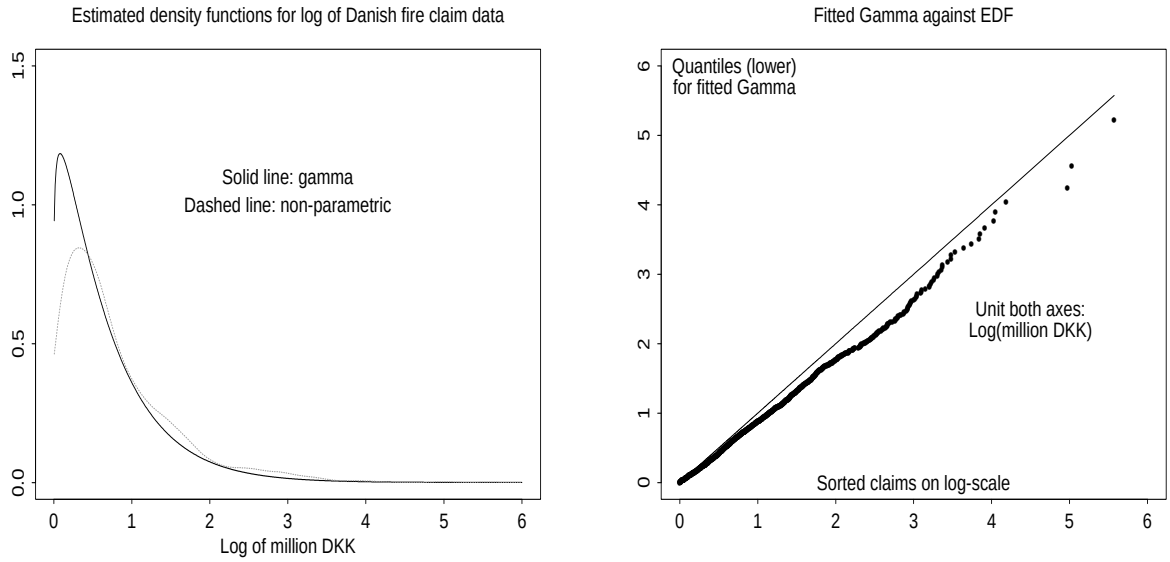


Figure 9.3. The **Gamma** *model fitted the Danish fire data on* **log***-scale. Density function with kernel density estimate (left) and Q-Q plot (right).*

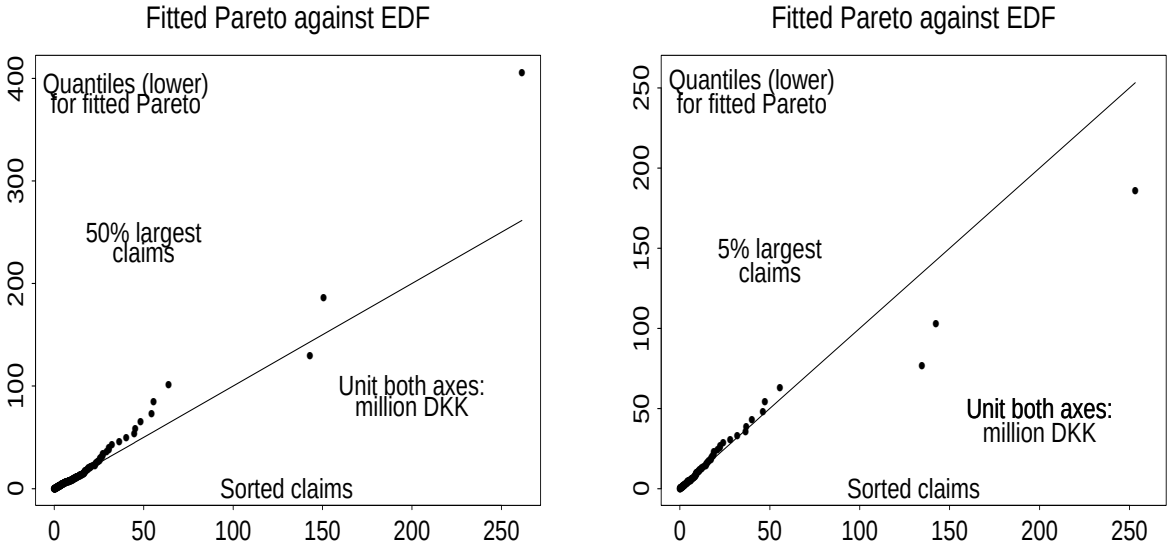


Figure 9.4 Q-Q plots of fitted **Pareto** *distributions against the empirical distribution function,* 50% *largest observations (left) and* 5% *largest (right).*

heavy-tailed with potential losses much larger than those on record? The underlying distribution can't be determined with much accuracy, yet somehow a model must be found. One possibility is geophysical modelling. Natural disasters are then simulated in the computer and their cost counted from detailed, physical descriptions of houses and installations. Evaluations of this kind are carried out around the world, but they are outside out natural range of topics, and we shall concentrate on what can be extracted from historical losses.

If you fit the Gamma and Pareto family to the natural disasers by maximum likelihood, the results look like this:

Shape	Mean	5%	1%
0.72	179	603	978
Gamma family			

Shape	Mean	5%	1%
1.71	200	658	1928
Pareto family			

These are *very* different families of distributions, yet their discrepancies, though considerable, are not enormous in the central region (say up to the upper 5% percentile). For the *very* large claims that changes, and the Pareto 1% percentile is twice that of the Gamma. There is a lesson here. Many families fit reasonably well up to some moderate threshold. *That makes modelling easier when there are strong limits on responsibilities.* If it isn't, the choice between parametric families becomes a more delicate one.

The right family: Impossible?
Incidentally, how impossible is it to determine the family from small amounts of data? Suppose a Q-Q plot is used. A given family such as Gamma or Pareto is then evaluated by comparing their estimated percentiles $\hat{q}_i$ to empirical ones $z_{(i)}$ where the former correspond to distributions fitted the data. What is actually done when the two sequences are matched, is unclear (different ways for different people), but perhaps some try to minimize

$$Q = \sum_{i=1}^{n} |\hat{q}_i - z_{(i)}|. \tag{1.23}$$

This criterion has been proposed as basis for formal goodness of fit tests in Devroye (1971). It could be that humans do it a little better, but results using other critera didn't deviate that much from those in Table 9.2.

Unit: Million DKK

	Part of data fitted			
	All	50% largest	10% largest	5% largest
Threshold (b)	1.00	1.77	5.56	10.01
Shape (α)	1.64	1.42	1.71	2.05
Scale (β)	1.52	1.82	7.75	14.62

Table 9.1 Pareto parameters for the over threshold distribution of the fire claims.

Shapes in true models: 1.71 in Pareto, 0.72 in Gamma. 1000 *repetitions.*

True model	*Historical record:* $n = 21$ *Models found*			*Historical record:* $n = 80$ *Models found*		
	Pareto	Gamma	log-normal	Pareto	Gamma	log-normal
Pareto	**.49**	.29	.22	**.72**	.12	.16
Gamma	.44	**.51**	.05	.34	**.66**	0

*Table 9.2 Probabilites of selecting given models (***Bold face***: Correct selection).*

Monte Carlo experiments were run with $m = 1000$ replications according to the following scheme:

$$\begin{array}{ccccccc} \textbf{True model} & & & & \text{Parametric family tried} & & \\ \text{Pareto} & & & & \textit{fitting } \hat{q}_1^* \geq \ldots \geq \hat{q}_n^* & & \\ \text{or} & \longrightarrow & z_1^*, \ldots, \geq z_n^* & \longrightarrow & & \longrightarrow & Q^* = \sum_i |z_{(i)}^* - \hat{q}_i^*|. \\ \text{Gamma} & & \textit{historical data} & & \textit{sorting } z_{(1)}^* \geq \ldots z_{(n)}^* & & \end{array}$$

Simulated historical data were drawn from the Pareto or Gamma model on the left and the model (possibly a different one!) fitted. That gave estimated percentiles $\hat{q}_i^*$ which could be compared to purely empirical ones $z_{(i)}^*$ and a value of the criterion Q^* computed for the parametric model tried. When Pareto, Gamma and log-normal were fitted to the same historical data, we obtain three different evaluations Q^*, and the distribution corresponding to the smallest, best-fitting one was picked. The selection statistics is shown in Table 9.2. It is clearly impossible to choose between the three models when there are only $n = 21$ claims. The chance is improved with $n = 80$ and with $n = 400$ (not shown) the success probability was about $0.90 - 0.95$.

Data in short supply: What then?

The preceding experiment showed the futility of trying to identify models from small amounts of historical data, but when faced with such situations, how should they be attacked? Here are some tentative suggestions. A good deal hinges on the maximum responsibility b per claim. If it is *smaller* than the *largest* observation $z_{(1)}$, it could be a case for the empirical distribution function. That doesn't help us much with the Norwegian natural disasters from Section 7.4 where b is *much larger* than $z_{(1)}$, and risk would be grossly under-estimated by that method. Surely the Pareto distribution is one of the leading contenders now. It is a conservative choice (which seems sensible), possibly estimation errors undermine some of that caution.

These points are illustrated by the experiment in Table 9.3 where the issue is the consequences of being wrong. For example, if the underlying distribution is a member of the Gamma family, how does a Pareto fit perform? Or what about estimated Gamma percentiles when the true model is Pareto? Clauses of maximum payments have much bearing on this (as mentioned), but these problems can also be inspected through

$$\hat{\theta} = \frac{\hat{q}_\varepsilon}{q_\varepsilon} \qquad \text{for} \qquad \varepsilon = 1\%.$$

Patterns in how $\hat{\theta}$ deviate from 1 reveal the impact of model and estimation error jointly. Suppose the Gamma family is fitted to claims that are actually Pareto distributed. It then emerges from Table 9.3 (Line two from bottom) that the 90% percentile of $\hat{\theta}$ is at most one; i.e. $q_{0.01}$ is

$m = 1000$ *replications*

	True model: **Pareto**, shape = 1.71						True model: **Gamma**, shape = 0.72					
	Record: n=21			*Record: n=80*			*Record: n=21*			*Record: n=80*		
Percentiles (%)	25	75	90	25	75	90	25	75	90	25	75	90
Fitted Pareto	**0.4**	**1.5**	**2.9**	**0.7**	**1.3**	**1.7**	0.8	1.4	2.2	0.9	1.3	1.6
Fitted Gamma	0.3	0.6	1.0	0.4	0.7	0.9	**0.8**	**1.1**	**1.3**	**0.9**	**1.1**	**1.2**
Model selected	0.4	1.2	2.3	0.6	1.2	1.6	0.8	1.2	1.5	0.9	1.1	1.3

Table 9.3 The distribution (as 25 70 *and* 90 *percentiles) of* $\hat{\theta} = \hat{q}_{0.01}/q_{0.01}$ *where* $\hat{q}_{0.01}$ *is fitted and* $q_{0.01}$ *true* 1% *percentiles of claims.* **Bold face***: Correct parametric family used.*

almost certain to be *under*-estimated! The tendency is reversed when the Pareto model is applied to Gamma-distributed losses. Now the percentile is *over*-estimated. Certainly, we *are* doing something silly, and yet in practice we might not know. The method that comes on top in Table 3 is the last one where the percentiles are computed form the best-fitting of both the Gamma and Pareto distributions, i.e. the alternative minimizing (1.23) has been picked. Now the the distribution of $\hat{\theta}$ varies around one, though with huge errors.

In summary it seems sensible to try determine the family empirically even for small data sets (though we often guess wrong). If we go for conservatism and caution, the Pareto model may be the answer despite the huge uncertainty of the fitted parameters.

1.7 Further reading

1.8 Exercises

Section 9.2

Exercise 9.2.1 Let I be a rate of inflation. A new price is then $(1+I)P$ if the old one was P. **a)** Suppose claim size Z is Gamma(α, ξ) in terms of the *old* price system. What are the parameters under the new, inflated price? **b)** The same same quation when the old price is Pareto(α, β). **c)** Again the same question when Z is log-normally distributed; i.e when $\log(Z)$ ~Normal(θ, τ). What are θ and τ under the inflated prices system?

Exercise 9.2.2 Consider the shifted distribution (1.5). What happens to the shift parameter b when prices changes from P to $(1+I)P$ as in the preceding exercise?

Exercise 9.2.3 The empirical distribution function (1.1) has mean and variance coinciding with average and sample variance of the historical data $z_1, \ldots, z_n$. What is the skewness coefficient of this distribution?

Exercise 9.2.4 Let a and b be fixed coefficients. Consider a linear transformation $Y = a + bZ$ of a claim Z. **a)** Show that

$$\text{skew}(Y) = \text{skew}(Z).$$

b) What happnes to the skewness coefficient when currency is changed? **c)** The same question when prices inflate? **d)** Suppose Z is shifted by b as in (1.5). Is the skewness coefficient changed?

Exercise 9.2.5 Redo Exercise 9.2.4, but now for the kurtosis of the distribution; see Section 2.2 for the definition of the kurtosis.

Exercise 9.2.6 We shall in this exericise consider *simulated*, log-normal historical data. We know all answers since the data are generated in the computer laboratory, and shall compare skewness estimated thro fitted distribution with the true one. **a)** Generate $n = 30$ log-normal claims using $\theta = -0.5$ and $\tau = 2$. **b)** Compute the skewness coefficient (1.5). Redo four times. **c)** Compare the five estimated skewness coefficients and percentiles in c) with the true ones you compute from the underlying log-normal distributions. Are there patterns? **d)** Redo a),b) and c) when $\tau = 0.5$. What about the patterns now?

Exercise 9.2.7 This exericise works with *simulated*, log-normal historical data as in Exercise 9.2.6, but we are now assuming that we know the correct family distributions. The parameters are still unknown, and the issue is the impact of the errors in their estimates. **a)** Generate $n = 30$ log-normal claims using $\theta = -0.5$ and $\tau = 2$. **b)** Fit a log-normal distribution and compute its skewness coefficient and its upper 5% percentile. Redo four times. **c)** Compare the five estimated skewness coefficients and percentiles in c) with the true ones. Are there patterns? **d)** Redo a),b) and c) when $\tau = 0.5$. Have the patterns changed?

Exercise An example is the log-normal

$$Z = \exp(\theta + \tau\varepsilon) \qquad \text{for which} \qquad \text{skew}(Z) = \frac{\exp(3\tau^2) - 3\exp(\tau^2) + 2}{(\exp(\tau^2) - 1)^{3/2}},$$

which approaches 0 as $\tau \leftarrow 0$ and increases rapidly as τ is raised. It is around 8 for $\tau = 1$, corresponding to the highly asymmetric density function in Figure 2.4 right.

Exercise

$$\text{skew}(\hat{Z}) = \frac{n^{-1}\sum_{i=1}^{n}(z_i - \bar{z})^3}{s^3}, \tag{1.24}$$

Section 9.3
Exercise 9.3.1

Section 9.4

$$\mathcal{L}(b_1, \ldots, b_v, \alpha) = \sum_{j=1}^{n}\left(\alpha\log(\alpha/\xi_j) - \log\Gamma(\alpha) + (\alpha - 1)\log(z_j) - \frac{\alpha}{\xi_j}z_j\right). \qquad \text{where} \qquad \log(\xi) = b_0 + b_1 x$$