USING WEIGHTED DISTRIBUTIONS TO MODEL OPERATIONAL RISK

BY

LOURDES B. AFONSO AND PEDRO CORTE REAL

ABSTRACT

The quantification of operational risk has to deal with various concerns regarding data, much more than other types of risk which banks and insurers are obliged to manage. One of the main questions that worries both researchers and practitioners is the bias in the data on the operational losses amounts recorded. We support the assertions made by several authors and defend that this concern is serious when modeling operational losses data and, typically, is presented in all the databases. We show that it's possible, based on mild assumptions on the internal procedures put in place to manage operational losses, to make parametric inference using loss data statistics, that is, to estimate the parameters for the losses amounts, taking in consideration the bias that, not being considered, generates a two fold error in the estimators for the mean loss amount and the total loss amount, the former being overvalued and the last undervalued. In this paper, we do not consider the existence of a threshold for which, all losses above, are reported and available for analysis and estimation procedures. In this sense, we follow a different approach to the parametric inference. Here, we consider that the probability that a loss is reported and ends up recorded for analysis, increases with the size of the loss, what causes the bias in the database but, at the same time, we do not consider the existence of a threshold, above which, all losses are recorded. Hence, no loss has probability one of being recorded, in what we defend is a realist framework. We deduce the general formulae, present simulations for common theoretical distributions used to model (operational reported) losses amounts, estimate the impact for not considering the bias factor when estimating the value at risk and estimate the true total operational losses the bank incurred.

KEYWORDS

Operational risk, loss data, bias, weighted distributions, value at risk.

1. INTRODUCTION

This paper presents an approach to estimate the distribution of the true losses based on the reported losses that can be used to correct the bias in the V@R

Astin Bulletin 46(2), 469-485. doi: 10.1017/asb.2016.4 © 2016 by Astin Bulletin. All rights reserved.

and in the total operational losses. In Section 2, we define our sample (reported losses) and sampling frame (occurred losses). In Section 3, we recall the definition of weighted distributions and prepare it to apply it to operational risk in Section 4. In this Section, we deduce formulae for Exponential and Pareto distributions. In Section 5, we apply our formulae to four distributions (Exponential, Pareto, Lognormal and Weibull) with detailed explanation for the Exponential distribution, we check the accuracy of our methodology and obtain V@R, TV@R and the expected value of the true operational losses that the bank incurred. Section 6 presents some remarks and conclusions.

The quantification of operational risk has to deal with various concerns regarding data, much more than other types of risk which banks and insurers are obliged to manage. Several studies, at first more empirical and at present already more theoretical and mathematical supported, document several of those concerns. First of all, the lack of internal or external data on operational losses. Although this problem has, in the last years, been dealt with by researchers and practitioners, by using data collected by commercial vendors, these commercial databases still have various handicaps that, more or less, summarize the problems regarding operational loss data and, at the same time, drives the motivation to our approach to the problem of making parametric inference, using loss data statistics, in some cases aggregated data, e.g. totals or mean values.

We can summarize the main problems for operational losses data by:

a. Some of the databases reported in several papers, contain data only for big banks. Shih *et al.* (2000) and Chapelle *et al.* (2005) measured the dependence between a bank size and operational losses in different studies. When they regresses log-losses on a bank's log-sizes Shih *et al.* (2000) estimated a coefficient of 0.25 and Chapelle *et al.* (2005) estimated 0.15. We can say that they are related but there is no fix coefficient, so part of the industry (small banks) is not represented. Another question is raised by the methods applied to compile the databases that, usually, have to depend on public disclosed losses. See, for instance de Fontnouvelle *et al.* (2003), where the authors compare results for two commercial databases collected this way, raising some interesting questions about the data or de Fontnouvelle *et al.* (2005) where some concerns about a completely different collection method and database are reported.
b. Usually these vendors can collect only data for losses that exceed some threshold, 1 million USD being common.
c. Deciding if a loss is an operational loss or not, is another problem posed to data compilers and, once decided that the loss can be classified as such, they have to define in which *business line* and *type* to classify the loss. The common classification being eight business lines, *Corporate Finance; Trading and Sales; Retail Banking; Payment and Settlement; Agency Services; Commercial Banking; Asset Management; and Retail Brokerage* and seven loss types, *Internal Fraud; External Fraud; Employment Practices and Workplace Safety; Clients, Products and Business Practices; Damage to Physical*

https://doi.org/10.1017/asb.2016.4 Published online by Cambridge University Press

Assets; Business Disruption and System Failure; and Execution, Delivery and Process Management.

Hence, when considering the caveats above we can say that (a) and (b) poses problems of bias. In the first case, we have a *structural bias* due to the large size of companies that supply the data, leaving us with a potentially biased database of institutions. For several reasons, mainly because the data is compiled from publicly available sources, only large institutions are considered, this should motivate not so large institutions to compile their own data resulting from their specific experience. In the second case, we have to deal with a confirmed biased sample of operational losses since, the vendors or the data collectors, only report data above a predefined threshold. Again, small companies will not be represented if only large losses are recorded. In the last case (c), we can have misclassification of operational losses, where some losses will not be reported as operational losses, or end up wrongly labeled among the line of business or loss type.

Our motivation is to propose a method to deal with the bias posed not only by the references (a)-(b) above, but also by our experience when dealing with small size insurers and banks. Our experience tells us that it is unlikely that all operational losses end up reported. Even when the institutions have in place methods to detect and document operational losses, intending to be exhaustive and error free, not every operational loss ends us reported. There are two main reasons for that relative small losses, unless all the process is automated, will tend not to be reported (see also, Moscadelli (2004)[page 17]). First, it usually implies cumbersome work and the time used is perceived by professionals not to provide a good cost/benefit relation. Secondly, more usually than not, it implies to recognize an own or a colleagues' error. So that, we are lead to consider that there is a size bias, making more likely to report bigger losses than small losses. However, mainly due to protect the company image and reputation, even some of the largest losses can end up not being reported. Perry and Fontnouvelle (2005) and Cummins *et al.* (2004) study the reputational consequences of operational loss announcements on the value of a bank.

This final consideration being our leitmotif. We are lead to believe that, when dealing with loss data reported due to operational risk, we are always in the presence of a biased sample, no matter if the data comes from a commercial vendor or it is provided by internal procedures to manage operational losses. Even in the situation where there is no threshold for the losses being recorded, that is, even when the institutions try to record all operational losses, we think that the probability of a loss being reported, is still positively correlated with value of the loss, but, at the same time, not all the largest losses are reported. Meaning that, even for high thresholds, there is a chance that a loss will not be reported. The framework for this paper is that, the probability that a loss is reported and ends up recorded for analysis, increases with the size of the loss, what causes the bias in the database but, at the same time, we do not consider that a threshold exists, above which all losses are recorded and available for analysis, hence, no loss has probability one of being recorded.

https://doi.org/10.1017/asb.2016.4 Published online by Cambridge University Press

In our application, we make use of some data, collected by a small Portuguese retail bank that, due to disclosure concerns we will not identify. For instance, in this case, the risk department estimated a probability of 1/250, for an operation to generate a operational loss and of 85%, for the loss ending up reported and documented. The remain data for our application, namely, the sample descriptive statistics comes from Table 6.3 of Chernobai *et al.* (2007) based on the original data set available from Cruz (2002). We will use V@R as the standard risk measure used to evaluate exposure to risk since, although Artzner *et al.* (1999) [page 216] shows that is not a coherent measure of risk, failing to verify the subadditivity property, not only it's a very common measure, almost standard due to Solvency II and Basel III directives, but also because in Danielsson *et al.* (2005) the authors show that for most practical applications V@R is subadditive.

Our approach is different from the usual nonstandard approaches: the class of heavy-tailed, alpha-stable distributions (extensive analysis of this distributions and their properties can be found for instance in Rachev and Mittnik (2000)), the extreme value theory as applied by Moscadelli (2004), by Embrechts *et al.* (2007) or by Dutta and Perry (2007) (also with a comprehensive evaluation of commonly used methods) or truncated distributions that also try to lead with the missing data in the databases and the reporting bias problem (see for instance Chernobai *et al.* (2007)).

2. Sampling frame and sample

We consider that the original random process we want do model is represented by the random variable *(rv)* X with a cumulative distribution function *(cdf)* $F_X(\cdot)$. In our case, the *rv* will be the individual operational loss amount.

We follow the usual model and consider that this random process originated a random sample of the operational losses occurred over a period (usually a year or several years), that is, $S_X = \{X_i, i = 1, \dots N\}$ with the X_i independent and identically distributed ($i.i.d.$) with $F_X(\cdot)$.

Now, consider that, due to several reasons, some presented in Section 1, it is possible that not all the observations originated by X are to be registered and considered in future analysis, that is, not all the observations presented in the original sample S_X, will be available to model operational losses and for statistical inference, namely, parametric estimation. We call *sample* the observations available for estimation and represent them by $S_Y = \{Y_j, j = 1, \dots M\}$, with $M \leq N$. We call *sampling frame* the unobservable S_X, produced by the original random process. Here, we make use of the usual denominations from sampling theory, that we will be using in our results.

Let us now suppose that, each individual loss presented in S_X has a probability, say $p_i, i = 1, \dots, n$, of being recorded and, in that case, belonging to the sample S_Y, the data that is available to us to study the phenomenon.

https://doi.org/10.1017/asb.2016.4 Published online by Cambridge University Press

If all the observations in S_X have the same probability of being recorded, the distribution of the Y_j would not depart from the distribution of the original random process. If not, the recorded observations will not have the original distribution. In this case, the sample will have a different distribution from the sampling frame.

We suppose that the researcher of operational losses ends up with a biased sample of all the operational losses that should have been reported. The bias is originated due to the positive correlation between the loss amount and the probability of being reported.

Let us now consider that each element in the sampling frame S_X, has probability of inclusion in the sample S_Y, depending on the quality of the mechanism put in place to filter the sampling frame and on the size of the element, with largest elements having bigger probabilities. If the mechanism is perfect, all the elements in the sampling frame would be selected and end up in the sample, so that we would have no loss of information and no biased sample.

At the same time, we need a sampling scheme that takes in consideration the rarity of the largest elements, without giving probability one to all the elements above some threshold. That is, we want to put the probability of sampling the elements in S_X in perspective not only to their absolute values. For instance, if a loss of below 500.000$ is almost as common as a loss of below 1.000.000$, we want to preserve this relative relation. On the contrary, if a loss of below 500.000$, is unlikely but below 1.000.000$, is very likely, we want to have a much higher probability to select 1.000.000$, than 500.000$.

That is, once the sampling frame is defined, we want the sampling scheme, representing the mechanism put in place to record operational losses, to take in consideration the random process origination the sampling frame and not only if a loss amount is twice another loss amount.

Let us consider that, after realization, the probability for an operational loss to be reported (or recorded, using the terminology of the probability theory) is, somehow, dependent on the quality of the mechanism put in place to record operational losses, and if the mechanism is not perfect, proportional to its likelihood.

The imperfections could arise for several reasons, for instance, due to the relative small size of some losses, that the staff do not consider worthwhile to report, due to managerial decisions, misclassification and, ultimately, because perfect control systems are difficult to implement, if at all possible.

3. Weighted distributions

It's well known that, if the recording probabilities of the sampling units are not equal, then the distribution of the Y_i (sample) may differ from the distribution of X_i (sampling frame).

https://doi.org/10.1017/asb.2016.4 Published online by Cambridge University Press

In our model N (and of course M) is a random variable, although, depending on the sampling scheme used, the distribution of M conditional on N may be a degenerated random variable.

We propose an approach considering a sampling scheme proportional to size and depending on the likelihood, in this case, proportional to the size of the loss, should be considered when dealing with loss data reported due to operational risk. In this framework, somehow contrary to the approach that makes the trend at the moment to deal with problems of modeling and making inference for operational risk, the Extreme Value Theory and Peeks Over Thresholds, not all the largest values have to be recorded and available to the researcher. In this case, we are not sure that all the big losses are available for study or even took part in the aggregated figures reported, e.g. total losses; mean loss, that the institution produce for accounting support.

We consider that the observations appear in the frame in a given order $\{X_1, \ldots, X_N\}$ and that the sample membership indicator, $\mathbb{I}_k$, are independent with $k = 1, \ldots, N$. The sampling scheme implies naturally that the sampling is made without replacement. The sample membership indicator are distributed relating to size according to

$$\mathrm{P}(\mathbb{I}_k = 1 \mid X_k) = F_X^{\xi}(x_k), \ \ \xi \in [0, +\infty[. \tag{3.1}$$

So, $\mathbb{I}_k \mid X_k \sim B(F_X^{\xi}(x))$ has a Bernoulli distribution with $F_X^{\xi}(x)$ the probability of success (in this case the probability to report the loss). We can say that this is a particular case of a Poisson sampling design, with inclusion probabilities proportional-to-size, about it see, for instance, Sarndal *et al.* (2003).

It's possible to think of ξ as a censorship parameter (other possible analogies can be a disclosure or a quality parameter). If $\xi = 0$ (implying no censorship, total disclosure of all losses or a system so effective that all losses end up reported) we would have $\mathrm{P}(\mathbb{I}_k = 1 \mid X_k) = 1$, so that $S_Y = S_X$, and we would be in the usual situation of a random sample from $F_X(\cdot)$.

However, when $\xi > 0$, we are in the presence of some degree of censorship in our sample, making more likely that big losses are included in the sample than small losses.

The following proposition helps us in establishing the framework for our model.

Proposition 3.1. *Let $X_1, \ldots, X_N$ be a random sample of individual losses, with X_i independent of N a random variable with support on $\mathbb{N}$. If we consider $S_X = \{X_1, \ldots, X_N\}$ as our sampling frame (or simply frame) and apply on S_X a sampling scheme proportional-to-size with no replacement, such that, $P(\mathbb{I}_i = 1 \mid X_i = x) = F_X^{\xi}(x)$, with $i = 1, \ldots N$, where $F_X(\cdot)$ is the cdf of X_i and $\xi \in [0, +\infty[$ is the censorship parameter, then:*

https://doi.org/10.1017/asb.2016.4 Published online by Cambridge University Press

a. *Not conditional on knowledge of the frame, the inclusion variables are i.i.d. Bernoulli with* $\pi = \frac{1}{\xi+1}$ *the probability of success;* $B\left(\frac{1}{\xi+1}\right) = B(\pi)$*, that is,* $P(\mathbb{I}_i = 1) = \dfrac{1}{\xi+1} = \pi,\ i = 1, \ldots N.$

b. *Since* $\sharp S_Y = \sum_X \mathbb{I}_k = \sum_{i=1}^{N} \mathbb{I}_{X_i}$*, we have that,* $\mathbb{E}(\sharp S_Y \mid N) = N\pi = \frac{N}{\xi+1}$.

c. $P(S_Y = s) = \left(\dfrac{1}{\xi}\right)^{\sharp s} \sum_{j \geq \sharp s} \left(\dfrac{\xi}{\xi+1}\right)^{j} P(N = j)$*, where s is the observed sample.*

d. $P(X_j = x \mid \mathbb{I}_j = 1) = F_X^{\xi}(x) f_X(x)(\xi+1)$*, j=1,... N,* $\xi \in [0, +\infty[$.

Proof. a. The independence follows from the sampling scheme. The Bernoulli distribution from:

$$\begin{aligned}
\mathrm{P}(\mathbb{I}_k = 1) &= \int_{\mathbb{R}} \mathrm{P}(\mathbb{I}_k = 1 \mid X = x)\mathrm{P}(X = x)dx \\
&= \int_{\mathbb{R}} F_X^{\xi}(x) f_X(x) dx = \mathbb{E}\left(F_X^{\xi}(X)\right) \\
&= \frac{1}{\xi+1}\left[F_X^{\xi+1}(x)\right]_{-\infty}^{+\infty} = \frac{1}{\xi+1}. \qquad (3.2)
\end{aligned}$$

b. It follows directly from (a).

c. Conditional on the knowledge of the frame X, we have for the probability of observing the specific samples s, $\mathrm{P}(S_Y = s \mid X) = \prod_{k \in s} \pi_k \prod_{j \in S_X - s} (1 - \pi_j)$, so that, due to the independence of the inclusion variables, we have:

$$\begin{aligned}
\mathrm{P}(S_Y = s) &= \int_{\mathbb{R}^N} \prod_{k \in s} \pi_k f_X(x_k) \prod_{j \in S_X - s} (1 - \pi_j) f_X(x_j) d \prod_{i=1}^{N} x_i \\
&= \prod_{k \in s} \int_{\mathbb{R}} \pi_k f_X(x_k) dx_k \prod_{j \in S_X - s} \int_{\mathbb{R}} (1 - \pi_j) f_X(x_j) dx_j \\
&= \prod_{k \in s} \int_{\mathbb{R}} F_X^{\xi}(x_k) f_X(x_k) dx_k \prod_{j \in S_X - s} \int_{\mathbb{R}} \left(1 - F_X^{\xi}(x_j)\right) f_X(x_j) dx_j \\
&= \prod_{k \in s} \left[\frac{1}{\xi+1} F_X^{\xi+1}(x)\right]_{-\infty}^{+\infty} \prod_{j \in S_X - s} \left(1 - \left[\frac{1}{\xi+1} F_X^{\xi+1}(x)\right]_{-\infty}^{+\infty}\right)
\end{aligned}$$

https://doi.org/10.1017/asb.2016.4 Published online by Cambridge University Press

$$= \left(\frac{1}{\xi+1}\right)^{\sharp s}\left(1-\frac{1}{\xi+1}\right)^{N-\sharp s}\mathbb{I}_{\{\sharp s, \sharp s+1, \ldots\}}(N)$$

$$\equiv \left(\frac{1}{\xi+1}\right)^{\sharp s}\left(1-\frac{1}{\xi+1}\right)^{N-\sharp s}\mathbb{I}_{\geq \sharp s}(N).$$

Now integrating in order to N, we have:

$$\mathrm{P}(S_Y = s) = \sum_{j \geq \sharp s}\left(\frac{1}{1+\xi}\right)^{\sharp s}\left(1-\frac{1}{1+\xi}\right)^{j-\sharp s}\mathrm{P}(N=j).$$

d.

$$\begin{aligned}\mathrm{P}(X_j = x \mid \mathbb{I}_j = 1) &= \mathrm{P}(\mathbb{I}_j = 1 \mid X = x)\frac{\mathrm{P}(X=x)}{\mathrm{P}(\mathbb{I}_j = 1)} \\ &= F_X^{\xi}(x)\frac{f_X(x)}{\frac{1}{\xi+1}} \\ &= F_X^{\xi}(x) f_X(x)(\xi+1). \qquad (3.3)\end{aligned}$$

From this result it follows immediately that:

$$\mathrm{P}(X_j \leq x \mid \mathbb{I}_j = 1) = F_X^{\xi+1}(x). \qquad (3.4)$$

■

Proposition 3.2. *With the assumptions of Proposition 3.1, the distribution of the observations in the sample, that is, the distribution of the losses recorded, hence, the distribution of the observations available to the researcher to make inference, is a weighted distribution on $f_X(\cdot)$ with weight function $w(x) = F_X^{\xi}(x)$.*

Before we start the proof, we introduce the definition of *weighted distribution*. Following Rao (1965) we have,

Definition 3.3. *Assume that interest is in a random variable X, with probability density function (pdf) (or probability mass function (pmf)) $f(x)$, with parameters $\theta \in \Theta$ a given parameter space. Also, assume that the values x and y are observed and recorded in the ratio of $w(x)/w(y)$, where $w(x)$ is a non-negative weight function, such that $\mathbb{E}(w(X))$ exists. If the relative probability that x will be observed and recorded is given by $w(x) \geq 0$, then the pdf of the observed data is*

$$f_w(x) = \frac{w(x)}{\omega}f(x), \text{ where } w(x) \geq 0 \text{ and } \omega = \int_{\mathbb{R}} w(x) f_X(x)dx = \mathbb{E}(w(X)).$$

The pdf $f_w(x)$ is denominated the weighted pdf corresponding to $f(x)$.

https://doi.org/10.1017/asb.2016.4 Published online by Cambridge University Press

We can read the early work on weighted distributions in Fisher (1934). The problem of parameter estimation using non equally probable sampling scheme was first addressed by Rao (1965), Patil and Rao (1977) and Patil and Rao (1978). In these papers, the authors identified various sampling situations which can be modeled using weighted distributions and calculated the Fisher information for certain exponential families, focusing primarily on $w(x) = x$, for nonnegative random variables, denominating this weighted distributions by the *size-based form* of the original distribution.

Proof. (Proposition 3.2): By considering (d) in Proposition 3.1 and equation (3.2), we've:

$$\mathrm{P}(X_j = x \mid \mathbb{I}_j = 1) = F_X^{\xi}(x) f_X(x)(\xi + 1) = \frac{F_X^{\xi}(x)}{\frac{1}{(\xi+1)}} f_X(x)$$

$$= \frac{F_X^{\xi}(x)}{\mathbb{E}\left(F_X^{\xi}(X)\right)} f_X(x),$$

and obviously $F_X^{\xi}(\cdot)$ is non-negative, so the conclusion follows: ■

The most common situation studied in the specialized literature deals with the size-biased weighted distribution, so that $f_w(x) = \dfrac{x}{\mathbb{E}\,(w(X))} f(x) = \dfrac{x}{\mathbb{E}(X)} f(x)$, where X is a non-negative random variable with first order moment.

In this paper, we propose that this weight function, originating the denominated *sized-biased distribution*, gives to much weight to the larger losses or, if you prefer, is to light on the smaller losses, not allowing the recording of to much of smaller losses and, at the same time, does not take in consideration the original process X for the operational losses.

The introduction of the $\xi(\geq 0)$ parameter, allows us to define in a natural way the quality of the in place mechanism to record operational losses, since we have that $\mathbb{E}\,(\mathbb{I}_X = 1) = 1/(\xi + 1)$, being possible for the people involved in the process of controlling and managing the operational risk, to have a good "informed guess" for the value of ξ, for instance if $\xi = 1/2$ then $2/3$ of all the operational losses end up recorded, or even, through some specific methods to estimate the parameter ξ. For instance, by inserting in the system erroneous impacts, that should be detected and document by the control system in place, with the objective of estimating the success rate $1/(\xi + 1)$. Naturally, the usual statistical inference methods can and should be applied here.

From (3.3), we can write $f(x)$ as a function of $f_w(x)$ and $F(x)$:

$$f(x) = \frac{1}{\xi + 1} F(x)^{-\xi} f_w(x),$$

Observing that $F_w(x) = \int_{\infty}^{x} F^{\xi}(y)(\xi+1)f(y)dy = F^{\xi+1}(x)$ we can also write $f(x)$ as a function of $f_w(x)$ and $F_w(x)$:

$$f(x) = \frac{1}{\xi+1} F_w(x)^{-\frac{\xi}{\xi+1}} f_w(x). \tag{3.5}$$

4. Weighted distributions applied to model operational risk

In this section, we will consider two distribution models for the individual operation loss amounts, the Exponential and Pareto (type I). We will deduce the impact in the parameters estimates, when using aggregate data, and not considering the bias presented in the sample, produced by a mechanism to record operational losses that is not perfect, that is by not considering a $\xi > 0$, in Proposition 3.1.

We will consider that the operational losses, X_i in $S_X = \{X_i, i = 1, \dots N\}$, the sampling frame, have *pdf* $f(x)$ and the recorded operational losses, Y_j in $S_Y = \{Y_j, j = 1, \dots M\}$, the sample available to make inference, have *pdf* $f_w(x)$. We will analyze the impact for not considering the bias presented in the sample S_Y and estimating the parameters as if the distribution is the original distribution $f(x)$.

For this distributions, we will consider that we know $f_w(x)$ and we want to estimate $E(X)$ and $V(X)$, the more frequent scenario.

4.1. The exponential model

Consider $Y_j \sim f_w(x)$ as the Exponential density with parameters λ and β by (3.5) we have $f(x) = \frac{1}{\xi+1}(1 - e^{-\frac{x-\lambda}{\beta}})^{-\frac{\xi}{\xi+1}} \frac{1}{\beta} e^{-\frac{x-\lambda}{\beta}} \mathbb{I}_{]\lambda,+\infty[}(x)$, considering $x = \beta \ln(y) + \lambda$, noting that $\frac{\partial}{\partial x} B(x, y) = \int_0^1 t^{x-1} \ln(t)(1-t)^{y-1} dt$, where $B(x, y) = \int_0^1 t^{x-1}(1-t)^{y-1} dt$ is the beta function. Note that $\frac{\partial}{\partial x} B(x, y) = B(x, y)(\psi(x) - \psi(x+y))$, being $\psi(z) = \frac{\partial}{\partial x} \ln \Gamma(x)$ the digamma function. Consider also that $\psi(n) = H_{n-1} - \gamma$ and $\psi(1) = -\gamma$ where H_n is the nth harmonic number or in the generic form, $H_x = \int_0^1 \frac{t^x - 1}{t-1} dt$, with γ the Euler–Mascheroni constant we have:

$$E(X) = \lambda + \beta H_{\frac{1}{\xi+1}}, \tag{4.1}$$

$$V(X) = \beta^2 \left(\frac{\pi^2}{6} - \psi' \left(\frac{\xi+2}{\xi+1} \right) \right). \tag{4.2}$$

https://doi.org/10.1017/asb.2016.4 Published online by Cambridge University Press

4.2. The Pareto (Type I) model

Consider now that $f_w(x)$ is Pareto with parameters α and β by (3.5) we have

$$f(x) = \frac{1}{\xi+1}\left(1-\left(\frac{\beta}{x}\right)^{\alpha}\right)^{\frac{-\xi}{\xi+1}} \frac{\alpha}{x}\left(\frac{\beta}{x}\right)^{\alpha} \mathbb{I}_{]\beta,\infty[}(x).$$

Considering $y = \left(\frac{\beta}{\alpha}\right)^{\alpha}$ and observing that $B\left(1-\frac{1}{\alpha}, \frac{1}{\xi+1}\right) = \int_0^1 y^{-\frac{1}{\alpha}}(1-y)^{-\xi/(\xi+1)}dy$ the moments for X are:

$$E(X) = \frac{\beta}{\xi+1} B\left(1-\frac{1}{\alpha}, \frac{1}{\xi+1}\right),$$

$$V(X) = \frac{\beta^2}{\xi+1} B\left(1-\frac{2}{\alpha}, \frac{1}{\xi+1}\right) - \frac{\beta^2}{(\xi+1)^2}\left(B\left(1-\frac{1}{\alpha}, \frac{1}{\xi+1}\right)\right)^2.$$

5. Application

We consider that the reported losses S_Y, have a known distribution (Exponential, Pareto, Lognormal or Weibull). To allow the comparison of results we used the Sample Description of Table 6.3 of Chernobai *et al.* (2007) based on the original data set available from Cruz (2002) that give us a mean value of 439.725, 99\$ and standard deviation of 538.403, 93\$ (in dollars). We also consider that the risk department estimated a probability of 1/250 for an operation to generate a operational loss and of 85% for the loss ending up reported and documented, ($\xi = (1–85\%)/85\%$). So that, for every 2.941.176 transactions made, we expect that 11.765 operations originate a loss and 10.000 of this losses are reported. We used the method of moments to estimate the parameters for each of the distributions considered.

For a given density function $f_w(x)$ (Exponential, Pareto, Lognormal or Weibull) for the reported losses S_Y, in our simulation we generated 1.000 samples of occurred losses S_X, each with 11.765 losses with density function $f(x)$ obtained from formula (3.5). For each sample, we selected the reported losses S_Y, according to (3.1), so that, the observations in our sample, that is, the data available to make inference and take decisions have distribution $F_w(x)$. The observed average sample size of the 1.000 reported losses is 10.020,53\$.

We used the simulation to compare the results obtained by the sampling process, originating S_Y, and the ones really experienced, S_X. We explain the procedure for the Exponential distribution, the first case presented below and repeat the process for the Pareto, Lognormal and Weibull distributions, just presenting the results. For Lognormal and Weibull distributions we do not derive the formulas for $E(X)$ and $V(X)$, we just compute the numerical approximation for the integrals to illustrate the results for this two distributions and allow the reader to compare results with Chernobai *et al.* (2007).

https://doi.org/10.1017/asb.2016.4 Published online by Cambridge University Press

5.1. The exponential model

Following Section 4.1 consider that the reported losses, Y, has an Exponential density function, $f_w(x)$, with $\lambda = 0$ and $\beta = 439.725, 99$. Usually the available data are the recorded losses. In this example, $E(Y) = 439.725, 99$, $\sigma(Y) = 439.725, 99$. From (4.1) and (4.2) $E(X) = 395.056, 71$ and $\sigma(X) = 424.803, 42$. From our 1.000 simulated universes we have estimated $\widehat{E}(X) = 394.967, 87$ and $\widehat{\sigma}(X) = 424.566, 34$. Using (3.1) we collected 1.000 samples from the universes with $\widehat{E}(Y) = 439.309, 36$ and $\widehat{\sigma}(Y) = 439.141, 89$. The results from the simulation are very closed to the theoretical moments. We present results for the Kolmogorov–Smirnov and the Chi-squared tests, to check the agreement of the simulated data with the Exponential $\lambda = 0$ and $\beta = 439.725, 99$. At a significance level of 5% we do not reject 955 (949 for Chi-squared) samples. We also tested if the universe, X, is Exponential $\lambda = 0$ and β using the maximum likelihood estimators, rejecting all the universes at the significance level of 5%. Minimum, average and maximum p values for each test can be observed on Table 5.2.

When considering a Value at Risk analysis, to answer the question "What is the maximum amount that I can expect to lose with a certain probability over a given horizon?" for a one year period with a confidence level of 0.1%, we obtain for the 99.9% percentiles of the individual losses,

$$F_X^{-1}(99, 9\%) = 2.966.094, 54\$ \text{ versus } F_Y^{-1}(99, 9\%) = 3.037.519, 53\$.$$

We calculated the empirical TV@R with the same confidence level for X, being $TV@R(X) = 3.408.439, 46\$$ and for the reported data $TV@R(Y) = 3.479.634, 14\$$.

Estimating the true total operational losses the bank incurred, we have

$$\mathbb{E}\left(\sum_{i=1}^{N} X_i\right) = (1 + \xi_0) \times 10.000 \times 395.056, 71 = 4.647.726.036, 27\$,$$

obtained with the data from the universe versus

$$\mathbb{E}\left(\sum_{i=1}^{M} Y_i\right) = 10.000 \times 439.725, 99 = 4.397.259.900, 00\$,$$

obtained with the data reported. Böcker *et al.* (2005) developed an approximated closed-form to compute $V@R$ for the aggregated loss, based on the distribution of a single loss and the expected value of the losses frequency. Since we do not define a threshold for the reported losses we have very high frequencies. Degen (2010) improves on Böcker *et al.* (2005) and shows that, for high frequencies, the relative error associated to the single-loss approximation is very large, even for very high α levels. For this reason, we only report the expected value of the aggregate losses.

https://doi.org/10.1017/asb.2016.4 Published online by Cambridge University Press

All the results can be seen at Table 5.1. We also report the V@R and TV@R for the 99% percentile.

5.2. The Pareto (Type I) model

For the Pareto model, we will consider the case where $\alpha = 2,29114$ and $\beta = 247.801$. Using Section 4.2 and the same methodology as the previous example, the results can be seen at Table 5.1.

5.3. The Lognormal model

Considering that $f_w(x) = \frac{1}{\sqrt{2\pi}\sigma x}\exp(-\frac{(\log(x)-\mu)^2}{2\sigma^2})$, $x > 0, \sigma > 0, \mu \in \mathbb{R}$. In this example, we obtained $\mu = 12.5359$ and $\sigma = 0.957058$. The results can be seen at Table 5.1.

5.4. The Weibull model

Considering that $f_w(x) = \frac{\alpha}{\beta}(\frac{x}{\beta})^{\alpha-1}\exp(-(\frac{x}{\beta})^{\alpha})$, $x \geq 0, \alpha > 0, \beta > 0$. In this example, we obtained $\alpha = 0.821926$ and $\beta = 395.464, 03$. The results can also be seen at Table 5.1.

5.5. Results

Table 5.1 shows that apart from $\widehat{\sigma}$ for the Pareto distribution, the empirical results are in great accordance with the theoretical ones. As expected, the capital requirement using $V@R$ is higher if we consider the reported losses than if we consider the occurred losses, 2% for Exponential, 3% for Weibull, 5% for Lognormal and 7% for Pareto. By the other side, as expected, we have the expected total amount of losses higher if we consider the occurred losses, 5% for Weibull case, 6% for Exponential and Lognormal and 12% for Pareto.

The Lognormal distribution has the interesting feature of the sampling method preserving the distribution, that is, S_Y and the S_X are both Lognormal, what does not happen with the others tested distributions. We didn't investigate much on this particularity of the Lognormal, but we think it relates to the entropy properties of this distribution. The average of the maximum likelihood estimates for the occurred losses are $\widehat{\overline{\mu}} = 12.39161$ and $\widehat{\overline{\sigma}} = 1.005097$.

In Table 5.2, we use a single column for the samples S_Y, because we've used the same set of pseudo-random numbers to generate all the distributions, hence all the non-parametric tests will produce the same test statistics, originating the same results and decisions.

https://doi.org/10.1017/asb.2016.4 Published online by Cambridge University Press

TABLE 5.1

RESULTS FOR ALL THE EXAMPLES.

Occurred	Reported	Exponential		Pareto		Lognormal		Weibull	
X	Y	X	Y	X	Y	X	Y	X	Y
				Theorical					
$E(X)$	$E(Y)$	395.056,71	439.725,99	417.366,59	439.725,19	396.844,27	439.714,09	391.177,00	439.725,99
$\sigma(X)$	$\sigma(Y)$	424.803,42	439.725,99	499.950,66	538.399,43	510.258,76	538.389,17	514.422,01	538.403,93
$V@R_{99\%}(X)$	$V@R_{99\%}(Y)$	1.953.938,46	2.025.013,02	1.723.425,99	1.849.400,38	2.430.724,95	2.577.549,87	2.427.557,18	2.535.410,19
$V@R_{99,9\%}(X)$	$V@R_{99,9\%}(Y)$	2.966.094,54	3.037.519,53	4.706.580,84	5.052.366,95	5.111.227,25	5.354.412,53	4.033.829,50	4.152.317,56
$E\left(\sum_{i=1}^{N} X_i\right)$	$E\left(\sum_{i=1}^{M} Y_i\right)$	4.647.726.036	4.397.259.900	4.910.195.194	4.397.251.910	4.668.756.082	4.397.140.937	4.602.082.353	4.397.259.900
				Empirical					
$\widehat{E}(X)$	$\widehat{E}(Y)$	394.967,87	439.309,36	417.237,35	439.358,53	396.697,70	439.210,72	391.046,61	439.173,74
$\widehat{\sigma}(X)$	$\widehat{\sigma}(Y)$	424.566,34	439.141,89	432.830,30	464.561,01	509.877,14	537.529,65	514.103,86	537.646,01
$TV@R_{99\%}(X)$	$TV@R_{99\%}(Y)$	2.395.587,73	2.467.134,25	3.061.988,15	3.286.507,23	3.576.296,68	3.765.141,86	3.123.561,05	3.236.677,66
$TV@R_{99,9\%}(X)$	$TV@R_{99,9\%}(Y)$	3.408.439,46	3.479.634,14	8.316.902,89	8.912.776,77	6.888.269,49	7.367.649,69	4.787.108,09	4.908.582,01

TABLE 5.2

p VALUES FOR KOLMOGOROV–SMIRNOV AND CHI2 TESTS.

	X				
	Exponential	Pareto	Lognormal	Weibull	*Y*
			K-S *p* values		
min	0	0	3,9212E-02	1,5616E-04	3,4932E-04
average	4,0669E-04	4,0265E-04	6,7963E-01	1,1667E-02	5,0641E-01
max	2,3191E-01	2,2921E-01	9,9999E-01	6,4403E-02	9,9952E-01
# not reject	0	0	999	5	955
			Chi2 *p* values		
min	0	0	2,5390E-04	4,7621E-11	6,0648E-05
average	4,0529E-23	1,1096E-24	4,8810E-01	7,4905E-03	5,0170E-01
max	2,8235E-21	8,0041E-22	9,9846E-01	2,4295E-01	9,9934E-01
# not reject	0	0	948	44	949

6. REMARKS AND CONCLUSIONS

As expected using the reported losses the capital requirement (using V@R) is overestimated (from 2% to 7% in our examples) and the total amount of losses is underestimated (from 5% to 12%). If the bank is in the presence of a heavy tail distribution for the reported losses, the values are of considerable amount to be ignored.

Even when the institutions have in place methods to detect and document operational losses, intending to be exhaustive and error free, not every operational loss ends up reported. We are lead to believe that, when dealing with loss data reported due to operational risk, we are always in the presence of a biased sample, no matter if the data used to model the individual losses and total losses, comes from a commercial vendor or it is provided by internal procedures to manage operational losses.

Using weighted distributions, we are able to consider that the probability of a loss to be reported and ends up recorded for analysis, increases with the size of the loss but, at the same time, we do not consider that a threshold exists, above which all losses are recorded and available for analysis, hence, no loss has probability one of being recorded.

Since operational risk management relies more on qualitative approaches than on quantitative ones, more work is needed to better understand and model the exposure to operational risk. The bias presented in operational losses data, mainly due to the natural emphasis given to (public) very large losses, makes it more challenging.

Our model takes in consideration the sample bias towards the largest losses by defining a weight function functional dependent on the distribution of the original random process and on the reliability recording the operational losses. In this way, we can infer how the bias affects the original distribution and the estimators of the parameters.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge to Magentakoncept — Consultores Lda, for the computational support. This work was partially supported by the Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) through the project PEst-OE/MAT/UI0297/2014(CMA).

REFERENCES

ARTZNER, P., DELBAEN, F., EBER, J.-M. and HEATH, D. (1999) Coherent measures of risk. *Mathematical Finance,* **9**(3), 203–228.

BÖCKER, K. and KLÜPPELBERG, C. (2005) Operational VAR: A closed-form approximation. *Risk,* **18**(12), 90–93.

CHAPELLE, A., CRAMA, Y., HUBNER, G. and PETERS, J. (2005) Measuring and managing operational risk in the finantial sector: An integrated framework. Technical report, National Bank of Belgium.

CHERNOBAI, A., RACHEV, S.T. and FABOZZI, F. (2007) *Operational Risk: A Guide to Basel II Capital Requirements, Models, and Analysis*. John Wiley & Sons, Inc., Hoboken, New Jersey.

CRUZ, M.G. (2002) *Modeling, Measuring and Hedging Operational Risk*. John Wiley & Sons, New York, Chichester.

CUMMINS, J.D., LEWIS, C.M. and WEI, R. (2004) The market value impact of operational risk events for u.s. banks an insurers. Technical report, University of Pennsylvania.

DANIELSSON, J., JORGENSEN, B. N., MANDIRA, S., SAMORODNITSKY, G. and DE VRIES, C. G. (2005) Subadditivity re–examined: the case for value-at-risk. Financial Markets Group, FMG Discussion Papers, dp549.

DE FONTNOUVELLE, P., JORDAN, J. and ROSENGREN, E. (2003) Using loss data to quantify operational risk. *SSRN Electronic Journal*.

DE FONTNOUVELLE, P., JORDAN, J. and ROSENGREN, E. (2005) Implications of alternative operational risk modeling techniques. Working Paper 11103, National Bureau of Economic Research.

DEGEN, M. (2010) The calculation of minimum regulatory capital using single-loss approximations. *Journal of Operational Risk,* **5**(4), 3–17.

DUTTA, K. and PERRY, J. (2007) An empirical analysis of loss distribution models for estimating operational risk capital. Working paper no. 06–13, Federal Reserve Bank of Boston.

EMBRECHTS, P., DEGEN, M. and LAMBRIGGER, D.D. (2007) The quantitative modeling of operational risk: between g-and-h and evt. *ASTIN Bulletin,* **37**(2), 265–291.

FISHER, R.A. (1934) The effects of methods of ascertainment upon the estimation of frequencies. *Annals Eugenics,* (6), 13–25.

MOSCADELLI, M. (2004) The modelling of operational risk: Experience with the analysis of the data collected by basel committee. Technical report, Bank of Italy.

PATIL, G. and RAO, C. (1977) Weighted distributions: A survey of their application. *Applications of Statistics*, 383–405.

PATIL, G. P. and RAO, C. R. (1978) Weighted distributions and size-biased sampling with applications to wildlife populations and human families. *Biometrics,* **34**(2), 179–189.

PERRY, J. and FONTNOUVELLE, P. (2005) Measuring reputational risk: The market reactions to operational loss announcements. Technical report, Federal Reserve Bank of Boston.

https://doi.org/10.1017/asb.2016.4 Published online by Cambridge University Press

RACHEV, S. and MITTNIK, S. (2000) *Stable Paretian models in Finance*. New York: John Wiley & Sons.

RAO, C. (1965) On discrete distributions arising out of methods of ascertainment. *Sankhyā: The Indian Journal of Statistics, Series A,* **27**(2/4), 311–324.

SARNDAL, C., SWENSSON, B. and WRETMAN, J. (2003) *Model Assisted Survey Sampling*. Springer Series in Statistics. Springer-Verlag, New York.

SHIH, J., SAMAD-KHAN, A.J. and MEDAPA, P. (2000) Is the size of an operational risk related to firm size? *Operational Risk Magazine*, February, 2000.

LOURDES B. AFONSO (Corresponding author)
Centro de Matemática e Aplicações, CMA,
Faculdade Ciências e Tecnologia,
Universidade Nova de Lisboa, 2829-516 Caparica, Portugal
E-Mail: lbafonso@fct.unl.pt

PEDRO CORTE REAL
Depart. de Matemática, Faculdade Ciências e Tecnologia,
Universidade Nova de Lisboa, 2829-516 Caparica, Portugal
E-Mail: parcr@fct.unl.pt

https://doi.org/10.1017/asb.2016.4 Published online by Cambridge University Press