Composite Goodness-of-Fit Tests for Left-Truncated Loss Samples

Anna Chernobai[1], Svetlozar Rachev[2,1,*] and Frank Fabozzi[3]

June 4, 2005

[1]Department of Statistics and Applied Probability
University of California, Santa Barbara, CA 93106, USA

[2]Institut für Statistik und Mathematische Wirtschaftstheorie
Universität Karlsruhe
Kollegium am Schloss, D-76128 Karlsruhe, Germany

[3]Department of Finance
Yale University, New Haven, CT 06520, USA

Corresponding author: Anna Chernobai.
E-mail address: annac@pstat.ucsb.edu
Tel.: +1 (805) 893 4857

Abstract

In many loss models the recorded data are left-truncated with an unknown number of missing data. We derive the exact formulae for several goodness-of-fit statistics that should be applied to such models. We additionally propose two new statistics to test the fit in the upper quantiles, applicable to models where the accuracy of the upper tail estimate is crucial, as in models addressing the Value-at-Risk and ruin probabilities. We apply the tests on a variety of distributions fitted to the external operational loss and the natural catastrophe insurance claim data, subject to the recording thresholds of $1 and $25 million.

Keywords:

Truncated Data, Goodness-of-Fit Tests, Loss Distribution, Operational Risk, Insurance.

JEL classification:

C24, G21, G22, C14, C12, C15, C19.

1 Introduction

In most loss models, the central attention is devoted to studying the distributional properties of the loss data. The shape of the dispersion of the data determines the vital statistics such as the expected loss, variance, and ruin probability, Value-at-Risk or Conditional Value-at-Risk where the shape in the right tail is crucial. Parametric procedures for testing the goodness of fit (GOF) include the Likelihood Ratio test and Chi-squared test. A standard semi-parametric procedure to test how well a hypothesized distribution fits the data involves applying the in-sample GOF tests that provide a comparison of the fitted distribution to the empirical distribution. These tests, referred to as EDF tests (i.e. empirical distribution function tests), include the Kolmogorov-Smirnov test, Anderson-Darling test and the Cramér-von Mises tests.[1]

In many applications, the data set analyzed is incomplete, in the sense that the observations that are present in the loss database only if they exceed a pre-determined threshold level. This problem is usually absent in risk models involving market risk and credit risk. However, it is a common problem in operational risk or insurance claims models. In operational risk, banks' internal databases are subject to a minimum recording thresholds of roughly \$6,000-\$10,000, and external databases usually collect operational losses starting from \$1 million, BIS (2003). Similarly, in non-life insurance models, the thresholds are set at \$5 million, \$25 million or other levels. Consequently, in the analysis of operational losses, recorded loss data are left-truncated, and, as a result, it is inappropriate to employ standard GOF tests.

GOF tests for truncated and censored data have been studied by Dufour and Maag (1978), Gastaldi (1993), Gyulbaud (1998), among others. In this paper, we derive the exact formulae for several GOF test statistics that should be applied where there exist incomplete samples with an unknown number of missing data and propose two new statistics to determine the goodness of fit in the upper tail that can be used for loss models where the accuracy of the upper tail estimate is of central concern. The paper is organized as follows. In Section 2 we describe the problem of left-truncated samples and explain the necessary adjustments that are required to the GOF tests to make them applicable for the truncated samples. In Section 3 we review the widely used test statistics for complete samples and derive the exact formulae for the statistics to be used for left-truncated samples. We propose in Section 4 two new EDF statistics to be used for the situations when the fit in the upper tail is of the central concern. Application of the modified EDF tests to operational loss data, obtained from Zurich IC^2 FIRST Database, and the USA natural catastrophe insurance claims data, obtained from Insurance Services Office Inc. Property Claim Services, is presented in Section 5, with final remarks in Section 6. Necessary derivations are provided in the Appendix.

2 Problem Setup

Suppose we have a left-truncated sample, with the data below a pre-specified threshold level H not recorded (not observable). The observable data sample $\mathbf{x} = \{x_1, x_2, ..., x_n\}$ has each data point at least as great as H, and includes a total of n observations. Let $\{X_{(j)}\}_{1 \leq j \leq n}$ be a vector of the order statistics, such that $X_{(1)} \leq X_{(2)} \leq ... \leq X_{(n)}$. The empirical distribution function of the sample is defined as

$$F_n(x) := \frac{\#\, observations \leq x_{(j)}}{n} = \begin{cases} 0 & x < x_{(1)} \\ \frac{j}{n} & x_{(j)} \leq x < x_{(j+1)}, \;\; j = 1, 2, ..., n-1 \\ 1 & x \geq x_n, \end{cases} \tag{1}$$

Graphically, the empirical distribution function of an observed data sample is represented as a step function with a jump of size $1/n$ occurring at each recorded sample value. On the other hand, with left-truncated data which is a part of a larger complete data set, the true size of jumps of the EDF at each value of the *complete* data sample would be of size $1/n^c$, $n^c = n + m$ rather than $1/n$, where n^c is the total number of points of the complete data set and m is the unknown number of missing points. In the GOF tests the null hypothesis states that the observed loss sample belongs to a family of truncated distributions, with the parameter specified (simple test) or unspecified (composite test).

We fit a continuous *truncated* distribution F to the data, given that the data exceed or equal to H, and estimate the conditional parameters θ with the Maximum Likelihood (or an alternative method) by

$$\widehat{\theta}_{\text{MLE}} \quad = \quad \arg\max_{\theta} \log \left(\prod_{k=1}^{n} \frac{f_\theta(x_k)}{1 - F_\theta(H)} \right). \tag{2}$$

Given that this is the true distribution, then the estimated number of missing points m and the number of observed points n are related as [2]:

$$\frac{m}{n} = \frac{z_{\text{H}}}{(1 - z_{\text{H}})} \tag{3}$$

resulting in

$$n^c = n \frac{z_{\text{H}}}{(1 - z_{\text{H}})} + n = \frac{n}{(1 - z_{\text{H}})} \tag{4}$$

where $z_{\text{H}} := \widehat{F}_\theta(H)$.

The empirical distribution function $F_{n^c}(x)$ of complete data sample is:

$$F_{n^c}(x) := \frac{\#\, observations \leq x_{(j)}}{n^c}, \tag{5}$$

where the numerator refers to the total number of observations of the complete data sample, not exceeding in magnitude the j^{th} order statistic of the incomplete (observed) data sample. By equations (3) and (4), equation (5) becomes:

$$F_{n^c}(x) \quad = \quad \frac{m + j}{n/(1 - z_{\text{H}})} = \frac{n z_{\text{H}} + j(1 - z_{\text{H}})}{n} = z_{\text{H}} + \frac{j}{n}(1 - z_{\text{H}}), \;\; j = 1, 2, ..., n.$$

Rearranging terms leads to the fitted distribution function of the observed sample of the following form:

$$\widehat{F^*}(x) = \begin{cases} \frac{\widehat{F}_\theta(x) - \widehat{F}_\theta(H)}{1 - \widehat{F}_\theta(H)} & x \geq H \\ 0 & x < H, \end{cases} \tag{6}$$

so that $\widehat{F}_\theta(X) \sim \mathcal{U}[\widehat{F}_\theta(H), 1]$ and $\widehat{F^*}(X) \sim \mathcal{U}[0,1]$ under the null that the fitted distribution function is true. Therefore, the empirical distribution function of the observed part of the missing data, using equation (1), is represented by

$$F_n(x)(1 - \widehat{F}_\theta(H)) + \widehat{F}_\theta(H) = \begin{cases} \widehat{F}_\theta(H) & x < x_{(1)} \\ \frac{j}{n}(1 - \widehat{F}_\theta(H)) + \widehat{F}_\theta(H) & x_{(j)} \leq x < x_{(j+1)}, \\ & \quad j = 1, 2, ..., n-1 \\ 1 & x \geq x_n, \end{cases} \tag{7}$$

Figure 6 gives a visual illustration of the idea we just described. With these modifications, the in-sample GOF tests can be applied to the left-truncated samples.

In this paper we consider tests of a *composite* hypothesis that the empirical distribution function of an observed incomplete left-truncated loss data sample belongs to a family of hypothesized distributions (with parameters not specified), i.e.

$$\mathbf{H_0} : F_n(x) \in \widehat{F^*}(x) \qquad \text{vs.} \qquad \mathbf{H_A} : F_n(x) \notin \widehat{F^*}(x), \tag{8}$$

where $\widehat{F^*}(x)$ follows equation (6). Under the null equation (8), $\widehat{F^*}(X) \sim \mathcal{U}\,[0,1]$, and the null is rejected if the p-value is lower than the level α, such as α from 5% to 10%. Letting D be the observed value of a GOF statistic (such as Kolmogorov-Smirnov or Anderson-Darling) and d the critical value for a given level α, the p-value is computed as p-value $= P(D \geq d)$. Since the distribution of the statistic is not parameter-free, one way to compute the p-values and the critical values is by means of Monte Carlo simulation, for each hypothesized fitted distribution, Ross (2001) . Under the procedure, the observed value D is computed. Then, for a given level α the following algorithm is applied:

1. Generate large number of samples (e.g. I =1,000) from the fitted truncated distribution of size n equal to the number of observed data (such that all these points are above or equal to H);

2. Fit *truncated* distribution and estimate conditional parameters $\hat{\theta}$ for each sample $i = 1, 2, ...I$;

3. Estimate the GOF statistic value D_i for each sample $i = 1, 2, ...I$;

4. Calculate p-value as the proportion of times the sample statistic values exceed the observed value D of the original sample;

5. Reject $\mathbf{H_0}$ if the p-value is smaller than α.

A p-value of, for example, 0.3, would mean that in 30% of same-size samples simulated from the same distribution with the same parameter estimation procedure applied, the test statistic value was higher than in the original sample.

3 EDF Statistics for Left-Truncated Loss Samples

The EDF statistics are based on the vertical differences between the empirical and fitted (truncated) distribution function They are divided into two classes: (1) the supremum class (such as Kolmogorov-Smirnov and Kuiper statistics), and (2) the quadratic class (such as Anderson-Darling and Cramér-von Mises statistics). In this section, we derive the exact computing formulae for a number of EDF test statistics, modified so that they can be applied to left-truncated loss samples. For the left-truncated samples, $\widehat{F^*}(x)$ denotes the null distribution function for left-truncated sample values (equation (6)). The corresponding variable $\widehat{F^*}(X)$ is distributed uniformly over the $[0,1]$ interval. The variable $\widehat{F}_\theta(X)$ is distributed uniformly over the $[\widehat{F}_\theta(H),1]$ interval. We reserve some other notations: $z_{\text{H}} := \widehat{F}_\theta(H)$ and $z_j := \widehat{F}_\theta(x_{(j)})$ for truncated samples. In this section we discuss the Kolmogorov-Smirnov, Kuiper, Anderson-Darling and the Cramér-von Mises statistics, and using an asterisk (*) to denote their left-truncated sample analog.

3.1 Supremum Class Statistics

3.1.1 Kolmogorov-Smirnov Statistic

A widely used supremum class statistic, the Kolmogorov-Smirnov (KS) statistic, measures the absolute value of the maximum distance between the empirical and fitted distribution function, and puts equal weight on each observation. Let $\{X_{(j)}\}_{1<j<n}$ be the vector of the order statistics, and $X_{(1)} < X_{(2)} < ... < X_{(n)}$, such that strict inequalities hold. Usually, such distance is the greatest around the median of the sample. For the left-truncated data samples the KS statistic is expressed as

$$KS^* = \sqrt{n} \sup_x \left| F_n(x) - \widehat{F^*}(x) \right|. \tag{9}$$

The KS^* statistic can be computed from:

$$\begin{aligned} KS^{+*} &= \sqrt{n} \sup_j \left\{ F_n(x_{(j)}) - \widehat{F^*}(x_{(j)}) \right\} = \frac{\sqrt{n}}{1-z_{\text{H}}} \sup_j \left\{ z_{\text{H}} + \frac{j}{n}(1-z_{\text{H}}) - z_j \right\}, \\ KS^{-*} &= \sqrt{n} \sup_j \left\{ \widehat{F^*}(x_{(j)}) - F_n(x_{(j)}) \right\} = \frac{\sqrt{n}}{1-z_{\text{H}}} \sup_j \left\{ z_j - \left(z_{\text{H}} + \frac{j-1}{n}(1-z_{\text{H}}) \right) \right\}, \end{aligned}$$

and becomes

$$KS^* = \max\{KS^{+*}, KS^{-*}\}. \tag{10}$$

3.1.2 Kuiper Statistic

The KS statistic gives no indication of whether the maximum discrepancy between $F_n(x)$ and $\widehat{F^*}(x)$ occurs when $F_n(x)$ is above $\widehat{F^*}(x)$ or when $\widehat{F^*}(x)$ is above $F_n(x)$. The Kuiper statistic (V) is closely related to the KS statistic. It measures the total sum of the absolute values of the two largest vertical deviations of the fitted distribution function from $F_n(x)$, when $F_n(x)$ is above $\widehat{F^*}(x)$ and when $\widehat{F^*}(x)$ is above $F_n(x)$. For left-truncated data samples it is computed as

$$V^* = KS^{+*} + KS^{-*}, \tag{11}$$

with KS^{+*} and KS^{-*} defined in Section 3.1.1.

3.1.3 Anderson-Darling Statistic

There are two versions of the Anderson-Darling (AD) statistic. In its simplest version it is a variance-weighted KS statistic and belongs to the supremum class. A weight $\psi(\widehat{F^*}(x)) = \left\{\sqrt{\widehat{F^*}(x)\big(1-\widehat{F^*}(x)\big)}\right\}^{-1}$ is attached to each observation in the definition of the KS statistic (equation (9)). Under this specification, the observations in the lower and upper tails of the truncated sample are assigned a higher weight. Let $\{X_{(j)}\}_{1\leq j\leq n}$ be the vector of the order statistics, such that $X_{(1)} \leq X_{(2)} \leq ... \leq X_{(n)}$. Then the AD statistic is defined for left-truncated samples as

$$AD^* = \sqrt{n}\sup_x \left| \frac{F_n(x) - \widehat{F^*}(x)}{\sqrt{\widehat{F^*}(x)\big(1-\widehat{F^*}(x)\big)}} \right|, \tag{12}$$

For the left-truncated samples, the computing formula is derived from :

$$\begin{aligned} AD^{+*} &= \sqrt{n}\sup_j \left\{ \frac{F_n(x_{(j)}) - \widehat{F^*}(x_{(j)})}{\sqrt{\widehat{F^*}(x_{(j)})\big(1-\widehat{F^*}(x_{(j)})\big)}} \right\} = \sqrt{n}\sup_j \left\{ \frac{z_{\text{H}} + \frac{j}{n}(1-z_{\text{H}}) - z_j}{\sqrt{(z_j - z_{\text{H}})(1-z_j)}} \right\}, \\ AD^{-*} &= \sqrt{n}\sup_j \left\{ \frac{\widehat{F^*}(x_{(j)}) - F_n(x_{(j)})}{\sqrt{\widehat{F^*}(x_{(j)})\big(1-\widehat{F^*}(x_{(j)})\big)}} \right\} = \sqrt{n}\sup_j \left\{ \frac{z_j - z_{\text{H}} - \frac{j-1}{n}(1-z_{\text{H}})}{\sqrt{(z_j - z_{\text{H}})(1-z_j)}} \right\}, \end{aligned}$$

and becomes

$$AD^* = \max\{AD^{+*}, AD^{-*}\}. \tag{13}$$

3.2 Quadratic Class Statistics

The quadratic statistics for complete data samples are grouped under the Cramér-von Mises family as

$$Q = n\int_{-\infty}^{\infty} \big(F_n(x) - \widehat{F}(x)\big)^2 \psi(\widehat{F}(x)) d\widehat{F}(x), \tag{14}$$

in which the weight function $\psi(\widehat{F}(x))$ is assigned to give a certain weight to different observations, depending on the purpose. For the left-truncated samples, we denote the Cramér-von Mises family as Q^* and $\widehat{F}(x)$ is replaced by $\widehat{F^*}(x)$:

$$Q^* = n \int_H^\infty \big(F_n(x) - \widehat{F^*}(x)\big)^2 \psi(\widehat{F^*}(x)) d\widehat{F^*}(x). \tag{15}$$

Depending on the form of the weighting function, the sample observations are given a different weight, and $\psi(\widehat{F^*}(x)) = \big\{\widehat{F^*}(x)\big(1 - \widehat{F^*}(x)\big)\big\}^{-1}$ denotes the quadratic Anderson-Darling statistic, and $\psi(\widehat{F^*}(x)) = 1$ corresponds to the Cramér-von Mises statistic.

Derivation of the computing formulae makes use of equation (1), and involves the Probability Integral Transformation (PIT) technique. For the left-truncated samples this leads to

$$Q^* = n \int_0^1 \big(F_n(z^*) - z^*\big)^2 \psi(z^*) dz^* = \frac{n}{1 - z_{\text{H}}} \int_{z_{\text{H}}}^1 \big(F_n(z^*) - \frac{z - z_{\text{H}}}{1 - z_{\text{H}}}\big)^2 \psi(\frac{z - z_{\text{H}}}{1 - z_{\text{H}}}) dz, \tag{16}$$

where $Z^* = \widehat{F^*}(X) = \frac{Z - z_{\text{H}}}{1 - z_{\text{H}}} \sim \mathcal{U}[0,1]$ and so $Z = \widehat{F}_\theta(X) \sim \mathcal{U}[z_{\text{H}}, 1]$ under the null. $F_n(Z^*) = F_n(X) = F^*(F_n(X))$ is the empirical distribution function of Z^*, $F^*(\cdot) \sim \mathcal{U}[0,1]$. $z_{\text{H}} = \widehat{F}_\theta(H) = F(\widehat{F}_\theta(H))$, $F(\cdot) \sim \mathcal{U}[z_{\text{H}}, 1]$.

3.2.1 Anderson-Darling Statistic

The supremum version of the AD statistic was described in Section 3.1.3. A more generally used version of this statistic belongs to the quadratic class defined by the Cramér-von Mises family (equation (15)) with the weight function for the left-truncated samples $\psi(\widehat{F^*}(x)) = \big\{\widehat{F^*}(x)\big(1 - \widehat{F^*}(x)\big)\big\}^{-1}$. Again, with this specification, most weight is being put on the outer left and right quantiles of the distribution, proportional to the appropriate tails. Let $\{X_{(j)}\}_{1 \le j \le n}$ be the vector of the order statistics, such that $X_{(1)} \le X_{(2)} \le ... \le X_{(n)}$. The computing formula for the AD statistic for the left-truncated samples becomes (the derivation is given in Appendix)

$$\begin{aligned} AD^{2*} &= -n + 2n \log(1 - z_{\text{H}}) - \frac{1}{n} \sum_{j=1}^{n} \Big(1 + 2(n - j)\Big) \log\big(1 - z_j\big) + ... \\ &+ \frac{1}{n} \sum_{j=1}^{n} \Big(1 - 2j\Big) \log\big(z_j - z_{\text{H}}\big). \end{aligned}$$

3.3 Cramér-von Mises Statistic

Cramér-von Mises (denoted as W^2) statistic belongs to the Cramér-von Mises family (equation (15)) with the weight function $\psi(\widehat{F^*}(x)) = 1$. Let $\{X_{(j)}\}_{1 \le j \le n}$ be the vector of the order statistics, such that

$X_{(1)} \leq X_{(2)} \leq ... \leq X_{(n)}$. The computing formula for the statistic for the left-truncated samples becomes (the derivation is given in Appendix)

$$W^{2*} = \frac{n}{3} + \frac{n\, z_{\rm H}}{1 - z_{\rm H}} + \frac{1}{n(1 - z_{\rm H})} \sum_{j=1}^{n} (1 - 2j)\, z_j + \frac{1}{(1 - z_{\rm H})^2} \sum_{j=1}^{n} \big(z_j - z_{\rm H}\big)^2. \tag{17}$$

4 "Upper Tail" Anderson-Darling Statistic

In practice, there are often cases when it is necessary to test whether a distribution fits the data well mainly in the upper tail and the fit in the lower tail or around the median is of little or less importance. Examples include operational risk and insurance claims modelling, in which goodness of the fit in the tails determines the Value-at-Risk (or Conditional Value-at-Risk) and the ruin probabilities. Given the Basel II Capital Accord's recommendations, under the Loss Distribution Approach the operational risk capital charge is derived from the Value-at-Risk measure, which requires an accurate estimate of the upper tail of the loss distribution. Similarly, in insurance, the upper tail of the claim size distribution is vital to obtain the right estimates of ruin probability. For this purpose, we introduce a statistic, which we refer to as the *"upper tail"* Anderson-Darling statistic and denote by AD_{up}. We propose two different versions of it.

4.1 Supremum Class "Upper Tail" Anderson-Darling Statistic

The first version of AD_{up} belongs to the supremum class EDF statistics. For the complete data samples, each observation of the KS statistic is assigned a weight of $\psi(\widehat{F}(x)) = \Big\{\big(1 - \widehat{F}(x)\big)\Big\}^{-1}$. Under this specification, the observations in the upper tail are assigned a higher weight than those in the lower tail. Let $\{X_{(j)}\}_{1 \leq j \leq n}$ be the vector of the order statistics, such that $X_{(1)} \leq X_{(2)} \leq ... \leq X_{(n)}$. Then we define the $AD_{\rm up}$ statistic for complete data samples as

$$AD_{\rm up} = \sqrt{n} \sup_x \left| \frac{F_n(x) - \widehat{F}(x)}{1 - \widehat{F}(x)} \right|, \tag{18}$$

Denoting $z_j := \widehat{F}_{\theta^u}(x_{(j)})$, the computing formula is derived from :

$$\begin{aligned} AD^{+}_{\rm up} &= \sqrt{n} \sup_j \left\{ \frac{\frac{j}{n} - z_j}{1 - z_j} \right\}, \\ AD^{-}_{\rm up} &= \sqrt{n} \sup_j \left\{ \frac{z_j - \frac{j-1}{n}}{1 - z_j} \right\}, \end{aligned}$$

and becomes

$$AD_{\rm up} = \max\{AD^{+*}_{\rm up}, AD^{-*}_{\rm up}\}. \tag{19}$$

For the left-truncated samples, denoting $z_j := \widehat{F}_\theta(x_{(j)})$, the counterpart of the AD_{up} statistic can be similarly computed by

$$\begin{aligned} AD_{\text{up}}^{+*} &= \sqrt{n}\sup_j \left\{ \frac{F_n(x_{(j)}) - \widehat{F^*}(x_{(j)})}{1 - \widehat{F^*}(x_{(j)})} \right\} = \sqrt{n}\sup_j \left\{ \frac{z_{\text{H}} + \frac{j}{n}(1 - z_{\text{H}}) - z_j}{1 - z_j} \right\}, \\ AD_{\text{up}}^{-*} &= \sqrt{n}\sup_j \left\{ \frac{\widehat{F^*}(x_{(j)}) - F_n(x_{(j)})}{1 - \widehat{F^*}(x_{(j)})} \right\} = \sqrt{n}\sup_j \left\{ \frac{z_j - z_{\text{H}} - \frac{j-1}{n}(1 - z_{\text{H}})}{1 - z_j} \right\}, \end{aligned}$$

and

$$AD_{\text{up}}^{*} = \max\{AD_{\text{up}}^{+*}, AD_{\text{up}}^{-*}\}. \tag{20}$$

4.2 Quadratic Class "Upper Tail" Anderson-Darling Statistic

Another way to define the *"upper tail"* Anderson-Darling Statistic is by an integral of the Cramér-von Mises family (equation (14)) with the weighting function of the form $\psi(\widehat{F}(x)) = \{1 - F(x)\}^{-2}$ for complete samples, and $\psi(\widehat{F^*}(x)) = \{1 - \widehat{F^*}(x)\}^{-2}$ for left-truncated samples, that gives a higher weight to the upper tail and a lower weight to the lower tail. We define this statistic as AD_{up}^2.

Let $\{X_{(j)}\}_{1 \le j \le n}$ be the vector of the order statistics, such that $X_{(1)} \le X_{(2)} \le ... \le X_{(n)}$. Its general form for complete samples can be expressed as:

$$AD_{\text{up}}^2 = n \int_H^\infty \frac{\left(F_n(x) - \widehat{F}(x)\right)^2}{\left(1 - \widehat{F}(x)\right)^2} d\widehat{F}(x) \tag{21}$$

First, for complete data samples and $\widehat{F}(X) \sim \mathcal{U}\ [0, 1]$ under the null hypothesis. If we denote $z_j := \widehat{F}_{\theta^u}(x_{(j)})$, then straightforward calculations lead to the computing formula

$$AD_{\text{up}}^2 = 2\sum_{j=1}^{n} \log\left(1 - z_j\right) + \frac{1}{n}\sum_{j=1}^{n} \Big(1 + 2(n - j)\Big) \frac{1}{1 - z_j}.$$

Second, for incomplete left-truncated samples, $\widehat{F^*}(X)$ is distributed $\mathcal{U}\ [z_{\text{H}}, 1]$ under the null. We now denote $z_j := \widehat{F}_\theta(x_{(j)})$. Applying the PIT technique leads to the computing formula of the AD_{up}^{2*} statistic for left-truncated samples of the following form (the derivation is given in Appendix):

$$AD_{\text{up}}^{2*} = -2n\log(1 - z_{\text{H}}) + 2\sum_{j=1}^{n} \log\left(1 - z_j\right) + \frac{1 - z_{\text{H}}}{n}\sum_{j=1}^{n} \left(1 + 2(n - j)\right) \frac{1}{1 - z_j}.$$

Tables 1 and 2 summarize the EDF statistics and their computing formulae for complete and left-truncated samples.

5 Application to Loss Data

In this section we apply the GOF testing procedure to (1) operational loss data, extracted from an external database and (2) catastrophe insurance claims data. The operational loss data set was obtained from Zurich IC Squared (IC^2) FIRST Database of Zurich IC Squared (IC^2), an independent consulting subsidiary of Zurich Financial Services Group. The external database is comprised of operational loss events throughout the world. The original loss data cover losses in the period 1950-2002. A few recorded data points were below \$1 million in nominal value, so we excluded them from the analysis, to make it more consistent with the conventional threshold for external databases of \$1 million. Furthermore, we excluded the observations before 1980 because of relatively few data points available (which is most likely due to poor data recording practices). The final dataset for the analysis covered losses for the time period between 1980 and 2002. It consists of five types of losses: "Relationship" (such as events related to legal issues, negligence and sales-related fraud), "Human" (such as events related to employee errors, physical injury and internal fraud), "Processes" (such as events related to business errors, supervision, security and transactions), "Technology" (such as events related to technology and computer failure and telecommunications) and "External" (such as events related to natural and man-made disasters and external fraud). The loss amounts have been adjusted for inflation using the Consumer Price Index from the U.S. Department of Labor. The numbers of data points of each type are n =849, 813, 325, 67, and 233, respectively.

The insurance claims data set covers claims resulting from natural catastrophe events occurred in the United States over the time period from 1990 to 1996. It was obtained from Insurance Services Office Inc. Property Claim Services (PCS). The data set includes 222 losses. The observations are greater than \$25 million in nominal value.

Left-truncated distributions of four types were fitted to each of the data set: Exponential, Lognormal, Weibull, and Pareto (GPD). Table 3 presents the observed statistic values and the p-values for the six data sets (five operational losses and insurance claims), obtained with the testing procedure described in Section 2.

The results reported in Table 3 suggest that fitting heavier tailed distributions, such as Pareto and Weibull, results in lower values of the GOF statistics, which leads to acceptance of the null for practically all loss types and all criteria of the goodness of fit, viewed from the high p-values. Since the analysis of operational losses deals with estimating the operational Value-at-Risk (VaR), it is reasonable to determine the ultimate best fit on the basis of the AD_{up} and AD^2_{up} statistics, introduced in this paper. As can be seen from Table 3, these proposed measures suggest a much better fit of the heavier-tailed distributions. Moreover, for the Pareto distribution, while the statistics that focus on the center of the data (Kolmogorov-

Smirnov, Kuiper, Cramér-von Mises) do not show a good fit, the AD_{up} and AD^2_{up} statistics indicate that the fit in the upper tail is very good. It should be noted that the Pareto and Weibull distributions very often suggest a superior fit in the upper tail to the Lognormal distribution. Yet is it the Lognormal distribution that was suggested in 2001 by the Basel Committee, BIS (2001).

6 Conclusions

In this paper we present a technique for modifying the existing goodness-of-fit test statistics so that they can be applied to loss models in which the available data set is incomplete and is truncated from below. Such left-truncation is often present in loss data when the data are being recorded starting from a fixed amount, and the data below are not recorded at all. Exact computing formulae for the Kolmogorov-Smirnov, Kuiper, Anderson-Darling, and Cramér-von Mises for the left-truncated samples are presented.

In risk management, it is often vital to have a good fit of a hypothesized distribution in the upper tail of the loss data. It is important in loss models that deal with Value-at-Risk, Conditional Value-at-Risk and ruin probability. We suggest using two other versions of the Anderson-Darling statistic (which we refer to as the *"upper tail"* Anderson-Darling statistic) in which the weighting function is proportional to the weight of only the upper tail of the distribution. A supremum and quadratic versions of the statistic are proposed. Such statistic is convenient to use when it is necessary to examine the goodness of fit of a distribution in the right tail of the data, while the fit in the left tail is unimportant.

The technique is applied to check the goodness of fit of a number of distributions using operational loss data and catastrophe insurance claims data sets. From the empirical analysis we conclude that heavier-tailed distributions better fit the data than Lognormal or thinner-tailed distributions in many instances. In particular, the conclusion is strongly supported by the *"upper tail"* Anderson-Darling tests.

Footnotes

* S. Rachev gratefully acknowledges research support by grants from Division of Mathematical, Life and Physical Sciences, College of Letters and Science, University of California, Santa Barbara, the Deutschen Forschungsgemeinschaft and the Deutscher Akademischer Austausch Dienst. Authors are grateful to R. Jammalamadaka of University of California Santa Barbara, and K. Burnecki and R. Weron of Wrocław University of Technology, for helpful comments and remarks.

[1] Related works on the discussion of these widely used tests include Anderson and Darling (1952), Anderson and Darling (1954), D'Agostino and Stephens (1986), Schorack and Wellner (1986).

[2] More accurately, m should be estimated as $m = \lceil n\frac{z_{\mathrm{H}}}{(1-z_{\mathrm{H}})} \rceil$, but it can be ignored for the purpose of this paper.

Figure Legends

Figure 1: Illustration of empirical distribution function and fitted cumulative distribution function with missing data below threshold H.

Bibliography

Anderson, T.W. and D.A. Darling. (1952).“Asymptotic Theory of Certain ‘Goodness of Fit’ Criteria Based on Stochastic Processes.” *The Annals of Mathematical Statistics* **23, no. 2, 193-212.**

Anderson, T.W. and D.A. Darling. (1954). “A Test of Goodness of Fit.” *Journal of the American Statistical Association* **49, no. 268, 765-769.**

BIS. (2001). “Working paper on the regulatory treatment of operational risk”. `http://www.bis.org`**.**

BIS. (2003). “The 2002 loss data collection exercise for operational risk: summary of the data collected”. `http://www.bis.org`**.**

D’Agostino, Ralph and Michael Stephens (eds.). (1986). *Goodness-of-Fit Techniques***. New York and Basel: Marcel Dekker, Inc.**

Dufour, R. and U.R. Maag. (1978). “Distribution Results for Modified Kolmogorov-Smirnov Statistics for Truncated or Censored Samples.” *Technometrics*, **20, 29-32.**

Gastaldi, T. (1993). “A Kolmogorov-Smirnov Test Procedure Involving a Possibly Censored or Truncated Sample.” *Communications in Statistics: Theory and Method*, **22, 31-39.**

Guilbaud, Olivier. (1998). “Exact Kolmogorov-Type Test for Left-Truncated and/or Right-Censored Data.” *Journal of American Statistical Association*, **83, 213-221.**

Ross, Sheldon. (2001). *Simulation***. 3rd ed. Boston, MA: Academic Press.**

Shorack, Galen and Jon Wellner. (1986). *Empirical Processes with Applications to Statistics***. New York, MA: John Wiley & Sons.**

Appendix

Derivation of AD^{2*} Computing Formula

$$AD^{2*} = n \int_{H}^{+\infty} \frac{\big(F_n(x) - \widehat{F^*}(x)\big)^2}{\widehat{F^*}(x)\big(1-\widehat{F^*}(x)\big)} d\widehat{F^*}(x) \overset{PIT}{=} n^c \int_{z_{\mathbf{H}}}^{1} \frac{\big(F_n(z^*)(1-z_{\mathbf{H}}) + z_{\mathbf{H}} - z\big)^2}{(z-z_{\mathbf{H}})(1-z)} dz$$

by the PIT technique, where in the original integral $F_n(Z^*) = F^*(F_n(X)) = F_n(X)$ is the empirical distribution function of $Z^* := \widehat{F^*}(X) = F^*(\widehat{F^*}(X))$ so that $F^*(\cdot) \sim \mathcal{U}[0,1]$. Changing variable and using equation (7), the integral becomes expressed in terms of $z_{\mathbf{H}} := \widehat{F}_\theta(H) = F(\widehat{F}_\theta(H))$ and $Z := \widehat{F}_\theta(X) = F(\widehat{F}_\theta(X))$ so that $F(\cdot) \sim \mathcal{U}[z_{\mathbf{H}}, 1]$. We denote $n^c = \frac{n}{1-z_{\mathbf{H}}}$.

Using equation (7), the computing formula is expressed in terms of $z_j := \widehat{F}_\theta(x_{(j)}) = F(\widehat{F}_\theta(x_{(j)}))$, $j = 1, 2, ..., n$ as

$$\frac{1}{n^c} AD^{2*} = \underbrace{\int_{z_{\mathbf{H}}}^{z_1} \frac{\big(z-z_{\mathbf{H}}\big)^2}{(z-z_{\mathbf{H}})(1-z)} dz}_{\mathbf{A}} + \underbrace{\sum_{j=1}^{n-1} \int_{z_j}^{z_{j+1}} \frac{\big(\frac{j}{n}(1-z_{\mathbf{H}}) + z_{\mathbf{H}} - z\big)^2}{(z-z_{\mathbf{H}})(1-z)} dz}_{\mathbf{B}} + \underbrace{\int_{z_n}^{1} \frac{\big(1-z\big)^2}{(z-z_{\mathbf{H}})(1-z)} dz}_{\mathbf{C}}.$$

Separately solving for A, B and C,

$$\mathbf{A} = z_{\mathbf{H}} - z_1 + (1-z_{\mathbf{H}})\Big(\log(1-z_{\mathbf{H}}) - \log(1-z_1)\Big);$$

$$\begin{aligned}
\mathbf{B} &= z_1 - z_n + \frac{1-z_{\mathbf{H}}}{n^2} \sum_{j=1}^{n-1} (n-j)^2 \big(\log(1-z_j) - \log(1-z_{j+1})\big) - ... \\
&- 2\frac{1-z_{\mathbf{H}}}{n^2} \sum_{j=1}^{n-1} j^2 \big(\log(z_j - z_{\mathbf{H}}) - \log(z_{j+1} - z_{\mathbf{H}})\big) \\
&= z_1 - z_n + (1-z_{\mathbf{H}})\log(1-z_1) - \frac{1-z_{\mathbf{H}}}{n^2} \sum_{j=1}^{n} \big(1 + 2(n-j)\big)\log(1-z_j) + ... \\
&+ (1-z_{\mathbf{H}})\log(z_n - z_{\mathbf{H}}) + \frac{1-z_{\mathbf{H}}}{n^2} \sum_{j=1}^{n} (1-2j)\log(z_j - z_{\mathbf{H}});
\end{aligned}$$

$$\mathbf{C} = z_n - 1 + (1-z_{\mathbf{H}})\Big(\log(1-z_{\mathbf{H}}) - \log(z_n - z_{\mathbf{H}})\Big).$$

Summing the terms A, B and C, multiplying by n^c, and simplifying yields the final computing formula

$$\begin{aligned}
AD^{2*} &= -n + 2n\log(1-z_{\mathbf{H}}) - \frac{1}{n} \sum_{j=1}^{n} \Big(1 + 2(n-j)\Big)\log\big(1-z_j\big) + ... \\
&+ \frac{1}{n} \sum_{j=1}^{n} \Big(1-2j\Big)\log\big(z_j - z_{\mathbf{H}}\big).
\end{aligned}$$

Derivation of W^{2*} Computing Formula

$$W^{2*} = n \int_{H}^{+\infty} \big(F_n(x) - \widehat{F^*}(x)\big)^2 d\widehat{F^*}(x) \overset{PIT}{=} n^c \int_{z_{\mathbf{H}}}^{1} \frac{\big(F_n(z^*)(1-z_{\mathbf{H}}) + z_{\mathbf{H}} - z\big)^2}{(1-z_{\mathbf{H}})^2} dz$$

by the PIT technique, where in the original integral $F_n(Z^*) = F^*(F_n(X)) = F_n(X)$ is the empirical distribution function of $Z^* := \widehat{F^*}(X) = F^*(\widehat{F^*}(X))$ so that $F^*(\cdot) \sim \mathcal{U}[0,1]$. Changing variable and using equation (7), the integral becomes expressed in terms of $z_{\mathbf{H}} := \widehat{F}_\theta(H) = F(\widehat{F}_\theta(H))$ and $Z := \widehat{F}_\theta(X) = F(\widehat{F}_\theta(X))$ so that $F(\cdot) \sim \mathcal{U}[z_{\mathbf{H}}, 1]$. We denote $n^c = \frac{n}{1-z_{\mathbf{H}}}$.

Using equation (7), the computing formula is expressed in terms of $z_j := \widehat{F}_\theta(x_{(j)}) = F(\widehat{F}_\theta(x_{(j)}))$, $j = 1, 2, ..., n$ as

$$\frac{(1-z_{\mathbf{H}})^2}{n^c} W^{2*} = \underbrace{\int_{z_{\mathbf{H}}}^{z_1} \big(z_{\mathbf{H}} - z\big)^2 dz}_{\mathbf{A}} + \underbrace{\sum_{j=1}^{n-1} \int_{z_j}^{z_{j+1}} \big(\frac{j}{n}(1-z_{\mathbf{H}}) + z_{\mathbf{H}} - z\big)^2 dz}_{\mathbf{B}} + \underbrace{\int_{z_n}^{1} \big(1-z\big)^2 dz}_{\mathbf{C}}.$$

Separately solving for A, B and C,

$$\mathbf{A} = -\frac{{z_{\mathbf{H}}}^3}{3} + {z_{\mathbf{H}}}^2 z_1 - z_{\mathbf{H}} {z_1}^2 + \frac{{z_1}^3}{3};$$

$$\begin{aligned}
\mathbf{B} &= \frac{{z_n}^3}{3} - \frac{{z_1}^3}{3} + z_{\mathbf{H}} {z_1}^2 - z_{\mathbf{H}} {z_n}^2 - {z_{\mathbf{H}}}^2 z_1 + {z_{\mathbf{H}}}^2 z_n + ... \\
&+ \frac{(1-z_{\mathbf{H}})^2}{n^2} \sum_{j=1}^{n-1} j^2 \big(z_{j+1} - z_j\big) + \frac{1-z_{\mathbf{H}}}{n} \sum_{j=1}^{n-1} j\big({z_j}^2 - {z_{j+1}}^2\big) + 2z_{\mathbf{H}} \frac{1-z_{\mathbf{H}}}{n} \sum_{j=1}^{n-1} j\big(z_{j+1} - z_j\big) \\
&= \frac{{z_n}^3}{3} - \frac{{z_1}^3}{3} + z_{\mathbf{H}} {z_1}^2 - z_{\mathbf{H}} {z_n}^2 - {z_{\mathbf{H}}}^2 z_1 + {z_{\mathbf{H}}}^2 z_n + \frac{(1-z_{\mathbf{H}})^2}{n^2} \left(n^2 z_n + \sum_{j=1}^{n} (1-2j) z_j \right) + ... \\
&+ \frac{1-z_{\mathbf{H}}}{n} \left(\sum_{j=1}^{n} {z_j}^2 - n{z_n}^2 \right) + 2z_{\mathbf{H}} \frac{1-z_{\mathbf{H}}}{n} \left(n z_n - \sum_{j=1}^{n} z_j \right) \\
&= (1-z_{\mathbf{H}}) z_n + z_{\mathbf{H}} (1-z_{\mathbf{H}}) z_n - (1-z_{\mathbf{H}}) {z_n}^2 + \frac{(1-z_{\mathbf{H}})^2}{n^2} \sum_{j=1}^{n} (1-2j) z_j + ... \\
&+ \frac{1-z_{\mathbf{H}}}{n} \sum_{j=1}^{n} {z_j}^2 - 2z_{\mathbf{H}} \frac{1-z_{\mathbf{H}}}{n} \sum_{j=1}^{n} z_j;
\end{aligned}$$

$$\mathbf{C} = \frac{1}{3} + {z_n}^2 - z_n - \frac{{z_n}^3}{3}.$$

Summing the terms A, B and C, multiplying by $\frac{n^c}{(1-z_{\mathbf{H}})^2}$, and simplifying yields the final computing formula

$$W^{2*} = \frac{n}{3} + \frac{n\, z_{\mathbf{H}}}{1-z_{\mathbf{H}}} + \frac{1}{n(1-z_{\mathbf{H}})}\sum_{j=1}^{n}(1-2j)z_j + \frac{1}{(1-z_{\mathbf{H}})^2}\sum_{j=1}^{n}\big(z_j - z_{\mathbf{H}}\big)^2.$$

Derivation of $AD^2_{\mathbf{up}}$ Computing Formula

$$AD^2_{\mathbf{up}} = n\int_{-\infty}^{+\infty}\frac{\big(F_n(x)-\widehat{F}(x)\big)^2}{\big(1-\widehat{F}(x)\big)^2}d\widehat{F}(x) \overset{PIT}{=} n\int_0^1\frac{\big(F_n(z)-z\big)^2}{(1-z)^2}dz$$

by the method of Probability Integral Transformation (PIT), where $F_n(Z) = F(F_n(X)) = F_n(X)$ is the empirical distribution function of $Z = \widehat{F}_{\theta^u}(X) = F(\widehat{F}_{\theta^u}(X))$ so that $F(\cdot) \sim \mathcal{U}[0,1]$.

Using equation (1), the computing formula is expressed in terms of $z_j := \widehat{F}_{\theta^u}(x_{(j)}) = F(\widehat{F}_{\theta^u}(x_{(j)})), j = 1, 2, ..., n$ as

$$\frac{1}{n}AD^2_{\mathbf{up}} = \underbrace{\int_0^{z_1}\frac{z^2}{(1-z)^2}dz}_{\mathbf{A}} + \underbrace{\sum_{j=1}^{n-1}\int_{z_j}^{z_{j+1}}\frac{\big(\frac{j}{n}-z\big)^2}{(1-z)^2}dz}_{\mathbf{B}} + \underbrace{\int_{z_n}^{1}\frac{\big(1-z\big)^2}{(1-z)^2}dz}_{\mathbf{C}}$$

Separately solving for A, B and C,

$$\mathbf{A} = z_1 - 1 + \frac{1}{1-z_1} + 2\log(1-z_1);$$

$$\begin{aligned}
\mathbf{B} &= z_n - z_1 - \frac{1}{n^2}\sum_{j=1}^{n-1}(n-j)^2\big(\frac{1}{1-z_j} - \frac{1}{1-z_{j+1}}\big) - ... \\
&- 2\frac{1}{n}\sum_{j=1}^{n-1}(n-j)\big(\log(1-z_j) - \log(1-z_{j+1})\big) \\
&= z_n - z_1 - \frac{1}{1-z_1} + \frac{1}{n^2}\sum_{j=1}^{n}\big(1+2(n-j)\big)\frac{1}{1-z_j} - 2\log(1-z_1) + 2\frac{1}{n}\sum_{j=1}^{n}\log(1-z_j);
\end{aligned}$$

$$\mathbf{C} = 1 - z_n.$$

Summing the terms A, B and C, multiplying by n and simplifying yields

$$AD^2_{\mathbf{up}} = 2\sum_{j=1}^{n}\log\big(1-z_j\big) + \frac{1}{n}\sum_{j=1}^{n}\Big(1+2(n-j)\Big)\frac{1}{1-z_j}.$$

Derivation of $AD^{2*}_{\mathbf{up}}$ Computing Formula

$$AD^{2*}_{\mathbf{up}} = n \int_{H}^{+\infty} \frac{\big(F_n(x) - \widehat{F^*}(x)\big)^2}{\big(1 - \widehat{F^*}(x)\big)^2} d\widehat{F^*}(x) \overset{PIT}{=} n^c \int_{z_{\mathbf{H}}}^{1} \frac{\big(F_n(z^*)(1 - z_{\mathbf{H}}) + z_{\mathbf{H}} - z\big)^2}{(1-z)^2} dz$$

by the PIT technique, where in the original integral $F_n(Z^*) = F^*(F_n(X)) = F_n(X)$ is the empirical distribution function of $Z^* := \widehat{F^*}(X) = F^*(\widehat{F^*}(X))$ so that $F^*(\cdot) \sim \mathcal{U}[0,1]$. Changing variable and using equation (7), the integral becomes expressed in terms of $z_{\mathbf{H}} := \widehat{F}_\theta(H) = F(\widehat{F}_\theta(H))$ and $Z := \widehat{F}_\theta(X) = F(\widehat{F}_\theta(X))$ so that $F(\cdot) \sim \mathcal{U}[z_{\mathbf{H}}, 1]$. We denote $n^c = \frac{n}{1 - z_{\mathbf{H}}}$.

Using equation (7), the computing formula is expressed in terms of $z_j := \widehat{F}_\theta(x_{(j)}) = F(\widehat{F}_\theta(x_{(j)}))$, $j = 1, 2, ..., n$ as

$$\frac{1}{n^c} AD^{2*}_{\mathbf{up}} = \underbrace{\int_{z_{\mathbf{H}}}^{z_1} \frac{\big(z - z_{\mathbf{H}}\big)^2}{(1-z)^2} dz}_{\mathbf{A}} + \underbrace{\sum_{j=1}^{n-1} \int_{z_j}^{z_{j+1}} \frac{\big(\frac{j}{n}(1 - z_{\mathbf{H}}) + z_{\mathbf{H}} - z\big)^2}{(1-z)^2} dz}_{\mathbf{B}} + \underbrace{\int_{z_n}^{1} \frac{\big(1 - z\big)^2}{(1-z)^2} dz}_{\mathbf{C}}$$

Separately solving for A, B and C,

$$\mathbf{A} = z_1 - z_{\mathbf{H}} - (1 - z_{\mathbf{H}}) + (1 - z_{\mathbf{H}})^2 \frac{1}{1 - z_1} - 2(1 - z_{\mathbf{H}}) \log(1 - z_{\mathbf{H}}) + 2(1 - z_{\mathbf{H}}) \log(1 - z_1);$$

$$\begin{aligned}
\mathbf{B} &= z_n - z_1 - \frac{(1 - z_{\mathbf{H}})^2}{n^2} \sum_{j=1}^{n-1} (n-j)^2 \big(\frac{1}{1 - z_j} - \frac{1}{1 - z_{j+1}}\big) - ... \\
&- 2\frac{1 - z_{\mathbf{H}}}{n} \sum_{j=1}^{n-1} (n-j) \big(\log(1 - z_j) - \log(1 - z_{j+1})\big) \\
&= z_n - z_1 - (1 - z_{\mathbf{H}})^2 \frac{1}{1 - z_1} + \frac{(1 - z_{\mathbf{H}})^2}{n^2} \sum_{j=1}^{n} \big(1 + 2(n-j)\big) \frac{1}{1 - z_j} - ... \\
&- 2(1 - z_{\mathbf{H}}) \log(1 - z_1) + 2\frac{1 - z_{\mathbf{H}}}{n} \sum_{j=1}^{n} \log(1 - z_j);
\end{aligned}$$

$$\mathbf{C} = 1 - z_n.$$

Summing the terms A, B and C, multiplying by n^c and simplifying yields the final computing formula

$$AD^{2*}_{\mathbf{up}} = -2n \log(1 - z_{\mathbf{H}}) + 2 \sum_{j=1}^{n} \log\big(1 - z_j\big) + \frac{1 - z_{\mathbf{H}}}{n} \sum_{j=1}^{n} \big(1 + 2(n-j)\big) \frac{1}{1 - z_j}.$$

Tables

$\mathbf{H_0}: F_n(x) \in \widehat{F}(x)$ vs. $\mathbf{H_A}: F_n(x) \notin \widehat{F}(x), \quad \widehat{F}(x) := \widehat{F}_{\theta^u}(x)$ Notations: $z_j := \widehat{F}_{\theta^u}(x_{(j)}), \quad j = 1, 2, ..., n$	
Statistic	Description & computing formula
KS	$KS = \sqrt{n} \sup_x \left\lvert F_n(x) - \widehat{F}(x) \right\rvert$ Computing formula: $KS = \sqrt{n} \max \left\{ \sup_j \left\{ \frac{j}{n} - z_j \right\}, \sup_j \left\{ z_j - \frac{j-1}{n} \right\} \right\}$
V	$V = \sqrt{n} (\sup_x \left\{ F_n(x) - \widehat{F}(x) \right\} + \sup_x \left\{ \widehat{F}(x) - F_n(x) \right\}$ Computing formula: $V = \sqrt{n} \left(\sup_j \left\{ \frac{j}{n} - z_j \right\} + \sup_j \left\{ z_j - \frac{j-1}{n} \right\} \right)$
AD	$AD = \sqrt{n} \sup_x \left\lvert \frac{F_n(x) - \widehat{F}(x)}{\sqrt{\widehat{F}(x)\left(1-\widehat{F}(x)\right)}} \right\rvert$ Computing formula: $AD = \sqrt{n} \max \left\{ \sup_j \left\{ \frac{\frac{j}{n} - z_j}{\sqrt{z_j\left(1-z_j\right)}} \right\}, \sup_j \left\{ \frac{z_j - \frac{j-1}{n}}{\sqrt{z_j\left(1-z_j\right)}} \right\} \right\}$
AD_{up}	$AD_{\text{up}} = \sqrt{n} \sup_x \left\lvert \frac{F_n(x) - \widehat{F}(x)}{1-\widehat{F}(x)} \right\rvert$ Computing formula: $AD_{\text{up}} = \sqrt{n} \max \left\{ \sup_j \left\{ \frac{\frac{j}{n} - z_j}{1-z_j} \right\}, \sup_j \left\{ \frac{z_j - \frac{j-1}{n}}{1-z_j} \right\} \right\}$
AD^2	$AD^2 = n \int_{-\infty}^{\infty} \frac{\left(F_n(x) - \widehat{F}(x)\right)^2}{\widehat{F}(x)\left(1-\widehat{F}(x)\right)} d\widehat{F}(x)$ Computing formula: $AD^2 = -n + \frac{1}{n} \sum_{j=1}^{n} \left(1-2j\right) \log z_j - \frac{1}{n} \sum_{j=1}^{n} \left(1+2(n-j)\right) \log(1-z_j)$
W^2	$W^2 = n \int_{-\infty}^{\infty} \left(F_n(x) - \widehat{F}(x)\right)^2 d\widehat{F}(x)$ Computing formula: $W^2 = \frac{n}{3} + \frac{1}{n} \sum_{j=1}^{n} \left(1-2j\right) z_j + \sum_{j=1}^{n} {z_j}^2$
AD^2_{up}	$AD^2_{\text{up}} = n \int_{-\infty}^{\infty} \frac{\left(F_n(x) - \widehat{F}(x)\right)^2}{\left(1-\widehat{F}(x)\right)^2} d\widehat{F}(x)$ Computing formula: $AD^2_{\text{up}} = \frac{1}{n} \sum_{j=1}^{n} \left(1+2(n-j)\right) \frac{1}{(1-z_j)} + 2 \sum_{j=1}^{n} \log(1-z_j)$

Table 1: Description of EDF statistics for complete data samples.

$\mathbf{H_0}: F_n(x) \in \widehat{F^*}(x)$ vs. $\mathbf{H_A}: F_n(x) \notin \widehat{F^*}(x)$, $\widehat{F^*}(x) := \frac{\widehat{F}_\theta(x)-\widehat{F}_\theta(H)}{1-\widehat{F}_\theta(H)}$

Notations: $z_j := \widehat{F}_\theta(x_{(j)})$, $z_{\mathrm{H}} = \widehat{F}_\theta(H)$, $j = 1, 2, ..., n$

Statistic	Description & computing formula
KS^*	$KS^* = \sqrt{n} \sup_x \left\lvert F_n(x) - \widehat{F^*}(x) \right\rvert$ Computing formula: $KS^* = \frac{\sqrt{n}}{1-z_{\mathrm{H}}} \max \left\{ \sup_j \left\{ z_{\mathrm{H}} + \frac{j}{n}(1-z_{\mathrm{H}}) - z_j \right\}, \sup_j \left\{ z_j - \left(z_{\mathrm{H}} + \frac{j-1}{n}(1-z_{\mathrm{H}}) \right) \right\} \right\}$
V^*	$V^* = \sqrt{n}\left(\sup_x \left\{ F_n(x) - \widehat{F^*}(x) \right\} + \sup_x \left\{ \widehat{F^*}(x) - F_n(x) \right\} \right)$ Computing formula: $V^* = \frac{\sqrt{n}}{1-z_{\mathrm{H}}} \left(\sup_j \left\{ z_{\mathrm{H}} + \frac{j}{n}(1-z_{\mathrm{H}}) - z_j \right\} + \sup_j \left\{ z_j - \left(z_{\mathrm{H}} + \frac{j-1}{n}(1-z_{\mathrm{H}}) \right) \right\} \right)$
AD^*	$AD^* = \sqrt{n} \sup_x \left\lvert \frac{F_n(x)-\widehat{F^*}(x)}{\sqrt{\widehat{F^*}(x)\left(1-\widehat{F^*}(x)\right)}} \right\rvert$ Computing formula: $AD^* = \sqrt{n} \max \left\{ \sup_j \left\{ \frac{z_{\mathrm{H}} + \frac{j}{n}(1-z_{\mathrm{H}}) - z_j}{\sqrt{\left(z_j - z_{\mathrm{H}}\right)\left(1-z_j\right)}} \right\}, \sup_j \left\{ \frac{z_j - z_{\mathrm{H}} - \frac{j-1}{n}(1-z_{\mathrm{H}})}{\sqrt{\left(z_j - z_{\mathrm{H}}\right)\left(1-z_j\right)}} \right\} \right\}$
AD^*_{up}	$AD^*_{\mathrm{up}} = \sqrt{n} \sup_x \left\lvert \frac{F_n(x)-\widehat{F^*}(x)}{1-\widehat{F^*}(x)} \right\rvert$ Computing formula: $AD^*_{\mathrm{up}} = \sqrt{n} \max \left\{ \sup_j \left\{ \frac{z_{\mathrm{H}} + \frac{j}{n}(1-z_{\mathrm{H}}) - z_j}{1-z_j} \right\}, \sup_j \left\{ \frac{z_j - z_{\mathrm{H}} - \frac{j-1}{n}(1-z_{\mathrm{H}})}{1-z_j} \right\} \right\}$
AD^{2*}	$AD^{2*} = n \int_H^\infty \frac{\left(F_n(x)-\widehat{F^*}(x)\right)^2}{\widehat{F^*}(x)\left(1-\widehat{F^*}(x)\right)} d\widehat{F^*}(x)$ Computing formula: $AD^{2*} = -n + 2n\log(1-z_{\mathrm{H}}) - \frac{1}{n}\sum_{j=1}^n \left(1 + 2(n-j)\right)\log\left(1-z_j\right) + \frac{1}{n}\sum_{j=1}^n \left(1-2j\right)\log\left(z_j - z_{\mathrm{H}}\right)$
W^{2*}	$W^{2*} = n \int_H^\infty \left(F_n(x) - \widehat{F^*}(x)\right)^2 d\widehat{F^*}(x)$ Computing formula: $W^{2*} = \frac{n}{3} + \frac{n\, z_{\mathrm{H}}}{1-z_{\mathrm{H}}} + \frac{1}{n(1-z_{\mathrm{H}})} \sum_{j=1}^n \left(1-2j\right) z_j + \frac{1}{(1-z_{\mathrm{H}})^2} \sum_{j=1}^n \left(z_j - z_{\mathrm{H}}\right)^2$
AD^{2*}_{up}	$AD^{2*}_{\mathrm{up}} = n \int_H^\infty \frac{\left(F_n(x)-\widehat{F^*}(x)\right)^2}{\left(1-\widehat{F^*}(x)\right)^2} d\widehat{F^*}(x)$ Computing formula: $AD^{2*}_{\mathrm{up}} = -2n\log(1-z_{\mathrm{H}}) + 2\sum_{j=1}^n \log\left(1-z_j\right) + \frac{1-z_{\mathrm{H}}}{n} \sum_{j=1}^n \left(1+2(n-j)\right)\frac{1}{1-z_j}$

Table 2: Description of EDF statistics for left-truncated (threshold=H) data samples.

	KS	V	AD	AD_{up}	AD^2	AD^2_{up}	W^2
				Exponential			
"Relationship"	11.0868	11.9973	$1.3{\cdot}10^{7}$	$1.2{\cdot}10^{23}$	344.37	$1.2{\cdot}10^{14}$	50.5365
	[<0.005]	[<0.005]	[<0.005]	[<0.005]	[<0.005]	[<0.005]	[<0.005]
"Human"	14.0246	14.9145	$2.4{\cdot}10^{6}$	$1.1{\cdot}10^{22}$	609.15	$3.0{\cdot}10^{12}$	80.3703
	[<0.005]	[<0.005]	[<0.005]	[<0.005]	[<0.005]	[<0.005]	[<0.005]
"Processes"	7.6043	8.4160	$3.7{\cdot}10^{6}$	$1.7{\cdot}10^{22}$	167.61	$6.6{\cdot}10^{5}$	22.5762
	[<0.005]	[<0.005]	[<0.005]	[<0.005]	[<0.005]	[<0.005]	[<0.005]
"Technology"	3.2160	3.7431	27.6434	$1.4{\cdot}10^{6}$	27.8369	780.50	2.9487
	[<0.005]	[<0.005]	[<0.005]	[<0.005]	[<0.005]	[<0.005]	[<0.005]
"External"	6.5941	6.9881	$4.4{\cdot}10^{6}$	$2.0{\cdot}10^{22}$	128.35	$5.0{\cdot}10^{7}$	17.4226
	[<0.005]	[<0.005]	[<0.005]	[<0.005]	[<0.005]	[<0.005]	[<0.005]
Catastrophe	5.5543	5.9282	$9.0{\cdot}10^{6}$	$4.1{\cdot}10^{22}$	72.2643	$6.1{\cdot}10^{13}$	13.1717
	[<0.005]	[<0.005]	[<0.005]	[<0.005]	[<0.005]	[<0.005]	[<0.005]
				Lognormal			
"Relationship"	0.8056	1.3341	2.6094	875.40	0.7554	4.6122	0.1012
	[0.082]	[0.138]	[0.347]	[0.593]	[0.043]	[0.401]	[0.086]
"Human"	0.8758	1.5265	3.9829	1086.16	0.7505	4.5160	0.0804
	[0.032]	[0.039]	[0.126]	[0.462]	[0.044]	[0.408]	[0.166]
"Processes"	0.6584	1.1262	2.0668	272.61	0.4624	4.0556	0.0603
	[0.297]	[0.345]	[0.508]	[0.768]	[0.223]	[0.367]	[0.294]
"Technology"	1.1453	1.7896	2.8456	41.8359	1.3778	6.4213	0.2087
	[<0.005]	[0.005]	[0.209]	[0.994]	[<0.005]	[0.067]	[<0.005]
"External"	0.6504	1.2144	2.1702	316.20	0.5816	2.5993	0.0745
	[0.326]	[0.266]	[0.469]	[0.459]	[0.120]	[0.589]	[0.210]
Catastrophe	0.6854	1.1833	5.3860	$1.1{\cdot}10^{4}$	0.7044	27.4651	0.0912
	[0.243]	[0.307]	[0.064]	[0.053]	[0.068]	[0.023]	[0.111]
				Weibull			
"Relationship"	0.5553	1.0821	3.8703	$2.7{\cdot}10^{4}$	0.7073	13.8191	0.0716
	[0.625]	[0.514]	[0.138]	[0.080]	[0.072]	[0.081]	[0.249]
"Human"	0.8065	1.5439	4.3544	$3.2{\cdot}10^{4}$	0.7908	8.6610	0.0823
	[0.093]	[0.051]	[0.095]	[0.068]	[0.053]	[0.112]	[0.176]
"Processes"	0.6110	1.0620	1.7210	2200.75	0.2069	2.2340	0.0338
	[0.455]	[0.532]	[0.766]	[0.192]	[0.875]	[0.758]	[0.755]
"Technology"	1.0922	1.9004	2.6821	52.5269	1.4536	4.8723	0.2281
	[<0.005]	[<0.005]	[0.216]	[0.944]	[<0.005]	[0.087]	[<0.005]
"External"	0.4752	0.9498	2.4314	4382.68	0.3470	5.3662	0.0337
	[0.852]	[0.726]	[0.384]	[0.108]	[0.519]	[0.164]	[0.431]
Catastrophe	0.8180	1.5438	5.6345	$1.5{\cdot}10^{4}$	1.3975	15.8416	0.1965
	[0.096]	[0.041]	[0.043]	[0.028]	[0.007]	[0.025]	[0.006]
				Pareto (GPD)			
"Relationship"	1.4797	2.6084	3.5954	374.68	3.7165	22.1277	0.5209
	[<0.005]	[<0.005]	[0.172]	[>0.995]	[<0.005]	[0.048]	[<0.005]
"Human"	1.4022	2.3920	3.6431	374.68	2.7839	23.7015	0.3669
	[<0.005]	[<0.005]	[0.167]	[>0.995]	[<0.005]	[0.051]	[<0.005]
"Processes"	1.0042	1.9189	4.0380	148.24	2.6022	13.1082	0.3329
	[<0.005]	[<0.005]	[0.104]	[>0.995]	[<0.005]	[0.087]	[<0.005]
"Technology"	1.2202	1.8390	3.0843	33.4298	1.6182	8.8484	0.2408
	[<0.005]	[<0.005]	[0.177]	[>0.995]	[<0.005]	[0.067]	[<0.005]
"External"	0.9708	1.8814	2.7742	151.94	1.7091	8.6771	0.2431
	[0.009]	[0.005]	[0.284]	[0.949]	[<0.005]	[0.106]	[<0.005]
Catastrophe	0.4841	0.8671	2.4299	1277.28	0.3528	4.3053	0.0390
	[0.799]	[0.837]	[0.369]	[0.239]	[0.490]	[0.235]	[0.645]

Table 3: Goodness-of-fit tests for operational loss data. P-values (in square brackets) were obtained via 1,000 Monte Carlo simulations.

Figures

Figure 1