Journal of Operational Risk **10**(1), 1–50

Research Paper

Improved goodness-of-fit measures

Peter Mitic

Santander UK, 2 Triton Square, Regent's Place, London NW1 3AN, UK;
email: peter.mitic@santander.co.uk

(Received ?; revised ?; accepted ?)

ABSTRACT

New goodness-of-fit measures which are significant improvements on existing measures are described. They use the intuitive geometrical concept of the area enclosed by the curve of a fitted distribution and the profile of the empirical cumulative distribution function. A transformation of this profile simplifies the geometry and provides three new goodness-of-fit tests. The integrity of this transformation is justified by topological arguments. The new tests provide a quantitative justification for qualitative judgements on goodness-of-fit, are independent of population size and provide a workable way to objectively choose a best fit distribution from a group of candidate distributions.

Changes to this sentence and the next OK?

Keywords: goodness-of-fit; transformed normal; cumulative distribution; significance level; topology.

1 INTRODUCTION AND MOTIVATION

Established goodness-of-fit (GoF) methods, such as the Anderson–Darling (AD) and Kolmogorov–Smirnoff (KS) tests, are the default tests when assessing goodness-of-fit in the context of operational risk. Indeed, using these tests is recommended as part of the Basel regulatory scheme (Basel Committee on Banking Supervision 2011).

Minor unmarked changes to this paper have been made according to US English idiom and spelling, journal style, etc. In addition, marginal queries have been added where major changes have been made or the meaning is unclear. Please check all text carefully throughout.

Print ISSN 1744-6740 | Online ISSN 1755-2710

However, they have severe limitations, even when applied to the fat-tailed distributions for which they were intended. An account of these shortcomings is given in Section 2.

As an example, consider the empirical and fitted distributions shown in Figure 1 on the facing page. The empirical cumulative distribution function (CDF) is shown by the dotted profile, and a fitted lognormal distribution is shown by the solid line. Change OK? Assessing GoF using the AD, KS or Cramér–von Mises (CvM) tests resulted in zero p-values in all cases. In these cases no fit to the data could be found. Although it is easy to isolate regions in which the fit is less good (for example, the part of the profile that represents low-value losses with high cumulative probability), the overall fit is not totally unsatisfactory. Change OK? "a small number of"?

The profiles shown in Figure 1 on the facing page are quite typical of those encountered in operational risk, and the proposed new GoF tests are specifically geared to profiles of this type. The new tests are intended to rectify the shortcomings of established GoF tests listed in Section 2. The requirements of the new tests are Changes to sentence OK?

(1) to be independent of the number of points in the empirical data,

(2) to reflect intuitive qualitative views of goodness-of-fit based on comparing fitted and empirical CDFs,

(3) to discriminate effectively between the common distributions used in modeling operational risk severity,

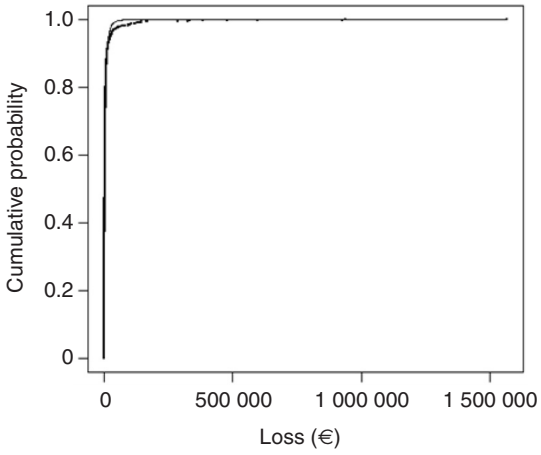
(4) if possible, to be deterministic.

The proposed tests in this paper also have a geometric interpretation that is intuitive, and is directly related to the test formulations. A transformation of the geometry makes the tests equally easy to interpret, and makes calculations much easier.

1.1 Structure of the paper

This introductory section has given a brief overview of the motivation for this paper. Minor changes to this paragraph OK? No mention of appendixes – OK? Section 2 gives an account of the shortcomings of "traditional" GoF tests in the context of the distributions found in operational risk and reviews alternative GoF tests that are applicable in particular contexts. Geometric aspects of the GoF test proposed in this paper are covered in Section 3, as is the geometric transformation that forms a basis for the method. The topological properties of the transformation are described in Section 4. These are used to derive an expression for the p-value of one of the new tests. Details of the new tests are given in Section 5. Section 6 assesses the significance of the current work, and gives some suggestions for future extensions.

FIGURE 1 Example GoF test with zero p-values.



2 SHORTCOMINGS OF TRADITIONAL TESTS FOR DISTRIBUTIONS USED IN OPERATIONAL RISK

The KS, AD and CvM tests are the GoF tests most commonly used for testing goodness-of-fit in the context of operational risk. Problems in using them are described in this section. These methods are a subset of the more general class of empirical distribution functions (EDFs), which generally provide more powerful tests than the "default" χ^2 test, (ie, they are less likely to reject a null hypothesis when the null hypothesis is true). Good accounts of the KS, AD and CvM tests within the EDF context are given in Stephens (1972, 1974).

Cramér (1928) and von Mises (1928) proposed their GoF tests nearly a century ago. They compare an empirical CDF $F(x)$ with n observations, and a proposed CDF $F^*(x)$, in conjunction with a weight function $w(x)$, using the statistic

All mathematical notation has been converted to LaTeX or been rekeyed. Please check all symbols and equations have been typeset correctly throughout your paper and mark any for amendment.

$$W^2 = n \int_{-\infty}^{\infty} (F(x) - F^*(x))^2 w(x)\,\mathrm{d}x, \quad \text{where } w(x) = 1.$$

This expression illustrates the problem of comparing vertical distances, $F(x) - F^*(x)$, for the typical distributions encountered in operational risk data. The problem arises because these distributions often have a large number of low-value losses. The empirical CDF for such losses is nearly vertical, so the vertical distances to the corresponding empirical points are large. Once many such distances are squared and summed, the result is likely to exceed a critical value.

Changes to sentence OK?

The same problem persists with the AD test, for which $w(x) = x^{-1}(1-x)^{-1}$. The original formulation for the AD test may be found in Anderson and Darling (1952). Words added – OK? This weight function has the effect of emphasizing points in the distribution tail, but this emphasis is often outweighed by the large number of points in the distribution body. The quadratic AD test described by Chernobai *et al* (2005) and implemented in the R package `truncgof`, is an improvement, but is still affected adversely by an excess of points.

The KS test (Massey 1951) also uses the vertical distance from the empirical to the proposed distributions with the statistic $\sup_x(F(x) - F^*(x))$. The KS test tends to be more sensitive near the center of the distribution than at the tails, so, although it is frequently used in operational risk calculations, it can fail to adequately reflect tail dependency. Changes to sentence OK?

The CvM, AD and KS tests all have the advantage that they are distribution free, since any trial distribution is discretized. The tests proposed in this paper retain this property through a distribution transformation.

The CvM, AD and KS tests can also exhibit one of two undesirable tendencies. Changes to sentence OK? Either they reject the null hypothesis for all distributions other than lognormal (for example, the `truncgof` package for the R application) or they accept null hypothesis for all distributions or none.

All three tests are subject to the problem of rejecting the null hypothesis if the data sets used are large. The problem is discussed by Lin *et al* (2013). Their argument is summarized in the following statements.

(a) Having stated a null hypothesis with respect to a population parameter b, any deviation from b, however small, has a significant effect on a test statistic.

(b) Under the null hypothesis, the limiting p-value as the sample size tends to infinity is the probability that the absolute difference, $|\beta - \hat{\beta}|$, where $\hat{\beta}$ is an estimator of β, is arbitrarily small.

(c) If the population parameter is exactly equal to the value of the population parameter set by the null hypothesis with an infinite number of decimal places, the p-value will approach 1. Otherwise it will approach 0. In other words, the estimator $\hat{\beta}$ has all its mass on the population parameter.

(d) In reality, the measured population parameter b is typically not a round figure, "an integer"? so a large sample will yield a p-value that is near to 0.

The problem highlighted by Lin *et al* is particularly acute for much of the data that we use. Population sizes are often greater than 1000, and can exceed 500 000.

Use of the χ^2 test tends to be avoided in the context of distribution fitting for operational risk because it does not weight the distribution tail and therefore does not find favor when there is so much stress on the importance of the tail.

The discussion by Lemeshko *et al* (2007) is interesting, in that they recommend transforming losses $x_1, x_2, \ldots, x_n$ using $x_i \to F(x_i)$, where $F(\cdot)$ is the CDF of the fitted curve. They then test GoF using a KS statistic $D = \sup_{0 \leqslant u \leqslant 1}(\sqrt{n}|F(u) - u|)$. The value of D can then be compared with a critical value. This transformation is essentially the transformation proposed in this paper to simplify domain geometry. However, the factor $\sqrt{n}$ ensures that the calculated values of D have very little chance of being less than a critical value for large n. Conover (1999) gives $1.224/\sqrt{n}$ as an approximation for the 5% one-tail critical KS value for $n > 35$. (The corresponding 5% two-tail value is $1.358/\sqrt{n}$.) For large n these critical values are very small (typically between 0.001 and 0.1), but the calculated values of D are large (typically between 1 and 90).

2.1 Examples of the failure of established GoF tests

This section cites typical examples of the failure of established GoF tests (AD, KS and CvM) to differentiate between candidate distributions.

Guégan and Hassani have repeatedly reported that KS p-values are inadequate. In Guégan and Hassani (2014) they calculated p-values in the context of fitting time series models. Very few KS p-values were nonzero. In most cases, only one distribution per model had a nonzero p-value, and many of those were less than 0.05. In three cases, according to the KS tests, no distributions were adequate, and Guégan and Hassani comment that they have observed this in previous analyses. One of these is Guégan and Hassani (2012), in which they use a peak-over-threshold method to thicken the right tail of their loss distributions in order to achieve an acceptable goodness-of-fit for GPD fits. Another is Guégan *et al* (2011), in which they attempt to fit lognormal, exponential, Weibull, Gumbel and Fréchet distributions. It proved impossible to distinguish between the fits for any of them, since all p-values were zero.

Changes to sentence OK?

Changes to sentence OK?

The same problem is reiterated in Lehérissé and Renaudin (2013). They use AD, KS and CvM tests in the context of quantile distance estimations. These tests provide inconsistent results for the same data set. In particular, there are examples where all the proposed models are rejected by all the GoF tests. Lehérissé and Renaudin comment that the GoF tests they used do not capture tail dependence.

Gourier *et al* (2009) also report that their AD test did not capture the heavy-tailedness of their data. In many cases, all p-values were zero. They comment that statistical tools to analyze rare and extreme events were lacking, and suggest that alternative distributions are required. They also note that, by construction, the AD test yields infinite values when the theoretical distribution has a finite endpoint that is below that of the empirical distribution, which is a very unsatisfactory situation.

2.2 Review of dedicated GoF tests

This section gives a brief description of some dedicated GoF measures that are applicable in certain circumstances only. In general they cannot be used successfully for operational risk purposes, and reasons for this are given.

Goldmann *et al* (2015) suggest GoF measures that are applicable to right-censored data. Censoring in this way is an immediate problem in the current context, because the data sets we use always contain large outliers that cannot be ignored. Indeed, the GoF tests proposed here are specifically geared to account for extremely large data. Nonetheless, Goldmann *et al* report some success when fitting lognormal, Weibull and Gamma distributions. Their process is to form the order statistics for a random sample of the losses, transform them using a random sample of order statistics drawn from a $U(0, 1)$ distribution, thereby generating values that have an approximate normal distribution, compute a standard error for each transformed sample and then use the CvM and AD tests to assess the GoF. It is hard to see how their method would work with our data, as the largest population size they use is 100.

Should Table 1 be cited here for comparison of data populations in this sentence

The transformation $x_i \to F(x_i)$, where $F(\cdot)$ is the CDF of a fitted curve, is taken further by Quesenberry (1986), who discusses the conditional probability integral transform (CPIT) class of transforms. These transformations also map losses $\{x_i > 0\}$ to the interval $(0, 1)$, but are much more complex than the transformation considered in this paper and do not have any immediate geometric appeal. Quesenberry attempts to fit normal and exponential distributions, and there are no results for long- or fat-tailed distributions. He also uses relatively small populations, and assesses GoF using the CvM and AD tests. It is therefore doubtful that CPIT would be a useful technique for operational risk. In addition, he cites several other cases of mapping to $(0, 1)$, all aimed at particular distributions.

Doray and Huard (2001) propose a novel GoF test which has similarities to the tests proposed here. Their test is applicable to a different context: frequency modeling. They consider the Poisson, negative binomial and binomial distributions. The first two are often used to model frequency in operational risk. Null hypotheses based on them are nearly always rejected, due to very high empirical frequencies, expressed as losses per year, with data spanning insufficient years. The first similarity to our tests is that they search for outliers, and reject the null hypothesis if a sufficient number of outliers are found. The second is that they use a distance estimator; Doray and Huard show that a quadratic distance between data and a target Poisson distribution has a χ^2 distribution. Population sizes were again 100 or less: too low for our data. They considered frequencies between 1 and 10, which would also be inappropriate for our data, where frequencies range from 50 to 20000 (per year).

Changes to this sentence and the next OK?

Rizzo (2009) has developed a GoF test specifically for Pareto distributions, using a statistic based on the expected value of $\|X_i - X\|^b$, where the X_i are losses, X is an

ordinate on a Pareto CDF, b is a stability index that ensures that X^b has finite variance, and $\|\cdot\|$ is the Euclidean norm. As with previous examples, the populations used for testing are small. Therefore, even applied to a Pareto distribution, this method would be hard to apply to our data.

Goegebeur and Giullou (2010) developed a GoF test for Weibull and generalized Pareto distributions (GPDs) by extending the concept of a Q–Q plot. They describe two test statistics. Both contain ratios of terms resembling the Weibull quantile function. They derive very simple asymptotic normal approximations for these statistics. Unfortunately, they only test using simulated data, although they use a satisfactory sample size of 5000. Changes to sentence OK?

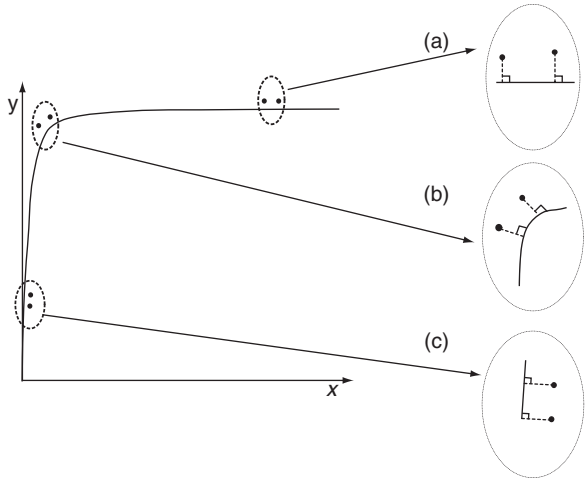
Lastly, we mention some GoF studies on extreme value distributions. Stephens (1977) modifies existing EDF statistics, and applies his tests to small samples only. Similarly, the study by Fard and Holmquist (2013) uses EDF statistics and sample sizes of 20 or 30. It does, however, make use of order statistics, which is of potential use in our context because losses are ordered in the empirical CDF. Kinnison (1989) bypasses the EDF set by using the correlation coefficient as a GoF statistic. He simulates data for a target distribution, and calculates the product moment correlation coefficient for the simulated data and the empirical data. That procedure is repeated many times in order to simulate a distribution for the test statistic. This is an interesting idea that might be usefully applied to other distributions. However, the correlation coefficient is also problematic in that if the sample size is large, the null hypothesis is more likely to be rejected. Changes to sentence OK?

3 TRANSFORMED-NORMAL (TN) TESTS

We consider an alternative suite of GoF tests based on the perpendicular distance of points in an empirical CDF from a fitted distribution. Their purpose is to satisfy the requirements set out in Section 1 in a more satisfactory way than "traditional" GoF tests. Three variants of this test will be discussed. The first is based on an enclosed area, and will be referred to as the TN-A test. The second employs a bootstrap technique for the CDF of the fitted distribution, and will be referred to as the TN-B test. The third uses sampling, and will be referred to as the TN-S test. Their formulation is based on two principal innovations, which will be described in Sections 3.2 and 3.3. Word added – OK?

Section 3.1 describes the geometry of the region in two-dimensional space surrounding empirical losses and the CDF of the fitted distribution. Subsequent sections describe how to transform that region into a unified region that is independent of the fitted distribution, and in which calculations based on geometry are much simpler. Change OK?

FIGURE 2 Normals to the fitted CDF curve.



(a) High-value losses. (b) Losses at the fitted CDF "angle". (c) Low-value losses.

3.1 The geometric basis of the TN tests

Consider the empirical distribution for loss data, comprising a list of n losses initially in no particular order. Then order the losses in increasing order of size.

DEFINITION 3.1 (Empirical cumulative loss distribution) Denote the n ordered losses by x_i $(i = 1, \ldots, n)$. These can be considered a single observation of n independent and identically distributed (iid) random variables $L_1, L_2, \ldots, L_n$. To each ordered loss x_i, assign a probability $y_i = (i - 0.5)/n$. The set $\Delta = \{x_i, y_i\}$ then defines the empirical cumulative loss distribution (referred to in this paper as the empirical CDF).

One or more candidate distributions (typically lognormal, Burr, Weibull, etc) are fitted to this empirical distribution, and the task is then to decide which, if any, is a "best" fit. Denote a fitted curve (it need not be an optimal fit) by ϕ. For every point (x_i, y_i) in the empirical distribution Δ, there exists a unique normal to ϕ through that point. These normals form the basis of the alternative GoF tests. Figure 2 illustrates three cases, typical for the empirical distributions encountered in operational risk. For low-value losses the normals are approximately horizontal. For high-value losses they are approximately vertical. In between (the "angle" of the distribution), there is a transition from horizontal to vertical, which can be rapid. Change OK?

3.2 A GoF statistic based on normals to the fitted CDF curve

The first innovation for the methods proposed in this paper is that the goodness-of-fit is measured as a geometrically optimal quantity, namely the shortest distance of each point to a curve. This eliminates the errors that result from measuring (either directly or implicitly) the vertical or horizontal distances from a point to a curve, which are hugely exaggerated in the case of extreme-valued distributions. As a first step, the following statistic, S_1, is proposed as a GoF measure:

$$S_1 = \sum_{i=1}^{n} |d_i|, \tag{3.1}$$

where n is the number of empirical losses and d_i is the Euclidean perpendicular distance from a point (x_i, y_i) to the curve. This quantity is calculable, but finding a measure of the significance of the result is harder. This task is considered later in the paper. For now, we concentrate on a simplification that leads to a simpler statistic, for which the significance is much easier to calculate.

3.2.1 Comments on the geometry of the empirical CDF

In practice, the construction of the empirical CDF ensures that the points comprising it do not appear as a "random" scatter on either side of the fitted CDF. The general pattern is one of groups of points all on one side or the other of the fitted CDF. The empirical CDF crosses the fitted CDF infrequently.

This property is important because it provides a degree of stability for the calculations that follow. The distances d_i in (3.1) do not vary rapidly between positive and negative, and in most cases do not vary much from one calculation to the next. The profile of the empirical CDF often appears quite smooth, which encourages accuracy.

Word added – OK?

3.3 A GoF statistic based on transformed geometry

Instead of dealing directly with the statistic S_1, the geometry of the problem can be simplified greatly by transforming the empirical losses and the fitted curve to a more convenient domain. The domain of the geometry described above, incorporating the empirical CDF and the curve fitted to the points within it, will be referred to as the loss space.

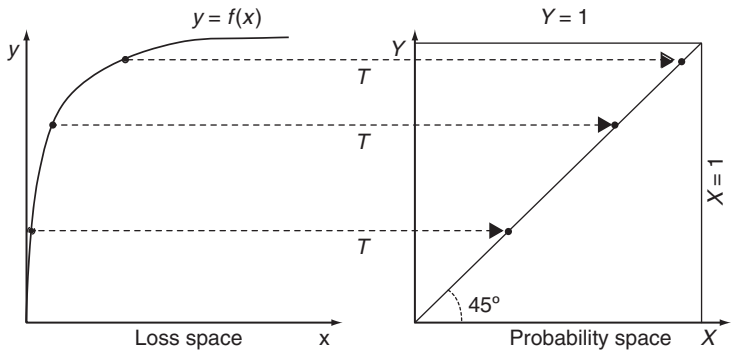
Change OK?

DEFINITION 3.2 (Loss space) The loss space is the region $\Lambda = \{(x, y)\colon x > 0;\ 0 < y < 1\}$. Given any point (x_i, y_i) in loss space, where $x_i > 0$ and $0 < y_i < 1$ for all $i = 1, \ldots, n$, with $x_{i+1} > x_i$ for all $i = 1, \ldots, n-1$, consider the transformation T, defined by

$$T\colon \{\mathrm{Re}^{+} \otimes (0,1)\} \to (0,1)^2, \qquad T(x, y) \to (F(x), y), \tag{3.2}$$

Change to Re^{+} OK according to style or do you mean $\mathbb{R}^{+}$ (see also query in Appendix A)?

FIGURE 3 Loss space mapped to probability space.



where $F(\cdot)$ is the CDF of the fitted curve.

Under this transformation, the probability measure in loss space is unchanged and the loss measure transforms to a uniform distribution on $(0, 1)$. For convenience, we will refer to the transformed region as probability space.

DEFINITION 3.3 (Probability space) Probability space is the image of loss space under the transformation T defined in (3.2), namely the region $\{(X, Y)\colon 0 < X < 1;\ 0 < Y < 1\}$.

In order to distinguish points in loss space and probability space, lowercase symbols will be used for elements in loss space and uppercase symbols will be used for elements in probability space. Hence, an alternative way to express the transformation in (3.2) is to write

$$\left.\begin{aligned} &X = F(x), \qquad Y = y, \\ &(x, y) \in \text{loss space}, \\ &(X, Y) \in \text{probability space}. \end{aligned}\right\} \tag{3.3}$$

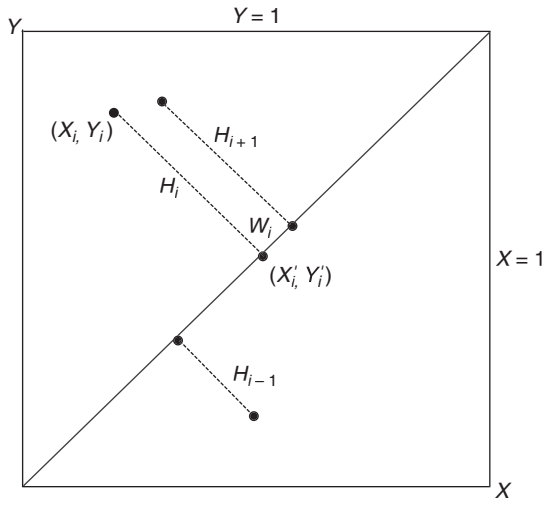
PROPOSITION 3.4 *The fitted CDF transforms to the* 45° *line in probability space.*

Under T, $X = F(x)$. *Therefore,* $X = y$, *since, by the definition of a CDF,* $F(x) = y$. *Hence,* $X = Y$.

Furthermore, this result applies for all CDF functions F. The transformation is shown in Figure 3.

Points not on the empirical CDF in loss space are not, with possibly a few exceptions, on the fitted CDF. They map to points (X, Y) in probability space that are not on the 45° line.

FIGURE 4 Normals in probability space.



The transformation from loss space to probability space is the second innovation in this paper, for two reasons: it provides a simple geometrical basis for a new GoF test, and the geometry of probability space is distribution independent, so a single treatment covers all distributions.

Changes to sentence OK?

3.4 The geometry of the transformed normal GoF tests

The intuition behind the TN GoF tests is simple. Points in probability space that are "near" to the 45° line represent a "good" fit, and points in probability space that are "far" from the 45° line represent a "poor" fit. Furthermore, if each adjacent pair of transformed points in probability space is joined by a straight line, the union of all such straight line segments can be used to define an area. A small area indicates a good fit and a large area indicates a poor fit.

Scare quotes necessary here and later? Please mark any that may be deleted.

Consider a sequence of n losses (X_i, Y_i) in probability space, originating from a corresponding sequence of n losses (x_i, y_i) in loss space (where the x_i are in increasing order of size and the y_i correspond to the ordered x_i) by the transformation of (3.2) and (3.3). A selection is shown in Figure 4, which also shows the transformed fitted CDF (the 45° line, $Y = X$). Three adjacent typical normals are shown, one associated with a loss below the line $Y = X$ and the other two associated with losses above that line. Let (X_i', Y_i') be the point of intersection of the normal from a point (X_i, Y_i) to the line $Y = X$. Clearly, $X_i' = Y_i'$ since (X_i', Y_i') is on the line $Y = X$.

The Euclidean distance from (X_i, Y_i) to (X_i', Y_i') is H_i, and, writing Y_i' in place of X_i', the Euclidean distance from (Y_i', Y_i') to (Y_{i+1}', Y_{i+1}') is W_i.

It is easy to show that

$$Y_i' = \frac{X_i + Y_i}{2}, \quad i = 1, \ldots, n, \tag{3.4}$$

and that the distance from (X_i, Y_i) to (X_i', Y_i') is

$$H_i = \frac{|X_i - Y_i|}{2}, \quad i = 1, \ldots, n. \tag{3.5}$$

Furthermore, the Euclidean distance from any point (Y_i', Y_i') to the adjacent point (Y_{i+1}', Y_{i+1}') on the 45° line is

$$W_i = \sqrt{2}(Y_{i+1}' - Y_i'), \quad i = 1, \ldots, n. \tag{3.6}$$

The distances H_i are the absolute normal deviations from the 45° line. The maximum possible value of any H_i is $1/\sqrt{2}$, which is the distance from the center of the domain $(0, 1)^2$ to a vertex $(0, 1)$ or $(1, 0)$.

3.5 The transformed normal area

In loss space, a sequence of straight line segments can be defined by joining all pairs of adjacent points $\{(x_i, y_i), (x_{i+1}, y_{i+1})\}$. This sequence, with linear normal boundaries defined by the maximum and minimum losses, and with the fitted CDF curve, defines a region that has a well-defined area. The smaller the area, the better the fit. The same intuition applies in probability space: the smaller the transformed area, the better the fit. This area transforms under T (defined in (3.2) and (3.3)) to a well-defined area in probability space. The enclosed area, EA$(\boldsymbol{X}, F)$, is the sum of the area enclosed by joining points above the 45° line and the area enclosed by joining points below the 45° line.

Bold font necessary here and in similar expressions below? Journal style is to use bold face for vector quantities.

Thus, in probability space, the boundaries of this area are the 45° line, the sequence of straight line segments joining adjacent points $\{(X_i, Y_i), (X_{i+1}, Y_{i+1})\}$, and perpendiculars to the 45° line corresponding to the minimum and maximum (transformed) losses. Figure 5 on the facing page shows the areas in L- and probability space.

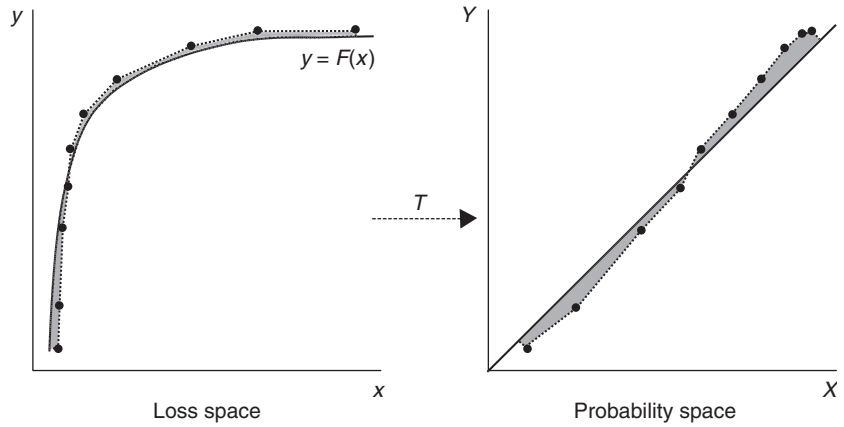
Does L denote "loss space" here and should be defined at first mention according to journal style, or do you mean Lebesgue L^P space?

The area enclosed between the n transformed losses and the 45° line in probability space forms the basis of the TN-A and TN-B tests, as described in Section 5. The TN-S test uses the heights H_i. The area enclosed is measured by a simple application of the trapezium rule.

Consider the trapezium defined by the points (X_i, Y_i), (X_{i+1}, Y_{i+1}), (Y_{i+1}', Y_{i+1}') and (Y_i', Y_i'). Its area A_i is given by

Words added – OK?

FIGURE 5 Areas in L- and probability space.



$$A_i = W_i \frac{H_i + H_{i+1}}{2}, \quad i = 1, \ldots, n-1. \tag{3.7}$$

Summing all such areas, the total area $A[n]$ enclosed between the line segment pairs $\{(X_i, Y_i), (X_{i+1}, Y_{i+1})\}$, $i = 1, \ldots, n$, and the line $Y = X$ is

$$A[n] = \sum_{i=1}^{n-1} A_i = \sum_{i=1}^{n-1} W_i \frac{H_i + H_{i+1}}{2}. \tag{3.8}$$

Calculating $A[n]$ is easy in the geometry of probability space. In loss space, finding the equivalent enclosed area would necessitate calculating the length of normals to the fitted CDF at each point of the empirical CDF, which is a much longer process. The flow diagram in Figure 6 on the next page gives a step-by-step process for deriving $A[n]$.

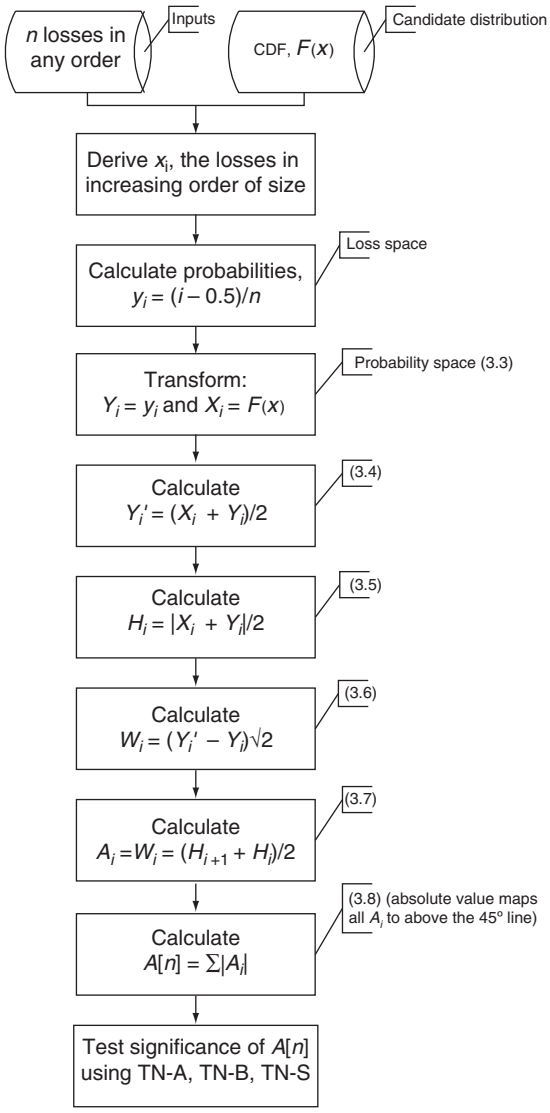
In practice, it is easier to visualize and to analyze if all quantities H_i are mapped to the upper region defined by $Y > X$. This avoids dealing with cases where two adjacent H_i occur on opposite sides of the diagonal line $Y = X$. The area of the upper region defined by $Y = X$ is, trivially, 0.5, and is referred to as the transformed area (TA).

"whether"? Otherwise what do you mean by "it"?

3.5.1 Difficulties with calculations in loss space

The concept of the enclosed area is the same in both loss space and probability space. Calculating it is, however, harder in loss space, which is why it is done in probability space. In order to do the calculation in loss space, normals to the fitted CDF through

FIGURE 6 Calculation of enclosed area, $A[n]$.



each point (x_i, y_i) must be calculated. The standard way of doing this is to calculate the gradient of tangents to the CDF. This is an extensive calculation if the population size is large. Also, calculating the enclosed area in loss space is more awkward, and would probably have to be done by using straight line approximations to regions

FIGURE 7 Example of a good fit.

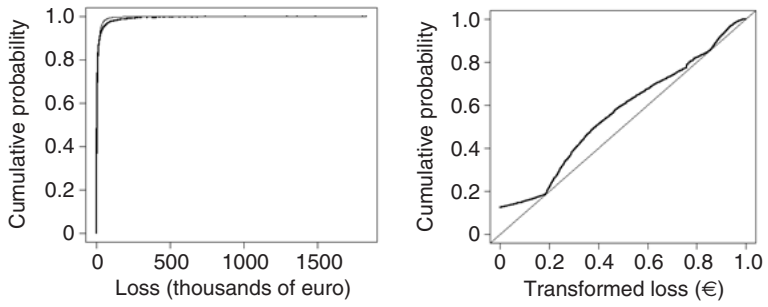
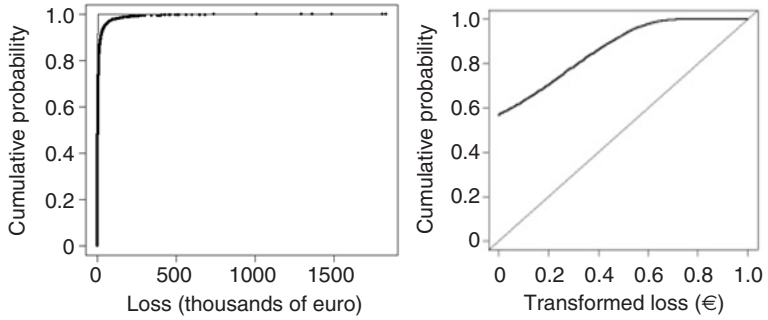


FIGURE 8 Example of a bad fit.



delimited by the normal. In probability space, the gradients of normals to the 45° line are always -1 and the regions delimited by those normals are always trapeziums.

3.6 Examples of the transformed normal and the transformed normal area

This section contains some examples of loss distributions that have been transformed using (3.2). They show both loss space and probability space (Figure 7, Figure 8 and Figure 9 on the next page) and illustrate the enclosed areas, $\overline{\mathrm{EA}}(\boldsymbol{X}, F)$, in both spaces. The main feature to note is that, in probability space, if the EA is large, there is evidence for rejecting the null hypothesis, and if it is small, there is no evidence for rejecting the null hypothesis.

Are vector graphic files available for these figures?

FIGURE 9 Example of a borderline fit.

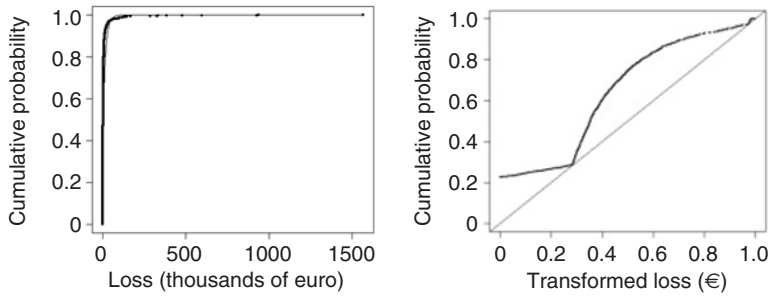


TABLE 1 Population sizes of data sets in this analysis.

Data set	Size	Comments
B1	142	Contains a comprehensive range of low- and high-value losses
B2	675 418	Distribution mainly comprises low-value losses
B3	454 570	Distribution mainly comprises low-value losses
B4	5 192	Contains a comprehensive range of low and high losses
B5	1 958	Contains few very high-value losses
B6	1 521	Contains few very high-value losses
B7	52 918	Mainly low-value losses
B8	28 322	Mainly low-value losses but with some significant tail losses

3.6.1 Data used

The transformation from loss space to probability space was tested on eight representative data sets covering a range of large and small losses. The population sizes range from about 150 to more than 600 000. In every case, the data used is not filtered and no thresholds have been applied. All losses are strictly positive. The data sets are labelled B1, B2, …, B8. Collectively, they are referred to as the B# data. The actual population sizes are given in Table 1.

Journal style is to place tables and figures where they are first mentioned in text, particularly if they appear in appendixes with no supporting text. Minor changes to comments for consistency – OK?

In each case, the value of the transformed area, TA, is given. These are consistent with intuitive views of the enclosed areas: low TA values represent "good" fits, and high TA values represent "bad" fits.

3.6.2 Examples

In the examples that follow, and in subsequent parts of the paper, the following abbreviations for distribution names are used.

- LN: lognormal.
- WB: Weibull.
- GPD: generalized Pareto distribution.
- LNMix: mixture of two lognormal distributions.
- LL: log logistic.

EXAMPLE 3.5 (Data set B4) This appears to be a good fit to a lognormal distribution. Population size = 5192. TA = 0.0549 (see Figure 7 on page 15).

EXAMPLE 3.6 (Data set B4) This appears to be a bad fit to a GPD. The fitted distribution is too severe for small losses that represent a high cumulative probability. Population size = 5192. TA = 0.2866 (see Figure 8 on page 15).

EXAMPLE 3.7 (Data set B6) This appears to be a borderline fit to a Gamma distribution. The fit looks reasonable in loss space, but not so in probability space. Changes to sentence OK? Population size = 1521. TA = 0.128 (see Figure 9 on the facing page).

3.7 Independence of the enclosed area of the number of losses

The principal aim of the current analysis is to formulate a GoF test that does not depend on the number of losses analyzed, n. In this section we show that the calculated area $A[n]$ ((3.8)) satisfies this aim, provided that n is large enough. We first give a theoretical Changes to sentence OK? basis for this claim of independence, and then show a practical example.

PROPOSITION 3.8 *$A[n]$ is independent of n for sufficiently large n.*

The proof of this proposition is given in Appendix C.

As an illustration, Table 2 on the next page shows the result of fitting a lognormal distribution to random samples of increasing size, generated from a lognormal$(8, 2)$ distribution. Each random sample has added random noise, in order to better simulate a genuine empirical distribution. The results are consistent within the limits of stochastic variation. When $n < 50$, $A[n] \sim 0.05$. At this level the fit process is less reliable, and we would consider augmenting the data with data from other sources.

TABLE 2 Variation of $A[n]$ with n.

Sample size, n	$A[n]$
50	0.0359
100	0.0281
1 000	0.0327
10 000	0.0337
100 000	0.0354
1 000 000	0.0351

4 TOPOLOGY OF LOSS SPACE AND PROBABILITY SPACE

Although $A[n]$ (in (3.8)) can be calculated easily, the significance of the value obtained is not immediately clear. Topology provides a toolkit for determining how a p-value (or equivalent, such as a critical value) may be derived using $A[n]$. Therefore, in this section we provide a theoretical basis for assessing the significance of the area calculation by considering the topologies of loss space and probability space.

The broadest overview of the argument developed in this section is that a measure of "closeness" in loss space corresponds to a measure of "closeness" in probability space via a continuous mapping. The result is a distribution function for the enclosed area, from which a p-value can be derived.

It is helpful to consider the steps in the argument for deriving the p-value at three levels:

(a) the argument can be presented as a broad overview of the process with few technicalities;

(b) topological concepts can be added;

(c) detailed proofs can be given.

Section 5 contains details of GoF tests based on the geometry of probability space.

4.1 p-value derivation: a broad overview

The steps in the argument are as follows.

(1) The CDF of the fitted curve can be completely covered by ellipses, which are chosen because their properties are particularly suited to the geometry of loss space. This cover defines a region surrounding the CDF, which we refer to as a "band".

(2) T (equation (3.2)) maps the band in loss space to a simpler band in probability space, such that the boundaries of the band in loss space map to lines parallel to the 45° line in probability space.

(3) Derive an expression for $A[n]$, and deduce that the p-value is the width of the band in probability space, measured perpendicular to the 45° line.

4.2 p-value derivation: the topological view

The details of steps 1 and 2 in the previous section are given in Appendix D. The band in loss space is shown in Figure D.2 on page 44, and the way it maps to a corresponding band in probability space is shown in Figure D.5 on page 47. There are three main results of the discussion in Appendix D. Each refers to a corresponding band in probability space, which

Do you mean "Section 3" or "Section 4.1"?

(i) is interpreted as a region such that the fitted CDF is a good fit for points within it,

(ii) resembles a parallelogram so closely that it can be approximated by a parallelogram without significant error for all subsequent calculations, and is termed an "almost-parallelogram" (see Section D.2.1 for a formal definition),

(iii) comprises shapes that resemble ellipses, and are called "almost-ellipses" (see Section D.2.1 for a formal definition),

(iv) is parameterized by a height $2r$, shown in Figure D.5 on page 47.

Step 3 above is discussed in Section 5; in that section, the expression for $A[n]$ is derived, is shown to be independent of n and is expressed in terms of a p-value (see (5.3)).

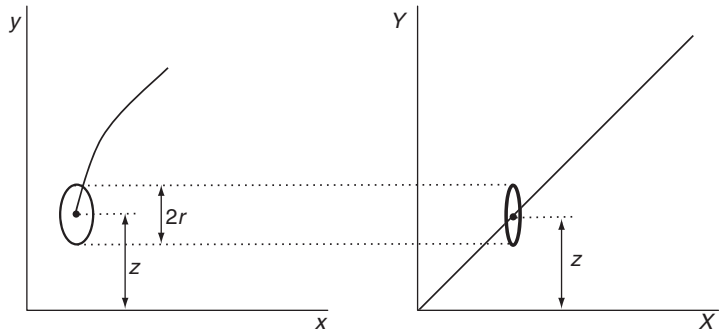
5 GOODNESS-OF-FIT TESTS BASED ON THE TRANSFORMED NORMAL

Step 3

Heading hierarchy OK? Should this be numbered and "The TN-A test" become Section 5.1.1? There are no subitems 3.1 onwards in the list in Section 4.1 or in Appendix D. Please clarify heading hierarchy and lists being referred to throughout this section.

Step 3, the basis of which is discussed in Appendix D, is covered in this section, which contains details of the three proposed GoF tests. Two are based on the EA, defined in (3.7). The third is based on the heights that form the area calculation, defined in (3.5) and illustrated in Figure 4 on page 11.

FIGURE 10 End correction applied to an almost-parallelogram.



5.1 The TN-A test

The TN-A test depends on the area ("A" denotes area) of an acceptable almost-parallelogram in probability space. This test has a very simple intuitive interpretation, which is to compare a measured area, defined ultimately by the data, with a fixed reference area. The goodness-of-fit is determined by the deviation of the measured area from zero. A small deviation indicates a good fit, and a large deviation indicates a poor fit.

Change OK?

Step 3.1

The almost-parallelogram in Figure D.5 on page 47 is shown in a near symmetric state, in which it nearly touches both the lines $Y = 0$ and $Y = 1$. In practice this will rarely be the case. If the open set centered on the maximum loss touches the line $Y = 1$, the open ball centered on the minimum loss may not touch the line $Y = 0$, and vice versa. There is an "end correction", z, shown in Figure 10. The distance z is the vertical distance from the midpoint on a vertical face of an almost-parallelogram to a horizontal boundary of probability space. Figure 10 shows the case when an open set surrounding the minimum loss does not touch the boundary. There is an equivalent case for the maximum loss. In practice z is very small.

"nearly"?

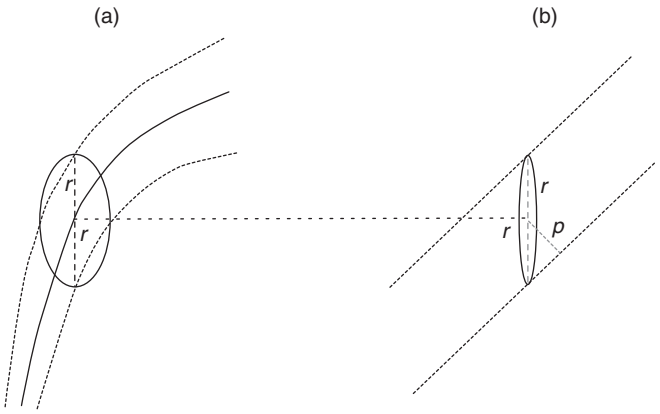
"cannot"?

Changes to sentence OK?

Figure D.5 on page 47 shows an almost-parallelogram with height $2r$. If it is positioned with the end correction shown in Figure C.1 on page 40, so that the almost-ellipse in Figure C.1 forms the lower end of the almost-parallelogram, then the area of the almost-parallelogram, $\mathrm{AP}(r)$, is

$$\mathrm{AP}(r) = 2r(1 - r - z). \tag{5.1}$$

FIGURE 11 Determination of p-value.



(a) Loss space. (b) Probability space.

The result for the area of the almost-parallelogram is the same if the correction z is near the maximum loss.

This result is significant because it does not depend on the number of losses in the calculation.

Step 3.2

Area AP(r) will facilitate calculation of a critical value for the TN-A test. We now interpret the p-value of the test as the dimension of the almost-parallelogram that is perpendicular to the 45° line. This is the length p, as in Figure 11: the distance of transformed points from the 45° line in probability space.

Clearly, $p = r/\sqrt{2}$.

In terms of p, the area in (5.1) is

$$\mathrm{AP}(p) = 2\sqrt{2}p(1 - \sqrt{2}p - z). \tag{5.2}$$

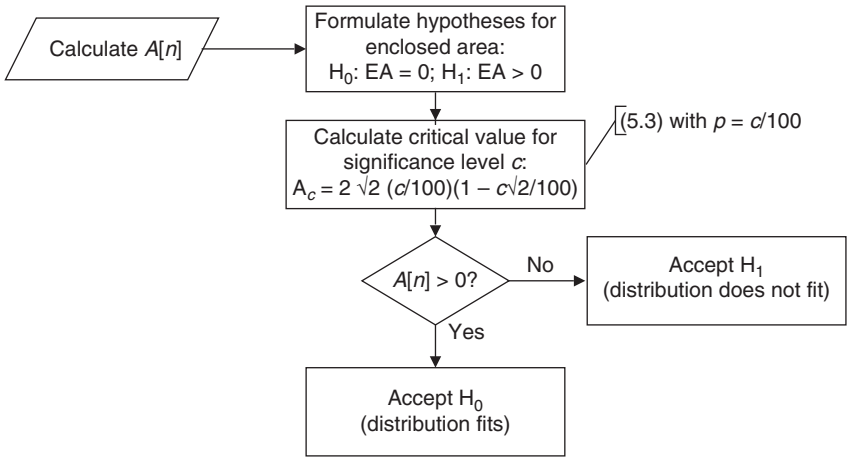
In practice the approximation for small z,

$$\mathrm{AP}(p) = 2\sqrt{2}p(1 - \sqrt{2}p), \tag{5.3}$$

is perfectly adequate.

The TN-A test proceeds as follows. (See also Figure 12 on the next page for a step-by-step guide for using this test.)

FIGURE 12 Application of the TN-A test.



ALGORITHM 5.1 (TN-A)

(1) The test statistic is the enclosed area $\mathrm{EA}(\boldsymbol{X}, F)$ of (3.8).

(2) The null and alternative hypotheses are, respectively,

($\mathrm{H_0}$) $\mathrm{EA}(\boldsymbol{X}, F) = 0$;

($\mathrm{H_1}$) $\mathrm{EA}(\boldsymbol{X}, F) > 0$ (treating the enclosed area as an absolute value in a one-tailed test).

(3) For a significance level c (%), calculate the critical value $\mathrm{AP}(c/100)$ using (5.3).

(4) Calculate the actual enclosed area A' using (3.8).

(5) If $A' < \mathrm{AP}(c/100)$, accept H_0; otherwise reject $\mathrm{H_0}$.

For example, at 5% significance the critical value is $\mathrm{AP}(0.05) \sim 0.1314$, and we accept the null hypothesis at 5% significance if any calculated area is less than this value.

At 5% significance, the following results were obtained using the B# data. They are consistent with an intuitive qualitative view of a "good" fit, and provide a range of GoF results per distribution, so that there is a choice in every case. The "winning" distribution has the lowest TN-A value.

5.2 The TN-B test

The TN-B test attempts to elucidate the distribution of the EA by a bootstrap method ("B" denotes bootstrap). Starting from a fixed set of losses and a fixed set of distributions fitted to those losses, there are no experimental results that could indicate what the distribution of the EA could be. We therefore use a modified bootstrap method, originally developed by Efron and discussed in Efron and Tibshirani (1986). The method used in this analysis generates a range of "feasible curves" to which the data can be fitted. The method proposed here reverses the traditional way in which bootstrap methods are applied, which is to resample data. Here we resample fitted curves, generated in a precise way.

Changes to this sentence and the next OK?

The algorithm for generating feasible curves is as follows. The starting point is the CDF of a distribution for which a fit is sought. Denote this CDF by $F(x, \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a vector of parameters determined by the fitting process.

ALGORITHM 5.2 (TN-B1)

(1) Decide on the number, R, of feasible curves.

(2) For each component ϕ of $\boldsymbol{\theta}$, generate a set of R alternative values of ϕ within the range $(0, 2\phi)$. These values, together with the "base" value $\phi = (\phi R)/R$, are then given by the set

$$S_\phi = \phi\left(0, \frac{2}{R}, \frac{4}{R}, \ldots, \frac{R-2}{R}, \frac{R}{R}, \frac{R+2}{R}, \frac{R+4}{R}, \ldots, \frac{2R}{R}\right).$$

(3) To generate a single resample for component ϕ of $\boldsymbol{\theta}$, draw a random sample of size $1 + R$, with replacement, from S_ϕ. Repeat for all ϕ in $\boldsymbol{\theta}$.

Having generated samples, the bootstrap process proceeds by calculating the EA for each bootstrap sample. The algorithm is as follows.

ALGORITHM 5.3 (TN-B2)

(1) Calculate the "base area" A_{base} using the original parameters $\boldsymbol{\theta}$.

(2) For each resample k (as generated above), calculate the enclosed area EA_k using (3.8).

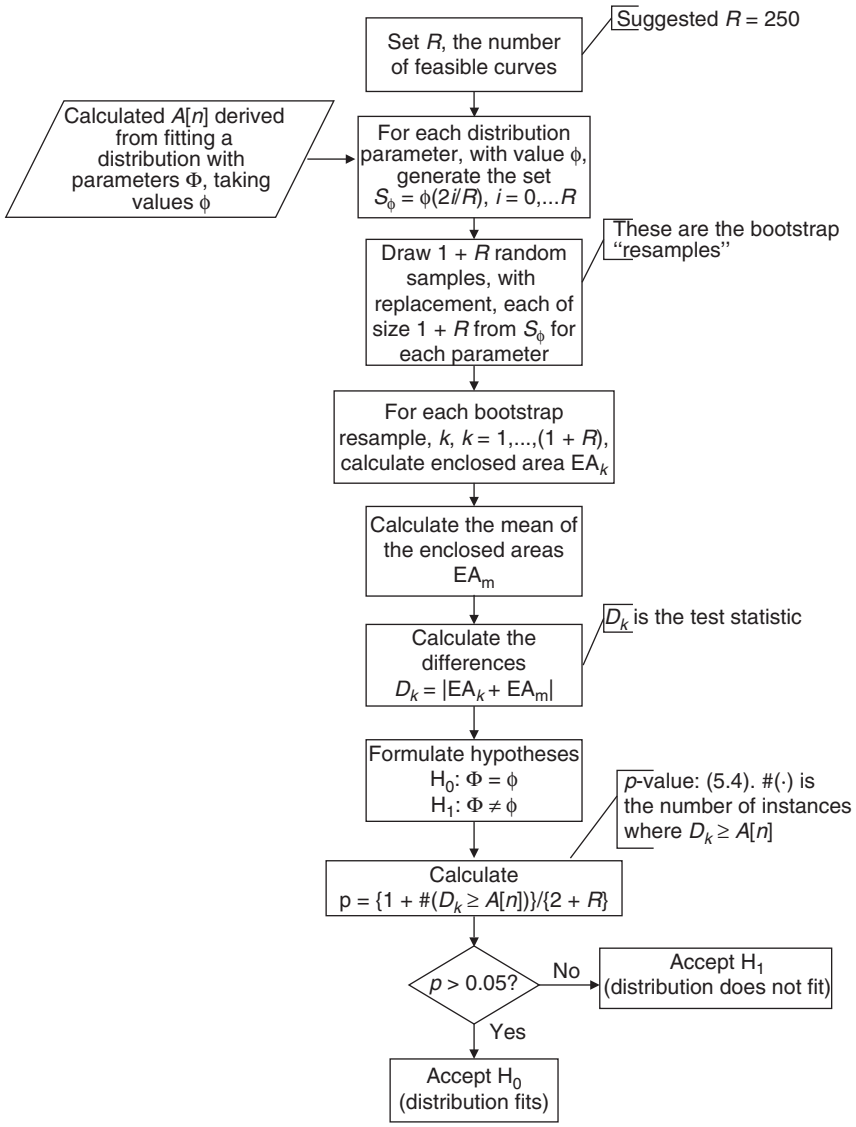
Clarify sentence?

(3) Calculate the mean enclosed area.

(4) Calculate the test statistic, D_k, the absolute difference between each enclosed area EA_k and the mean enclosed area, EA_m.

(5) Calculate the p-value using the method below.

FIGURE 13 Application of the TN-B test.



Davison and Hinkley (1995) describe a method of calculating the p-value of a bootstrap test by counting all trials with a result greater than a "base" result, and expressing the count as a proportion of the sample size. We apply this method, calculating $1 + R$ enclosed areas and counting the number that are greater than the "base

area". The "1+" in the numerator and denominator ensure that the p-value is always nonzero.

With the definition $D_k = |\mathrm{EA}_k - \mathrm{EA}_\mathrm{m}|$,

$$p\text{-value} = \frac{1 + \#\{D_k \geqslant A_\mathrm{base}\}}{1 + (1 + R)}. \tag{5.4}$$

This is an estimate of the probability that an observed area exceeds the base area, given that the null hypothesis is true, ie, $p_\mathrm{obs} = \Pr(A_k > A_\mathrm{base} \mid \mathrm{H_0})$. Figure 13 on the facing page gives a step-by-step guide for applying the combination of algorithms TN-B1 and TN-B2.

Changes to sentence OK?

Several by-products are available from the TN-B process. The first is the variance of the p-values. Davison and Hinkley show that the variance of the p-value can be calculated as follows:

$$\begin{aligned}\mathrm{var}(p\text{-value}) &= \mathrm{var}\left(\frac{1 + \#\{D_k \geqslant A_\mathrm{base}\}}{1 + (1 + R)}\right)\\ &= \frac{1}{(R+1)}(R+1)\,p_\mathrm{obs}(1 - p_\mathrm{obs})\\ &= \frac{p_\mathrm{obs}(1 - p_\mathrm{obs})}{R+1}.\end{aligned}$$

So if $p_\mathrm{obs} = 0.05$, we require $R \geqslant 1900$ to get a 10% relative error. In practice, a much lower figure, $R = 250$, appears to be satisfactory.

The second by-product is an estimate of a c% confidence interval for mean EA. This can be done by calculating $c/2$% upper and lower limits of the empirical distribution. Using a 5% two-tailed significance level, this confidence interval tends to be wide.

Lastly, a histogram of EA_k shows that the distribution of areas is usually non-Gaussian. A few examples (for data B1) are shown in Figure 14 on the next page.

Note that, since the TN-B test produces a p-value rather than a standard error, the null hypothesis will be rejected if the p-value is less than the required critical value.

The results from the B# test data are shown in Table 4 on page 27.

Referred to before Table 3 – OK?

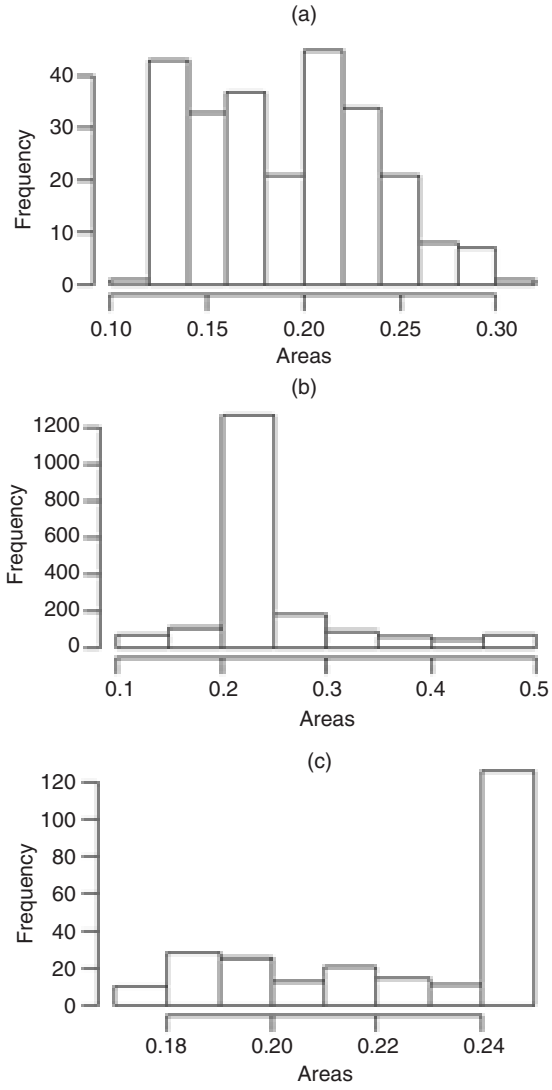
Comparing the results of the TN-A and TN-B tests in Table 3 on page 27 and Table 4 on page 27, respectively, it is clear that they agree very closely. However, the TN-B test does contain subjective elements, eg, the number of feasible curves and the range used to select their parameters, that could affect the result. In particular, varying the fitted parameters between 0 and 100% of their fitted values worked well for all distributions apart from the GPD. For the GPD, variation up to 400% was necessary in order to not reject the null hypothesis in all cases.

Changes to this sentence and the next OK?

Words added – OK?

As a potential extension of the TN-B test, a sample size less than the population size could be used. This is discussed by Bickel and Freedman (1981). The sampling distribution used, as determined by the central limit theorem, is subject to problems

FIGURE 14 Non-Gaussian area distributions.



(a) Weibull fit. (b) Lognormal fit. (c) GPD fit.

of determining a sample size that does not result in a rejection of the null hypothesis for all distributions due to a multiplicative factor $\sqrt{\text{sample size}}$ in the standard error. Therefore, we did not pursue this strategy.

TABLE 3 Results for TN-A test.

Risk category	LN	WB	LL	Gamma	Burr	GPD	LNMix
B1	0.0403	0.0615	0.039	0.0788	0.0474	**0.2468**	0.0255
B2	0.0492	0.054	0.0265	**0.1535**	0.0815	**0.2095**	0.0306
B3	0.0629	0.0881	0.053	**0.1743**	0.076	0.0544	0.0557
B4	0.0549	0.0755	**0.142**	0.1161	0.0696	**0.2866**	0.0184
B5	0.0537	0.0637	**0.1476**	0.1087	0.0744	**0.21**	0.0292
B6	0.0605	0.0648	**0.1437**	*0.1279*	0.0812	**0.2306**	0.05
B7	0.03	0.0517	0.0254	0.0505	0.0515	0.0631	0.0302
B8	0.028	0.0453	**0.1816**	0.0524	0.0351	0.0339	0.0144

Normal text means accept H_0. Italic text means narrowly accept H_0. Bold text means reject H_0.

TABLE 4 p-values for the TN-B test.

Risk category	LN	WB	LL	Gamma	Burr	GPD	LNMix
B1	0.13	0.11	0.25	0.17	0.22	**0.004**	0.41
B2	0.46	0.37	0.46	**0.024**	0.33	**0.004**	0.32
B3	0.42	0.2	0.23	**0.018**	0.31	0.13	0.29
B4	0.25	*0.075*	**0.004**	*0.08*	0.25	**0.004**	0.34
B5	0.31	0.16	**0.004**	0.14	0.3	**0.004**	0.32
B6	0.28	0.17	**0.004**	0.1	0.25	**0.004**	0.3
B7	0.24	0.24	0.4	0.23	0.37	0.18	0.33
B8	0.2	0.15	**0.004**	0.25	0.27	0.24	0.29

Normal text means accept H_0. Italic text means narrowly accept H_0. Bold text means reject H_0.

5.2.1 An alternative bootstrap distribution

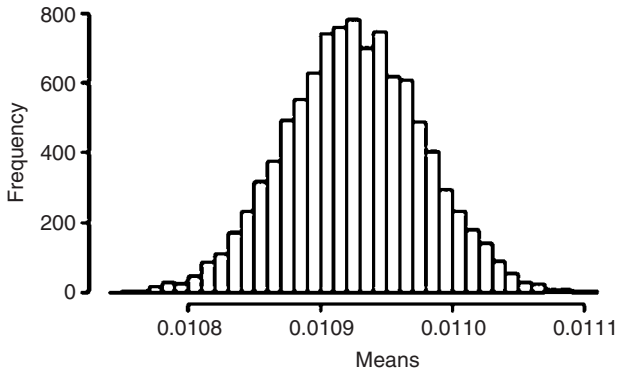
We have also examined an alternative bootstrap strategy whereby, instead of generating "feasible" curves from the distribution for which a fit is sought, a lognormal distribution is used. If the empirical losses are x_i ($i = 1, \ldots, n$), we first calculate the statistics $m = \text{mean}(\log(x_i))$ and $s = \text{SD}(\log(x_i))$. We then use a lognormal(m, s^2) distribution to generate feasible curves by varying m and s using Algorithm 5.2. The results are similar to those in Table 4.

Change OK?

5.3 The TN-S test

The TN-S test uses sampling ("S" denotes sample) to generate a Gaussian distribution, from which an easy significance test follows immediately. This test is a variant of the TN-A GoF test in which the same principles are used but without an area calculation.

FIGURE 15 Typical sampling distribution.



Means of samples of size 28 322.

Instead, sampling provides a distribution which is approximately normal if a large number of samples are taken. Goodness-of-fit can then be assessed from the sampling distribution.

This analysis starts from n losses x_i $(i = 1, \ldots, n)$ with corresponding probabilities y_i $(i = 1, \ldots, n)$ in loss space, transformed under T to probability space using the fitted CDF $F(\cdot)$. The transformed points are X_i and Y_i (where $Y_i = y_i$ and $X_i = F(x_i)$), as described in Section 3.4.

From the set $\{X_i, Y_i;\ i = 1, \ldots, n\}$, the lengths of perpendiculars, H_i (3.5), from each $\{X_i, Y_i\}$ to the 45° line in probability space can be derived.

We regard the losses x_i as observations of iid random variables. The result of transforming these losses to probability space is also a set of observations of iid random variables. A further transformation, the calculation of the H_i, generates a further set of observations of iid random variables. Once the H_i are used to calculate an area, the independence property no longer applies as the H_i must then be ordered. The iid property (or lack of it) is of prime importance in this context, and we refer to the discussion of it in the context of the central limit theorem given by Hogg *et al* (2013).

The sampling test then proceeds by drawing samples of size N $(N \leqslant n)$ from the parent population of H_i (before the H_i are ordered) and calculating the mean of each sample. The central limit theorem then provides the distribution of means of samples of size N:

$$\bar{X} \sim \text{normal}\left(\mu, \frac{\sigma^2}{N}\right),$$

where μ and σ^2 are the mean and variance, respectively, of the parent population of H_i. An example of a typical sampling distribution (data B8, lognormal mix, sample size = 100% of population size, $n = 28\,322$) is shown in Figure 15 on the facing page.

For a random sample of values $\{x_1, x_2, \ldots, x_N\}$ from a normal probability distribution with population mean m and population variance σ^2, the probability distribution of the sample mean is normal$(\mu, \sigma^2/N)$. The measured mean, m, of samples of size N is therefore an approximation for μ. The population variance can be estimated using $s^2 = (\sum(x_i - m)^2)/(N-1)$, and s^2/N is an approximation for the sample variance. Thus, the population variance can also be estimated using $\sigma^2 \approx Ns^2$.

The significance test is then a simple test using the normal distribution. Under the null hypothesis, $\mu = 0$, and against the alternative hypothesis $\mu \neq 0$, at $c\%$ significance (where $\Phi^{-1}(\cdot)$ is the inverse normal density function):

- accept the null hypothesis if

$$\left|\frac{m-\mu}{s\sqrt{N}}\right| < \Phi^{-1}\left(\frac{c}{2}\right);$$

- reject the null hypothesis if

$$\left|\frac{m-\mu}{s\sqrt{N}}\right| \geqslant \Phi^{-1}\left(\frac{c}{2}\right).$$

A number of practicalities arise when using this test.

(1) In order to achieve consistency of results, about 10 000 samples have to be taken. A histogram of the means of samples then consistently resembles a normal distribution.

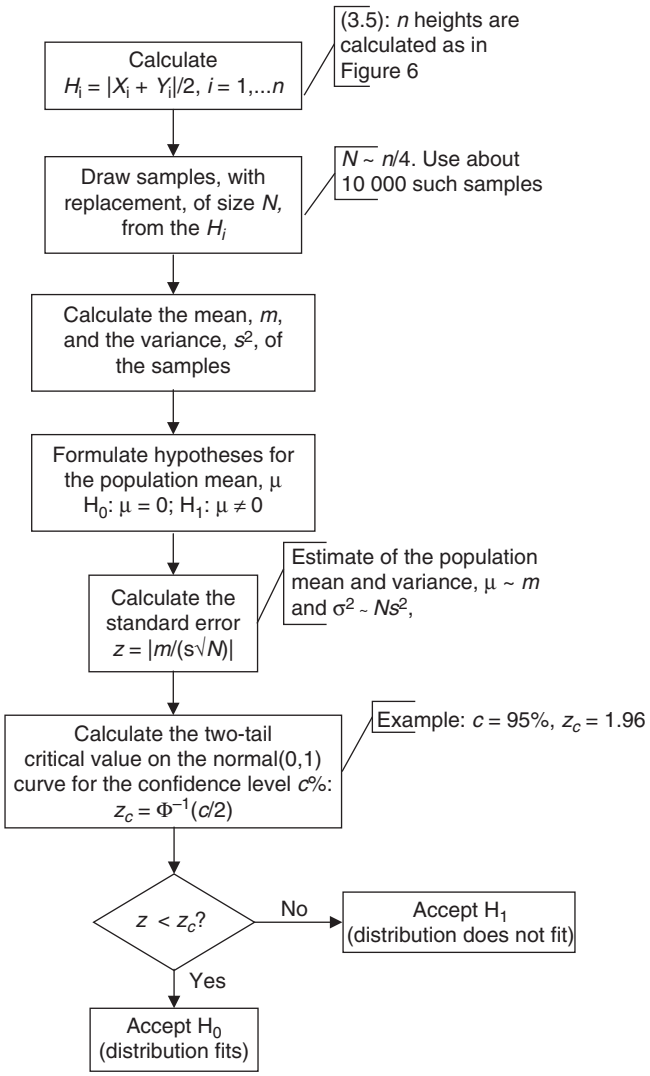
(2) The sample size N is a significant factor in the workability of this test. Conditioning by fitting a lognormal distribution, which should be a sufficiently good fit for all loss data sets considered, it was found that two ranges of N were useful. The first is to use all the H_i. The second is to use $0.2n < N < 0.3n$. In practice $N = 0.25n$ gave reasonable results in an acceptable running time. Using all the H_i took an excessive time to complete the calculations for large data sets ($> 400\,000$). Also, using all the H_i is insufficiently discriminating: very few distributions result in rejections of the null hypothesis.

A step-by-step guide to using this test is in shown in Figure 16 on the next page.

Table 5 on page 31 shows results of the sampling significance test, based on a sample size of 25% of the population. The figures given are the standard errors based on the normal distribution. The null hypothesis is accepted if the standard error is less than 1.96, based on two-tailed 95% significance. It provides a good degree of

Change OK?

FIGURE 16 Application of the TN-S test.



differentiation between the distributions tested, and it is in broad agreement with previous results obtained (see Table 3 on page 27 and Table 4 on page 27). It is clear that the TN-S test is more discriminatory than either of the TN-A and TN-B tests. It also has two negative aspects. The first is that it rejects the null hypothesis in all cases

Changes to sentence OK?

TABLE 5 Results for the TN-S test.

Risk category	LN	WB	LL	Gamma	Burr	GPD	LNMix
B1	1.37	**2.21**	1.2	1.41	1.41	**14**	1.89
B2	1.32	1.47	1.29	**2.52**	**3.3**	**12.52**	1.42
B3	1.47	1.79	1.53	**2.71**	**2.74**	**2.84**	1.49
B4	1.33	**2.46**	**4.98**	1.78	**2.17**	**2.85**	1.84
B5	1.34	1.6	**4.99**	1.84	**2.67**	**24.3**	1.56
B6	1.42	1.57	**5.08**	**2.16**	**2.92**	**14.8**	1.62
B7	1.78	1.83	1.53	1.45	1.83	**5.81**	1.79
B8	1.82	1.72	**9.81**	1.35	1.36	**3.39**	1.34

Standard error test results: sample size 25% of total population. Normal text means accept H_0. Bold text means reject H_0.

when fitting a GPD, despite some qualitatively good GPD fits. The second is that the test effectively has to be calibrated using a test lognormal distribution.

5.4 Comparison with established GoF tests

In this section we present results corresponding to the preceding TN tests for the KS and AD tests. The AD results were obtained using the R package `ADGofTest`. The KS results were obtained using the native function ks.test in R. They are given in Table 6 on the next page and Table 7 on the next page. Zero entries indicate that the p-values obtained are less than 10^{-6}.

Compared with the results in Table 3 on page 27, Table 4 on page 27 and Table 5, the AD and KS p-values are strikingly different. Table 6 on the next page and Table 7 on the next page show that the null hypothesis can be accepted in one case only: a lognormal mixture distribution using B1 data. This data set has 143 losses. All others are at least ten times as large. There is some differentiation for this data set among the distributions considered, but even then the null hypothesis is rejected for most distributions. For the large data sets both the KS and AD tests are totally ineffective. These results are a good illustration of the "large sample = small p-value" phenomenon described in Lin *et al* (2013). They also confirm observations noted in Section 2.1.

Words added – OK?

5.5 Comments on the p-values returned by the TN tests

The argument of Lin *et al* (2013) in Section 2 presents a paradox with respect to the TN tests. They argue that p-values necessarily ought to be near to zero for large populations, but they are not for the TN tests. The paradox is resolved as follows.

The empirical CDFs for data sets used are relatively uniform, in the sense that the deviation from any point (x_i, y_i) to a neighboring point (x_{i+1}, y_{i+1}) or (x_{i-1}, y_{i-1})

TABLE 6 AD p-values.

Risk category	LN	WB	LL	Gamma	Burr	GPD	LNMix
B1	0.0321	0.00052	0.0453	0.00009	0.0054	0	0.1467
B2	0	0	0	0	0	0	0
B3	0	0	0	0	0	0	0
B4	0	0	0	0	0	0	0
B5	0	0	0	0	0	0	0
B6	0	0	0	0	0	0	0
B7	0	0	0	0	0	0	0
B8	0	0	0	0	0	0	0

TABLE 7 KS p-values.

Risk category	LN	WB	LL	Gamma	Burr	GPD	LNMix
B1	0.0298	0.00001	0.0421	0.0007	0	0	0.0734
B2	0	0	0	0	0	0	0
B3	0	0	0	0	0	0	0
B4	0	0	0	0	0	0	0
B5	0	0	0	0	0	0	0
B6	0	0	0	0	0	0	0
B7	0	0	0	0	0	0	0
B8	0	0	0	0	0	0	0

in loss space is small. The deviation could be measured by the Euclidean norm of Section 4, for example. Variance of these deviations is constrained by construction of the empirical CDF. In particular, ordering the points (x_i, y_i) ensures that they change by small increments. Indeed, the increments become smaller with an increasing number of points. The TN tests depend on an area calculation or components thereof, and the accuracy of the estimation of the enclosed area also increases with an increasing number of points for the same reason.

Trefethen and Weideman (2014) demonstrate that the trapezium rule applied to an analytic function on the real line or a subset of the real line converges geometrically. If the points (x_i, y_i) are modeled by an analytic function on $(0, \sqrt{2})$, for example, a cubic spline, we conclude that the enclosed area defined by n points (x_i, y_i) converges to a limit A. (See also the discussion in Section 3.7.) Since the limiting value of the estimator of p-value, $\hat{\beta}$, is determined by limiting area, $\hat{\beta}$ does not tend to either 0 or 1, as discussed in Section 2.

OK? Is it the rule or the function that converges here?

Change OK or do you mean Section 3.7 of Trefethen and Weideman (2014)?

FIGURE 17 Power test against a Weibull distribution with constant scale and variable shape.

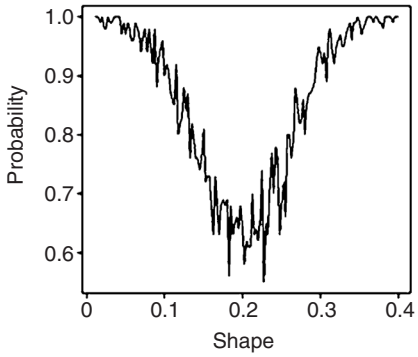
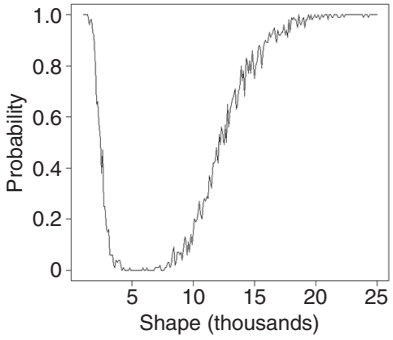


FIGURE 18 Power test against a Weibull distribution with constant shape and variable scale.



5.6 Power of the TN tests

This section gives a brief indication of the power of the TN tests. Since all the TN tests use the same enclosed area calculation, it is sufficient to consider only the TN-A test.

The power of a statistical test is the probability of rejecting a null hypothesis H_0 when it is false. Loosely, this definition amounts to "making a correct decision", and a high probability is expected. Changes to sentence OK?

It is not possible to give a complete analysis of the power of a TN test because

the possibilities of varying fitted distributions and parameters of those distributions are too numerous. Therefore, we give two examples that serve as brief indicators of power.

A distribution that we have often used as a starting point for modeling is lognormal$(8, 2)$. Consider the power of the TN-A test with respect to a random sample generated from this distribution by computing the TN-A statistic for a range of fitted Weibull distributions. The following steps explain the method used.

(1) Let Z be a random variable used to model data.

(2) Propose a null hypothesis H_0: $Z \sim$ lognormal$(8, 2)$.

(3) Propose an alternative hypothesis H_1: $Z \sim$ Weibull(k, l) where k is a shape parameter and l is a scale parameter.

(4) The rejection region for the TN-A statistic is $\{\text{tn-a}\colon \text{tn-a} > 0.1314\}$.

(5) The power is estimated as follows:

   (a) generate a lognormal$(8, 2)$ random sample;

   (b) repeat the next step 100 times, recording whether or not the test statistic falls within the critical region;

   (c) for each pair (k, l), apply the transformation T of (3.2) to the sample, using a Weibull(k, l) distribution;

   (d) return the proportion of trials where the test statistic falls within the critical region.

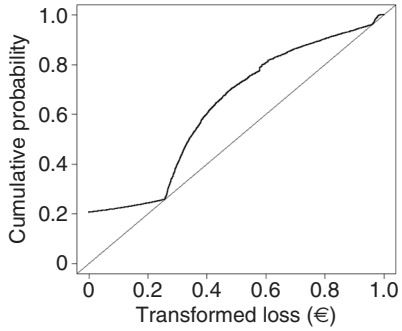
The results are shown in Figure 17 on the preceding page and Figure 18 on the preceding page. Figure 17 shows the result of varying the shape parameter, keeping the scale parameter constant at 50 000. The Weibull distribution is very sensitive to the shape parameter, and the volatility in the measured proportion is clear. The power exceeds a nominal value of 0.8 for all but a reasonably narrow range of shape parameter (approximately $k \in (0.13, 0.27)$).

If the shape parameter is held constant (at 0.8) and the scale parameter is varied, the resulting power function, shown in Figure 18 on the preceding page, is smoother, as the Weibull distribution is less sensitive to variation in the scale parameter. Using the nominal power $= 0.8$ figure, a wider range for the scale parameter leads to acceptance of the alternative hypothesis.

The two Weibull examples show that, in the cases considered, the TN-A test accepts the false null hypothesis for a limited but possibly wide parameter range. This is to be expected, as a Weibull fit is qualitatively acceptable. The test appears to be immune from the two significant problems of population size: if too many observations are

Changes to sentence OK?

FIGURE 19 Distinct body and tail fit.



used, trivial effects can register as significant; if too few observations are used, the hypothesis test may not be able to detect a meaningful effect, even if there is one.

5.7 Q–Q/P–P plots: similarities and differences

A Q–Q plot is usually used as a qualitative tool to compare quantiles of two distributions. Commonly, the first distribution comprises empirical data, and the second is a distribution that has been fitted to the data. Similarly, a P–P plot is a scaled version of a Q–Q plot such that the quantiles are mapped to [0, 1]. The empirical component of these plots corresponds to what is plotted on the y-axis in loss space and the Y-axis in probability space. The difference is that there is no Q–Q or P–P equivalent of the transformation T of (3.2). In Q–Q or P–P plots, equal distributions correspond to the 45° line, and deviations from this line indicate skewness and tail heaviness. Change OK?

Such deviations could be quantified by calculating the area enclosed by the ordinates Change OK?
of the Q–Q or P–P plot and the 45° line. However, the advantage of the TN process is that losses are mapped to a standardized image that is independent of any distribution fitted to the data. Therefore, all distributions are standardized such that a single TN test is applicable for all of them.

5.8 Distinct body and tail fit

The profile of the transformed losses provides an opportunity to examine informally the goodness-of-fit of the distribution body and tail separately. The profile often intersects the 45° line once only, thereby partitioning the distribution into a body and a tail, but without regard to the position of the partition. Nevertheless, it is often possible to identify a region of poor fit and a region of good fit. Identifying regions of poor and good fit can be quantified by formulating a formal definition of "tail" and

"body", and calculating the enclosed area for both regions. Indeed, more than two regions can be defined and analyzed in the same way.

Figure 19 on the preceding page shows an example of probability space using data B4, and fitting a Gamma distribution. Overall this fit passes the TN-A test since the calculated area 0.1161 is less than the 5% critical value 0.1314. However, the fit for the extreme tail (approximately the top 5% of losses) is much better than the body fit since the transformed loss profile is relatively near to the 45° line.

6 DISCUSSION AND SUMMARY

The aims set out in Section 2 of this paper have been largely met in formulating the TN tests.

"original list of requirements in Section 1"? Otherwise please clarify where these aims are explicitly stated.

The TN tests are independent of the number of points in the empirical data, provided that there are sufficient points (at least fifty). This is the most significant result presented in this paper. As discussed in Section 5, a large number of empirical data points can be an advantage because it produces a more accurate estimation for the enclosed area.

Changes to sentence OK?

"estimate of"? Other changes to sentence OK?

This independence property is probably the most important aspect of the TN tests, and solves two associated problems in quantifying operational risk. The first concerns the modeling threshold. One "solution" to the problem of the failure of the AD and KS tests is to introduce a lower threshold, so that only losses above the threshold are modeled. As the threshold increases, the number of points modeled decreases, and the AD and KS tests have more chance of accepting the null hypothesis. However, increasing a threshold in order to force an acceptable GoF result is open to severe criticism. The threshold should be set objectively in advance. The TN tests require no remedial action to achieve acceptable results. The second associated problem is what to do if no distributions appear to fit. Using AD and KS, common practice is to use a default distribution, often lognormal. This is not necessary for the TN tests.

New paragraph here OK?

Change OK?

For large population sizes, the AD and KS tests are far too sensitive to reflect fits that are qualitatively appealing. In most cases the TN tests satisfy the criterion "if a fit looks good, the test should say so". This is clearly a subjective measure, but its is important in the context of rejection of all distributions by the AD and KS tests. In all cases, and in particular for borderline cases, the TN tests provide a workable objective measure for deciding whether or not to reject a null hypothesis. It is noteworthy that the TN tests are successful for both low- and high-value losses.

The TN tests successfully discriminate between the common fat- and long-tailed distributions. In all cases, at least two distributions allow the null hypothesis to be accepted. This allows a credible competition between distributions. No cases are such that the null hypothesis is rejected or accepted in all cases. Fur-

"There are no tests where the null hypothesis is either rejected or accepted in all cases"? Otherwise please clarify sentence.

thermore, there is no contradiction between TN GoF results for extreme distributions and less extreme distributions. One type or the other is a good fit, not both.

The requirement for a deterministic GoF test is partly met. The TN-A test is deterministic. The TN-B and TN-S tests use stochastic elements, so are not. However, the results for TN-A and TN-B agree closely, which gives some confidence in the stochastic methods used. The TN-S test is subject to conditioning on the sample size, which is less desirable. It is also more stringent than TN-A or TN-B.

Change OK?

An added bonus to the original list of requirements in Section 1 is that the TN-A test allows a direct comparison of candidate distributions. The distribution with the lowest TN-A value wins, because the TN-A value is a direct measure of enclosed area, and the smaller the enclosed area, the better the fit. This assumes, of course, that the TN-A value is less than a critical value. The TN-A test can therefore be used to positively select a winning distribution, and, indeed, to place candidate distributions in order of goodness-of-fit. This is in marked contrast to using AD or KS tests, where the p-value should only be used to reject a distribution.

Changes to sentence OK?

6.1 Further work

We suggest the following as extensions of the ideas in this paper.

(1) Curve fitting using a minimum enclosed area criterion; an efficient algorithm to parse the parameter space would be needed.

(2) Weighting tail losses more than body losses in order to accentuate the influence that tail losses have on subsequent capital calculations.

(3) The use of order statistics as seems appropriate for further analysis. A natural ordering is imposed on losses, treated as observations from a fixed distribution, by construction of an empirical CDF.

Word added OK? Otherwise I'm not sure how this sentence is an extension of the ideas.

(4) Application of the methods suggested in this paper to distributions that are not usually used as operational risk severity distributions (eg, Gaussian).

APPENDIX A

This appendix gives standard definitions for the Euclidean metric, a metric space and an open ball, taken from Sutherland (1975).

DEFINITION A.1 (Euclidean metric in $\mathbb{R}^2$) For two points $\boldsymbol{z}_1 = (x_1, y_1) \in \mathbb{R}^2$ and $\boldsymbol{z}_2 = (x_2, y_2) \in \mathbb{R}^2$, the Euclidean metric $d(\boldsymbol{z}_1, \boldsymbol{z}_2)$ is defined as

Change to notation for the set of real numbers here and elsewhere – OK?

$$d(\boldsymbol{z}_1, \boldsymbol{z}_2) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

and is often denoted by $d(\boldsymbol{z}_1, \boldsymbol{z}_2) = \|\boldsymbol{z}_1, \boldsymbol{z}_2\|$.

DEFINITION A.2 (Metric space) The metric $d(\boldsymbol{x}, \boldsymbol{y})$, $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^2$, satisfies the following conditions, showing that $\{\mathbb{R}^2, d\}$ is a metric space:

$$
\begin{aligned}
&d(\boldsymbol{x}, \boldsymbol{y}) \geqslant 0,\\
&d(\boldsymbol{x}, \boldsymbol{y}) = 0 \iff \boldsymbol{x} = \boldsymbol{y},\\
&d(\boldsymbol{x}, \boldsymbol{y}) = d(\boldsymbol{x}, \boldsymbol{y}),\\
&d(\boldsymbol{x}, \boldsymbol{y}) = d(\boldsymbol{x}, \boldsymbol{z}) + d(\boldsymbol{z}, \boldsymbol{y}).
\end{aligned}
$$

DEFINITION A.3 (Open ball) Given any point $z \in \mathbb{R}^2$ and a real number $r > 0$, an open ball is the set $B_r(z) = \{z' \in \mathbb{R}^2 : d(z', z) < r\}$.

DEFINITION A.4 (Open set in a metric space) A subset U of the metric space $\{\mathbb{R}^2, d\}$ is open in $\{\mathbb{R}^2, d\}$ if, given any $u \in U$, there exists $\varepsilon(u). > 0$ such that $B_{\varepsilon(u)}(u) \subset U$. So an open set in a metric space means that you can always draw a ball around any point in the set.

I have assumed the full stop denoted a "place holder" in the following expression and typeset it as a centered dot accordingly. OK? Or is it a typo and should be deleted?

APPENDIX B

This appendix contains details of the transformation of particular points of an open ellipse in loss space centered on (x, y), with semi-axes r and s. Three particular cases are notable. These are the transformations of points in an open ellipse in loss space that

(1) represent extremities of cumulative probability,

(2) represent extremities of loss,

(3) are a maximal distance from the base point (x, y).

"the"?

Extremities of cumulative probability in an open ellipse are represented by the two points that are on the intersection of the boundary of the open ellipse and a vertical line through the center. This vertical line transforms to a vertical line of the same length in probability space. It measures constant loss with extremities of cumulative probability.

The mappings of the extreme points for this case are

$$
\begin{aligned}
(x, y + r) &\to (F(x), y + r),\\
(x, y - r) &\to (F(x), y - r).
\end{aligned}
$$

Extremities of loss in an open ellipse are represented by the two points that are on the intersection of the boundary of the open ellipse and a horizontal line through the center. This horizontal line transforms to a horizontal line of much smaller length

in probability space. It measures constant cumulative probability with extremities of loss.

The mappings of the extreme points for this case are

$$(x+s, y) \to (F(x+s), y),$$
$$(x-s, y) \to (F(x-s), y).$$

The third case necessitates a calculation of the points on the boundary of the open set that are on a normal to the fitted CDF. In order to do this, denote the gradient of the fitted curve at the point (x, y) by m. The gradient of the normal to the fitted CDF at the same point is then $-1/m$. The two extreme points are given by the solution (u, v) of the following equations: Word added – OK?

$$\left(\frac{u-x}{s}\right) + \left(\frac{v-y}{r}\right) = 1, \qquad v = \left(\frac{-1}{m}\right)u + \left(y + \frac{x}{m}\right).$$

APPENDIX C

The coordinates and variables used in this appendix are defined in Section 3.4.

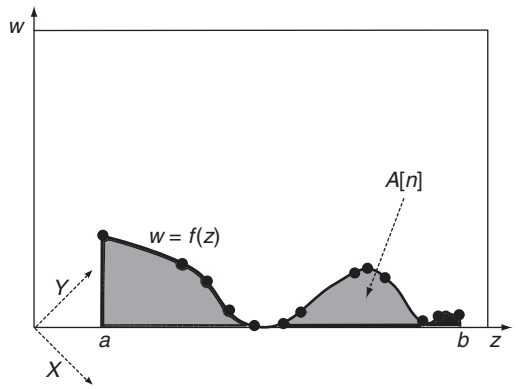
PROOF OF PROPOSITION 3.8 First, define an axis, z, aligned along the 45° line, and another axis, w, perpendicular to it. The z- and w-axes are 45° counterclockwise rotations of the X- and Y-axes. z then ranges from 0 to $\sqrt{2}$, and the heights H_i of (3.5) are aligned along the positive w-axis. Then rename points (X_i, Y_i) $(i = 1, \ldots, n)$, in probability space to (z_i, w_i) $(i = 1, \ldots, n)$.

Next, approximate two measures, $A[n]$ and $A[m]$, where $m > n$, of the enclosed area by the integral of a function $f(z)$ between two limits $z = a$ and $z = b$ $(0 < a, b < \sqrt{2})$. The coordinate systems and $f(z)$ are shown in Figure C.1 on the next page. The requirement on $f(z)$ is that is must be at least twice differentiable in the range (a, b).

The absolute truncation errors when applying the trapezium rule for $A[n]$ and $A[m]$ are then given by (see Burden and Faires 2000)

$$\left.\begin{aligned} \left|A[n] - \int_a^b f(z)\,\mathrm{d}z\right| &= \frac{(b-a)^3}{12n^2} f''(\xi_n), \quad \xi_n \in [a, b], \\ \left|A[m] - \int_a^b f(z)\,\mathrm{d}z\right| &= \frac{(b-a)^3}{12m^2} f''(\xi_m), \quad \xi_m \in [a, b]. \end{aligned}\right\} \tag{C.1}$$

FIGURE C.1 Trapezium rule estimation in the (z, w)-plane.



Hence,

$$\begin{aligned} |A[n] - A[m]| &\leqslant \left| A[n] - \int_a^b f(z)\,\mathrm{d}z \right| + \left| A[m] - \int_a^b f(z)\,\mathrm{d}z \right| \\ &< \frac{(b-a)^3}{12}\left(\frac{f''(\xi_n)}{n^2} + \frac{f''(\xi_m)}{m^2} \right) \\ &= \frac{k_n}{n^2} + \frac{k_m}{m^2}, \end{aligned} \tag{C.2}$$

where

$$k_n = \frac{(b-a)^3}{12} f''(\xi_n) \quad \text{and} \quad k_m = \frac{(b-a)^3}{12} f''(\xi_m).$$

Given a small real number $\mu > 0$, there exists an integer $N > 0$ such that

$$\frac{\varepsilon}{2} < \frac{k_n}{n^2} \quad \forall n > N \quad \text{and} \quad \frac{\varepsilon}{2} < \frac{k_m}{m^2} \quad \forall m > N.$$

Therefore,

$$|A[n] - A[m]| < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon, \qquad \min(n, m) > N. \tag{C.3}$$

Therefore, the sequence $\{A[n]\}$ is a real-valued Cauchy sequence that is convergent. Change OK?
So, for sufficiently large n, the trapezium approximation does not depend on n. □

APPENDIX D

In this appendix, the topological derivation of the formulas for p-values (for the TN-A test) is advanced.

D.1 p-value derivation: topological view

The process in this section expands on the points in Section 4.1, and introduces the necessary topological concepts.

(1) CDF cover.

   (a) Define elliptic open sets in $\mathbb{R}^2$.

   (b) Define a metric space in $\mathbb{R}^2$ based on these open sets and an appropriate metric.

   (c) Define a topological space in $\mathbb{R}^2$ based on the metric space.

   (d) Define a restriction of the topological space to loss space.

   (e) Demonstrate that there is an open cover of the CDF, parameterized by the length, r, of the semi-axis of an ellipse that is parallel to the y-axis. This open cover constitutes the "band" surrounding the CDF.

(2) Mapping.

   (a) Map the open cover of the CDF using transformation T to probability space.

   (b) Deduce that mapped open cover in probability space is a topological space, since T is continuous.

   (c) Deduce that open sets in the probability space topology resemble ellipses, but are not actually ellipses.

   (d) Derive the boundaries for the result of mapping a band surrounding the CDF in loss space.

   (e) Describe the image of the mapped band in probability space.

(3) p-value.

   (a) Determine an expression for $A[n]$ in terms of r.

   (b) Deduce that the p-value is the width of the band in probability space, measured perpendicular to the 45° line.

D.2 p-value derivation: detailed topological view

This section contains details of the steps outlined in Section D.1.

D.2.1 The topology of loss space

The definitions and proofs in this and subsequent sections can be found in any introductory text on topology. We refer to Sutherland (1975) as a recommended text, and reiterate some definitions therein to fit with the notation used in this paper. In this present analysis we prefer not to use Sutherland's definition of an open ball because it fits less well with the scale of loss space, in which the y-axis is confined to $(0, 1)$, but the x-axis ranges from 0 to ∞. For reference, Sutherland's definitions for an open ball, the Euclidean metric and construction of a metric space using them are given in Appendix A. The details of the steps appear below.

Change OK?

Changes to sentence OK?

The CDF cover in loss space

STEP 1(a). Given a point (x, y) in $\mathbb{R}^2$, and two real constants a and b, an open set B is the region

$$B(x, y, a, b) = \left\{(u, v)\colon \left(\frac{u-x}{a}\right)^2 + \left(\frac{v-y}{b}\right)^2 < 1\right\}.$$

Such sets will be referred to as "open ellipses".

STEP 1(b). The metric $d(\boldsymbol{p}, \boldsymbol{q}) = \|\boldsymbol{p}-\boldsymbol{q}\|$, where $\|\cdot\|$ denotes the Euclidean norm, will be applied to such open ellipses, with the proviso that both $\boldsymbol{p}$ and $\boldsymbol{q}$ are points in the same open ellipse. With a collection S of open ellipses as defined in the previous step, the combination $\{S, d\}$ satisfies the conditions of a metric space (see Appendix A).

Bold (vector) notation OK in this section? Please mark any $\boldsymbol{p}$ and $\boldsymbol{q}$ that should be p and q.

STEP 1(c). Now let $\Im$ be the collection of open ellipses of the metric space $\{S, d\}$. Then $\{S, \Im\}$ is a topological space. This is a very general result: any metric space gives rise to a topological space. A loose interpretation of $\{S, \Im\}$ is that open ellipses, as defined above, can be used to cover the plane $\mathbb{R}^2$.

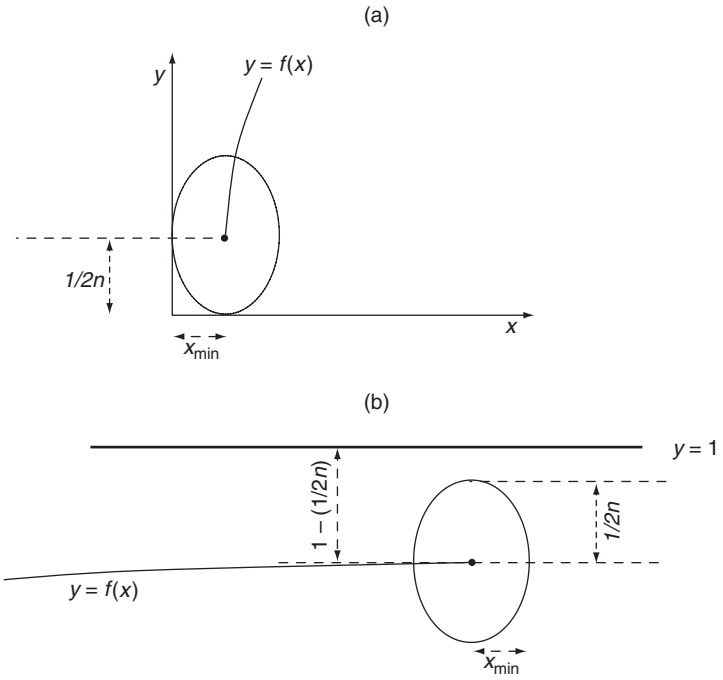
STEP 1(d). We now consider open ellipses in $\{S, \Im\}$ that cover the fitted CDF in loss space. This cover will map to an open cover of the 45° line in probability space. Let $\{S, \Im \mid \Lambda\}$ be the restriction of S to loss space Λ. Then $\{S, \Im \mid \Lambda\}$ is also a topological space. At this stage it is useful to demonstrate that S is not empty by considering the minimum and maximum empirical losses.

Suppose there are n empirical losses. The minimum loss, $x_{\min}$, is strictly positive and is assigned the minimum probability $1/2n$ when constructing the empirical CDF. An open ellipse can be drawn with the point $(x_{\min}, 1/2n)$ at its center, as in part (a) of Figure D.1 on the facing page. This ellipse is the largest possible without breaching the boundary of loss space, and a smaller one would be a sufficient cover.

The maximum loss, $x_{\max}$, is assigned the maximum probability $1 - (1/2n)$. Since

$$1 - \frac{1}{2n} > \frac{1}{2n} \quad \text{as } n > 1,$$

FIGURE D.1 Open ellipse surrounding the minimum and maximum losses.

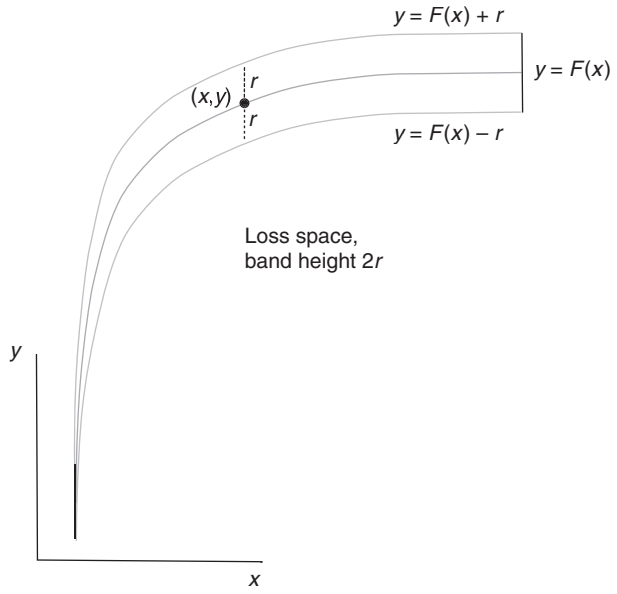


(a) Open ellipse covering the minimum loss. (b) Open ellipse covering the maximum loss.

an open ellipse of exactly the same size can surround the point $(x_{\max}, 1-1/2n)$ when constructing the empirical CDF, as part (b) of Figure D.1.

STEP 1(e). The empirical CDF can be covered by taking the open ellipse in Figure D.1 and sliding it along the CDF from one end to the other. This cover constitutes the "band" that covers the CDF. All points in the band can be covered by at least one open ellipse, so the band has no holes. The band has two boundaries that are translations of the empirical CDF. They are the lines $y = F(x)+r$ and $y = F(x)-r$, where r is the length of the vertical semi-axis of an open ellipse through a point (x, y) on the CDF. In order to make the mapping to probability space tractable, the other two boundaries are taken to be a vertical line through the minimal-loss point, $(x_{\min}, 1/2n)$, and a vertical line through the maximal-loss point, $(x_{\max}, 1-(1/2n))$. The band comprises all points within the boundary, and no points on the boundary. The noninclusion of the two points $(x_{\min}, 1/2n)$ and $(x_{\max}, 1-(1/2n))$ has no impact on the transformation to probability space. An alternative way to think about its construction is to consider

FIGURE D.2 Band covering the empirical CDF.



moving a straight line segment of length $2r$, centered on the CDF, along the CDF between the minimum and maximum losses. The band is illustrated in Figure D.2.

DEFINITION D.1 (Height of the band) Referring to Figure D.2, the band is parameterized by the CDF, $y = F(x)$, and by r. We define the height of the band as $2r$.

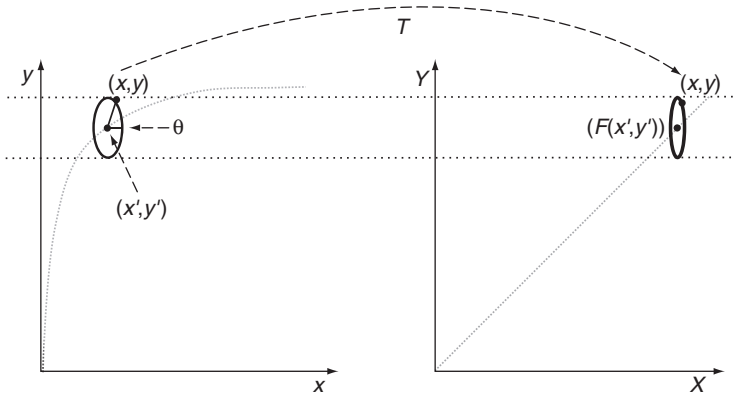
Transformation of the CDF cover to probability space

STEP 2(a). Apply transformation T to all points in a band of height $2r$ (ie, the open cover of the CDF), and also to the boundary of the band. The following steps show details of the map of an open ellipse and a band with parameter r.

STEP 2(b). Transformation T is continuous and so is its inverse T^{-1}. Therefore, T is a homeomorphism and the image of loss space under T, $T(\Lambda)$, is a topological space. As such, the image of loss space under T comprises a collection of sets, which are open in $T(\Lambda)$.

The geometric interpretation of the foregoing analysis is that overlapping open sets in loss space define a band with height $2r$ enclosing a CDF curve, where $r <$

FIGURE D.3 Open sets in loss space and probability space.



$\min(x_{\min}, 1/2n)$. This region maps under T to a region in probability space, and we will show that this shape resembles a parallelogram.

STEP 2(c). Open ellipses in loss space transform to open sets in probability space that resemble ellipses. An open set in probability space is not symmetric about a vertical axis through its center. One-half is stretched parallel to the X-axis. When graphed to scale, they appear very thin and tall.

Take any point (x, y) on the CDF of the fitted curve in loss space. An open ellipse with semi-axes r and s centered on this point is the locus

$$\{u, v\}\colon u = x + s\cos(\theta);\ v = y + r\sin(\theta); \quad 0 \leqslant \theta \leqslant 2\pi.$$

Under the transformation T, (u, v) maps to (X, Y), where

$$X = F(u) = F(x + s\cos(\theta)),$$
$$Y = v = y + r\sin(\theta).$$

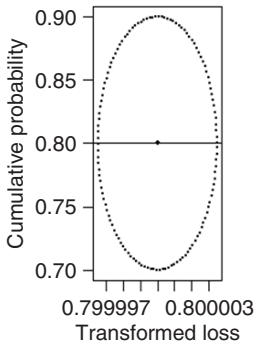
Eliminating θ,

$$\left(\frac{F^{-1}(X) - x}{s}\right) + \left(\frac{Y - y}{r}\right) = 1,$$

which is ellipse-like in probability space with axes aligned with the coordinate axes. We name these ellipse-like regions "almost-ellipses".

This mapping is shown in Figure D.3. The dimensions of the almost-ellipse in Figure D.3 are greatly exaggerated. The minor axis, $F(x + s) - F(x - s)$, is much smaller than the major axis, $2r$. Change OK?

FIGURE D.4 Example of an almost-ellipse.



As an example, consider the point $(7515, 0.8)$, which is a typical point on the CDF fitted to B6 data. The circle of radius 0.1 centered on this point defines the boundary of an open set S in loss space. Under the transformation T, S maps to an open set S' in probability space. S' is an almost-ellipse centered on $(0.8, 0.8)$ with semi-major axis of length 0.1 and a very much smaller semi-minor axis of approximate length 3.49×10^{-7}. This almost-ellipse and its center point are shown in Figure D.4, with the 45° line, which appears horizontal with the scales indicated.

Appendix B contains details of the transformation of particular points on an open ellipse in loss space.

STEP 2(d). The most significant result of the mapping of a band with parameter r is the transformation of the boundaries $y = F(x) + r$ and $y = F(x) - r$.

These boundaries are defined by considering the points $(x, y + r)$ and $(x, y - r)$:

$$T((x, y + r)) \mapsto (F(x), y + r) = (F(x), F(x) + r)$$

and

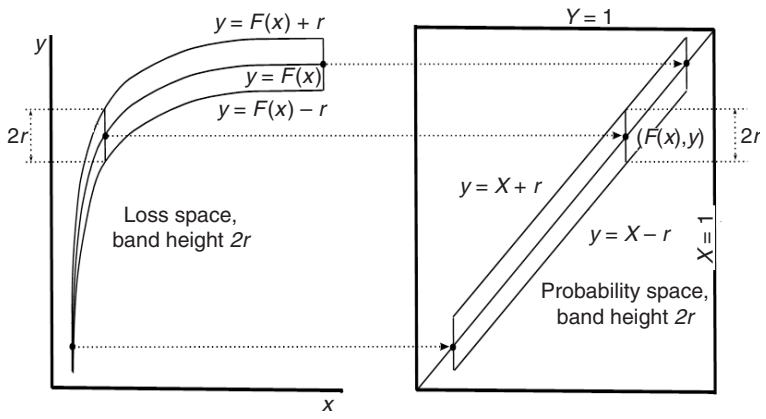
$$T((x, y - r)) \mapsto (F(x), y - r) = (F(x), F(x) - r).$$

These mapped boundaries are lines parallel to the 45° line in probability space. Their equations are $Y = X + r$ and $Y = X - r$.

The other two portions of the boundary in loss space map to two vertical lines in probability space, one through $(F(x_{\min}), F(x_{\min}))$ and the other through $(F(x_{\max}), F(x_{\max}))$.

STEP 2(e). Figure D.5 on the facing page shows the form of the mapped band in probability space.

FIGURE D.5 Transformation of the CDF envelope in loss space.



Referring to Figure D.5, as the center of an open ellipse in loss space moves along the fitted curve $y = F(x)$, the image of that open ellipse under T is an almost-ellipse, the center of which traces a path on the 45° line in probability space. The traced image is essentially a parallelogram. It has two sides parallel to the 45° line, and the other two are very slightly curved, deviating minimally from the vertical. We call this shape an "almost-parallelogram", and treat it as an actual parallelogram in calculations.

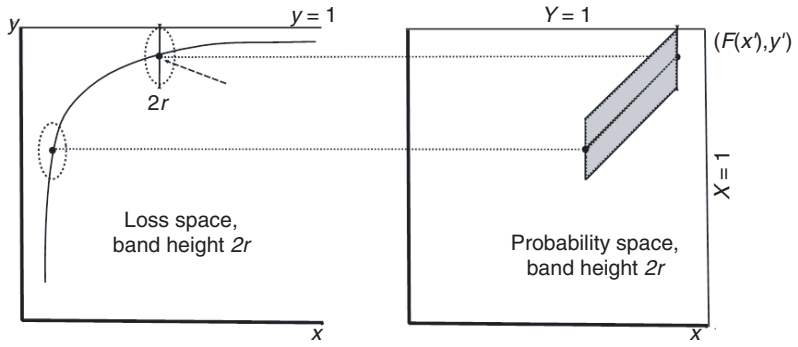
The almost-parallelogram defines a region within which the fitted CDF can be interpreted as a good fit for all points. Transformed points outside this region represent a poorer fit, and the GoF is measured using the area of the almost-parallelogram, derived from a band of height $2r$ in loss space, mapped to a region with height $2r$, as in Figure D.5.

Changes to sentence OK?

Recall that $2r$ represents a band height such that an open cover of the fitted curve lies entirely within the loss space. If a wider band with parameter $r' > r$ in loss space is mapped to probability space, all open ellipses that cover the fitted CDF must lie entirely within probability space. From the discussion in Step 1(d), the minimum value of r for the cover of the CDF to lie entirely within loss space is $\min(x_{\min}, 1/2n)$ (where n is the number of empirical losses). If r exceeds this minimum value, the image of the band is a shortened almost-parallelogram, as shown in Figure D.6 on the next page.

This restriction represents the case where we try to account for losses that fit less well. As r increases, the parallelogram-like shape of the image is retained, and its area can still be calculated. This area can then be used as a GoF measure.

FIGURE D.6 Map of a wide band.



D.3 Summary of results from Appendix D

There are two principal results.

"three"?

(i) The mapped band in probability space containing the diagonal 45° line represents a good fit for points within it.

(ii) The band in probability space can be approximated by a parallelogram.

(iii) The band in probability space is parameterized by a "height" $2r$, shown in Figure D.5 on the preceding page.

APPENDIX E

This appendix gives a practical guide to using the TN tests.

The first step is to calculate the value of the enclosed area test statistic $A[n]$ (3.8), using the steps in flow diagram Figure 6 on page 14. After that, the significance of the result can be assessed using all or some of the tests TN-A, TN-B and TN-S. They are summarized in flow diagrams for F-A, F-B and F-S, respectively (see Figure 12 on page 22, Figure 13 on page 24 and Figure 16 on page 30).

All diagrams have been rekeyed. Please check carefully, particularly cross-references to equations and figures.

E.1 F-A: application of the TN-A test

Use the process in Figure 6 on page 14 to calculate $A[n]$ (see Figure 12 on page 22).

E.2 F-B: application of the TN-B test

The inputs are the calculated $A[n]$ using the process in Figure 6 on page 14, and the parameters of the fitted distribution (see Figure 13 on page 24).

E.3 F-S: application of the TN-S test

This test used the calculated heights H_i from the process in Figure 6 on page 14, and calls this process repeatedly using amended distribution parameter values to calculate amended values of $A[n]$ (see Figure 16 on page 30). Changes to sentence OK?

DECLARATION OF INTEREST

The author reports no conflicts of interest. The author alone is responsible for the content and writing of the paper.

ACKNOWLEDGEMENTS

The author is grateful for the interest and assistance of colleagues at Santander (UK) in the preparation of this paper; Karl Rutlidge and Bertrand Hassani deserve particular mention. The author also thanks the referee for helpful comments.

REFERENCES

Anderson, T. W., and Darling, D. (1952). Asymptotic theory of certain "goodness-of-fit" criteria based on stochastic processes. *Annals of Mathematical Statistics* **23**, 193–212.

Basel Committee on Banking Supervision (2011). Operational risk: supervisory guidelines for the advanced measurement approaches. Bank for International Settlements, June. URL: www.bis.org/publ/bcbs196.pdf.

Bickel, P., and Freedman, D. (1981). Some asymptotic theory for the bootstrap. *Annals of Statistics* **9**(6), 1196–1217.

Burden, R., and Faires, J. (2000). *Numerical Analysis*, 7th edn. Brooks–Cole, Pacific Grove, CA.

Chernobai, A., Rachev, S., and Fabozzi, F. (2005). Composite goodness-of-fit tests for left-truncated loss samples. Technical Report, University of California Santa Barbara. Change OK?

Conover, W. J. (1999). *Practical Nonparametric Statistics*, 3rd edn. Wiley.

Cramér, H. (1928). On the composition of elementary errors. *Scandinavian Actuarial Journal* **1**, 13–74.

Davison, A. C., and Hinkley, D. V. (1995). *Bootstrap Methods and Their Application*. Cambridge University Press.

Doray, L. G., and Huard, L. (2001). On some new goodness-of-fit tests for the Poisson distribution. In *New Trends in Statistical Modelling, Proceedings of the 16th International Workshop on Statistical Modelling, Odense, Denmark*, Klein, B., and Korsholm, L. (eds), pp. 429–435. Statistical Modelling Society.

Efron, B., and Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science* **1**(1), 54–77.

Fard, M., and Holmquist, B. (2013). Powerful goodness-of-fit tests for the extreme value distribution. *Chilean Journal of Statistics* **4**(1), 55–67.

Goegebeur, Y., and Guillou, A. (2010). Goodness-of-fit testing for Weibull-type behaviour. *Journal of Statistical Planning and Inference* **140**(6), 1417–1436.

Goldmann, C., Klar, B., and Meintanis, S. (2015). Data transformations and goodness-of-fit tests for type-II right censored samples. *Metrica* **78**(1), 59–83.

Gourier, E., Farkas, W., and Abbate, D. (2009). Operational risk quantification using extreme value theory and copulas: from theory to practice. *The Journal of Operational Risk* **4**(3), 1–24.

Guégan, D., and Hassani, B. (2012). Operational risk: a Basel II++ step before Basel III. *Journal of Risk Management in Financial Institutions* **8**(13), 37–53.

Guégan, D., and Hassani, B. (2014). Using a time-series approach to correct serial correlation in operational risk capital calculation. *The Journal of Operational Risk* **8**(3), 31–58.

Guégan, D., Hassani, B., and Naud, C. (2011). An efficient threshold choice for operational risk capital computation. *The Journal of Operational Risk* **8**(4), 3–19.

Hogg, R. V., McKean, J., and Craig, A. T. (2013). *Introduction to Mathematical Statistics*, 7th edn. Pearson, London.

Kinnison, R. (1989). Correlation coefficient goodness-of-fit test for the extreme-value distribution. *American Statistician* **43**(2), 98–100.

Lehérissé, V., and Renaudin, A. (2013). Quantile distance estimation for operational risk: a practical application. *The Journal of Operational Risk* **8**(2), 73–102.

Lemeshko, B. Yu., Chimitova, E. V., and Kolesnikov, S. S. (2007). Nonparametric goodness-of-fit tests for discrete, grouped or censored data. In *Proc. XIIth International Conference on Applied Stochastic Models and Data Analysis (ASMDA 2007), Chania, Crete*, Skiadas, C. H. (ed). ASMDA International. URL: www.asmda.com/id30.html.

Lin, M., Lucas, H., and Shmueli, G. (2013). Too big to fail: large samples and the p-value problem. *Information Systems Research* **24**(4), 906–917.

Massey, F. (1951). The Kolmogorov–Smirnov test for goodness-of-fit. *Journal of the American Statistical Association* **46**(253), 68–78.

Quesenberry, C. P. (1986). Some transformation methods in goodness-of-fit. In *Goodness-of-Fit Techniques*, d'Agostino, R. B., and Stephens, M. (eds), Chapter 6, pp. 235–277. Marcel Dekker, New York.

Rizzo, M. L. (2009). New goodness-of-fit tests for Pareto distributions. *ASTIN Bulletin* **39**(2), 691–715.

Stephens, M. A. (1972). EDF statistics for goodness-of-fit: part I. Technical Report 186, Office of Naval Research, Stanford University. URL: www.dtic.mil/cgi-bin/GetTRDoc ?AD=AD0737653.

Stephens, M. A. (1974). EDF statistics for goodness-of-fit and some comparisons. *Journal of the American Statistical Association* **69**(347), 730–737.

Stephens, M. A. (1977). Goodness-of-fit for the extreme value distribution. *Biometrika* **64**(3), 583–588.

Sutherland, W. A. (1975). *Introduction to Metric and Topological Spaces*. Oxford University Press.

Trefethen, L. N., and Weideman, J. A. C. (2014). The exponentially convergent trapezoidal rule. *SIAM Review* **56**(3), 385–458.

von Mises, R. E. (1928). *Wahrscheinlichkeit, Statistik und Wahrheit*. Springer.