MOODY'S ANALYTICS

RESEARCH REPORT
AUGUST 2014
INSURANCE ERS

Percentile Estimation for Stable Capital Requirement

David Redfern PhD
Douglas McLean PhD
Moody's Analytics Research

Contact Us

Americas
+1.212.553.1658
clientservices@moodys.com

Europe
+44.20.7772.5454
clientservices.emea@moodys.com

Asia (Excluding Japan)
+85 2 2916 1121
clientservices.asia@moodys.com

Japan
+81 3 5408 4100
clientservices.japan@moodys.com

Overview

» When Monte Carlo methods are used to determine solvency capital requirements, the end result will depend on the way in which the set of net asset valuations are obtained. For example, if the samples are generated at random, then changing the random number seed would lead to a different capital requirement. This document examines the percentile estimation process in order to alleviate or lessen the variability of the capital requirement.

» Several different percentile estimators are considered for determining the capital requirement based on a set of random observations. These estimators are shown to have varying levels of bias, but all of them also exhibit much larger levels of variance. This variability dominates the capital estimation process, and hence there is no clear advantage of using any particular estimator. The recommendation is thus to use a simple estimator based on linear interpolation.

» The variance in percentile estimation is shown to be due to sampling error, implying that variance reduction techniques are candidates for improving percentile estimation. Low-discrepancy (Sobol) sampling is proposed as one way to reduce the sampling error and is shown to have a dramatic effect on the stability of the percentile estimation in the particular univariate example studied (see Figure 7).

» The methods presented also have application to the calculation of percentiles for real world scenario set validation.

Table of Contents

1. Introduction

Determining solvency capital requirement using Monte Carlo methods involves generating many thousands of possible net asset valuations for different states of the world. These observations are then analysed by finding a particular percentile, often the 99.5th percentile, to determine the capital required to cover the worst 1-in-200 year event. Possibly the simplest method of determining the 95th percentile value is to rank a set of observations into ascending order and choose the value closest to the 95th percentile rank. For example, if 1000 samples are ranked, the 95th percentile value might be taken to be the observation ranked at position 950 out of 1000. In a similar manner, the 99.5th percentile might be taken to be the 995th ranked observation. However if you simply take a single ordered value to represent a particular percentile you could be subjecting yourself to a certain amount of error depending on the way the samples are obtained. For example, if the samples are generated at random, then changing the random number seed would lead to a different set of net asset values, which might lead to a different percentile estimate. Estimates that vary significantly if the random number seed is changed are said to be quite *unstable*.

The goal of this study is to examine ways of estimating percentiles based on a set of observations, determine which estimators are most stable, and explore any methods that can be used to alleviate or lessen the variability of the percentile estimates.

The instability in percentile estimation will be shown to be due to *sampling error*, and is often perceived to be especially important in the tails of the distribution where there may be relatively few samples. For this reason, and also because this is often a key area of interest for risk management, it is important that the estimator behaviour in the tails is understood and sufficiently accurate for the desired application. However, the techniques mentioned in this document are applicable to a wide variety of situations in which the percentile of a distribution is to be determined. Some of these may be just as likely to require evaluation of percentiles in the bulk as well as in the tails. In particular, the following applications are targeted:

- Identification of the 1-year Value-at-Risk (VaR) based on a set of net asset-liability valuations/capital estimates generated via stochastic modelling;
- Determining percentiles of individual risk drivers or financial or economic variables such for the purposes of real world validation or analysis tests;
- Determination of percentiles for an empirical risk driver model designed to sample from the distribution implied by a finite set of observations.

This document firstly describes the features of a good estimator, and then goes on to describe the process of estimating percentiles based on random samples in Section 3. This highlights some of the important concepts including association of values with quantiles (probability levels), the link with the empirical inverse cumulative distribution function (CDF), and the effect of sampling error on localised regions of the empirical inverse CDF. In Section 4, a number of potential estimators are reviewed, many of which are compared in Section 0. Section 0 considers the effect of breaking down the observations into subsets, in order to reduce memory requirements when the number of observations is very large. Finally, Section 0 turns the problem on its head, asking whether the percentile estimates are more stable if we employ variance reduction methods (such as more sophisticated sampling methods) rather than attempting to use more sophisticated estimators.

2. Estimators

An estimator is essentially a rule or algorithm by which a single population estimate of a particular quantity can be derived from a set of observations. These observations represent a finite subset of the entire population of possible values. There are two considerations when choosing an estimator:

- Bias;
- Variance (or standard deviation).

Bias represents a systematic difference between the estimates and the true value. Generally we do not know the form of the underlying distribution and so the true percentile value is not known exactly. We rely on the estimator to produce estimates that reflect the true value and clearly any significant bias would be misleading.

Variance refers to the spread of estimates that are obtained when independent sets of observations are taken. A set of observations is often obtained by random sampling, such as in Monte Carlo techniques, or scientific measurement in which experimental error may be present. The value of a given quantity obtained from the set of observations will obviously depend on the random sampling. Amongst other things, sources of variablility can include the underlying random number generator, any random number seeds, and also the number of samples. The number of samples is related to the level of sampling error, since this will affect the degree of asymptotic convergence of the random sampling. There can only ever be a finite number of discrete values used to describe a continuous distribution. In the presence of such sources of variability in observation sets, a good estimator will have very low variance.

Some candidate estimators may have low bias but high variance. Others might have low variance but some significant amount of bias. The choice of an estimator should consider the best tradeoff between bias and variance given the particular application.

Measurement of the bias and variance of an estimator can be done by examining the statistics of repeated sampling. The following steps are performed.

1. Obtain a set of samples, use these to obtain a single estimate of the percentile in question.
2. Perform step 1 many times with independent sets of samples, obtaining an estimate for each set. Analysis of this set of estimates will give information as to the distribution of estimates such as mean (and hence bias, if the true value is known), variance, etc. Note that all these measures are subject to sampling error. That is, they will vary depending on the scheme being used to produce the independent samples, such as random number seeding.
3. Perform step 2 many times to get many values for mean, variance, etc. In effect this is a nested simulation scheme. These values can be analysed statistically to determine expected mean and expected variance (as well as confidence intervals) that account for the variability due to the random number generation.

This is the method that has been used to compare some of the candidate estimators in Section 0.

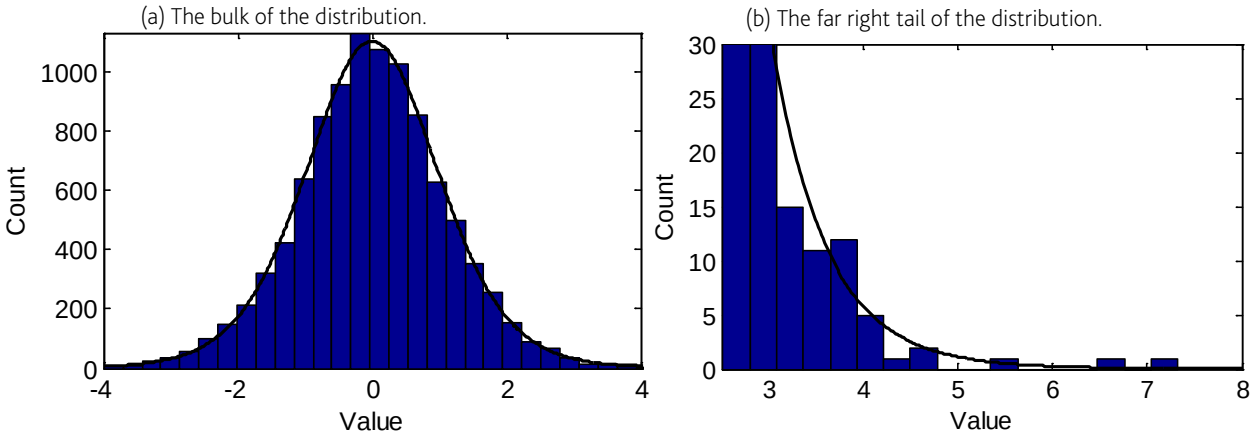
3. Application to Percentiles based on Randomly Sampling

Consider the problem of determining the 99.5th percentile of the student t distribution with 10 degrees of freedom based on a set of random samples. This distribution has tails that are fatter than the normal distribution, potentially exacerbating the variability of such percentile estimates. Figure 1(a) shows a histogram of a set of 10000 samples from a student t distribution with 10 degrees of freedom, overlayed with the probability density function (PDF) for this distribution. The PDF has been scaled to account for the total number of samples and the width of the bars in the histogram. This allows for comparison, giving us some idea about where there are more samples than expected and where there are fewer samples than expected. Figure 1(b) shows a close up of the right tail showing the largest values. In this particular set, there are two samples with value greater than 6.0. Judging by the overlayed PDF this would be quite unusual, since the height of the bars exceeds the PDF. Likewise, comparison with the PDF suggests that the total number of samples with value larger than 3.0 would appear to be somewhat less than what would be expected.

Order statistics

The standard procedure for estimating any particular percentile from a set of data such as this begins by sorting the data into ascending order. These sorted data values are known as *order statistics*. These sorted values are treated as if they are in an array in which the first element in the array is the smallest sample. This first element is referred to as the *first order statistic*. The second element in the array is the next smallest element and is referred to as the *second order statistic*, and so on. An estimate of the percentile can then be obtained as the value at the 99.5th percentile of indices into the array. For example, if we have 10000 samples sorted into ascending order, then the 9950th element of this sorted array can be used as an estimate of the 99.5th percentile.

Figure 1 Histogram of 10000 raw data points sampled at random from the student t distribution with 10 degrees of freedom. Also shown in black is the exact PDF of this distribution.



Relationship with the empirical inverse CDF

The approach just presented is really making use of the empirical form of the inverse CDF (otherwise known as the quantile function) for the (usually unknown) distribution from which the samples are taken. That is, there is an implicit assumption that a particular value (order statistic) is associated with a particular percentile level. The usual assumption when constructing such an empirical inverse CDF is that the samples are perfectly stratified. That is, when the CDF of the (usually unknown) distribution is applied to the samples, the resulting samples are perfectly uniformly distributed. The degree to which this is true depends on the sample generator, and in most cases there is some degree of non-uniformity. Some parts of the uniform distribution will be seen to be slightly undersampled, while other parts will be slightly oversampled in a compensating manner. Changing the random number seed will change where the undersampling and oversampling occur.

As mentioned earlier, points on the empirical CDF and empirical inverse CDF are constructed assuming perfectly stratified samples. The sample values, x_i, are sorted into ascending order $(x_1 \leq x_2 \leq \cdots \leq x_n)$ and then a quantile (probability level) is assigned to each of the ordered values as:

$$p_i = \frac{(i - 0.5)}{n}, \quad \text{for} \quad i = 1,2, \ldots, n.$$

This stratification of p_i is effectively partitioning the unit interval (0 to 1) into n equal subintervals and using the mid point of each. Viewed as (value, probability level) pairs, (x_i, p_i), this maps out the empirical CDF, and viewed as (probability level, value) pairs, (p_i, x_i), the points are on the empirical inverse CDF.

Figure 2 shows the empirical inverse CDF for the same data portrayed in Figure 1. Comparison is made with the ideal inverse CDF of the underlying distribution, which we know to be the student t distribution with 10 degrees of freedom. On a scale in which the entire distribution is considered (Figure 2(a)), the fit looks very good, as would be expected given the asymptotic properties of any good random number generator. Looking in detail at a small part shows the effect of the oversampling and undersampling. Figure 2(b) shows the right tail of the points on the empirical inverse CDF along with the ideal inverse CDF. The perceived undersampling in this region (observed in Figure 1(b)) manifests itself as a local deviation of the empirical inverse CDF from its ideal level. Such local deviations lead to variability in percentile estimates obtained from such randomly generates sample sets.

Figure 3 shows the right tail of the empirical inverse CDF for a number of different samples sets, generated with different random number seeds. In each subfigure we plot several different cases with connecting lines so that they can be distinguished. The local variability of the empirical CDF at the 99.5th percentile is seen to vary more for Figure 3(a), which has 1000 samples in each set, than for Figure 3(b), which has 10000 samples in each set. This reduction of variability with increasing size of the sample set is consistent with the notion that the underlying uniform random number generator is becoming closer to being asymptotically uniform.

Figure 2 Empirical distributional forms and their comparison with the ideal distributional form. Note the local deviation of the empirical inverse CDF from that of the known distribution.

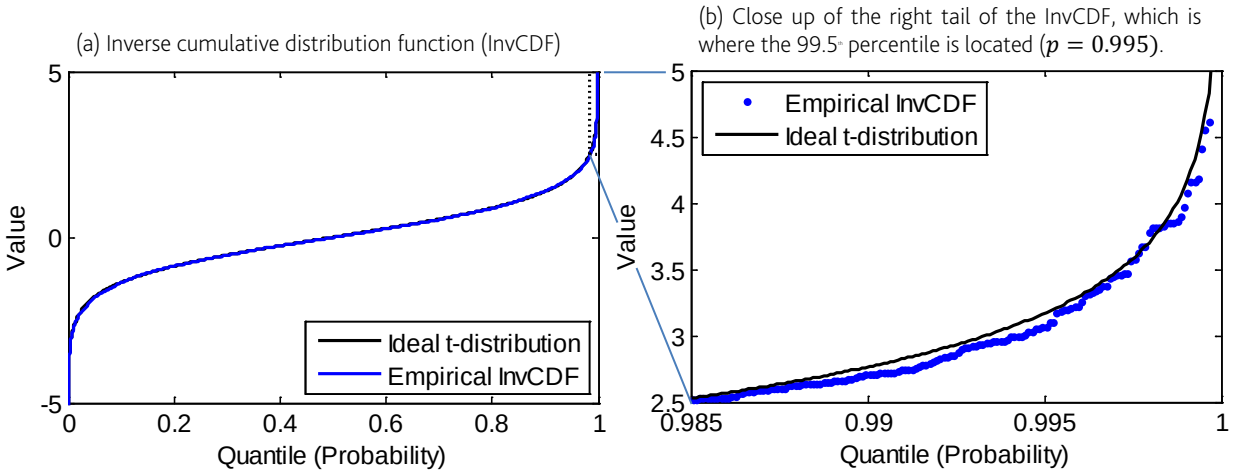
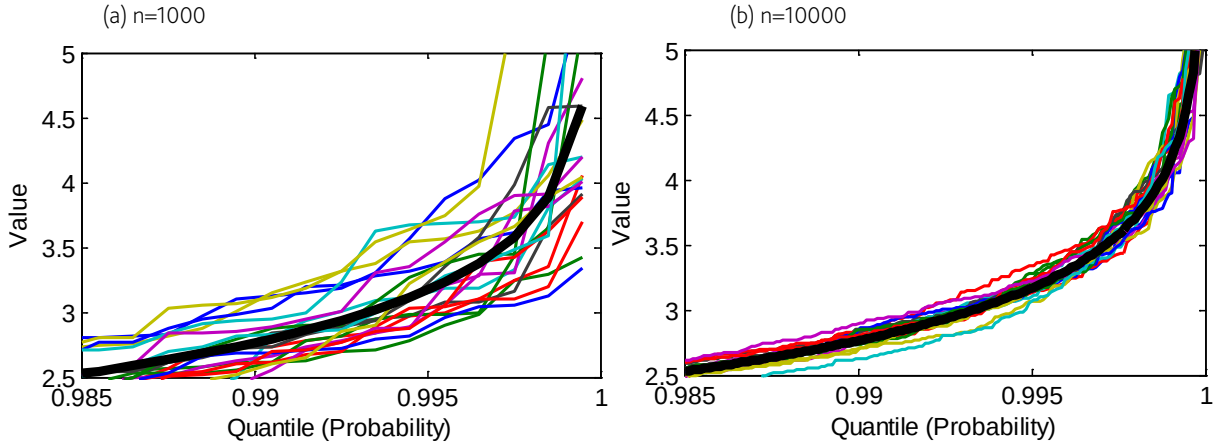


Figure 3 Close ups of the right tail of the empirical inverse CDF, which is where the 99.5th percentile is located, showing the variability with different random number seed. The number of samples is different for each of the subfigures. In both cases the thick black curve shows the exact inverse CDF for the t distribution with 10 degrees of freedom.



4. Candidate Estimators

The candidate estimators for determining the percentiles of a population distribution based on a set of samples can be split into two broad categories:

» Global estimators, which take into account all of the observations on a roughly equal basis;

» Local estimators, which only place a significant weight on a small number of order statistics, often defined by a window around the percentile level of interest.

In this section it will be claimed that there is no global estimator that is effective and simple to implement. It will also be shown, in conjunction with the local variability of the empirical inverse CDF outlined in the previous section, that any local estimator will be subject to significant variance in most cases.

In this section we briefly survey many of the candidate estimators, both global and local.

Fitting a functional form to the entire distribution (Global)

If the underlying distribution is known, fitting a functional form to the sample set may be the best way of fitting a distribution and determining population percentiles. However, in most situations the functional form of the distribution is not known.

Fourier transform methods (Global)

Converting the problem into the frequency domain allows particular frequency components to be isolated. Using simple filtering, the high frequency components can be removed, leaving the lower frequency components that are more likely to be real features of the underlying distribution. Despite being non-trivial to implement there are arbitrary decisions as to how to filter the frequency domain. That is, what frequencies are high enough to be considered "noise" stemming from the sampling error, and what frequencies constitute the real underlying distributional form.

Single order statistic estimators (Local)

There are a number of estimators that make use of a single order statistic. These methods effectively define the empirical inverse CDF as a staircase function (discrete steps). There are a few variants depending on exactly which ordered statistic is used. For example, if there are 1000 samples sorted to give 1000 order statistics, the 99.5th percentile may be chosen to be the 995th order statistic (empirical CDF or Kaplan-Meier estimator) or the 996th order statistic (upper empirical CDF). The nearest neighbour estimator takes the closest order statistic to the ranked order percentile, and types 1 to 3 in the study in Section 0 are also variants of single order statistic estimators. See Mausser (2001) and the references cited therein for more details.

Linear interpolation (Local)

As an interpolation method this method forces the empirical inverse CDF to pass through the identified (probability level, value) pairs. There are many slightly different ways of defining the (probability level, value) pairs, but then all linear interpolation methods effectively join the neighbouring points. Types 4 to 9 in the study in Section 0 are linear interpolation methods. We highlight the type 5 method in the following algorithm.

ALGORITHM – PERCENTILE ESTIMATION BY LINEAR INTERPOLATION

Let:

$\boldsymbol{x}$ be the samples;

$\boldsymbol{n}$ be the number of samples in total;

$\boldsymbol{p}$ be the probability corresponding to the target percentile, that is, $0 \leq p \leq 1$. For example, for the 99.5th percentile, we would set $p = 0.995$.

Then

1. Sort the samples into ascending order in the array $\boldsymbol{x}$. If there are any duplicates it does not matter what order they appear in the sorted array $\boldsymbol{x}$. Invoke an indexing on the elements of the sorted array $\boldsymbol{x}$, so that
2. Determine $c = p * n + 0.5$. This is a decimal representation of the index into the array $\boldsymbol{x}$ that produces the desired percentile. Note that this is a base 1 index, and should be adjusted to base 0 for some computer implementations by subtracting 1.0.
3. If $c \leq 0.5$ then the desired percentile is the smallest value in $\boldsymbol{x}$. That is, the first element in the sorted array.

   If $c \geq n - 0.5$ then the desired percentile is the largest value in $\boldsymbol{x}$. That is, the last element in the sorted array.

   If $0.5 < c < n - 0.5$ then

   a. Split c into an integer part, i, and a fractional part, d, with $0 \leq d < 1$.

   b. The desired percentile is:

$$percentile = x_i + d * (x_{i+1} - x_i)$$

Kernel smoothing (Local)

Kernel smoothing is a class of statistical techniques that can be used to estimate the value of a function from a discrete set of (possibly noisy) data points. Also called *moving average smoothing* or *moving average filtering*, the estimate is taken to be a weighted average of the values at nearby data points on the empirical inverse CDF.

These methods use a window to define a number of ordered statistics or, equivalently, a band of percentiles, that is centred on the percentile of interest. The kernel estimators are some kind of simple average or a weighted average with the weights corresponding to values from some kernel function. Gaussian, logistic, tricube and Epanechnikov kernels are all fairly common but arbitrary choices. In some respects, the use of more order statistics is naturally expected to give larger bias, but the extent to which this is true depends on the actual kernel/weights used.

The simplest kernel is the box (or uniform) kernel, which effectively applies an equal weighting to all ordered statistics within the window. That is, the percentile is an average of the order statistics within the window. However, due to the curvature of the inverse CDF function a box kernel is known to result in a clear bias, as demonstrated in Section 0, and in particular Figure 4(b).

Harrell-Davis Estimator (Local)

The Harrell-Davis estimator is a weighting scheme that is tailored to a specific percentiles levels. For a target percentile level of α with N samples, the weights, $w_{\alpha,N,k}$, of the ordered statistics indexed by k are determined by the following:

$$w_{\alpha,N,k} = I_{k/N}\big((N+1)\alpha, (N+1)(1-\alpha)\big) - I_{(k-1)/N}\big((N+1)\alpha, (N+1)(1-\alpha)\big),$$

where $I_x(a,b)$ is the regularised incomplete beta function. Weights are defined for all of the ordered statistics ($1 \le k \le N$), but the only significant weights are those in a localized region around the percentile in question. The weights are dependent on the number of data points and also the desired percentile and so if either of these quantities are changed the weights need to be recalculated. See Mausser (2001) for more details as well as Albin and Brodin (2006), which presents some modifications of the Harrell-Davis estimator and looks at bias-correcting schemes.

Local Regression Smoothing (Local)

This is a general class of smoothing method in which a functional form is fit to the points of the empirical inverse CDF contained within a certain localised window centred on the percentile level in question. The names "lowess" and "loess" are derived from the term "locally weighted scatter plot smoothing," as both methods use locally weighted regression to smooth the functional form data, in this case the empirical inverse CDF. The most common functional forms are simple linear and quadratic forms. The points within the window cane be weighted by any of a number of kernel weighting functions so that points close the the percentile level of interest are fit more closely. In addition to the regression weight function, you can use a robust weighting scheme, which makes the process more resistant to outliers.

Fitting a functional form to a part of the distribution (Local)

Fitting a particular functional form to the tail means that we can directly read off the percentile of interest. Generalised Pareto distributions are commonly used to fit to tails of distributions, but forms of the Weibull distributions and others (generalised extreme value distributions) are sometimes used as alternatives. See McNeil et al. (2005) for more details. If made to be robust, these methods may be particularly useful in evaluating conditional tail expectations, since the expectation can be taken analytically over the functional form. If maximum likelihood methods are used for this fitting process then this naturally provides confidence intervals for the percentile estimates.

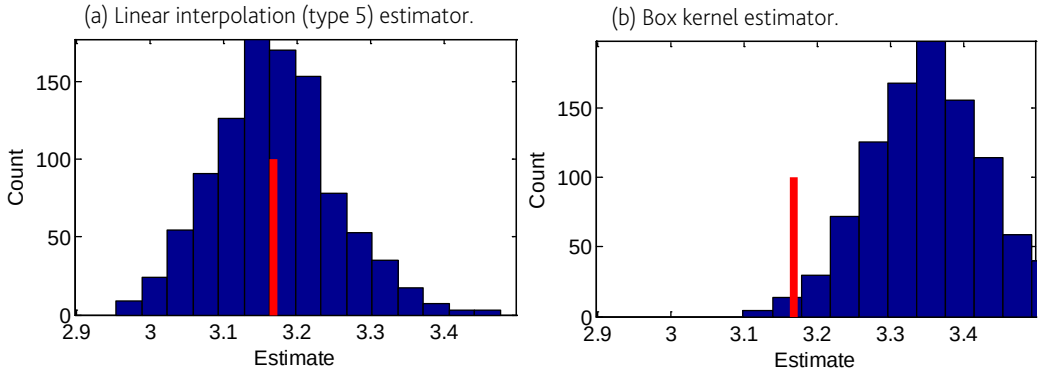
Obviously these are only applicable for target percentile levels in the tail (rather than the bulk). In addition, implementation of such a tail fitting method is not trivial. This is particularly the case when the number of samples is quite small and/or the underlying distribution only has short tails. In these cases the number of fitting points in the tail will be correspondingly smaller. In practice the underlying distribution is not known and so there are often arbitrary decisions to be made concerning the percentile level at which the tail should actually start (although the mean excess function can provide somewhat noisey information in this regard: See p279 of McNeil et al. (2005) for more details). Getting this wrong could lead to poor fit of the tail and hence misleading or biased percentile estimates.

5. A Comparison of Estimators

In this section we perform a statistical analysis of many of the candidate estimators described in the previous section. This comparison aims to provide evidence with which to choose a particular estimator. The method that we used here is described in Section 2.

Figure 4 shows results of a small part of the analysis. Estimates of percentiles are obtained from many independent sets of samples. The distribution of these estimates is analysed to provide, for example, an estimate of the bias by taking the mean of the samples. Figure 4(a) shows a histogram obtained by applying a linear interpolation estimator to estimate the 99.5th percentile of the t distribution with 10 degrees of freedom. Figure 4(b) shows a similar histogram obtained using a box kernel estimator with a reasonably wide window. The bulk of the distribution is clearly much larger than the true value (marked in red), indicating possible estimator bias.

Figure 4 Histograms illustrating part of the method used in comparing the estimators. These histograms show the estimates of the 99.5th percentile taken from 1000 independent sets of 10000 values randomly generated from the t distribution with 10 degrees of freedom. The exact percentile value is indicated by the thick red line.



We continue the example of determining a specific percentile of the t distribution with 10 degrees of freedom, but here we aim to determine the 95th percentile.

The specific estimators that are included in this study are those outlined in Table 1. Many of these algorithms are discussed in more detail in Hyndman and Fan (1996). All of the estimators are defined as weighted averages of consecutive order statistics. Types 1 to 3 are based (primarily) on single order statistics, types 4 to 9 are based on two consecutive order statistics and HD refers to the Harrell-Davis estimator as outlined in the previous section. The estimators of type 1 to 9 are generically defined to be linear interpolation between the points (p_k, x_k) on the empirical inverse CDF, where x_k is the kth order statistic and p_k is its allocated probability. Specific expressions for p_k are given in Table 1. That is:

$$Q_i(p) = (1-\gamma).x_j + \gamma.x_{j+1},$$

where i is the type, the value j is chosen depending on the value of p such that:

$$\frac{j-m}{n} \le p < \frac{j-m+1}{n}.$$

Also, x_j is the j^{th} order statistic, n is the sample size, the value of γ is a function of $j = \text{floor}(np+m)$ and $g = np + \delta - j$, and δ is a constant.

Table 1 Details of the estimators used in the comparison study.

	General Form	Details
Type 1	Single order statistic	Inverse of empirical distribution function. $\gamma = 0$ if $g = 0$, and $\gamma = 1$ otherwise.
Type 2	Single order statistic	Similar to type 1 but with averaging at discontinuities. $\gamma = 0$ if $g = 0$, and $\gamma = 1$ otherwise.
Type 3	Single order statistic	Nearest even order statistic. $\gamma = 0$ if $g = 0$ and j is even, and otherwise. This is used by SAS.
Type 4	Linear interpolation	$m = 0$, and $p_k = k/n$. That is, linear interpolation of the empirical cdf.
Type 5	Linear interpolation	$m = 0.5$, and $p_k = (k - 0.5)/n$. That is, a piecewise linear function where the knots are the values midway through the steps of the empirical cdf. This is used by Matlab.
Type 6	Linear interpolation	$m = p$, and $p_k = k/(n + 1)$. Thus $p_k = E[F(x_k)]$. This is used by Minitab and by SPSS.
Type 7	Linear interpolation	$m = 1 - p$, and $p_k = (k - 1)/(n - 1)$. In this case, $p_k = \mathrm{mode}[F(x_k)]$. This is used by S.
Type 8	Linear interpolation	$m = (p + 1)/3$, and $p_k = (k - 1/3)/(n + 1/3)$. Then $p_k = \sim\mathrm{median}[F(x_k)]$. The resulting quantile estimates are approximately median-unbiased regardless of the distribution of x.
Type 9	Linear interpolation	$m = p/4 + 3/8$, and $p_k = (k - 3/8)/(n + 1/4)$. The resulting quantile estimates are approximately unbiased for the expected order statistics if x is normally distributed.
HD	Harrell-Davis	The Harrell-Davis estimator as described in Section 4.

The results of the comparison study are shown in Table 2 and Figure 5. This study shows that the various estimators have slightly different bias characteristics. Since a smaller magnitude of bias is better, the type 2 and 5 estimators would be considered best for this particular study, closely followed by the type 8 and type 9 estimators.

The variance (standard deviation) part of this study indicates that almost all of the estimators exhibit comparable variance. The Harrell-Davis estimator is perhaps the pick of the bunch, but its advantage is only marginal. More importantly, they all exhibit very significant variance and so it would be reasonable to expect that percentile estimations from independent sets of samples (supposedly from the same underlying distribution) could show very significant variability. In this sense, any percentile estimation using any of these estimators will not be stable to changes in the underlying random sampling (or experimental error, etc). This is due to the local nature of these estimators.

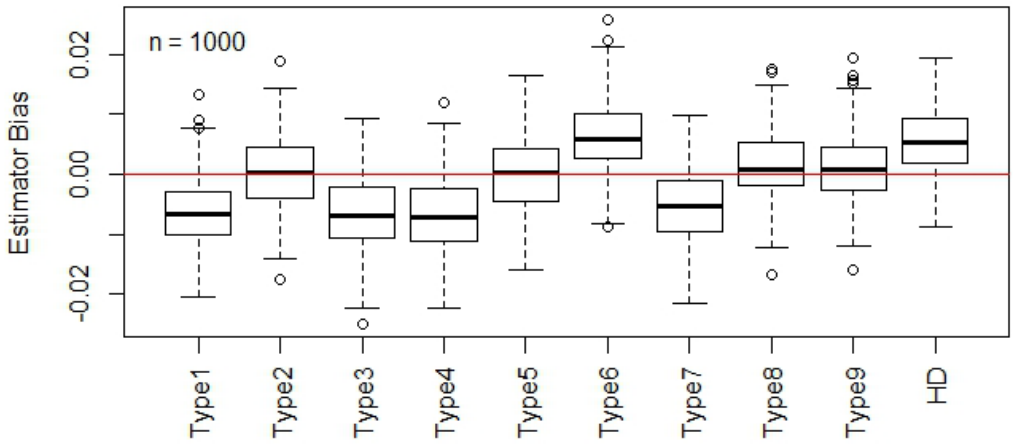
The next section suggests an alternative approach using a compound estimator that is more suitable for some applications.

Table 2 Numerical representation of the results of the comparison of the bias of the estimators. Equivalent graphical results are presented in Figure 5(b).

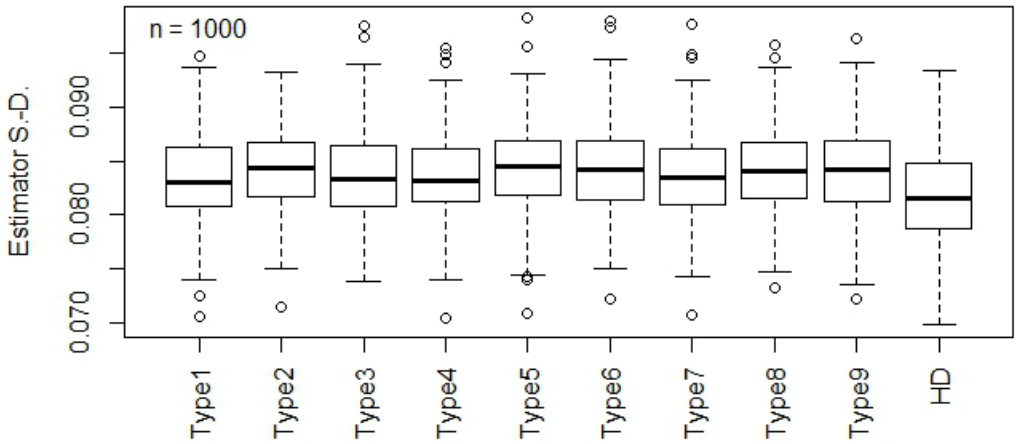
	Type 1	Type 2	Type 3	Type 4	Type 5	Type 6	Type 7	Type 8	Type 9	HD
Minimum	-0.0206	-0.0175	-0.0252	-0.0224	-0.0159	-0.0089	-0.0216	-0.0167	-0.0161	-0.0089
1st Quartile	-0.0101	-0.0039	-0.0105	-0.0112	-0.0047	0.0026	-0.0095	-0.0019	-0.0026	0.0019
Median	-0.0066	0.00019	-0.0069	-0.0071	0.00016	0.0058	-0.0054	0.00081	0.00070	0.0052
Mean	-0.0063	0.0002	-0.0064	-0.0069	-0.00004	0.0063	-0.0054	0.0014	0.0013	0.0054
3rd Quartile	-0.0030	0.0043	-0.0022	-0.0025	0.0043	0.0101	-0.0010	0.0053	0.0044	0.0093
Maximum	0.0132	0.0190	0.0094	0.0119	0.0165	0.0258	0.0099	0.0175	0.0194	0.0194

Figure 5 Graphical presentation of the results of the comparison. The box plots show the median (thick black line), interquartile range IQR (box) and whiskers indicating 1.5 IQR outside of the IQR. The types are as defined in Table 1, and the equivalent numerical details are presented in Table 2.

(a) The estimator bias, measured as a deviation from the exact value for the 95th percentile of the t-distribution with 10 degrees of freedom.



(b) The standard deviation of the estimator.



6. Estimation by Subsetting

The previous section compared the properties of a number of percentile estimators that were local in nature. The study showed that all of them produced significant variability with changes in random number generation. In this section we propose an estimator that uses a subsetting approach in conjunction with any of the local estimators mentioned previously. The main idea is that the total number of samples n is partitioned into m subsets of size d (where $n = md$). A percentile estimate is obtained from each subset. These are averaged to give the overall percentile estimate. The algorithm is as follows.

ALGORITHM – ESTIMATION OF PERCENTILES BY SUBSETTING

Let:

n be the number of samples in total;

m be the number of subsets, where (usually) m is a divisor of n;

$d = n/m$ is the number of samples in each subset;

x be the samples;

p be the probability corresponding to the target percentile, that is, $0 \leq p \leq 1$. For example, for the 99.5th percentile, we would set $p = 0.995$;

Then

1. Partition the n samples into m subsets of size d in a way that the subsets can be considered to be interchangeable. Note that this is critical to the success of the algorithm. One way to do this is by random allocation before any sorting is done.
2. Obtain an estimate of the desired percentile from each of the subsets. This can be done in any reasonable way although best results would be obtained using an unbiased estimator with small variance, such as the type 5 linear interpolation method.
3. Determine the mean of the estimates obtained from the subsets. This is the percentile estimate.

Note that it is critical that step 1 is done in the right way. It is very easy to perform the partition in such a way that the subsetting method would become very unreliable. An example of a very bad way to do this is to sort the elements in x, and then partition the elements so that all of the smallest elements go into the first partition, the next smallest elements go into the second partition, and so on. This method ensures that the samples in each subset are a distorted representation of the underlying distribution and this will not give rise to the correct estimate.

An example of an appropriate way to do the partitioning is by random sampling. In particular, when we have a very large number of samples it is not feasible to keep all of the samples in memory in order to sort them for the purposes of determining the order statistics of the whole sample set. Often such samples are generated at random, meaning that we can take the first d elements as they are generated to be the first subset. Once the percentile estimate of this subset is obtained, all of the elements in the subset can be deleted from memory. The next d elements that are generated can then be used to form the second subset, from which the second estimate can be obtained , and so on until all m subset estimates are obtained.

In terms of variance, there is no distinct advantage of using a subsetting approach over any of the approaches considered earlier, provided that the total number of samples, n, is unchanged. The variance of the estimates from the subsets will be larger than the variance of the estimate obtained by considering the whole sample set at once. However the averaging of the subset estimates reduces the variability and so the net confidence in estimates obtained via subsetting or not will be comparable.

7. Percentile Estimation using Variance Reduction Techniques

The previous sections have highlighted the role of sampling error in the variability of percentile estimation process. It is natural to consider ways in which sampling error can be reduced as such variance reduction techniques would likely improve percentile estimates also. The simplest way to reduce sampling error is to increase the number of samples, however there are limits to the practicality of this as more samples means longer run times and more analysis. An alternative is to use one or more variance reduction techniques that naturally incur lower sampling error, such as stratified sampling, low-discrepancy sampling or importance sampling.

Stratified sampling in one dimension can be extremely effective if the number of samples is fixed and known in advance. Extension to higher dimensions can lead to some practical limitations. In particular there are often special requirements on the number of samples before good uniformity is observed, and this is where sampling using low-discrepancy sequences can be advantageous. In some ways, low-discrepancy sampling can be viewed as an extension of stratified sampling into multiple dimensions. The uniformity of low-discrepancy sequences is far superior to randomly generated samples whilst maintaining low correlation between samples of different dimensions. In addition, there are no particular requirements in the number of samples that are required for these uniformity and correlation properties to be observed. Note that there is a notion of *scrambling* of low-discrepancy sequences, which produces an independent set of samples with the same uniformity and correlation properties. These scrambled low-discrepancy sequences can be used to determine confidence intervals on estimates. See Glasserman (2004), Jäckel (2002) or Redfern (2014) for more detailed compare-and-contrast studies of random sampling, stratified sampling and low-discrepancy sampling.

Even better are the potential benefits to be had from using importance sampling. This is essentially attempting to achieve greater Monte Carlo efficiency by changing the probability measure from which the samples are generated, giving more weight to samples that are important to the task at hand. In most cases effective importance sampling needs to be tailored to the particular situation, which limits its general applicability. Importance sampling is still a very active area of research. See Section 4.6 of Glasserman (2004) for more details and several good examples.

An example using low discrepancy sampling

Consider again the task of estimating a specific percentile of a population based on a given sample set. Compared to the random sampling examined in earlier sections, if the samples are derived from low-discrepancy sequences, the benefits of the reduced sampling error will be apparent in reduced variability of the estimators. To demonstrate this, we continue the example earlier of determining the 99.5th percentile of a population based on a set of samples taken from the t distribution with 10 degrees of freedom. Figure 6 shows the right tail of the empirical inverse CDF for a variety of cases. Each of the subfigures shows 10 examples with lines joining the empirical inverse CDF points. The left side (subfigures (a) and (c)) shows the effect of random sampling using 10 different random number seeds with either 1000 or 10000 samples, directly analogous with those shown in Figure 3. The right side of Figure 6 (subfigures (b) and (d)) shows the equivalent using low-discrepancy sampling, the one-dimensional Sobol sequence in this case. Results from ten different scrambles are shown, which is analogous to 10 different random number seeds (See Redfern (2014) for more details). At the 99.5th percentile (0.995 probability level) there is clearly much less variability in the empirical inverse CDF when low-discrepancy samples are used.

Presenting this data in a slightly different way, Figure 7 shows some histograms of estimates obtained using random sampling and low-discrepancy sampling. The linear interpolation estimator has been used here, but any percentile estimator can be used. In each subfigure, results from 1000 independent sets of 10000 samples are used, making use of 1000 different scrambles of the Sobol sequence. The true value of the 99.5th percentile is shown in red. Clearly, these results indicates that the estimates using low-discrepancy sampling are much more tightly clustered around the true value than the estimates obtained via random sampling. This highlights the advantage of the reduced sampling error that low-discrepancy sequences can bring to the modelling process.

Increasing the number of random samples will reduce the sampling error and hence reduce the spread in estimates. In this particular case, the standard deviation of estimates obtained via random sampling is about an order of magnitude greater than that using low-discrepancy sampling. This can be put in perspective by using the well known $1/\sqrt{N}$ convergence rate due to random sampling. To reduce the standard deviation of the estimates obtained via random sampling by an order of magnitude (to approximately match the level of those obtained via low-discrepancy sampling with this particular number of samples, i.e. 10000) would require about 100 times the number of random samples (i.e. one million). Note that this is a simple case requiring only one dimension of a Sobol sequence, and the factor of about 100 is specific to the number of samples that have been used. When two or more dimensions are required, meaning that the quantity of interest is a function of more than one risk factor, the advantage of using low-discrepancy sampling may be less pronounced.

Figure 6 Close ups of the right tail of the empirical inverse CDF, which is where the 99.5th percentile is located, showing the variability that can result using a different random number seed ((a) and (c)) or using a different low-discrepancy scramble ((b) and (d)). In all cases there are 10 curves plotted, with the (thick) smooth black curve showing the exact inverse CDF for the t distribution with 10 degrees of freedom.

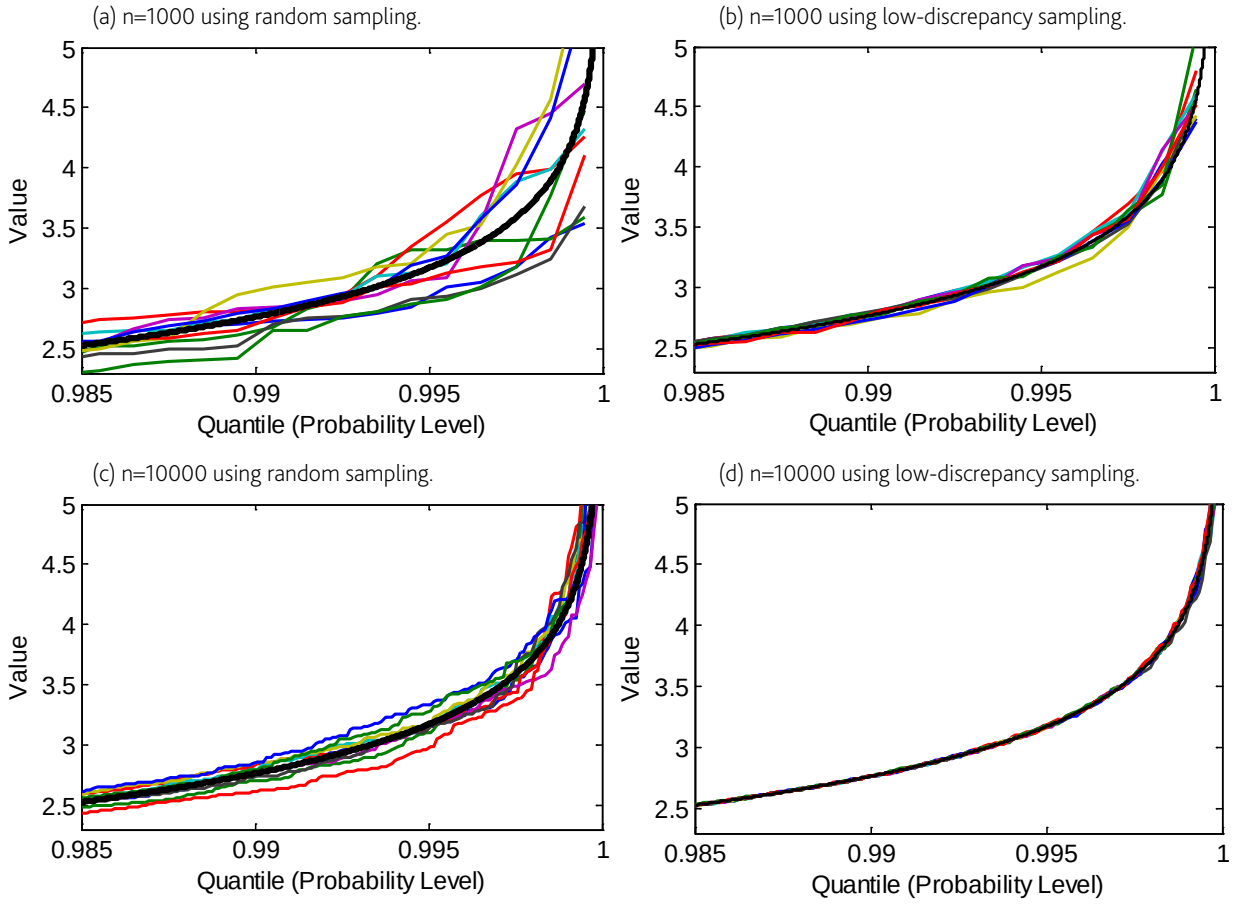
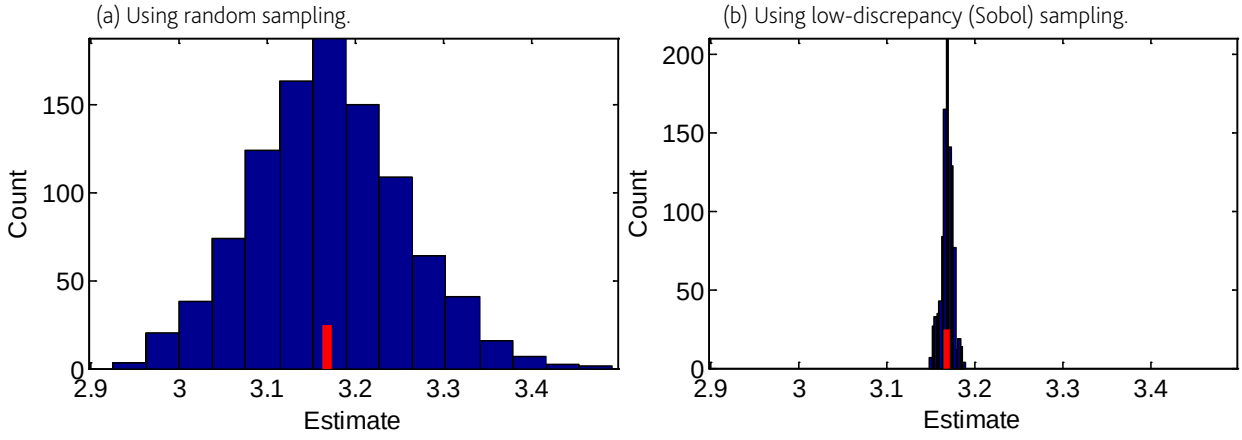


Figure 7 Histograms of estimates of the 99.5th percentile of the t distribution with 10 degrees of freedom, obtained using the linear interpolation estimator (type 5) estimator. Each estimate was obtained from independent sets of 10000 samples, using either random sampling or low-discrepancy (Sobol) sampling. The true value of the 99.5th percentile is shown as the red line.



8. Summary and Conclusions

This document has focused on the percentile estimation process which is a major part of the procedure for determining the solvency capital requirement. Reducing the variability (or error) in the estimation of percentiles will lessen the variability of the capital requirement.

We started by presenting some of the features of a good estimator, being low bias and low variance. We also noted that it is common that a tradeoff between bias and variance needs to be made when choosing an estimator.

In Section 3, the percentile estimation process was described by considering random samples from a known distribution so that the true value of percentiles could be determined. The link between the ordered values of the samples (order statistics) and probabilities/quantiles was established, and the local nature of deviations from the true percentile values was related to sampling error. This is due to the local non-uniformity in the underlying "uniform" random number generation.

Specific percentile estimators were presented in Section 4. It was claimed that there is no appropriate global estimator since, in general, the underlying distribution is not known. The remaining estimators are local in that they take into account only a subset of the order statistics in a localised region around the percentile level of interest. In Section 0 it was shown that these local estimators exhibit varying degrees of bias, but all have very significant levels of variance. This variance is due to the aforementioned sampling error and dominates the percentile estimation process for many of the better local estimators.

The analysis in Section 0 led to some more specific conclusions:

- The box kernel estimator is known to give rise to significant bias in percentile estimates.
- There is some evidence of bias in the Harrell-Davis estimator.
- Since there are many estimators that have roughly the same performance, we may as well use a simple estimator. Of the estimators studied in the comparison in Section 0, the type 2 and type 5 linear interpolation estimators performed best with respect to the desire for a small bias. Since the type 2 estimator is effectively a discrete function of probability, whereas the type 5 estimator is continuous, we would recommend the type 5 linear interpolation estimator as the preferred (local) percentile estimator.

Section 0 considers percentile estimation by partitioning the observations into subsets in a suitable way. This can be useful when large numbers of observations and/or memory limitations mean that it is not feasible to store all of the observations in memory. However this subsetting cannot reduce the variability in the final percentile estimate if the same total number of samples is being used.

The bottom line is that there is no quick way to improve stability in percentile estimation based on random sampling. Reducing the sampling error will make the percentile estimates and hence capital requirement more stable. One way to do this is to increase the number of random samples, but this also increases the calculation burden.

In an attempt to reduce the effects of sampling error, the impact of variance reduction techniques was considered in Section 0. As one example of a variance reduction technique, the estimation of percentiles from data sets obtained using low-discrepancy (Sobol) sequences was studied. It was shown that if you use low-discrepancy sampling rather than random sampling, it is possible that the required number of samples to achieve a particular level of certainty in the percentile estimate could be reduced. For the simple example considered, involving only one risk factor, the impact was shown to be dramatic, as shown in Figure 7. This should flow on to a much greater degree of confidence in the solvency capital requirement. Note that when the quantity of interest is a function of more than one risk factor, the advantage of using low-discrepancy sampling may be less pronounced.

References

J. M. P. Albin and Erik Brodin, "Smoothing the Harrell-Davis Quantile Estimator", Chambers University of Technology, 2006. Available from http://www.math.chalmers.se/~palbin/newHDHD.pdf.

Paul Glasserman, *Monte Carlo Methods in Financial Engineering*, Ch 5, Springer, 2004.

Peter Jäckel, *Monte Carlo Methods in Finance*, John Wiley & Sons, Chichester, 2002.

Helmut Mausser, "Calculating Quantile-based Risk Analytics with L-estimators", *Algo Research Quarterly*, 4(4), pp33-47, 2001.

McNeil, A.J., Embrechts, P. and Frey, R., *Quantitative Risk Management Concepts, Techniques and Tools*, Princeton University Press, 2005.

David Redfern, "Low discrepancy generators", Moody's Analytics technical document, (to appear 2014).